

На правах рукописи

БАСКИН Игорь Иосифович

МОДЕЛИРОВАНИЕ СВОЙСТВ ХИМИЧЕСКИХ СОЕДИНЕНИЙ С
ИСПОЛЬЗОВАНИЕМ ИСКУССТВЕННЫХ НЕЙРОННЫХ СЕТЕЙ И
ФРАГМЕНТНЫХ ДЕСКРИПТОРОВ

02.00.17 – математическая и квантовая химия

ДИССЕРТАЦИЯ
на соискание ученой степени
доктора физико-математических наук

Москва – 2009

СОДЕРЖАНИЕ

Содержание	2
Введение	9
Глава 1. Искусственные нейронные сети	13
1.1. Введение	13
1.2. Основные принципы нейросетевого моделирования	14
1.2.1. Общая терминология	14
1.2.2. Нейрон МакКаллока-Питтса	15
1.2.3. Персептрон Розенблатта	17
1.2.4. Нейросети обратного распространения (backpropagation)	19
1.2.5. Другие архитектуры нейронных сетей	34
1.3. Основные принципы применения искусственных нейронных сетей для прогнозирования свойств химических соединений	60
1.4. Ограничения искусственных нейронных сетей	62
Глава 2. Фрагментные дескрипторы в поиске зависимостей структура-свойство	63
2.1. История фрагментных дескрипторов	63
2.2. Типы фрагментных дескрипторов	67
2.2.1. Классификация по типам молекулярных графов	67
2.2.2. Классификация по типам молекулярны структур	87
2.2.3. Классификация по типам значений дескрипторов	90
2.2.4, Класификация по типам дескрипторных наборов	91
2.2.5. Классификация по связности фрагментов	96

2.2.6. Классификация по уровням детализации молекулярных графов	97
2.2.7. Фрагментные дескрипторы с выделенными атомами.....	100
2.3. Ограничения фрагментных дескрипторов.....	101
Глава 3. Математическое обоснование выбранного подхода	103
3.1. Химическая значимость поиска базиса инвариантов помеченных графов	103
3.2. Две основные теоремы о базисе инвариантов графов.....	105
3.3. Теоретические основы сочетания искусственных нейронных сетей и фрагментных дескрипторов	108
Глава 4. Разработка нейросетевых подходов	111
4.1. Подход к решению проблемы «переучивания» нейронных сетей.....	111
4.1.1. Суть эффекта «переучивания» нейросетей	111
4.1.2. Методы предотвращения «переучивания» нейросетей	113
4.1.3. Трехвыборочный подход.....	115
4.1.4. Процедура двойного скользящего контроля	116
4.1.5. Быстрая пошаговая множественная линейная регрессия	118
4.2. Подход к интерпретации нейросетевых моделей	119
4.3. Концепция обучаемой симметрии.....	129
Глава 5. Разработка фрагментных подходов	142
5.1. Принципы построения и генерации фрагментных дескрипторов	142
5.1.1. Типы фрагментов	143
5.1.2. Иерархическая классификация атомов во фрагментах	145
5.1.3. Построение фрагментного дескриптора	153
5.1.4. Генерация кодов фрагментов с обобщенными типами атомов.....	154

5.1.5. Алгоритм генерации фрагментных дескрипторов.....	156
5.2. Примеры прогнозирования физико-химических свойств органических соединений с использованием фрагментных дескрипторов и линейно-регрессионных моделей.....	158
5.2.1. Прогнозирование поляризуемости органических соединений	159
5.2.2. Прогнозирование энтальпий образования алифатических полинитросоединений	161
5.2.3. Прогнозирование магнитной восприимчивости органических соединений	163
5.2.4. Прогнозирование энтальпии парообразования органических соединений	169
5.2.5. Прогнозирование энтальпии сублимации органических соединений.....	171
5.2.6. Прогнозирование температуры вспышки органических соединений	176
5.2.7. Прогнозирование сродства азо- и антрахиноновых красителей к целлюлозному волокну.....	180
5.3. Фрагментные дескрипторы с «выделенными» атомами.....	183
5.3.1. Прогнозирование химических сдвигов в ^{31}P ЯМР спектрах замещенных монофосфинов	185
5.3.2. Прогнозирование способности аналогов 1-[(2-гидроксиэтокси)-метил]-6(фенилтио)тимина (HEPT) ингибировать обратную транскриптазу вируса ВИЧ-1.....	187
5.3.3. Прогнозирование констант скорости гидролиза эфиров карбоновых кислот	189

5.4. Псевдофрагментные подходы. FRAGPROP. Прогнозирование физических свойств полимеров	191
Глава 6. Сочетание искусственных нейронных сетей и фрагментных дескрипторов	198
6.1. Первые свидетельства эффективности совместного использования искусственных нейронных сетей и фрагментных дескрипторов.....	198
6.2. Прогнозирование физико-химических свойств органических соединений с использованием фрагментных дескрипторов и нейросетевых моделей	203
6.3. Моделирование физических свойств органических жидкостей в рамках процедуры трехвыборочного скользящего контроля.....	204
6.3.1. Общая методология моделирования	205
6.3.2. Моделирование вязкости органических соединений	207
6.3.3. Моделирование плотности жидких органических соединений	213
6.3.4. Моделирование давления насыщенных паров.....	216
6.3.5. Моделирование температуры кипения разнородных органических соединений	218
6.4. Прогнозирование температуры плавления ионных жидкостей	223
Глава 7. Разработка интегрированных подходов.....	227
7.1. Совместное применение методологии искусственных нейронных сетей и методов молекулярного моделирования.....	227
7.1.1. Предсказание положения длинноволновой полосы поглощения симметричных цианиновых красителей.....	229

7.1.2. Оценка значений констант ионизации для различных классов органических соединений	233
7.1.3. Моделирование мутагенной активности замещенных полициклических нитросоединений с помощью искусственных нейронных сетей	238
7.1.4. Прогнозирование констант заместителей с использованием искусственных нейронных сетей и квантово-химических дескрипторов.....	245
7.2. Корреляции структура-условия-свойство	246
7.2.1. Концепция построения нейросетевых зависимостей структура – условия – свойство.....	246
7.2.2. Построение и анализ нейросетевых зависимостей структура-условие-свойство для физико-химических свойств углеводородов.....	248
7.2.3. Построение и анализ нейросетевых зависимостей структура – условия реакции – константы скорости для реакции кислотного гидролиза сложных эфиров карбоновых кислот	256
7.3. Индуктивный перенос знаний при интеграции моделей «структура-свойство».....	262
7.3.1. Многоуровневый принцип построения моделей «структура-свойство»	264
7.3.2. Параллельный принцип построения моделей «структура-свойство». Многозадачное обучение.	270
7.4. Нейронное устройство для проведения прямых корреляций «структура-свойство».....	274
7.4.1. Введение.....	274
7.4.2. Описание нейронного устройства	276

7.4.3. Примеры разных конфигураций нейронного устройства.....	283
7.4.4. Применение нейронного устройства в исследованиях «структура- свойство» для органических соединений	285
7.4.5. Выводы	292
Глава 8. Разработка программных средств.....	294
8.1. История разработки программных средств.....	294
8.2. Программный комплекс «NASAWIN»	297
8.2.1. Представление химической информации.....	298
8.2.2. Интеграция с программными компонентами, осуществляющими расчет дескрипторов химических структур.....	298
8.2.3. Химически-ориентированная визуализация	299
8.2.4. Модификация дескрипторов и свойств.....	299
8.2.5. Предварительный отбор дескрипторов.....	299
8.2.6. Построение классификационных моделей структура-активность	300
8.2.7. Нейросетевые парадигмы	301
8.2.8. Интерпретация нейросетевых моделей.....	301
8.2.9. Отбор дескрипторов в ходе обучения нейросети	301
8.2.10. Определение момента начала «переучивания» нейросети	302
8.2.11. Кластеризация баз данных	303
8.2.12. Динамическая визуализация хода обучения нейросети.....	303
8.2.13. Определение области применимости модели	304
8.2.14. Химически-ориентированный блок прогноза	304
8.3. Дескрипторный блок «FRAGMENT»	304

8.4. Дескрипторный блок «FRAGPROP».....	306
8.5. Автономные прогнозаторы свойств органических соединений	310
Выводы	312
Литература	315
Благодарности.....	364
Список обозначений и сокращений	365

ВВЕДЕНИЕ

На современном этапе развития химии, когда накоплен и организован в виде электронных баз данных огромный объем экспериментальных данных, особое внимание уделяется компьютерным методам обработки характеристик уже исследованных веществ с целью предсказания свойств, которыми обладают еще не исследованные соединения либо которыми будут обладать новые, еще не синтезированные вещества. Это, в свою очередь, открывает большие перспективы в решении одной из главных задач химической науки - целенаправленной разработке новых веществ и материалов с заранее заданными свойствами.

Тем не менее, несмотря на актуальность этой задачи, до последнего времени отсутствовала универсальная, строго обоснованная и, в то же время, легкая для понимания методология, которая позволила бы химику на основе обработки экспериментальных данных осуществлять прогнозирование всевозможных свойств химических соединений. Главной целью настоящей диссертационной работы была разработка универсальной методологии, позволяющей с единых позиций прогнозировать самые разнообразные свойства органических соединений на основе обработки эмпирических данных. В данной работе сначала математически обоснован, а потом и на множестве примеров проиллюстрировали центральный тезис диссертационной работы – такой универсальной методологией является сочетание многослойных искусственных нейронных сетей и фрагментных дескрипторов.

Искусственные нейронные сети в настоящее время являются одним из наиболее широко применяемых методов для восстановления по экспериментальным данным как разнообразных количественных зависимостей, так и для проведения качественной классификации. Благодаря уникальной возможности осуществлять построение нелинейных моделей любого уровня сложности, особенно в тех случаях, когда неизвестен общий вид аналитической зависимости, нейронные сети нашли широкое применение в рамках поиска зависимостей

между структурами органических соединений и их физико-химическими свойствами (QSPR) и биологической активностью (QSAR).

Несмотря на широкое использование искусственных нейросетей для получения зависимостей структура – свойство, до настоящего времени не существовало универсального программного комплекса, реализующего все необходимые этапы построения моделей и позволяющего исследователям-химикам комплексно, с учетом особенностей работы со структурной информацией, применять методологию нейронных сетей. Именно разработка такого программного комплекса, реализующего универсальную методологию построения моделей, предназначенных для количественного прогнозирования разнообразных свойства органических соединений на базе сочетания многослойных нейронных сетей и фрагментных дескрипторов, а также его апробация на различных примерах, и составляла важнейшую задачу диссертационной работы.

Следует отметить, что на период начала работы отсутствовало понимание основных принципов работы с нейронными сетями для построения QSAR/QSPR-моделей. В частности, не было ясно, как лучше всего предотвращать «переучивание» нейросетей, как объективно оценивать прогнозирующую способность полученных моделей, а также как эффективно отбирать дескрипторы для их построения, как их использовать для определения области применимости моделей. Кроме того, в рамках методологии QSAR/QSPR практически не предпринималось попыток учета влияния внешних условий (таких, например, как температура, давление, концентрация вещества, наличие и свойства того или иного растворителя и т.п.) на исследуемые свойства, а также прогнозировать свойства многокомпонентных систем. Не было также ясно, как применять аппарат нейронных сетей в сочетании с техникой молекулярного моделирования. Кроме того, ранее не существовало методов, позволяющих давать понятную химикам интерпретацию нейросетевым регрессионным моделям. На эти и ряд других важных вопросов, связанных с применением нейросетей для построения QSARQSPR-моделей, дан ответ в данной работе.

Следующая важная часть работы связана с разработкой универсального набора фрагментных дескрипторов, которые могли бы служить для как можно

более точного прогнозирования самых разнообразных свойств органических и металлоорганических соединений. Кроме специального дизайна самих дескрипторов, основанного на иерархической классификации типов атомов, эта цель была достигнута путем введения «выделенных» атомов, благодаря которым фрагментные дескрипторы удалось распространить на прогнозирование локальных свойств атомов в органических соединениях, кинетических констант органических реакций, физических свойств полимеров, а также на количественное прогнозирование биологической активности внутри рядов соединений. Кроме того, при помощи «выделенных» атомов можно преодолеть один из недостатков большинства фрагментных дескрипторов – игнорирование стереохимической информации.

Для преодоления другого недостатка фрагментных дескрипторов – проблемы «редких фрагментов» - нами разработаны «псевдофрагментные» дескрипторы, значения которых формируются путем комбинирования свойств атомов внутри фрагментов. Совместное использование фрагментных и псевдофрагментных дескрипторов обычно ведет к заметному повышению прогнозирующей способности построенных моделей за счет эффективной аппроксимации вкладов отсутствующих в обучающей выборке фрагментов. Кроме того, идея псевдофрагментных дескрипторов явилась отправной при разработке специальных архитектур нейронных сетей, позволяющих строить прямые корреляции между структурой химического соединения и его свойствами без предварительного вычисления каких-либо дескрипторов – нейронная сеть сама строит внутри себя наиболее оптимальные псевдофрагментные дескрипторы.

Дальнейшему повышению универсальности нейросетевым количественных моделей «структура-свойство» и повышению точности осуществляемого ими прогноза служат предложенные в данной работе «интегрированные» подходы: 1) концепция построения моделей «структура-условия-свойство»; 2) концепция построения моделей «структура-свойство» для многокомпонентных систем; 3) многоуровневый подход и многозадачное обучение как средства объединения различных моделей «структура-свойство» в единую сеть.

Диссертационная работа состоит из семи глав. Первые две главы, составляющие обзор литературы, посвящены математическому аппарату искусственных нейронных сетей и фрагментным дескрипторам. В третьей главе, составляющей начало обсуждения результатов, приводится математическое обоснование выбранного подхода, основанного на сочетании многослойных нейронных сетей и фрагментных дескрипторов. Следующие две главы посвящены, соответственно, разработкам нейросетевых и фрагментных подходов. Шестая глава посвящена сочетанию нейросетей с фрагментными дескрипторами, седьмая – вышеупомянутому интегрированному подходу. В последней восьмой главе диссертационной работы рассматриваются разработанные программные средства.

ГЛАВА 1. ИСКУССТВЕННЫЕ НЕЙРОННЫЕ СЕТИ

1.1. Введение

Первые исследования, посвященные применению нейронных сетей (или перцептронов) для решения химических задач, были осуществлены еще в начале 70-х годов в СССР [1, 2], но эти пионерные работы не были должным образом оценены и оказались практически забытыми. Лишь в конце 80-х годов возродился интерес химиков к подобному подходу, и он начал стремительно расти [3].

Нейронные сети (часто называемые искусственными нейронными сетями, вычислительными нейронными сетями или просто нейросетями) представляют собой упрощенную математическую модель обработки информации головным мозгом человека [4-9]. Однако большинство современных архитектур нейронных сетей не воспроизводят в точности биологическую модель мозга, скорее, они могут рассматриваться в рамках класса алгоритмов статистического анализа данных [10-24], объединенных под общим названием нейроинформатики. Кроме того, нейронные сети часто рассматривают как высоко-параллельные методы решения задач вычислительной математики в «нейросетевом базисе» (что составляет предмет особой области вычислительной математики – нейроматематики [25]), на базе которых работают основанные на пороговой логике высокопроизводительные высоко-параллельные вычислительные устройства – нейрокомпьютеры [26-29].

Благодаря своей способности обучаться и обобщать данные, нейросети начали успешно применяться в химии, особенно в тех случаях, когда неизвестен аналитический вид зависимости между структурой и свойствами соединений [30-40].

1.2. Основные принципы нейросетевого моделирования

1.2.1. Общая терминология

Все нейросетевые методы имеют в своей основе определенные идеи, отражающие те или иные аспекты обработки информации в человеческом мозгу. Искусственные нейронные сети (или просто нейросети) состоят из определенного количества «искусственных нейронов», являющихся упрощенной математической моделью биологических нейронов, и связей между ними, соответствующих контактам через синапсы между аксонами и дендритами биологических нейронов (см. Рис. 1). В процессе работы нейросети осуществляется преобразование сигналов (кодирующих обрабатываемые данные) внутри нейронов и их передача между соседними нейронами.

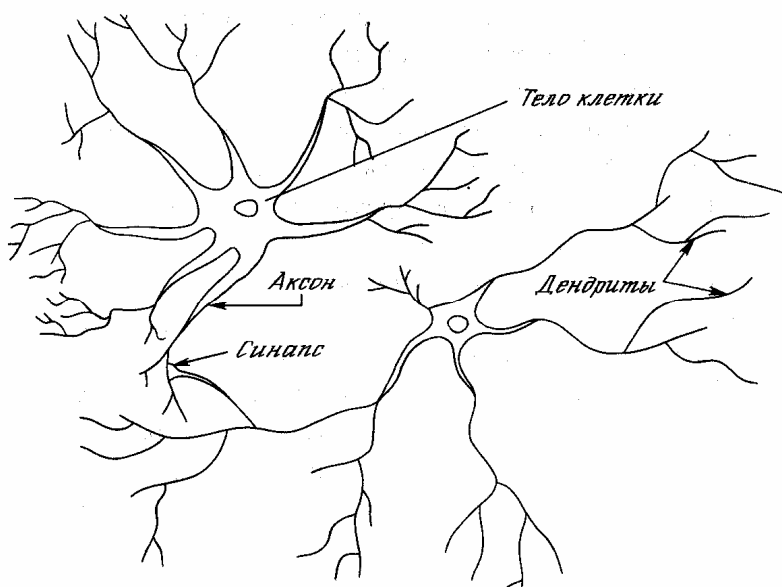


Рис. 1. Биологические нейроны

Архитектура нейронной сети определяется топологией соединений нейронов между собой. Нейроны внутри сети, как правило, организованы в группы, называемые слоями. Для всех нейронов, принадлежащих одному слою, характерно одинаковое число входных связей, соединяющих нейрон с предыдущим слоем или с внешними устройствами ввода и вывода данных. Нейроны, принимающие внешние данные для последующей обработки, называются

входными; нейроны, выводящие уже обработанные данные, называются выходными. Остальные же нейроны, участвующие в промежуточной обработке данных, называются скрытыми. В соответствии с типом нейронов, их слои также называются входными, выходными либо скрытыми.

1.2.2. Нейрон МакКаллока-Питтса

Впервые математическая модель искусственного нейрона была предложена в 1943 г. У.С.Мак-Каллоком и В.Питтсом [4]. Подобно тому, как биологические нейроны, вследствие наступающей под действием нейромедиаторов деполяризации мембраны, способны возбуждаться и проявлять спайковую активность, так и их искусственные аналоги (т.н. нейроны Мак-Каллока-Питтса) характеризуются определенным уровнем активности (обычно в интервале от 0, соответствующего нейрону в состоянии покоя, до 1, что соответствует возбужденному нейрону). Этот уровень активности передается в виде сигнала на соседние искусственные нейроны, что имитирует биологический процесс распространения деполяризации мембраны по аксону, выделения молекул нейромедиатора, их диффузии через синаптические щели и воздействия на рецепторы, расположенные на мембранах дендритов соседних нейронов. Весь этот сложный процесс передачи сигнала от одного нейрона к другому описывается в методологии искусственных нейронных сетей одним числом, называемым «весом связи», которое является аналогом понятия синаптической проводимости биологических нейронов. Обычно считается, что степень воздействия искусственного нейрона j на другой нейрон i равна произведению уровня активности первого нейрона o_j на вес связи (синаптическую проводимость) ω_{ji} между ними. Положительное значение синаптической проводимости соответствует прохождению через синаптические контакты возбуждающих нейромедиаторов, например, глутамата или ацетилхолина, а отрицательное – тормозящих, например, гамма-аминомасляной кислоты. В то же время абсолютная величина этого числа отражает легкость передачи сигнала, что в случае биологических нейронов определяется количеством и разветвленностью синаптических контактов, уровнем экспрессии и активности постсинаптических рецепторов, легкостью выде-

ления нейромедиаторов и многими другими факторами, управляемыми как генетически, так и при помощи разнообразных сигнальных систем.

В рамках методологии искусственных нейронных сетей функционирование отдельного нейрона обычно описывается уравнением (см. Рис. 2):

$$o_i = f(a_i), \quad a_i = \sum_j o_j w_{ji} - t_i \quad (1)$$

где: a_i – общий сетевой вход нейрона i ; o_j – выходной сигнал нейрона j ; w_{ji} – вес связи (синаптическая проводимость) между нейронами j и i ; t_i – порог активации нейрона i (превышение этого порога суммой воздействий со стороны соседних нейронов приводит его в возбужденное состояние); o_i – результирующий выходной сигнал, равный уровню активности данного нейрона i ; $f(x)$ – т.н. функция активации нейрона (или передаточная функция), которая в простейшем случае, к примеру, может быть определена как пороговая:

$$f(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0 \end{cases} \quad (2)$$

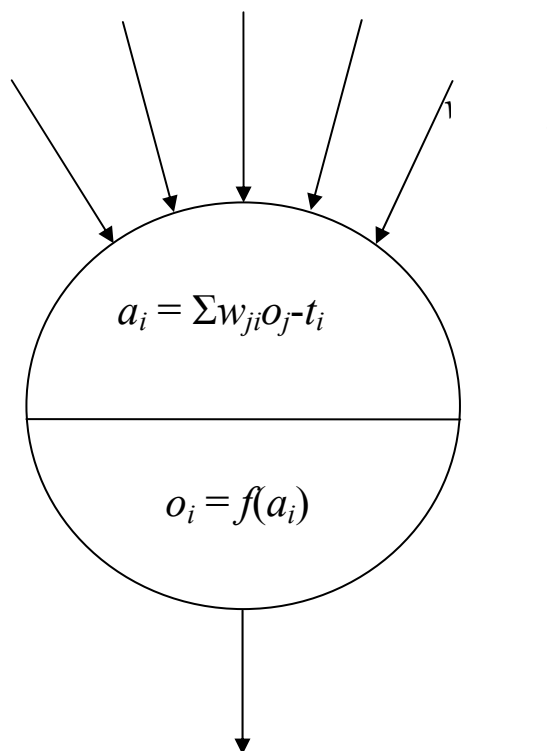


Рис. 2. Нейрон МакКаллока-Питтса

Таким образом, уравнение (1) в сочетании с определением функции (2) упрощенно описывает функционирование биологического нейрона, находящегося, в частности, в коре головного мозга человека.

Подобно своему биологическому прототипу, нейроны МакКаллока-Питтса способны обучаться путем настройки параметров w , описывающих синаптическую проводимость.

Как правило, вместо использования пороговых величин t_i в нейросеть добавляют так называемые «псевдонейроны смещения» (bias pseudoneurons) с постоянным выходным сигналом, равным 1.

1.2.3. Персептрон Розенблатта

На приведенном выше описании искусственного нейрона были основаны разработанные более 40 лет назад первые типы искусственных нейронных сетей, получивших название «персептроны» [5-7] (в русскоязычной литературе пишутся иногда как «перцептроны»), а вместе с ними и первые попытки создать искусственный интеллект путем имитации работы головного мозга человека на клеточном уровне. Название «персептрон» происходит от английского слова perception – восприятие. Оно было предложено в 1958 г. Фрэнком Розенблаттом в попытках имитировать с помощью нейронов МакКаллока-Питтса человеческое восприятие (прежде всего зрение) и распознавание с его помощью объектов внешнего мира. Персептрон Розенблатта имел многослойную архитектуру (см. Рис. 3), причем только последний (выходной) содержал нейроны с настраиваемыми весами, а формируемые ими выходные сигналы свидетельствовали о принадлежности анализируемого объекта к определенному классу. Само описание объекта в персептронах Розенблатта формировалось на входном слое нейронов, названном рецепторным полем по аналогии с биологическим прототипом. Сигналы с рецепторного поля поступали на необязательный скрытый слой нейронов по связям, веса которых инициировались случайными числами и в процессе обучения не менялись, а сформированные на нейронах скры-

того слоя сигналы уже, в свою очередь, поступали на выходной слой нейронов для дальнейшей обработки (см. Рис. 3).

Эти попытки имитации человеческого восприятия на нейронах МакКаллока-Питтса, однако, оказались не совсем удачными, поскольку они не оправдали всех возлагавшихся на них надежд [7]. Поскольку в то время был известен способ настройки весов связей, идущих лишь к нейронам одного (выходного) слоя, то на практике персептроны Розенблатта оказались неспособными обучаться распознаванию сложных образов, и их реальная распознающая способность оказалась не выше, чем у более простых и понятных стандартных методов дискриминатного анализа. Все это привело к разочарованию и, как следствие, прекращению практически всех проводившихся работ в области искусственных нейронных сетей.

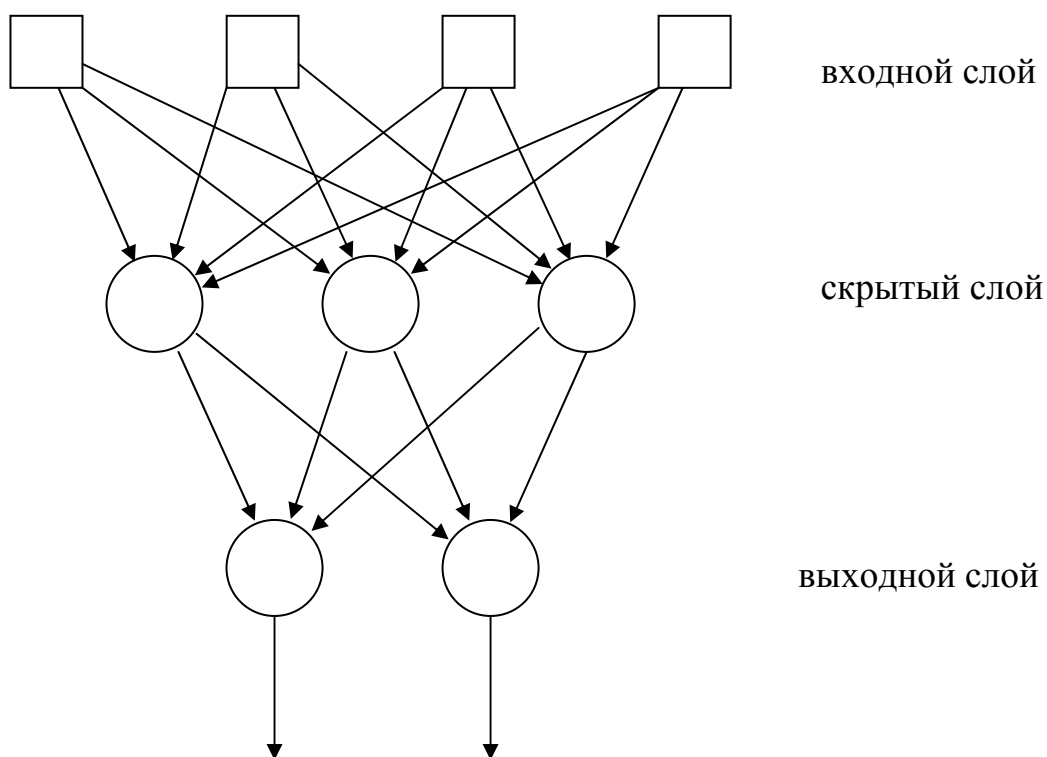


Рис. 3. Многослойный персептрон Розенблатта. Преобразования сигналов производится по формулам (1) и (2) на скрытых и выходных нейронах, изображенных кружками, тогда как изображенные квадратами входные псевдонейроны служат исключительно для ввода данных.

1.2.4. Нейросети обратного распространения (backpropagation)

1.2.4.1. Общая характеристика

К середине 80-ых годов стало ясно, что одна из причин неудач кроется в конкретном виде пороговой функции активации (2). Оказалось, что замена пороговой функции (2) на непрерывную, ограниченную и монотонно-возрастающую, например, сигмоидную функцию (3), способна привести к построению многослойных персептронов, все веса связей которых способны эффективно обучаться при помощи алгоритма обратного распространения ошибок (error backpropagation) [41, 42]. Именно благодаря открытию (точнее, переоткрытию) этого алгоритма, с конца 80-ых годов начался этап активного развития и использования аппарата искусственных нейронных сетей в разных областях науки и техники (см. книги и учебные пособия [10-24]), а с начала 90-ых – в различных областях химии (см. [30-34]) и, в частности, в области исследования зависимости структура-свойство для органических соединений [35-39].

$$f(x) = \frac{1}{1 + e^{-x}} \quad (3)$$

Кроме чисто математических причин, переход к подобным непрерывным дифференцируемым функциям имеет и определенное нейрофизиологическое обоснование. С точки зрения способа передачи информации, сигнал реальных биологических нейронов модулирован не по амплитуде, а по частоте, и, к тому же, является стохастическим, что вполне согласуется с уравнениями (1) и (3) при условии, что уровень сигнала (активации) o_i показывает, с какой вероятностью нейрон i переходит в возбужденное состояние.

Алгоритм обратного распространения ошибки (см. ниже) сыграл настолько важную роль в истории становления многослойных персептронов, что сами нейросети этого типа часто стали называть нейросетями с обратным распространением (backpropagation neural networks).

К основным достоинствам таких нейросетей можно отнести их способность находить нелинейные и многопараметрические линейные зависимости, характеризующиеся высокой точностью интерполяции, даже в тех случаях, ко-

гда экспериментальные данные сильно зашумлены. Для многослойных персептронов характерна послойная передача сигнала, от входа нейросети к ее выходу. В то же время при обучении нейросетей этого типа настройка весовых коэффициентов связей проводится последовательно, начиная со связей выходного слоя, поэтому методы обучения таких нейросетей носят название методов обратного распространения ошибки [41, 42].

1.2.4.2. Функционал ошибки нейросети

Суть обучения нейросети заключается в минимизации функционала ошибки для выборки $E(w)$ в пространстве ее настроечных параметров, каковыми являются веса связей (пороги нейрона здесь тоже рассматриваются как веса связей, ведущих от псевдонейронов смещения с постоянным значением выхода, равным единице, к этому нейрону):

$$E(w) = \sum_{p=1}^P v^p E^p(w), \quad (4)$$

где: v^p – вес p -ого объекта (например, химического соединения) из обучающей выборки; P – количество объектов в обучающей выборке; $E^p(w)$ – индивидуальный функционал ошибки для p -ого объекта из обучающей выборки, который обычно (но не всегда!) представляют как взвешенную сумму значений функции потерь $l(\cdot, \cdot)$ для каждого из выходных нейронов (т.е. для каждого из одновременно прогнозируемых свойств в случае QSAR/QSPR-анализа):

$$E^p(w) = \sum_{k=1}^K v_k l(d_k^p, o_k^p), \quad (5)$$

где: v_k – вес k -ого выходного нейрона; K – количество выходных нейронов (равное числу одновременно прогнозируемых свойств химических соединений в случае QSAR/QSPR-анализа). В большинстве случаев (но не всегда!) используется квадратичная функция потерь, что превращает нейронную сеть в вариант метода наименьших квадратов:

$$l(d, o) = \frac{1}{2} (d^2 - o^2). \quad (6)$$

Значения весов объектов v^p , отличные от единицы, берутся, главным образом, тогда, когда нейросеть обучается классифицировать объекты для придания большего веса тем из них, которые принадлежат к классам с меньшим числом представителей. В остальных же случаях (т.е. практически всегда) веса объектов считаются одинаковыми и равны единице. Аналогично, значения весов выходных нейронов v_k , отличные от единицы, берутся лишь в редких случаях многозадачного обучения, в остальных же случаях они принимаются равными единице. С учетом вышесказанного, индивидуальный функционал ошибки для p -ого объекта из обучающей выборки обычно имеет вид:

$$E^p(w) = \frac{1}{2} \sum_{k=1}^K (d_k^p - o_k^{p[N]})^2 \quad (7)$$

где: d_k^p - желаемый выход для k -ого выходного нейрона p -ого объекта (экспериментальное значение k -ого свойства для p -ого соединения) из обучающей выборки; $o_k^{p[N]}$ - вычисленный выход для k -ого выходного нейрона p -ого объекта (спрогнозированное значение k -ого свойства для p -ого соединения) из обучающей выборки; N – номер выходного слоя; K – число выходов нейросети, равное числу одновременно прогнозируемых свойств химических соединений в случае QSAR/QSPR-анализа. Функционал ошибки для всей выборки в этом случае имеет вид:

$$E(w) = \frac{1}{2} \sum_{p=1}^P \sum_{k=1}^K (d_k^p - o_k^{p[N]})^2 . \quad (8)$$

1.2.4.3. Вычисление производных функционала ошибки по методу обратного распространения

Для эффективной минимизации функционала необходимо уметь быстро вычислять его градиент, т.е. вектор первых производных по отношению ко всем настраиваемым параметрам. В случае индивидуального функционала ошибки для p -ого соединения из обучающей выборки элементы искомого вектора градиента можно выразить в следующем виде:

$$\frac{\partial E^p(w)}{\partial w_{ji}^{[n]}} = \frac{\partial E^p(w)}{\partial a_i^{p[n]}} \cdot \frac{\partial a_i^{p[n]}}{\partial w_{ji}^{[n]}} = \frac{\partial E^p(w)}{\partial a_i^{p[n]}} x_j^{p[n]} = \frac{\partial E^p(w)}{\partial a_i^{p[n]}} o_j^{p[n-1]} \equiv -\delta_i^{p[n]} o_j^{p[n-1]} \quad (9)$$

где: величина $-\delta_i^{p[n]}$, называемая иногда невязкой нейрона, обозначает частную производную функционала ошибки для p -ого объекта из обучающей выборки по отношению к сетевому входу нейрона i , находящегося в слое n (знак минуса взят для совместимости с принятыми в литературе обозначениями); $o_j^{p[n-1]}$ - выходной сигнал находящегося в слое $n-1$ нейрона j для p -ого объекта из обучающей выборки. Таким образом, частная производная функционала ошибки нейросети по отношению к весу связи равна произведению выхода находящегося в предыдущем слое нейрона, из которого выходит данная связь, на невязку нейрона следующего слоя, в который входит данная связь.

Из вышеизложенного следует, что для вычисления градиента ошибки необходимо рассчитать значения выходов и невязок всех нейронов. Поскольку нейросеть обратного распространения (многослойный персептрон) устроен таким образом, что каждый нейрон (кроме входных псевдонейронов и псевдонейронов смещения) получает сигнал из нейронов предыдущего слоя, то вычисление выходов нейронов производится по формулам (1) и (3) последовательно при движении от входного к выходному слою. Подобную последовательность вычислений называют прямым распространением сигнала. В противоположность этому, расчет невязок нейронов производится в обратном направлении при движении от выходного слоя к входному (обратное распространение ошибки).

Действительно, для нейронов выходного слоя, дифференцируя выражение (7), имеем:

$$\delta_i^{p[N]} = -\frac{\partial E^p(w)}{\partial a_i^{p[N]}} = -\frac{\partial}{\partial a_i^{p[N]}} \frac{1}{2} \sum_{k=1}^K (d_k^p - o_k^{p[N]})^2 = \frac{\partial o_i^{p[N]}}{\partial a_i^{p[N]}} (d_i^p - o_i^{p[N]}) \equiv f'(a_i^{p[N]}) (d_i^p - o_i^{p[N]}), \quad (10)$$

Для остальных нейронов, применяя цепное правило дифференцирования к формуле (1) и опуская некоторые тривиальные промежуточные преобразования, получаем:

$$\delta_i^{p[n]} = -\frac{\partial E^p(w)}{\partial a_i^{p[n]}} = -\sum_j \frac{\partial E^p(w)}{\partial a_j^{p[n+1]}} \cdot \frac{\partial a_j^{p[n+1]}}{\partial x_j^{p[n+1]}} \cdot \frac{\partial x_j^{p[n+1]}}{\partial a_i^{p[n]}} = f'(a_i^{p[n]}) \sum_j w_{ij}^{[n+1]} \delta_j^{p[n+1]} \quad (11)$$

Таким образом, значения невязок нейронов каждого скрытого слоя рассчитываются исходя из значений невязок нейронов последующего слоя, что можно условно описать процессом распространения ошибки в направлении, обратном распространению сигнала. Для сигмовидной передаточной функции (3) производная вычисляется по следующей формуле:

$$f'(a) = \frac{d}{da} \left(\frac{1}{1+e^{-a}} \right) = \frac{e^{-a}}{(1+e^{-a})^2} = o(1-o) \quad (12)$$

Производные суммарного функционала ошибки для всей обучающей выборки могут быть получены суммированием производных индивидуальных функционалов ошибки:

$$\frac{\partial E(w)}{\partial w_{ji}^{[n]}} = \sum_{p=1}^P \frac{\partial E^p(w)}{\partial w_{ji}^{[n]}}. \quad (13)$$

Формулы (9-13) составляют суть метода обратного распространения, который можно рассматривать как очень эффективный алгоритм расчета градиента функционала ошибки нейросети в пространстве весов связей (поскольку суммарное время вычисления всех производных, число которых может быть очень велико, не превышает времени расчета самого функционала).

1.2.4.4. Градиентные методы обучения

Исторически первым методом обучения сетей обратного распространения явился метод Уидроу-Хоффа, называемый чаще дельта-правилом [43], который традиционно записывается в виде:

$$\Delta w_{ji} = w_{ji}^{(t+1)} - w_{ji}^{(t)} = \eta \delta_i o_j, \quad (14)$$

где: $w_{ji}^{(t)}$ - текущий вес на t -ом шагу обучения связи, идущей от нейрона j к нейрону i ; δ_i - невязка i -ого нейрона, получаемая по методу обратного распространения (см. выше); o_j – выходное значение j -ого нейрона; η – параметр скорости обучения. Типичное значение параметра скорости обучения – 0.25, но оно может меняться в широких пределах, особенно в сторону уменьшения на окончательных этапах обучения.

Все весовые коэффициенты связей перед началом обучения инициализируются небольшими случайными числами. Правильный выбор границ инициализации, обеспечивающий удаленные от нуля значения производной передаточной функции нейронов (в противном случае происходит т.н. «паралич» нейронов), может сократить время обучения нейросети и улучшить качество получаемых нейросетевых моделей [44, 45]. На каждой итерации обучения производится корректировка значений весов по формуле (14) после предъявления очередного примера из обучающей выборки. Такой режим обучения называют последовательной адаптацией (online mode), в противоположность режиму группового обучения (batch mode), когда корректировка значений весов происходит после предъявления всей обучающей выборки. В классическом варианте обучение проводится до тех пор, пока не будет выполнено одно из возможных условий остановки обучения (например, когда значение функционала ошибки не опустится ниже заранее заданного порога, либо когда число итераций не превысит определенный лимит).

Хотя исторически дельта-правило возникло как обобщение алгоритма обучения персептрона Розенблатта на непрерывные входы и выходы и первоначально никак не было связано с представлениями о функционале ошибки нейросети, тем не менее оно оказалось математически эквивалентным применению метода скорейшего спуска к минимизации функционала ошибки нейросети в пространстве весов связей. Действительно, при подстановке формулы (9) в (14) получаем:

$$\Delta w_{ji}^{(t)} = w_{ji}^{(t+1)} - w_{ji}^{(t)} = -\eta \frac{\partial E^p(w)}{\partial w_{ji}} \quad (15)$$

Формула (15) определяет шаг, который делается в направлении, противоположном градиенту, и поэтому дельта-правило представляет собой метод минимизации функционала ошибки в пространстве весов связей при помощи простейшего варианта градиентного метода скорейшего спуска с фиксированным значением параметра скорости обучения.

В своем первоначальном виде дельта-правило представляет в настоящее время главным образом историческую ценность, поскольку именно с его изло-

жения в статье Румельхарта [42] начался современный этап развития всей методологии искусственных нейронных сетей. Между тем, будучи простейшим градиентным методом оптимизации нелинейных функций, дельта-правило обладает целым рядом серьезных недостатков. Во-первых, теория нелинейной оптимизации гарантирует возможность достижения локального минимума за конечное число шагов лишь при постепенном уменьшении параметра скорости по мере обучения, тогда как при фиксированном его значении алгоритм может зациклиться в окрестностях узкого минимума. Во-вторых, в тех случаях, когда производные по различным весам сильно различаются (а именно так обычно и бывает в нейросетях), рельеф функционала ошибки представляет собой узкий овраг, попав в который градиентные методы вместо движения по его дну начинают осциллировать по его стенкам (поскольку практически во всех точках кроме очень узкой области у самого дна оврага градиент направлен почти перпендикулярно направлению движения к минимуму), что часто приводит к чрезвычайному замедлению и даже к практической остановке процесса обучения (см. Рис. 4). В-третьих, градиентные методы оптимизации часто застревают в мелких локальных минимумах.

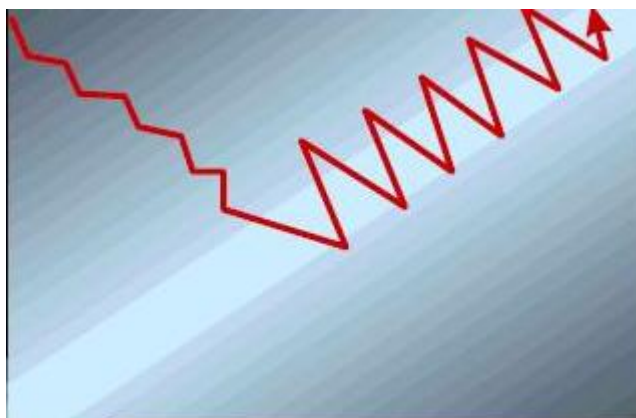


Рис. 4. Неэффективность метода скорейшего спуска: градиент направлен почти перпендикулярно необходимому направлению движения к минимуму

Осознание вышеприведенных проблем очень скоро привело к модификации метода и созданию расширенного варианта дельта-правила, в котором частично устранено или, по крайней мере, ослаблено влияние всех трех вышеперечисленных типов недостатков. Достигнуто это путем введения момента инер-

ции, приводящего, по мере обучения, к накоплению влияния градиента на изменение весов:

$$\Delta w_{ji}^{(t)} = -\eta \frac{\partial E^p(w)}{\partial w_{ji}} + \mu \Delta w_{ji}^{(t-1)}, \quad (16)$$

где: μ - параметр момента инерции. Типичное значение этого параметра 0.9, и оно не меняется по ходу обучения. Хотя формально допустимы любые значения в интервале $0 \leq \mu < 1$, наибольший эффект достигается при его значении, близкой к 1.

Объяснить влияние момента инерции на процесс обучения можно следующим образом [18]. Допустим, мы находимся вдалеке от дна оврага в области, где градиент меняется плавно, и потому на протяжении некоторого времени его изменением можно пренебречь. В этом случае изменения весов могут быть аппроксимированы как:

$$\Delta w_{ji}^{(t)} = -\frac{\partial E^p(w)}{\partial w_{ji}} (1 + \mu + \mu^2 + \dots) \approx -\frac{\eta}{1 - \mu} \cdot \frac{\partial E^p(w)}{\partial w_{ji}} \equiv -\eta' \frac{\partial E^p(w)}{\partial w_{ji}} \quad (17)$$

Из формулы (17) следует, что вдали от минимума и дна оврага эффективная скорость обучения η' довольно высока (поскольку параметр μ близок к единице). С другой стороны, вблизи минимума либо дна оврага, когда направление градиента из-за упомянутых выше осцилляций постоянно меняет направление, изменения весов могут быть аппроксимированы как:

$$\Delta w_{ji}^{(t)} = -\frac{\partial E^p(w)}{\partial w_{ji}} (1 - \mu + \mu^2 - \dots) \approx -\frac{\eta}{1 + \mu} \cdot \frac{\partial E^p(w)}{\partial w_{ji}} \equiv -\eta'' \frac{\partial E^p(w)}{\partial w_{ji}} \quad (18)$$

В этом случае эффективная скорость обучения η'' значительно понижается, что и дает возможность двигаться по дну оврага и подходить близко к минимумам. Таким образом, благодаря введению инерции в процесс обучения, появляется возможность адаптивно менять эффективную скорость обучения (см. Рис. 5).

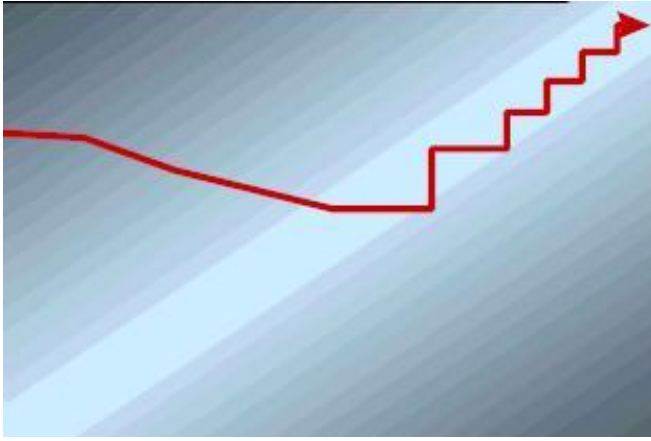


Рис. 5. Введение момента инерции позволяет благодаря появившейся возможности адаптивно менять эффективную скорость обучения значительно быстрее продвигаться к минимуму

Благодаря введению параметра μ , у нейросети появляется также способность преодолевать мелкие локальные минимумы на гиперповерхности функционала ошибки в пространстве весов. Чтобы понять причину этого, запишем разностное уравнение (15) в виде дифференциального:

$$\dot{w}_{ji} = -\eta \frac{\partial E(w)}{\partial w_{ji}} \quad (19)$$

Уравнение (19), описывающее обучение нейросети по дельта-правилу, математически эквивалентно дифференциальному уравнению движения неинерционного тела в вязкой среде. Введение момента соответствует появлению у такого тела инерции (т.е. массы μ), и процесс обучения при помощи расширенного дельта-правила уже описывается дифференциальным уравнением движения инерционного тела в вязкой среде:

$$\mu \ddot{w}_{ji} + (1 - \mu) \dot{w}_{ji} = -\eta \frac{\partial E(w)}{\partial w_{ji}} \quad (20)$$

Таким образом, гипотетическое тело, уравнение движения которого описывается уравнением (20), может, разогнавшись, преодолевать по инерции небольшие локальные минимумы, застревая лишь в относительно глубоких минимумах функционала ошибки, соответствующих статистически значимым нейросетевым моделям.

Тем не менее, не смотря на все успехи, достигнутые при помощи расширенного дельта-правила с включенным параметром момента, данный метод все равно не лишен недостатков. Прежде всего, в методе присутствуют две «магические» константы, обоснованный выбор точных значений которых сделать не-

возможно. Кроме того, эффективная скорость обучения, хотя и существенно выше, чем у первоначального дельта-правила, но все равно существенно уступает более совершенным методам обучения, рассмотренным ниже. Последнее обстоятельство отчасти связано еще и с тем, что методы обучения, основанные на последовательной адаптации, менее эффективны по сравнению с методами группового обучения, каковыми являются все рассматриваемые ниже алгоритмы.

1.2.4.5. Метод эластичного распространения (RPROP)

При обучении по методу эластичного распространения для настройки весовых коэффициентов используется только информация о знаках частных производных функции ошибки нейросети [46, 47]. Тем самым, методу RPROP удастся избежать замедление темпа обучения на плоских «равнинах» ландшафта функции ошибки, что характерно для схем, где изменения весов пропорциональны величине градиента. Величина, на которую изменяются весовые коэффициенты, вычисляется следующим образом:

$$\Delta w_{ij}^{(t)} = \begin{cases} -\Delta_{ij}^{(t)}, & \text{если } \frac{\partial E^{(t)}}{\partial w_{ij}} > 0 \\ +\Delta_{ij}^{(t)}, & \text{если } \frac{\partial E^{(t)}}{\partial w_{ij}} < 0 \\ 0, & \text{иначе} \end{cases}, \quad (21)$$

где

$$\Delta_{ij}^{(t)} = \begin{cases} \eta^+ * \Delta_{ij}^{(t-1)}, & \text{если } \frac{\partial E^{(t-1)}}{\partial w_{ij}} * \frac{\partial E^{(t)}}{\partial w_{ij}} > 0 \\ \eta^- * \Delta_{ij}^{(t-1)}, & \text{если } \frac{\partial E^{(t-1)}}{\partial w_{ij}} * \frac{\partial E^{(t)}}{\partial w_{ij}} < 0 \\ \Delta_{ij}^{(t-1)}, & \text{иначе} \end{cases}, \quad (22)$$

где: $\frac{\partial E}{\partial w_{ij}}$ - величина, характеризующая суммарный градиент для всех входных векторов из обучающей выборки; η^- и η^+ - факторы уменьшения и увеличения скорости обучения; t – счетчик итераций; $\Delta_{ij}^{(t)}$ - величина индивидуального

адаптивно настраиваемого темпа обучения на t -ой итерации для связи, соединяющей нейрон i с нейроном j .

Если знак производной по данному весу изменил направление, то это означает, что величина шага по данной координате слишком велика, и поэтому алгоритм уменьшает ее в η^- раз. В противном же случае шаг увеличивается в η^+ раз для ускорения обучения вдали от минимума.

1.2.4.6. Методы сопряженных градиентов

Методы сопряженных градиентов [48, 49], так же как и в случае расширенного дельта-правила, осуществляют на каждом шаге обучения движение в направлении, получаемом путем комбинирования направления антиградиента и направления движения на предыдущем шаге. Принципиальное же отличие методов сопряженных градиентов от последнего заключается в том, что размер шага в выбранном направлении не является фиксированным, а определяется на каждой итерации при помощи процедуры одномерного поиска минимума вдоль выбранного направления.

Все алгоритмы методов сопряженных градиентов на первой итерации начинают поиск в направлении антиградиента:

$$p^{(1)} = -g^{(1)} = -\nabla E(w) = - \begin{pmatrix} \frac{\partial E(w)}{\partial w_{j(1)i(1)}} \\ \vdots \\ \frac{\partial E(w)}{\partial w_{j(M)i(M)}} \end{pmatrix} \quad (23)$$

где: $p^{(t)}$ – вектор направления, вдоль которой ведется поиск на t -ой итерации; $g^{(t)}$ – вектор градиента функционала ошибки нейросети в пространстве весов связей на t -ой итерации; $j(m)$ – номер нейрона, из которого выходит связь m ; $i(m)$ – номер нейрона, в который входит связь m ; M – число связей в нейросети.

После выбора направления определяется оптимальный шаг поиска $\alpha^{(t)}$, на величину которого меняются все веса связей по формуле:

$$w_{j(m)i(m)}^{(t+1)} = w_{j(m)i(m)}^{(t)} + \alpha^{(t)} p_m^{(t)}, \quad (24)$$

где: $p_m^{(t)}$ - это m -ая компонента вектора направления $p^{(t)}$, соответствующая связи m . Оптимальное значение $\alpha^{(t)}$ определяется путем минимизации функционала ошибки вдоль направления $p^{(t)}$ при помощи одного из алгоритмов одномерного поиска.

Из большого арсенала алгоритмов одномерного поиска наилучшим образом себя зарекомендовал при обучении нейронных сетей при помощи методов сопряженных градиентов алгоритм Чараламбуca (Charalambous) [50], который использует кубическую интерполяцию в сочетании с методом деления интервала на части.

После оптимального шага, сделанного в выбранном направлении, методы сопряженных градиентов определяют следующее направление поиска как линейную комбинацию нового направления антиградиента и предыдущего направления движения:

$$p^{(t)} = -g^{(t)} + \beta^{(t)} p^{(t-1)} \quad (25)$$

Различные методы сопряженных градиентов различаются выбором коэффициента $\beta^{(t)}$. Так, в методе Флетчера-Ривса (Fletcher-Reeves) [51] он равен отношению квадрата нормы градиента к квадрату нормы градиента на предыдущей итерации:

$$\beta^{(t)} = \frac{(g^{(t)})^T g^{(t)}}{(g^{(t-1)})^T g^{(t-1)}} \quad (26)$$

В методе Полака-Рибьеры (Polak-Ribière) [49] искомый коэффициент равен скалярному произведению приращения градиента на текущий градиент, деленный на квадрат нормы градиента на предыдущей итерации:

$$\beta^{(t)} = \frac{(\Delta g^{(t)})^T g^{(t)}}{(g^{(t-1)})^T g^{(t-1)}} \quad (27)$$

Алгоритмы методов сопряженных градиентов требуют не многим больше памяти, чем градиентные алгоритмы, поэтому могут быть использованы для обучения нейронных сетей с большим количеством настраиваемых параметров.

1.2.4.7. Квазиньютоновские методы обучения

Эта группа методов базируется на Ньютоновском методе аппроксимации функций, но не требует вычисления вторых производных.

$$X_{t+1} = X_t - H_t^{-1} g_t, \quad (28)$$

где: X – матрица весовых коэффициентов; g – вектор градиента; t – счетчик итераций; H – матрица вторых частных производных (матрица Гессе).

$$H = \begin{pmatrix} \frac{\partial^2 E(w_{i(1)j(1)}, \dots, w_{i(M)j(M)})}{\partial w_{i(1)j(1)}^2} & \dots & \frac{\partial^2 E(w_{i(1)j(1)}, \dots, w_{i(M)j(M)})}{\partial w_{i(1)j(1)} \partial w_{i(M)j(M)}} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 E(w_{i(1)j(1)}, \dots, w_{i(M)j(M)})}{\partial w_{i(M)j(M)} \partial w_{i(1)j(1)}} & \dots & \frac{\partial^2 E(w_{i(1)j(1)}, \dots, w_{i(M)j(M)})}{\partial w_{i(M)j(M)}^2} \end{pmatrix}, \quad (29)$$

где: функция $i(m)$ показывает номер нейрона, из которого исходит связь m ; $j(m)$ показывает номер нейрона, в который входит связь m ; M – число связей (т.е. число настраиваемых параметров) в нейросети.

Идея квазиньютоновских методов базируется на возможности аппроксимации кривизны нелинейной оптимизируемой функции без явного формирования ее матрицы Гессе. Сама матрица при этом не хранится, а ее действие аппроксимируется скалярными произведениями специально подобранных векторов. Наиболее удачным методом из этой группы является метод Бroyдена-Флетчера-Гольдфарба-Шанно (BFGS) [52], согласно которому:

$$s_{t+1} = -g_{t+1} + \frac{(g_{t+1} - g_t)^T g_{t+1}}{(g_{t+1} - g_t)^T s_t} s_t, \quad (30)$$

где: s_t – направление, вдоль которого проводится одномерная оптимизация на t -ой итерации; g_{t+1} – вектор градиента на $t+1$ -ой итерации.

Для квазиньютоновских методов наилучшим алгоритмом поиска вдоль выбранного направления является, по-видимому, метод перебора с возвратами [49, 52]. На первой итерации этот алгоритм использует значения функционала ошибки и его производных, чтобы построить его квадратичную аппроксимацию вдоль направления поиска. Минимум этой аппроксимирующей функции выбирается в качестве приближения к оптимальной точке, в которой оценивается функционал ошибки. Если значение функционала недостаточно мало, то

строится кубическая интерполяция, и ее минимум выбирается в качестве новой оптимальной точки. Этот процесс продолжается до тех пор, пока не будет достигнуто существенное уменьшение функционала ошибки.

1.2.4.8. Метод Левенберга-Марквардта

Метод Левенберга-Марквардта (LM) (Levenberg-Marquardt) [53] реализует специальный способ аппроксимации матрицы Гессе для случая, когда функционал ошибки определяется как сумма квадратов ошибок, что как раз и имеет место при обучении нейросетей обратного распространения. В рамках данного метода матрица Гессе H аппроксимируется как

$$H \cong J^T J, \quad (31)$$

а вектор градиента g может быть рассчитан по формуле

$$g = J^T e, \quad (32)$$

где: J – матрица Якоби производных функционалов ошибки отдельно для каждого выходного нейрона (т.е. для каждого свойства) и для каждого объекта (т.е. химического соединения) в обучающей выборке по настраиваемым параметрам (т.е. весам нейросети); e – вектор ошибок нейросети. Матрицу Якоби можно записать в следующем виде:

$$J = \begin{pmatrix} \frac{\partial e_{11}}{\partial w_{i(1)j(1)}} & \dots & \frac{\partial e_{11}}{\partial w_{i(M)j(M)}} \\ \vdots & \dots & \vdots \\ \frac{\partial e_{K1}}{\partial w_{i(1)j(1)}} & \dots & \frac{\partial e_{K1}}{\partial w_{i(M)j(M)}} \\ \vdots & \dots & \vdots \\ \frac{\partial e_{1P}}{\partial w_{i(1)j(1)}} & \dots & \frac{\partial e_{1P}}{\partial w_{i(M)j(M)}} \\ \vdots & \dots & \vdots \\ \frac{\partial e_{KP}}{\partial w_{i(1)j(1)}} & \dots & \frac{\partial e_{KP}}{\partial w_{i(M)j(M)}} \end{pmatrix} \quad (33)$$

где: функция $i(m)$ показывает номер нейрона, из которого исходит связь m ; $j(m)$ – номер нейрона, в который входит связь m ; M – число связей (т.е. число настраиваемых параметров) в нейросети; e_{kp} – ошибка прогноза для k -го выходного нейрона и p -го объекта из обучающей выборки; K – число выходных нейронов (равное числу одновременно прогнозируемых свойств); P – число объектов

(химических соединений) в обучающей выборке. Отсюда видно, что элементы матрицы Якоби легко могут быть вычислены на основе метода обратного распространения ошибки по приведенной выше формуле (33), что существенно проще вычисления матрицы Гессе.

Метод Левенберга-Марквардта реализует итерационную схему настройки весов нейросети по формуле:

$$w_{ij}^{k+1} = w_{ij}^k - (J^T J + \mu I)^{-1} J^T e^k \quad (34)$$

где, как и прежде, w_{ij}^k - вес связи (на k -ой итерации) исходящей из нейрона i и входящей в нейрон j ; J – матрица Якоби; I – единичная матрица (т.е. содержащая единицы на диагонали и нули вне ее); e^k – вектор ошибок нейросети на k -ой итерации; μ - динамически изменяемый по ходу обучения нейросети коэффициент, называемый фактором демпинга. Когда μ приближается к 0, то метод Левенберга-Марквардта переходит в метод Ньютона с приближением матрицы Гессе в форме (31), когда же значение μ велико, то получается метод градиентного спуска с маленьким шагом. Поскольку метод Ньютона имеет большую точность и скорость сходимости вблизи локального минимума по сравнению с методом градиентного спуска, то задача состоит в том, чтобы в процессе минимизации как можно быстрее перейти к методу Ньютона. С этой целью параметр μ уменьшают после каждой успешной итерации (т.е. приводящей к уменьшению функционала ошибки) и увеличивают лишь тогда, когда пробный шаг показывает, что функционал ошибки возрастает.

Метод Левенберга-Марквардта в настоящее время является одним из самых эффективных методов (по крайней мере, в смысле скорости) обучения нейронных сетей обратного распространения, в связи с чем он приобрел большую популярность в области QSAR/QSPR-исследований [54]. Тем не менее, у него есть существенный недостаток: необходимо, чтобы число объектов (химических соединений) в обучающей выборке превышало число настраиваемых параметров (т.е. межнейронных связей) нейросети. В связи с этим, в QSAR/QSPR-исследованиях его можно применять только при относительно не-

большом числе дескрипторов, относительно большом объеме обучающей выборки и при относительно небольшом числе скрытых нейронов.

1.2.5. Другие архитектуры нейронных сетей

1.2.5.1. Самоорганизующиеся карты Кохонена и другие конкурирующие нейросети

Нейросети Кохонена (Kohonen) [14, 20, 55] (см. Рис. 6), широко используемые для кластерного анализа, позволяют получить такое отображение исходных данных, при котором близкие вектора входных значений отображаются в топологически близкие выходные нейроны.

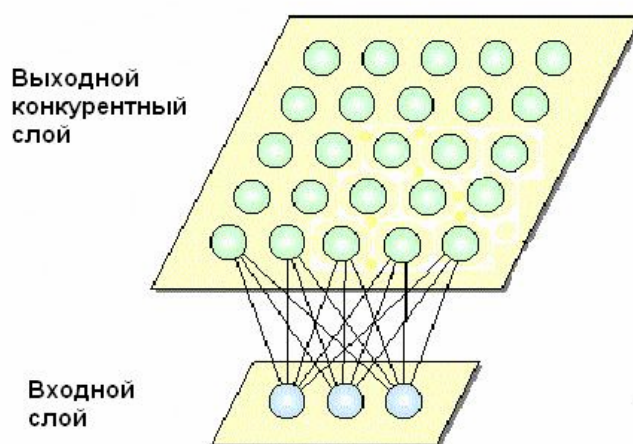


Рис. 6. Самоорганизующаяся карта Кохонена

Нейросеть Кохонена состоит из двух слоев – входного и конкурирующего выходного. В каждый нейрон конкурирующего выходного слоя поступают сигналы сразу со всех нейронов входного слоя. Одной из особенностей конкурирующего выходного слоя нейросетей Кохонена является то, что нейроны в нем организованы в виде 2-мерной решетки (решетки другой размерности для нейросетей Кохонена тоже возможны, но на практике используются редко). Чтобы избежать краевых эффектов и сделать все выходные нейроны равноправными,

на границах решетки часто вводятся периодические условия, что эквивалентно ее свертыванию в тор. Таким образом, выходной конкурирующий слой нейросетей Кохонена можно представить в виде 2-мерной решетки, натянутой на поверхность тора. Эта решетка может быть как тетрагональной (наиболее часто используемый вариант), так и гексагональной (классический вариант). Подобная организация выходного слоя имеет два последствия. Во-первых, вводится мера близости между нейронами в этом слое, называемая топологическим расстоянием, равная минимальному числу шагов, с помощью которых, двигаясь от одного узла решетки к ближайшему другому, можно перейти от одного из этих нейронов к другому. Во-вторых, вводится определенная «двухмерность» на множестве нейронов конкурирующего выходного слоя, что дает возможность отобразить их на плоскость, т.е. создать топологическую карту нейронов.

Еще одной особенностью нейронов конкурирующего выходного слоя является то, что они обладают весьма специфической функцией активации: только у одного нейрона, имеющего максимальный сетевой вход, уровень выходного сигнала равен единице (т.е. он активизируется), тогда как у других нейронов этого слоя он равен нулю. Таким образом, между нейронами происходит своеобразная конкуренция за право формирования единичного выходного сигнала, и лишь нейрон с наивысшим сетевым входом становится победителем. К его активации приводят лишь те вектора входных значений, которые ближе (более сходны) к вектору весов входящих связей данного нейрона по сравнению с аналогичными векторами других нейронов данного слоя. Это следует из того, что сетевой вход нейрона рассчитывается как скалярное произведение вектора входных значений (вектора дескрипторов в QSAR/QSPR-анализе) и вектора входящих в него весов (см. формулу 001), а скалярное произведение нормализованных векторов рассматривается как мера их близости. Для того, чтобы близкие вектора входных значений приводили к активации топологически близких выходных нейронов, нейросеть Кохонена необходимо этому обучить.

Алгоритм обучения нейросетей Кохонена можно представить как итерационную 4-шаговую процедуру:

1. Инициализировать все веса w_{ji} случайными числами. Нормировать все веса w_{ji} и входные сигналы x_j .
2. Для очередного входного вектора в конкурирующем выходном слое найти нейрон-победитель i^* с наименьшим до него расстоянием d_i :

$$d_i = \sqrt{\sum_{j=1}^M (x_j - w_{ji})^2}. \quad (35)$$

Заметим, что в случае нормированных w_{ji} и x_j минимизация d_i эквивалентна максимизации сетевого входа a_i :

$$a_i = \sum_{j=1}^M x_j w_{ji} \quad (36)$$

3. Для всех нейронов выходного конкурирующего слоя адаптировать веса связей, идущих к ним:

$$w_{ji}^{(t+1)} = w_{ji}^{(t)} + \alpha^{(t)} \cdot \gamma^{(t)} \cdot (x_j - w_{ji}), \quad (37)$$

где: $\alpha^{(t)}$ – параметр скорости обучения (в интервале от 0 до 1), который уменьшают по ходу обучения; $\gamma^{(t)}$ – функция соседства, которая кодирует понижение влияния нейрона при увеличении топологического расстояния до него:

$$\gamma^{(t)} = \exp\left(-\frac{r_{i^*i}}{(\sigma^{(t)})^2}\right), \quad (38)$$

где: r_{i^*i} – топологическое расстояние между нейроном-победителем i^* и текущим нейроном i ; $\sigma^{(t)}$ – радиус соседства, значение которого также уменьшают по мере обучения нейросети.

4. Пункты 1-3 повторять до тех пор, пока все вектора из обучающей выборки не будут предъявлены нейросети определенное число раз.

Таким образом, при обучении нейросети Кохонена происходит самоорганизация конкурирующего выходного слоя нейронов, в результате которой близкие входные вектора оказываются отображенными из исходного многомерного пространства в расположенную на плоскости (либо на поверхности тора) решетку нейронов таким образом, чтобы близким входным векторам соответствовали топологически близкие выходные нейроны. У обученной нейросети Кохонена веса связей, входящих в нейроны, практически совпадают с усреднен-

ными значениями соответствующих компонент входных векторов, приводящих к активации этих нейронов.

Перед началом обучения нейросети Кохонена веса связей инициализируются случайными числами, после чего они, а также входные вектора данных, нормируются. После подобной нормировки вектора весов связей и данных могут быть представлены векторами, идущими из центра координат на поверхность гиперсферы единичного радиуса, а процесс обучения нейросети может быть представлен как итерационный процесс вращения векторов весов связей по направлению к ближайшим векторам данных.

Следует отметить, что нейросети Кохонена не являются чисто математической конструкцией – они имеют очень солидный нейрофизиологический фундамент. Действительно, устройство некоторых отделов головного мозга очень напоминает строение и принцип функционирования указанных нейросетей. В качестве примеров можно привести: а) строение соматосенсорной коры головного мозга, в которой информация с сенсорных участков близких частей тела отображаются в топологически близкие нейроны; б) строение слуховой коры летучей мыши, в которой строится карта окружающих предметов за счет преобразования первичных данных ультразвуковой эхолокации.

Таким образом, нейросети Кохонена позволяют строить на плоскости карту, выявляющую топологическую структуру выборки в многомерном пространстве входных векторов. В связи с этим нейросети Кохонена часто называют самоорганизующимися картами (self-organizing maps - SOM). В том случае, когда число примеров в выборке значительно больше числа нейронов в сети Кохонена, и, следовательно, каждый из нейронов активируется по крайней мере несколькими примерами, то говорят о нейросети Кохонена низкого разрешения (low-resolution SOM). Если же число примеров в выборке сравнимо, либо даже меньше числа нейронов, то говорят о нейросети Кохонена высокого разрешения (high-resolution SOM). При наличии ассоциированного выходного свойства у примеров из выборки нейроны часто изображаются в виде ячеек, каждая из которых окрашена в цвет, кодирующий среднее значение этого свойства у всех примеров, приводящих к активации соответствующего нейрона. Получаемые

цветные карты представляют собой очень эффектный (а в эстетическом плане даже и красивый) способ визуализации и анализа данных.

К числу задач, решаемых при помощи нейросетей Кохонена, обычно относят следующие: визуализация, кластеризация и сжатие многомерных данных, а также аппроксимация плотностей вероятности и комбинаторная оптимизация.

Вышеупомянутое сжатие данных в нейросетях Кохонена происходит за счет понижения размерности данных до размерности решетки нейронов конкурирующего слоя, а так же за счет кодирования множества векторов, активирующих какой-либо нейрон, одним усредненным вектором, компоненты которого равны значениям весов связей, идущих к этому нейрону. Подобная операция кодирования множества векторов одним кодирующим вектором (codebook vector) называется квантованием векторов (vector quantization) [56, 57] и часто используется для аппроксимации плотности вероятности распределения векторов данных [58]. Поскольку алгоритмы обучения всех нейросетевых квантователей векторов неизменно включают стадию «конкурентной борьбы» между нейронами за право быть активированными текущим вектором входных сигналов, подобные нейронные сети часто называют конкурирующими.

Кроме рассмотренных выше нейросетей Кохонена, другими представителями этого же класса нейросетей, уже нашедшими применение при обработке химических данных, являются: нейронный газ [59-61], растущий нейронный газ [61, 62], а также целый набор обучающихся квантователей векторов (Learning Vector Quantizers - LVQ) [63, 64]: LVQ1, LVQ2, LVQ2.1, LVQ3. В нейронном газе, в отличие от нейросетей Кохонена, нейроны конкурирующего слоя не объединены в какую-либо решетку или другую графовую структуру, поэтому вместо топологического расстояния в функции соседства (38) используется обычное Эвклидово расстояние. Напротив, в растущем нейронном газе нейроны, как и в сетях Кохонена, уже объединены в решетку, однако, в отличие от сетей Кохонена, размерность решетки и число нейронов в ней не задается заранее, а определяется по ходу обучения путем постепенного наращивания нейросети. Обучающиеся квантователи векторов используют информацию о принадлежности векторов к определенным классам для того, чтобы вектора, активизи-

рующие один и тот же нейрон, относились по возможности к одному классу. Это достигается путем использования разного знака перед $\alpha^{(t)}$ в формуле (37) в зависимости от правильности или неправильности классификации текущего вектора. Поэтому нейросети последнего класса можно применять также для целей классификации.

Нейросети Кохонена могут использоваться непосредственно, а также как часть составных нейронных сетей, где они служат для предварительной обработки входных данных.

1.2.5.2. Нейросети встречного распространения (counterpropagation)

Нейросети встречного распространения (counterpropagation neural networks) [65] представляют собой пример составных нейронных сетей, включающих в свой состав самоорганизующуюся карту Кохонена (см. выше) и т.н. звезду Гроссберга [66]. В отличие от нейросетей Кохонена, они реализуют стратегию обучения «с учителем», и поэтому могут быть использованы как для классификации, так и для решения регрессионных задач.

Нейросети встречного распространения состоят из 3 слоев: входного, скрытого слоя Кохонена и выходного слоя Гроссберга (см. Рис. 7). В соответствии с особенностями архитектуры, обучение проводится в 2 этапа: сначала проводится обучение слоя Кохонена «без учителя» согласно рассмотренной выше стандартной схеме для этого класса сетей по формулам (35-38) с использованием только входных векторов, после чего идет настройка «с учителем» выходного слоя Гроссберга с использованием выходов нейронов Кохонена и векторов желаемых сигналов (т.е. экспериментальных значений прогнозируемых свойств в случае QSAR/QSPR-анализа) по формуле:

$$v_{ij}^{(t+1)} = v_{ij}^{(t)} + \beta(y_j - v_{ij})k_i, \quad (39)$$

где: $v_{ij}^{(t)}$ - вес связи, идущей из нейрона i в слое Кохонена на нейрон j в слое Гроссберга на t -ой итерации; k_i – выход i -ого нейрона Кохонена; y_j – желаемый выход для j -ого нейрона Гроссберга; β – параметр скорости обучения, который

первоначально берется равным ~ 0.1 , и затем постепенно уменьшается по ходу обучения.

Нейросети встречного распространения могут работать в режимах аккредитации и интерполяции. В наиболее часто используемом режиме аккредитации только для одного нейрона-победителя в слое Кохонена генерируется ненулевой выходной сигнал (как и должно быть в стандартном варианте нейросети Кохонена). В этом случае в результате обучения значение $v_{ij}^{(i)}$ устанавливается равным среднему значению желаемого выхода j по всем векторам, приводящим к активации нейрона Кохонена i , и поэтому для настройки слоя Гроссберга можно обойтись без итерационной процедуры по формуле (39). При решении регрессионной задачи нейросетями встречного распространения в режиме аккредитации происходит аппроксимация функциональной зависимости кусочными поверхностями постоянного уровня, что в случае небольших обучающих выборок приводит к слишком большим ошибкам.

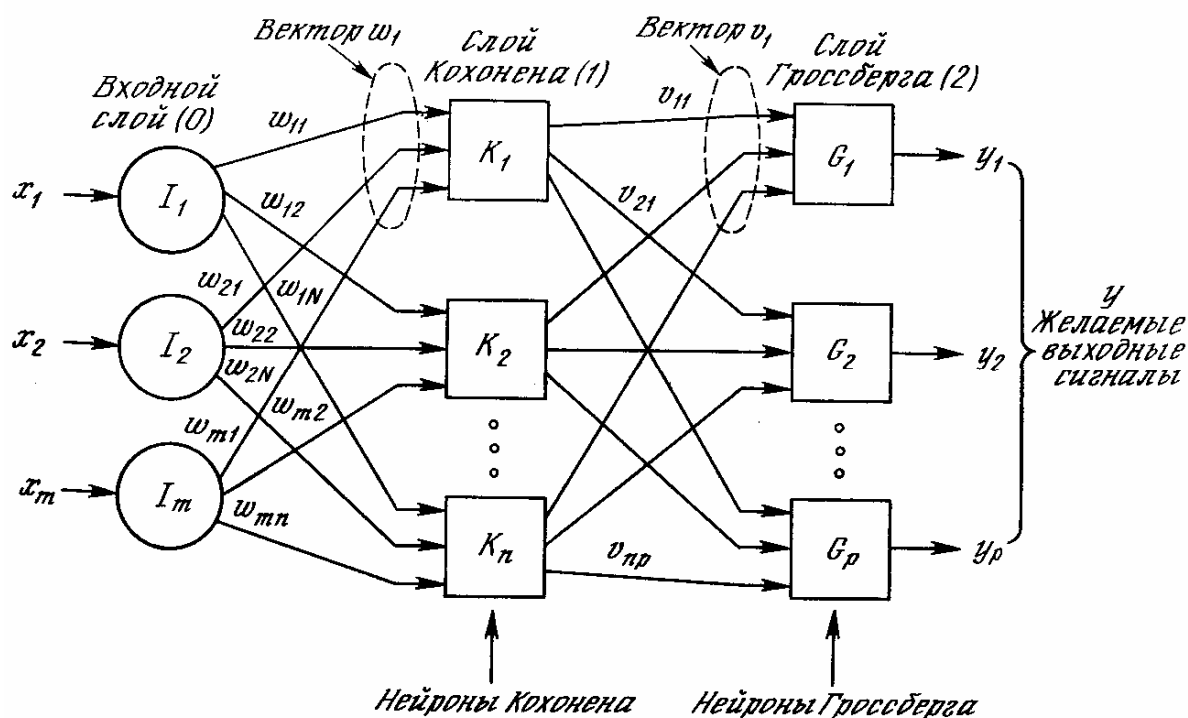


Рис. 7. Нейронная сеть встречного распространения

В режиме интерполяции входные сигналы генерируются как нейроном-победителем, так и некоторым количеством идущих за ним «призеров», нахо-

дящихся на 2-ом, 3-м и т.д. местах по уровню входного сетевого сигнала. Это приводит к более точной аппроксимации кусочными наклонными поверхностями. Подобный эффект может быть достигнут путем подмешивания в слой Кохонена дополнительных линейных либо нелинейных нейронов, латерально связанных с нейронами Кохонена (см. [21]). Нейросети встречного распространения обучаются значительно быстрее нейросетей обратного распространения, однако они не столь универсальны, менее точно аппроксимируют функциональные зависимости, слишком чувствительны к нерелевантным компонентам входных векторов и к большой их размерности.

Следует подчеркнуть, что составные нейросети, включающие в свой состав различные сетевые архитектуры и использующие различные методы обучения, более близки по принципам функционирования к человеческому мозгу, чем рассмотренные выше однородные структуры, подобные нейросетям обратного распространения.

1.2.5.3. Нейросети с радиальной базисной функцией

В отличие от многослойных персептронов, самоорганизующихся карт Кохонена и нейросетей на основе теории адаптивного резонанса (см. ниже), которые имеют под собой определенные нейрофизиологические основания, нейросети с радиальной базисной функцией (Radial Basis Function [RBF] neural networks) менее всего связаны с представлениями из биологии, базирясь в наибольшей степени на аппарате математической статистики. В сущности, их можно даже считать методами непараметрического статистического анализа, описанными при помощи терминологии аппарата нейронных сетей. Они тесно связаны с современными методами ядерного (kernel) статистического оценивания.

Нейронные сети с радиальной базисной функцией (RBF-сети) в определенном смысле можно считать дальнейшим развитием сетей встречного распространения (см. пункт 1.2.5.2). Предложенные рядом авторов в 1989 г. [67], они предназначаются, прежде всего, для решения задач аппроксимации функ-

ций и классификации (распознавания образов) [68, 69]. Как и нейросети встречного распространения, RBF-сети состоят из 3 слоев: входного, скрытого (служащего для кластеризации входных векторов) и выходного для формирования выходных сигналов (см. Рис. 8).

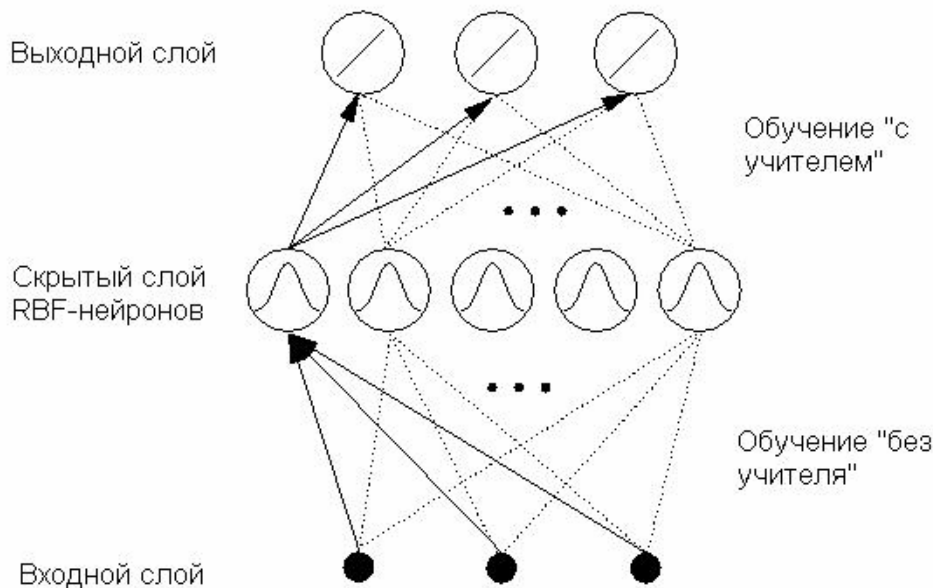


Рис. 8. Нейронная сеть с радиальной базисной функцией (RBF-сеть)

Скрытый слой у RBF-сетей состоит из RBF-нейронов, функционирование каждого из которых можно описать следующей формулой:

$$y_i = \exp\left(-\frac{\sum_{j=1}^M (x_j - w_{ji})^2}{2\sigma_i^2}\right), \quad (40)$$

где: x_j – j -ый компонент вектора входных значений; w_{ji} – j -ый компонент вектора весов RBF-нейрона i ; σ_i – дисперсия, характеризующая ширину радиально-базисной функции для RBF-нейрона i ; M – размерность входного вектора. Вектор весов i -ого RBF-нейрона $W_i = \{w_{1i}, w_{2i}, \dots, w_{Mi}\}$ задает положение центра его радиально-базисной функции. Выходные нейроны RBF-сети обычно берутся линейными, т.е. обладающими линейной активационной (передаточной) функцией.

Обучение RBF-сети проводится в два этапа. На первом, проходящем «без учителя», определяются положения центров радиально-базисных функций для

всех RBF-нейронов, а также их дисперсии. Для этого проводится кластерный анализ исходных данных либо при помощи нейросети Кохонена, либо, чаще всего, алгоритма k -means [70, 71], после чего центры найденных кластеров используются как центры радиально-базисных функций, ширины которых можно, в частности, определить как средние расстояния между центрами кластеров и его ближайшими соседями. Второй этап обучения RBF-сетей проводится «с учителем» - либо итерационно, в соответствии с алгоритмом обратного распространения ошибки, либо с использованием одного из алгоритмов построения линейных регрессионных моделей, в частности, при помощи регрессии на главных компонентах (SVD-регрессии) [72]. Различные варианты RBF-сетей различаются выбором: а) метода кластеризации (если она вообще проводится); б) способов определения положения центра и ширины радиально-базисной функции; в) способов построения линейно-регрессионной модели для обучения выходных нейронов. Ширина радиально-базисной функции иногда берется единой для всех RBF-нейронов, и ее значение, обеспечивающее наибольшую прогнозирующую способность нейронной сети, определяется с помощью процедуры скользящего контроля.

Важными модификациями RBF-сетей являются вероятностная нейронная сеть (Probabilistic Neural Network – PNN, P-нейросеть), предложенная Спехтом (Specht) в 1990 г. [73], и нейронная сеть обобщенной регрессии (Generalized Regression Neural Network – GRNN, GR-нейросеть), введенная этим же автором годом позже [74].

GR-нейросети. Функционирование GR-нейросетей основано на использовании математического аппарата непараметрической ядерной регрессии Надарая-Ватсона (Nadaraya-Watson) [75, 76], идея которой заключается в оценке функции плотности вероятности совместного распределения случайной векторной величины x и случайной скалярной величины y по методу Парзена (Parzen) [77]:

$$f(x, y) = \frac{1}{N(2\pi)^{(M+1)/2} \sigma^{(M+1)}} \cdot \sum_{i=1}^N \exp\left[-\frac{(x-x_i)^T(x-x_i)}{2\sigma^2}\right] \cdot \exp\left[-\frac{(y-y_i)^2}{2\sigma^2}\right], \quad (41)$$

где: N – количество примеров в обучающей выборке; M – размерность входных векторов (т.е. количество дескрипторов при QSAR/QSPR-анализе); x_i – входной вектор для i -ого примера из обучающей выборки (т.е. вектор дескрипторов для i -ого соединения); y_i – известное значение выходной величины y для i -ого примера (т.е. экспериментальное значение прогнозируемого свойства y для i -ого соединения); σ – единый параметр, соответствующий ширине Гауссовых функций, и называемый в контексте регрессионного анализа параметром сглаживания.

При известной функции $f(x,y)$ наиболее вероятное значение (т.е. математическое ожидание) y для произвольного вектора x может быть найдено по формуле:

$$\hat{y}(x) = E(y | x) = \frac{\int_{-\infty}^{+\infty} y f(x, y) dy}{\int_{-\infty}^{+\infty} f(x, y) dy}. \quad (42)$$

Подставляя (41) в (42) после некоторых преобразований можно получить окончательное выражение оценки y для произвольного x :

$$\hat{y}(x) = \frac{\sum_{i=1}^N y_i \exp\left[-\frac{(x-x_i)^T(x-x_i)}{2\sigma^2}\right]}{\sum_{i=1}^N \exp\left[-\frac{(x-x_i)^T(x-x_i)}{2\sigma^2}\right]}. \quad (43)$$

Легко заметить, что числители стоящих в экспоненте дробей представляют собой квадраты Эвклидовых расстояний между произвольным вектором x и вектором x_i из i -ого примера обучающей выборки:

$$\|x-x_i\|^2 \equiv (x-x_i)^T(x-x_i). \quad (44)$$

Заметим, однако, что при наличии существенных корреляций между компонентами входных векторов x более корректно в статистическом плане (хотя и более трудоемко в вычислительном плане) использовать в формуле (43) вместо квадратов расстояний Эвклида квадраты расстояний Махаланобиса $(x-x_i)^T \Sigma^{-1}(x-x_i)$, где Σ – матрица ковариации компонентов векторов x . Таким образом, согласно формуле (43), наиболее вероятное значение y для произвольного вектора x прогнозируется как взвешенная сумма значений y_i для всех примеров из обу-

чающей выборки, причем каждому примеру придается вес, экспоненциально убывающий при возрастании квадрата расстояния от него до вектора x , а скорость этого убывания контролируется параметром сглаживания σ .

Как архитектура, так и функционирование GR-нейросетей описывается формулой (43). GR-нейросеть состоит из 4 слоев: 1) входного; 2) скрытого; 3) слоя суммирования; 4) выходного слоя (см. Рис. 9). Число нейронов во входном слое равно количеству компонент входного вектора x . Скрытый слой GR-нейросети состоит из RBF-нейронов, функционирующих в соответствии с формулой (40). Число нейронов в скрытом слое равно количеству примеров в обучающей выборке, а вес связи w_{ji} между входным нейроном j и скрытым нейроном i устанавливается равным значению j -ой компоненты вектора x_i (т.е. значению j -ого дескриптора для i -ого соединения из обучающей выборки в случае QSAR/QSPR-анализа). Слой суммирования GR-нейросети состоит из двух линейных нейронов, причем первый из них вычисляет значение числителя в формуле (43), а второй – знаменателя. Вес связи, идущей от скрытого нейрона i к первому из нейронов суммирования, устанавливается равным y_i (т.е. экспериментальному значению прогнозируемого свойства y для i -ого соединения из обучающей выборки), а все веса связей, идущих от нейронов скрытого слоя ко второму нейрону слоя суммирования устанавливаются равными единице. Выходной слой GR-нейросети состоит из одного нейрона, который выполняет деление числителя на знаменатель в соответствии с формулой (43) (подобные нейроны, формирующие в процессе вычислений два сетевых входа и осуществляющие деление одного на другой, называют Паде-нейронами).

Таким образом, единственным настраиваемым параметром GR-нейросети является фактор сглаживания σ . Его оптимальное значение обычно подбирается исходя из критерия максимизации прогнозирующей способности нейросети, оцениваемой при помощи процедуры перекрестного скользящего контроля.

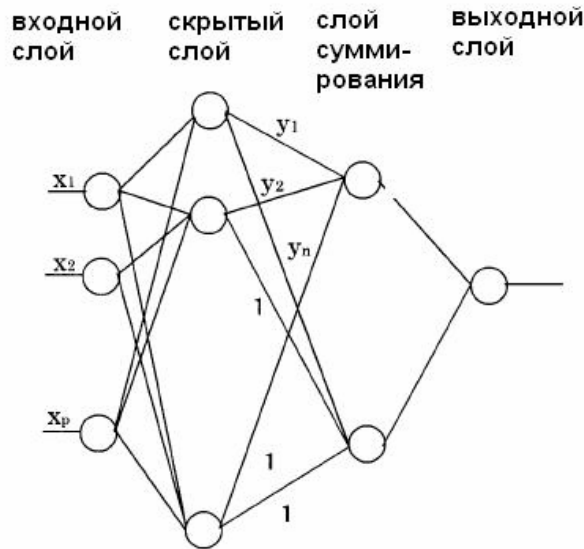


Рис. 9. Архитектура GR-нейросети

Р-нейросети. В отличие от GR-нейросетей, предназначенных для проведения регрессионного анализа, Р-нейросети служат для классификации входных векторов. В соответствии с этим, Р-нейросети оценивают функцию плотности вероятности распределения случайной векторной величины x отдельно для каждого из классов c по формуле:

$$f_c(x) = \frac{1}{N_c (2\pi)^{M/2} \sigma^M} \cdot \sum_{i: x_i \in C} \exp\left[-\frac{(x - x_i)^T (x - x_i)}{2\sigma^2}\right], \quad (45)$$

где суммирование идет только по N_c примерам из обучающей выборки, относящимся к классу C , остальные же обозначения те же, что и для формулы (41).

Р-нейросеть состоит из тех же четырех слоев, что и рассмотренная выше GR-нейросеть (см. Рис. 10). Структура и функционирование первых двух слоев (т.е. входного и скрытого, называемого в некоторых публикациях слоем образов [patterns]) упомянутых двух нейросетей также практически совпадают. В частности, RBF-нейроны скрытого слоя формируют следующие выходные сигналы:

$$p_i^c(x) = \frac{1}{(2\pi)^{M/2} \sigma^M} \cdot \exp\left[-\frac{(x - x_i)^T (x - x_i)}{2\sigma^2}\right]. \quad (46)$$

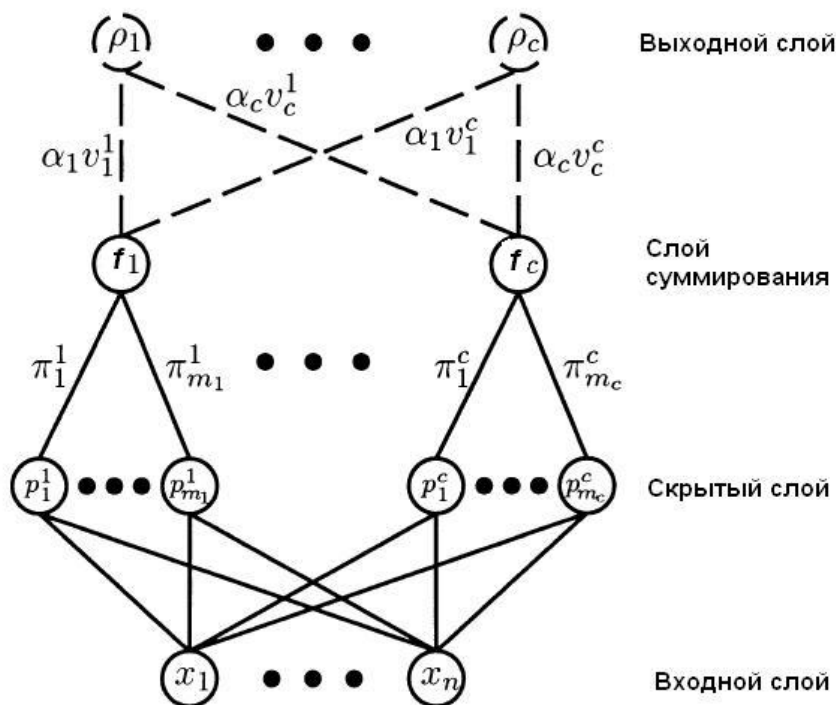


Рис. 10. Архитектура Р-нейросети

Отличия же начинаются в третьем слое (слое суммирования), который состоит из такого количества линейных нейронов, которое равно числу классов. В отличие от других RBF-сетей, в каждый нейрон c из слоя суммирования входят связи только с тех нейронов скрытого слоя, которые соответствуют примерам, принадлежащим классу C . Веса этих связей выбираются таким образом, чтобы выполнялось условие (47):

$$\sum_{i=1}^{N_c} \pi_i^c = 1, \quad c = 1, \dots, M_C, \quad (47)$$

где: π_i^c - вес связи, ведущей из i -ого нейрона скрытого слоя в c -ый нейрон слоя суммирования, отвечающий за формирование функции плотности вероятности $f_C(x)$ для класса C ; N_C - число примеров из обучающей выборки, относящихся к классу C ; M_C - число нейронов в слое суммирования, равное общему числу классов в классификационной задаче. Если нет других соображений, то все примеры, относящиеся к одному классу, считаются равнозначными, и поэтому веса всех связей, идущих к одному нейрону из слоя суммирования, берутся одинаковыми и равными $1/N_C$. В этом случае на нейронах слоя суммирования

формируются выходные сигналы, равные значениям функций плотности вероятности f , вычисляемым по формуле (45).

Выходной слой Р-нейросети также содержит M_C линейных нейронов по числу классов в решаемой классификационной задаче. Вес связи, идущей от нейрона c слоя суммирования к нейрону k выходного слоя, берется равным произведению α_k (априорной вероятности для класса k) на штраф v_k^c за ошибочное отнесение примера, относящегося к классу k , к классу c . В этом случае на нейроне k выходного слоя формируется сигнал, равный:

$$\rho_k(x) = \sum_{c=1}^{M_C} v_k^c \alpha_c f_c(x). \quad (48)$$

Если нет никаких дополнительных соображений, то веса всех штрафов v_k^c берутся равными единице, а величины априорных вероятностей для классов берутся равными доле примеров, относящихся к этим классам, в обучающей выборке. Поскольку величина формируемого на нейроне выходного слоя сигнала $\rho_k(x)$ равна значению функции риска отнесения текущего примера x к классу k , то наиболее вероятный класс l для данного примера может быть найден из условия минимального значения функции риска:

$$l = \arg \min_{1 \leq k \leq M_C} \{\rho_k(x)\}. \quad (49)$$

Так же, как и в случае GR-нейросети, при наличии сильных корреляций между компонентами входных векторов x более корректно (хотя это значительно усложняет расчеты) использовать в формуле (046) вместо квадратов расстояний Эвклида $(x-x_i)^T(x-x_i)$ квадраты расстояний Махаланобиса $(x-x_i)^T \Sigma^{-1}(x-x_i)$.

Так же, как и в случае GR-нейросети, единственным настраиваемым параметром Р-нейросети является фактор сглаживания σ , оптимальное значение которого подбирается исходя из критерия максимизации прогнозирующей способности нейросети, оцениваемой при помощи процедуры перекрестного скользящего контроля.

1.2.5.4. Нейросети на основе теории адаптивного резонанса

Базирующаяся на нейрофизиологической теории адаптивного резонанса процедура категоризации векторов (см. работы [78-81]) основана на сравнении очередного вектора с эталонными векторами, описывающими уже найденные ранее категории (кластеры). Если очередной вектор «похож» по определенному критерию близости на один из эталонных векторов, то он используется для его настройки, в противном же случае он сам объявляется представителем новой категории данных и запоминается в виде нового эталонного вектора. Описанная процедура реализуется в виде нейросети, состоящей из слоя сравнения, который оценивает «сходство» векторов, слоя распознавания (каждый нейрон его описывает свою категорию (кластер) данных) и нескольких дополнительных элементов. Подробность категоризации (а, значит, и количество категорий) контролируется специальным «параметром бдительности». Подобная архитектура получила название ART-1 для категоризации бинарных векторов и ART-2 для категоризации векторов вещественных чисел. ARTMAP представляет собой модульную нейросеть, состоящую из двух сетей типа ART для категоризации векторов (в случае корреляций структура-свойство эти вектора соответствуют дескрипторам и свойствам органических соединений), и модуля сравнения, в котором происходит запоминание «ассоциаций» между категориями дескрипторов и свойств. Основанный на использовании аппарата нечеткой логики вариант ARTMAP под названием fuzzy ARTMAP используется для решения регрессионных задач [82, 83]. Особенностью fuzzy ARTMAP является очень высокая устойчивость к переучиванию, что достигается благодаря автоматической настройке «параметра бдительности», контролирующего кластеризацию векторов дескрипторов, что позволяет строить модели с минимально возможной сложностью.

1.2.5.5. Нейросети с обратными связями (рекуррентные нейросети)

Все рассмотренные выше нейросети не имели обратных связей, т.е. связей, идущих от нейронов выходного слоя к псевдонейронам входного. Если ус-

ловно принять, что срабатывание искусственного нейрона происходит за один такт, то время работы всей нейронной сети, выраженное числом таких тактов, равно числу слоев вычислительных нейронов, т.е. общему числу слоев за вычетом входного слоя псевдонейронов (не производящих вычислений). Если же в нейросети ввести обратные связи, то время работы таких нейросетей ничем не ограничено и может продолжаться до бесконечности. Поскольку при стремлении времени такта к нулю разностные уравнения, описывающие работу нейросетей, переходят в дифференциальные, то, следовательно, функционирование этого вида нейросетей может быть альтернативно описано при помощи дифференциальных уравнений движения (т.е. движения условной псевдочастицы, координаты которой соответствуют значениям выходных сигналов нейросети). В связи с этим нейросети с обратными связями могут быть использованы для моделирования динамических процессов (т.е. происходящих во времени). Однако, сфера применения указанных нейросетей этим далеко не ограничивается.

Как и любая нелинейная динамическая система, нейронная сеть с обратными связями в процессе своей работы в конечном счете приходит к одному из трех состояний: 1) к стационарному состоянию, когда выходные сигналы нейросети перестают меняться во времени; 2) к периодически или квазипериодически изменяющемуся семейству состояний; 3) в особых случаях к неперiodическим изменениям в пределах определенного множества состояний. В дифференциальной топологии для описания этих конечных состояний используют термин «аттрактор», куда включаются: 1) устойчивые точки, называемые также стоками, соответствующие остановке движения; 2) замкнутые орбиты, в понятие которых включаются либо предельные циклы, соответствующие устойчивому периодическому движению, либо торы, соответствующие устойчивому квазипериодическому движению; 3) аперiodические кривые, соответствующие хаотическому движению (подобные аттракторы часто называют странными) [84, 85]. Нейронные сети, динамика работы которых соответствует первому случаю, т.е. достижению стационарного состояния, называются устойчивыми, в противном же случае их называют неустойчивыми. Математическим описанием процессов, характерных для динамики неустойчивых нейросетей с обратными

ми связями, занимается специальная наука – синергетика. Неустойчивые нейросети могут быть использованы в химии: 1) для описания колебательных химических процессов [86] типа реакции Белоусова-Жаботинского [87, 88] в случае нейросетей с периодическими состояниями (и наоборот, колебательные химические реакции могут быть использованы в качестве «элементной базы» для построения нейрокомпьютеров [89]); 2) для моделирования хаотических явлений, диссипативных процессов и связанных с ними явлений самоорганизации в случае нейросетей с хаотическим поведением. Следует также отметить, что именно функционирование нейросетей с обратными связями (и особенно неустойчивых!) наиболее точно соответствуют процессам, происходящим в коре головного мозга человека, о чем свидетельствует возможность использования нейросетей этого вида для описания реальных электроэнцефалограмм (см., например, [90]), а также наличие у них определенных свойств и особых режимов работы, традиционно приписываемых высшей нервной деятельности, таких как ассоциативный характер памяти, «творчество», «грезы» во время «сна» и т.д. – на некоторых из них мы остановимся ниже. Что же касается исследований связи «структура-свойство», то это пока что область применения устойчивых нейросетей.

Хотя первые вычислительные эксперименты с рекуррентными нейросетями проводились, по-видимому, еще на заре нейросетевой эры, лишь с начала 80-ых годов, с появлением адекватного математического аппарата для их анализа и, вместе с ним, нахождением условий стабильности [91], начался активный этап их изучения и практического использования. Несомненный приоритет в изучении простейших однородных нейросетей с обратными связями принадлежит Дж. Хопфилду (J. Hopfield), именем которого были названы эти нейронные сети и работы по которым сыграли огромную роль в формировании современных представлений об искусственных нейронных сетях [92-95].

Нейросети Хопфилда. Нейросети Хопфилда можно условно представить состоящими из двух слоев: 1) распределительного слоя псевдонейронов (в оригинальных статьях Хопфилда он отдельно не рассматривается), которые занимают лишь распределением сигналов; 2) вычислительного слоя нейронов,

выполняющего одновременно функции входного и выходного слоев, причем между слоями имеются как прямые связи, идущие от распределительного слоя к вычислительному, так и обратные связи, идущие от вычислительного слоя к распределительному (см. Рис. 11). В зависимости от решаемой задачи, сети Хопфилда работают с уровнями сигнала, находящимися в интервале как от 0 до 1, так от -1 до 1. В классическом варианте нейроны вычислительного слоя обладают пороговой активационной функцией (2) при уровнях сигнала от 0 до 1 либо (50) при уровнях сигнала от -1 до 1:

$$f(x) = \begin{cases} 1, & x \geq 0 \\ -1, & x < 0 \end{cases} \quad (50)$$

В этом случае нейросеть является бинарной, т.е. множество ее возможных состояний располагается на вершинах n -мерного гиперкуба, где n -число нейронов в вычислительном слое [92]. Имеется также и аналоговый вариант нейросети Хопфилда с сигмоидной активационной функцией (3) для уровней сигнала от 0 до 1 либо с функцией (51) для уровней сигнала от -1 до 1 [93]:

$$f(x) = th(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}. \quad (51)$$

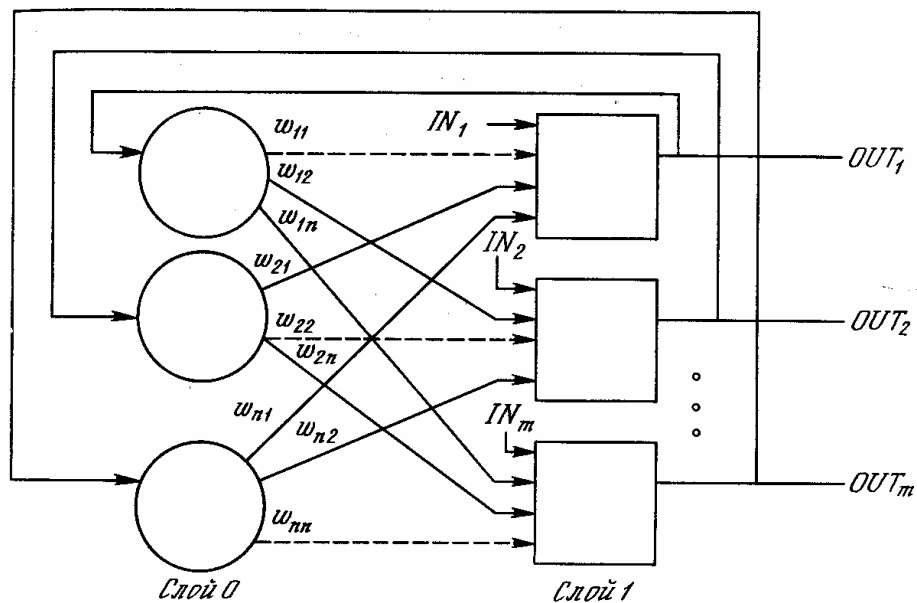


Рис. 11. Нейросеть Хопфилда. Пунктирная линия обозначает связь с нулевым весом. Слой 0 является распределительным, а слой 1 – вычислительным.

В этом случае множество возможных состояний располагается внутри вышеупомянутого n -мерного гиперкуба, однако абсолютное большинство

практически важных приложений связано с использованием именно бинарных нейросетей Хопфилда, которыми и будет ограничено дальнейшее изложение.

Условиями стабильности нейросетей Хопфилда являются $w_{ij} = w_{ji}$ и $w_{ii} = 0$, т.е. матрица весов связей должна быть симметрична и ее диагональные элементы должны быть нулевыми [91]. Анализ функционирования (в том числе и рассмотрение устойчивости) нейросетей Хопфилда обычно проводят при помощи математического аппарата, основанного на применении особой функции, называемой функцией Ляпунова, которая ограничена снизу и не возрастает при изменениях состояний сети. При работе нейросети значение функции Ляпунова будет уменьшаться до достижения ее минимума, пусть даже и локального, после чего изменения энергии прекратятся (такая сеть по определению является устойчивой). Таким образом, наличие функции Ляпунова является достаточным условием устойчивости нейросети с обратными связями. Ввиду стремления системы к постоянному уменьшению значений функции Ляпунова, эту функцию часто, пользуясь аналогиями из области физики, называют функцией энергии нейросети, а сами они называются нейросетями, минимизирующими свою энергию.

Функция энергии E нейросети Хопфилда в общем виде может быть записана как:

$$E = (-1/2) \sum_i \sum_j w_{ij} o_i o_j - \sum_i I_i o_i + \sum_i t_i o_i, \quad (52)$$

где: I_i – внешний вход в нейрон I ; w , o и t – как и прежде, веса связей, выходы и пороги активации нейронов. Внешний вход вычислительного нейрона используется только для начальной установки исходного значения выхода этого нейрона и убирается сразу же после его первого срабатывания, поэтому при рассмотрении работы нейросети значение внешнего входа можно принять равным нулю. Если (как в случае нейросетей без обратных связей) ввести условный псевдонейрон смещения *bias* с постоянным уровнем выходного сигнала, равным единице, и соединить его со всеми нейронами вычислительного слоя связью с весом, равным их порогам активации, взятым с обратным знаком, то выражение для энергии нейросети существенно упростится:

$$E = (-1/2) \sum_i \sum_j w_{ij} o_i o_j . \quad (53)$$

В этом случае при асинхронном срабатывании i -ого нейрона (т.е. при постоянстве выходов остальных нейронов) энергия нейросети изменится на δE :

$$\delta E = -\delta o_i \cdot \sum_{j \neq i} w_{ji} o_j = -\delta o_i \cdot a_i , \quad (54)$$

где: δo_i – изменение выхода i -ого нейрона при его срабатывании; a_i – сетевой вход i -ого нейрона, который при этом должен оставаться неизменным. Если сетевой вход a_i меньше нуля, то нейрон не срабатывает, его выход не меняется, и, следовательно, общая энергия сети не меняется. Если сетевой вход больше нуля, то нейрон активируется. Если он и до этого был активным, то его выход и общая энергия сети при этом не меняются. Если же он до этого был неактивен, то значение δo_i будет положительным, а значение δE будет отрицательным, следовательно, энергия сети при срабатывании нейрона будет уменьшаться. При точном равенстве сетевого входа нулю δE тоже равно нулю, т.е. общая энергия сети не меняется. Таким образом, во всех случаях энергия сети либо уменьшается, либо остается неизменной. Следовательно, в процессе работы нейросети ее энергия минимизируется.

С этим свойством нейросетей Хопфилда тесно связано их использование в нейроматематике для решения задач оптимизации, и прежде всего комбинаторной оптимизации, что в перспективе не может не найти применения при решении задач «структура-свойство». Для того, чтобы решить задачу оптимизации при помощи нейросети Хопфилда, надо переформулировать эту задачу в терминах минимизации функции Ляпунова, для которой нужно построить соответствующую нейросеть. Второй и не менее важной областью применения нейросетей Хопфилда является их использование для реализации ассоциативной памяти и осуществления распознавания образов (эти две функции тесно связаны друг с другом и поэтому они рассматриваются обычно вместе). Для этой цели используются сети, работающие с сигналами в интервале от -1 до 1.

Фаза обучения нейросетей Хопфилда очень проста и сводится к однократному применению имеющего серьезные нейрофизиологические обоснова-

ния классического правила Хебба (Hebb) [96] к каждому из векторов обучающей выборки, в результате чего веса связей примут следующие значения:

$$w_{ij} = \frac{1}{N} \sum_{k=1}^N x_i^k x_j^k, \quad i = 1, \dots, M, \quad j = 1, \dots, M \quad (55)$$

где: x_i^k - i -ый компонент k -ого вектора из обучающей выборки; N – количество векторов в обучающей выборке; M – число компонент в каждом из векторов. После того, как обучающие вектора занесены по формуле (55) в голографическую память нейросети (в оптических нейрокомпьютерах веса нейросетей Хопфилда кодируются участками реальных голограмм [97]), нейросеть может быть использована для извлечения занесенного в память вектора по предъявлению ей неполного или частично искаженного его варианта. Для этого выходы нейросети инициализируются значениями предъявляемого вектора, и сеть запускается на счет. В результате через определенное время она стабилизируется в ближайшем глубоком минимуме функции энергии, и на ее вычислительных нейронах формируются выходные значения, вектор которых совпадает с одним из занесенных в память векторов. Если предъявленный вектор представляет собой неполный вариант запомненного вектора, то в результате работы нейросети происходит его дополнение отсутствующими компонентами. В этом случае говорят об извлечении информации из ассоциативной памяти по заданной части этой информации (как это и происходит в случае человеческой памяти). Если же предъявляемый вектор представляет собой искаженный вариант запомненного вектора, то происходит исправление его, что фактически представляет собой распознавание образов (предполагается, что в голографическую память нейросети загружены типичные представители распознаваемых классов образов).

Нейросети Хопфилда имеют очень ограниченную емкость – безошибочное извлечение запомненных векторов возможно только в том случае, если число их не превышает 14% от числа нейронов. При превышении этого порога нейросеть начинает стабилизироваться в аттракторах, не совпадающих ни с одним из запомненных векторов. Подобные аттракторы получили негативные названия, такие как ложная (spurious) или паразитная память, химеры, русалки и

т.д., и до последнего времени от них старались избавиться. Так, в качестве эффективного средства борьбы с состояниями ложной памяти Хопфилдом с соавторами была предложена процедура разобучения (unlearning) [98], суть которой заключается в следующем: 1) в многократном предъявлении нейросети в качестве начальных состояний случайно сгенерированных векторов; 2) в прослеживании их эволюции до стационарных состояний o^* (которые могут отвечать как истинному запомненному вектору, так и ложной памяти); 3) модификации всех весов на вклад:

$$\delta w_{ij} = -\varepsilon o_i^* o_j^*, \quad i=1, \dots, M, \quad j=1, \dots, M, \quad (56)$$

где ε – небольшая положительная константа. Поскольку состояниям ложной памяти обычно соответствуют более мелкие локальные минимумы, то именно они в первую очередь и подвергаются разобучению. И действительно, в результате комбинирования процедур обучения и разобучения удается существенно повысить емкость нейросети Хопфилда с 14% от числа нейронов до значения, близкого к 100%, что, по-видимому, составляет теоретический максимум емкости нейросети этого типа [99].

Крик (Crick) и Митчисон (Mitchison) высказали предположение о том, что процесс, аналогичный разобучению, происходит в мозгу человека во время фазы быстрого (парадоксального) сна [100]. Как известно, во время сна возникают фантастические сюжеты, весьма далекие, хотя и чем-то напоминающие те, что человек видел наяву. Согласно гипотезе Крика-Митчисона, видения во время сна представляют собой состояния ложной памяти, в которых временно стабилизируются нейроны определенных участков коры головного мозга, возбуждаемые случайными воздействиями ствола мозга во время фазы быстрого сна. Происходящее при этом разобучение приводит к забыванию возникающих во время сна парадоксальных картин видений и тем самым упрощает доступ к запомненным образам, соответствующим объектам реального внешнего мира. Таким образом, нейросети Хопфилда присуща человеческая способность «грезить» и «видеть сны», причем, как было показано А.А. Ежовым, подобные «сны» могут быть «вещими», и во время них нейронная сеть, подобно челове-

ческому мозгу во время сна, способна проявлять удивительные «творческие способности» и даже делать открытия [18, 101, 102].

В качестве примера «творческих способностей» нейросетей Хопфилда можно упомянуть активную кластеризацию, в процессе которой не только формируются аттракторы, которые притягивают векторы из обучающей выборки (попавшие в один аттрактор векторы можно считать относящимися к одному классу), но и такие аттракторы (соответствующие состояниям ложной памяти), в которые не притягивается ни один вектор из обучающей выборки (в этом случае говорят о формировании пустых классов) [102]. Иными словами, за счет глубокой обработки информации нейросеть оказывается способной не только относить известные объекты к определенным классам, но и предсказывать наличие новых классов объектов и даже генерировать (описывать) эти объекты [101]. Известным примером такой предсказательной категоризации является создание Д.И. Менделеевым периодической системы элементов, в которой с самого начала были определены три пустые клетки (фактически новые классы объектов) и описаны свойства входящих в них гипотетических объектов – новых химических элементов, которые впоследствии и были обнаружены.

Аналогия с высшей нервной деятельностью человека на этом, однако, не заканчивается. Оказывается, для нейросети Хопфилда не чуждо даже эстетическое «чувство прекрасного». В процессе запоминания векторов нейросеть в своей голографической ассоциативной памяти «рассматривает» запоминаемые вектора как «зашумленные» версии некоего «идеала». Она старается восстановить его за счет отбрасывания шума (путем наложения друг на друга образов, приводящих к уменьшению вклада некоррелированных компонент, т.е. шума) и уже найденный идеальный образ запомнить, создав для него аттрактор памяти. Судя по всему, Крик и Митчисон правы лишь отчасти: нейросеть во время «сна» грезит не для того, чтобы забыть ложные видения, а для того, чтобы, забыв несущественные элементы действительности, сформировать внутри себя «прекрасный идеал», который необходимо как можно лучше запомнить и к которому «устремлять свои помыслы». Благодаря этому обстоятельству нейросети Хопфилда могут быть использованы как эффективный инструмент очище-

ния от шума, поиска скрытых закономерностей в «зашумленных» объектах и выделения из них исходного объекта-прототипа. В качестве характерных примеров можно привести осуществляемый при помощи нейросетей Хопфилда поиск промоторов в ДНК [18], скрытых повторов в ДНК и реконструкцию эволюционных изменений в них [103].

Машина Больцмана. Одним из недостатков нейросетей Хопфилда является их тенденция стабилизироваться в локальном, а не глобальном минимуме функции энергии. Одним из способов преодоления этой трудности является использование стохастического варианта нейросети Хопфилда, называемого обычно машиной Больцмана. Подобное название нейросетей этого класса обусловлено тесной связью методов их описания с математическим аппаратом статистической термодинамики (а также данью уважения к ее основателю Больцману).

Если в детерминированных нейронных сетях, к которым относятся нейросети Хопфилда, нейрон *всегда* возбуждается при превышении сетевым входом a_i определенного порогового значения (которое путем введения bias-псевдонейронов всегда можно сделать нулевым), то в стохастических нейросетях, к которым относится машина Больцмана, сетевой вход определяет лишь вероятность p_i перехода нейрона i в возбужденное состояние:

$$p_i = \frac{1}{1 + \exp(-a_i / T)}, \quad (57)$$

где T – искусственная температура. Заметим, что в знаменателе этого выражения находится фактор Больцмана, показывающий вероятность пребывания системы в условиях термодинамического равновесия при температуре T на энергетическом уровне, превышающем нулевой на $k \cdot a_i$ энергетических единиц (где k – постоянная Больцмана).

При запуске машины Больцмана на выходы вычислительных нейронов заносятся начальные значения, определяемые входным вектором. Машина запускается при высоком значении искусственной температуры, и сети предоставляется возможность самостоятельно минимизировать свою энергию при управляемом извне постепенном понижении указанной температуры. После ох-

лаждения системы и достижения термодинамического равновесия считаются выходные значения нейронов (при неполном охлаждении считаются вероятности пребывания нейронов в активном состоянии). Поскольку описанная процедура полностью соответствует известной процедуре нахождения глобального минимума по методу искусственного закаливания (simulated annealing), то всегда можно подобрать такую скорость охлаждения системы, чтобы можно было достигнуть глобального минимума энергии. В данном случае можно говорить о «кристаллизации мысли» у нейросети, поскольку как работа нейросети, так и реальный процесс кристаллизации из расплава, описывается одним и тем же математическим аппаратом статистической термодинамики. Более того, при анализе работы машины Больцмана часто используют те же самые фазовые диаграммы состояний и таким же образом рассматривают фазовые переходы, как и в физической химии для реальных веществ и материалов.

Процесс обучения машины Больцмана обычно включает стадии «активного обучения», «разобучения во время сна» и «коррекции весов» [104]. На стадии «активного обучения» поочередно закрепляют на нейронах выходные значения, задаваемые входными векторами, дают сети релаксировать до наступления равновесия и для каждой пары нейронов по всему множеству обучающих векторов определяют P_{ij}^+ - вероятность того, что нейроны i и j одновременно находятся в активном состоянии. На стадии «разобучения во время сна» нейросеть запускают множество раз, начав со случайных состояний, и в результате для каждой пары нейронов определяют P_{ij}^- - вероятность того, что нейроны i и j одновременно находятся в активном состоянии. И, наконец, на последней стадии проводят коррекцию весов по формуле:

$$\Delta w_{ij} = \eta(P_{ij}^+ - P_{ij}^-), \quad (58)$$

где η – коэффициент скорости обучения.

1.3. Основные принципы применения искусственных нейронных сетей для прогнозирования свойств химических соединений

В большинстве работ по применению нейросетей обратного распространения (многослойных персептронов) для поиска зависимостей структура-свойство используется следующая методология. Прежде всего, подготавливается база данных, содержащая структуры химических соединений и известные значения тех свойств, которые в дальнейшем предполагается при помощи обученной нейросети прогнозировать. Как правило, эта база разбивается на две части, по первой из которых, называемой обучающей выборкой, путем многократного ее предъявления нейросети производится обучение последней, а по второй, называемой контрольной выборкой, производится контроль прогнозирующей способности обученной нейросети. В качестве вариантов иногда используются две контрольные выборки, а также процедура скользящего контроля, при которой каждое из соединений при одной из разбинок попадает в контрольную выборку. На следующем этапе для всех химических соединений из выборок производится расчет дескрипторов, т.е. чисел, описывающих структуру химического соединения. Как правило, эти числа инвариантны к перенумерации вершин молекулярного графа, которым может быть описана структура химического соединения, т.е. являются инвариантами графов. Дескрипторы могут быть фрагментными (подструктурными), топологическими индексами, физико-химическими, квантово-химическими, характеристиками пространственных структур и т.д. Достаточно полный набор дескрипторов, используемых в современных исследованиях структура-свойство, описан в книге Р.Тодескини (R.Todeschini) и В.Консонни (V.Consonni) [105].

Далее, после необязательной стадии предварительного отбора либо преобразования дескрипторов следует этап построения нейронной сети. Число нейронов входного слоя обычно берется равным числу дескрипторов, и уровень выходного сигнала каждого из них устанавливается равным значению соответствующего дескриптора после его нормализации или масштабирования). Число выходных нейронов равно числу одновременно прогнозируемых свойств, при-

чем в качестве прогнозируемого значения каждого из свойств берется выходное значение соответствующего выходного нейрона (обычно после денормализации или демасштабирования). Скрытые нейроны служат для промежуточных вычислений, и их число часто подбирается, исходя из критерия максимизации прогнозирующей способности нейросети, а псевдонейроны смещения выполняют служебные функции и обладают постоянным выходным значением, равным единице.

В процессе обучения нейросети обучающая выборка предъявляется ей определенное число раз (обычно довольно большое). В процессе предъявления выборки значения дескрипторов каждого из соединений последовательно вводятся (обычно после нормализации или масштабирования) в качестве активности соответствующих входных нейронов. Далее запускается нейросеть на счет, и с выходных нейронов снимаются прогнозируемые значения свойств, которые (после денормализации или демасштабирования) сравниваются с экспериментальными. На основании найденной разницы по определенным алгоритмам производится подстройка весов связей между нейронами с целью уменьшения этой разницы. Таким образом, в процессе обучения происходит постепенное уменьшение ошибок прогнозирования свойств химических соединений, входящих в обучающую выборку.

На момент начала выполнения диссертационной работы в литературе имелось лишь несколько публикаций [106-108] по применению искусственных нейронных сетей для прогнозирования свойств химических соединений. В этих работах нейросети были использованы для моделирования биологической активности в узких рядах органических соединений в рамках классического подхода Ганча-Фуджиты на основе использования констант заместителей в качестве дескрипторов. Нейросетевое моделирование многочисленных физико-химических свойств, применение для этого фрагментных дескрипторов и демонстрация универсальности этого подхода была впервые осуществлена в рамках данной диссертационной работы.

К настоящему времени опубликовано уже больше тысячи работ, связанных с нейросетевым моделированием свойств химических соединений. Современ-

менное состояние дел в этой быстро развивающейся области науки подробно рассмотрено нами в нескольких обзорах [37, 39, 109].

1.4. Ограничения искусственных нейронных сетей

Как и любой метод машинного обучения, искусственные нейронные сети имеют свои ограничения, которых, однако, по мере развития теории нейросетевого моделирования и общей теории обучающихся систем, становится все меньше и меньше. В начале 90-ых годов прошлого века, т.е. на момент появления первых работ по их применению для прогнозирования свойств химических соединений, такими ограничениями или даже недостатками считались следующие:

- нейросеть – это «черный ящик», т.е. нейросетевые модели не поддаются интерпретации;
- нейросетевые модели не могут быть точно воспроизведены ввиду инициализации весов связей перед обучением случайными числами;
- нейронные сети не работают с большим числом дескрипторов;
- нейросети легко «переучиваются», и тогда они хорошо воспроизводят свойства соединений, содержащихся в обучающей выборке, но при этом плохо прогнозируют свойства любых других соединений;
- с помощью нейросетей ничего нельзя сделать такого, на что бы не были способны стандартные методы статистического анализа данных.

Хотя перечисленные выше утверждения не всегда справедливы, они, тем не менее, указывают на реальные проблемы, с которыми столкнулись исследователи в ходе первых работ по применению нейронных сетей для прогнозирования свойств химических соединений. Без их решения нейросети не могли бы быть использованы как составная часть универсальной методологии прогнозирования свойств химических соединений. Поэтому разработка эффективных методов решения этих проблем составила важную часть диссертационной работы (см. Главу 4).

ГЛАВА 2. ФРАГМЕНТНЫЕ ДЕСКРИПТОРЫ В ПОИСКЕ ЗАВИСИМОСТЕЙ СТРУКТУРА-СВОЙСТВО

Фрагментный дескриптор – это числовая характеристика химической структуры, показывающая, присутствует ли внутри нее определенный структурный фрагмент, либо специфицирующая сколько раз он в ней содержится. К преимуществам фрагментных дескрипторов обычно относят следующие (см. [110-116]): 1) простота и эффективность вычислений; 2) простота интерпретации со структурно-химической точки зрения; 3) базисный характер, выражающийся в возможности аппроксимировать с их помощью любую зависимость «структура-свойство».

2.1. История фрагментных дескрипторов

Среди множества дескрипторов, используемых в настоящее время в исследованиях SAR/QSAR/QSPR, (см. [105]), фрагментные дескрипторы занимают особое место. Ранние работы по их применению для предсказания разнообразных свойств химических соединений датируются 50-ми, 40-ми и даже 30-ми годами прошлого столетия, когда они использовались в рамках методологии *аддитивных схем*. Фогель (Vogel) [117], Цан (Zahn) [118], Саудерс (Souders) [119, 120], Франклин (Franklin) [121, 122], Татевский [123], Бернштейн (Bernstein) [124], Лаидлер (Laidler) [125], Бенсон (Benson) и Басс (Buss) [126] и Аллен (Allen) [127] были первопроходцами в этом направлении. В цитированных работах все вычислительные подходы были основаны на классической структурной теории в рамках представлений об атомах и химических связях. Е.А.Смоленский был, по-видимому, первым, кто применил еще в 1964 г. язык теории графов для прогнозирования физико-химических свойств органических соединений [128]. Первые аддитивные схемы постепенно эволюционировали и превратились в современный набор *методов групповых вкладов (group contribution methods)*. Основная отличительная черта аддитивных схем и методов групповых вкладов состоит в их тесной связи с физико-химической теорией, и по-

этому они применимы для прогнозирования только тех свойств, для которых подобная теория разработана.

Эпоха исследований QSAR (Quantitative Structure-Activity Relationships – количественных соотношений структура-свойство) началась в 1963-1964 гг. с появлением двух новаторских подходов. Первым из них - σ - ρ - π анализ Ганча (Hansch) и Фуджиты (Fujita) [14, 15], основанный на использовании констант заместителей и констант распределения в системе октанол-вода. Второй из них, метод Фри-Вильсона (Free-Wilson) [129], основан на предположении об аддитивности вкладов структурных фрагментов (которые представляют собой заместители, присоединенные в определенных положениях к единому в узком ряду соединений молекулярному остову) в общее значение биологической активности химического соединения. Таким образом, метод Фри-Вильсона можно рассматривать как расширение метода аддитивных схем на область прогнозирования биологической активности. Применимость обоих подходов ограничена узкими рядами соединений с одинаковым остовом, причем для метода Фри-Вильсона еще требуется, чтобы все рассматриваемые типы заместителей были хорошо представлены в обучающей выборке. Комбинация обоих подходов привела на практике к введению в модели QSAR т.н. *индикаторных переменных*, показывающих наличие определенных структурных фрагментов в молекуле.

Семидесятые годы прошлого века привели к созданию первых приложений SAR (non-quantitative Structure-Activity Relationships – неколичественные корреляции структура-активность), которые были разработаны под значительным влиянием таких научных направлений в вычислительной математике как искусственный интеллект, экспертные системы и теория распознавания образов. В рамках этих подходов химические структуры описываются набором индикаторов наличия определенных структурных фрагментов в молекулах, причем подобные фрагменты часто интерпретируются как *топологические* (или *2D*) *фармакофоры* (биофоры, токсофоры и т.д.) либо *фармакофобы* (биофобы, токсофобы и т.д.). Все эти подходы имеют целью классифицировать органические соединения как активные либо неактивные по отношению к определенно-

му типу биологической активности. Гиллер [2], Голендер и Розенблит [130, 131], Пирузян, Авидон и др. [131], Крамер (Cramer) [132], Бруггер (Brugger), Стюпер (Stuper) и Джурс (Jurs) [133, 134] и Хоудс (Hodes) и др. [135] были первопроходцами в этой области.

Современные методологии применения фрагментных дескрипторов в исследованиях QSAR и QSPR (Quantitative Structure-Property Relationships – количественные соотношения структура-свойство) не требуют введения явных ограничений на типы химических структур и прогнозируемых для них свойств, и поэтому их можно считать *универсальными*. Первый такой универсальный подход к использованию фрагментных дескрипторов в исследованиях QSAR/QSPR был разработан в 70-ые годы прошлого века Адамсоном (Adamson) с соавт. [136, 137]. Суть этого подхода заключается в расчете фрагментных дескрипторов для выборки химических соединений исходя из структур их молекулярных графов путем подсчета числа вложений в них простейших типов подграфов (порой получаемых напрямую из строк линейной нотации Висвессера [138]) с последующим введением этих дескрипторов в статистический анализ (обычно осуществляемый по методу множественной линейной регрессии) для поиска корреляций с экспериментальными значениями биологической активности [138, 139], физико-химических свойств [140] либо реакционной способности [141].

Важный класс фрагментных дескрипторов, т.н. *скрины* (или *структурные ключи, отпечатки пальцев*), также был разработан в 70-ые годы [142-146]. Их наборы образуют битовые строки, которые могут эффективно храниться и обрабатываться на компьютерах. Хотя первоначально им предназначалась лишь роль инструмента, позволяющего осуществлять подструктурный поиск в больших химических базах данных, в настоящее время они также активно используются при *поиске по подобию* (similarity searching) [147, 148], кластеризации больших баз данных, содержащих химические структуры, [149, 150], оценки их *разнородности* (diversity) [151], а также при проведении исследований SAR [152] и QSAR [153].

Следующий важный вклад в эту область был сделан Крамером (Cramer), который в 1980 г. определил параметры BC(DEF) путем проведения факторного анализа набора физических свойств для выборки разнородных органических жидкостей [154]. Эти параметры, с одной стороны, сильно коррелируют с разнообразными физическими свойствами разнородных органических соединений и, с другой стороны, их значения могут быть предсказаны с использованием фрагментных дескрипторов в рамках аддитивно-конституционных моделей [155]. Таким образом, Крамером был впервые разработан основанный на фрагментных дескрипторах набор моделей QSPR, позволяющий предсказывать целый набор физических свойств для разнородных органических соединений.

По-видимому, наиболее важным достижением 80-ых годов прошлого века в области применения фрагментных дескрипторов для прогнозирования биологической активности стала разработка Клопманым (Klopman) и др. компьютерной программы CASE (Computer-Automated Structure Evaluation) [156-159]. Эта программа, представленная авторами как «самообучающаяся система искусственного интеллекта» [159], способна распознавать активирующие и деактивирующие фрагменты (биофоры и биофобы) относительно определенного вида биологической активности, а также оценивать вероятность того, что произвольное тестовое соединение будет обладать этой активностью. Эта методология была успешно применена для предсказания множества видов биологической активности, в частности: мутагенности [157, 160, 161], канцерогенности [156, 159, 161-163], галлюциногенной активности [164], антиконвульсантной активности [165], ингибиторной активности по отношению к спартеиновой монооксигеназе [166], β -адренергической активности [167], способности связываться с μ -опиатным рецептором [168], антибактериальной активности [169], антилейкемической активности [170] и др. Она также позволяет строить количественные модели с использованием фрагментных дескрипторов при помощи статистического аппарата множественной линейной регрессии [162, 167].

Начиная с начала 90-ых годов, появляется большое число разнообразных подходов (наряду с соответствующими программами), основанных на использовании фрагментных дескрипторов в исследованиях SAR/QSAR/QSPR. На эту

тему опубликовано несколько концептуальных статей и мини-обзоров [110, 111, 113-116, 171]. Рассмотрим основные принципы классификации фрагментных дескрипторов.

2.2. Типы фрагментных дескрипторов

Структурные фрагменты и соответствующие фрагментные дескрипторы могут быть классифицированы: по типам молекулярных графов, по типам молекулярных структур, по типам значений дескрипторов, по типам дескрипторных наборов, по связанности фрагментов, по уровням детализации молекулярных графов и т.д.

2.2.1. Классификация по типам молекулярных графов

Структурные фрагменты, применяемые в исследованиях «структурасвойство», могут быть отнесены ко множеству типов молекулярных графов. В частности, можно выделить: простые фиксированные типы молекулярных графов; фрагменты WLN и SMILES; центрированные на атомах фрагменты; центрированные на связях фрагменты; фрагменты на основе максимальных общих подграфов; атомные пары и топологические мультиплеты; заместители и молекулярные остовы; фрагменты на основе базисных подграфов; фрагменты на основе «добытых» (mined) подграфов; библиотечные фрагменты и др. Рассмотрим каждый из вышеперечисленных типов.

2.2.1.1. Простые фиксированные типы молекулярных графов

Древнегреческий атомизм, согласно которому все вещества состоят из атомов, приводит к простейшему типу структурных фрагментов – атомам, т.е. вершинам молекулярных графов. Существует по крайней мере одно свойство, молекулярный вес, значение которого для всех химических соединений могут быть точно, если не принимать во внимание пренебрежительно малые реляти-

вистские поправки, представлены как сумма атомных вкладов, т.е. атомных весов:

$$MW = \sum_{i=1}^N n_i \cdot AW_i \quad (59)$$

где: MW = молекулярный вес; N - число типов атомов (в данном случае, типов химических элементов) в молекуле; n_i – число атомов типа i в молекуле; AW_i – атомный вес атома, относящегося к i -ому типу. Обобщение выражения (59) приводит к общему способу оценки свойств химических соединений с использованием основанных на атомных вкладах аддитивных схем по формуле:

$$P \approx \sum_{i=1}^N n_i \cdot A_i \quad (60)$$

где P обозначает произвольное молекулярное свойство, а A_i – соответствующие атомные вклады. В отличие от уникального случая с молекулярным весом, уравнение (60) дает лишь приблизительную оценку других свойств. Е.А.Смоленский [172], исходя из понятия о *химической дисперсии*, ввел специальный количественный показатель S , находящийся в интервале от 0 до 1, для описания способности какого-либо свойства быть представленным при помощи уравнения (60). Его численная величина равна наивысшему значению коэффициента детерминации, который для данного свойства в принципе может быть достигнут в рамках основанного на формуле (60) подхода 1D QSPR [173]. Для некоторых свойств, таких как парахор [117], молярная рефракция [174] и др., подобные качество 1D-QSPR-моделей вполне приемлемо, но для остальных свойств метод нуждается в улучшении. Наиболее распространенный путь достижения этого состоит во введении усовершенствованных классификационных схем для атомов, которые учитывают не только типы химических элементов, но и гибридизацию, число присоединенных атомов водорода, вхождение в состав определенных атомных групп или ароматических систем, и т.д. Следует, однако, отметить, что подобные подходы, которые в неявном виде учитывают молекулярную связность, не являются основанными на изолированных атомных вкладах (*separate-atom-based*).

В настоящее время подобные подходы, основанные на анализе атомных вкладов, широко используются для прогнозирования физико-химических свойств и биологических активностей органических соединений. Метод Гхоуза-Криппена (Ghose-Crippen) для предсказания коэффициента распределения в системе октанол-вода $\log P$ (ALOGP) [175-177], его усовершенствованные варианты, предложенные Гхоузом и др. [178, 179] и Вайлдманом (Wildman) и Криппеном [180], разработанный Сузуки (Suzuki) и Кудо (Kudo) метод CHEMICALC-2 для предсказания $\log P$ [181], программа SMILOGP, разработанная Конвардом (Convard) с соавторами для предсказания этого же свойства [182], метод XLOGP, разработанный Вангом (Wang) с соавт. для $\log P$ [183, 184], метод прогнозирования растворимости в воде, разработанный Хоу (Hou) и др. [185], - это лишь небольшое число примеров прогнозирования физико-химических свойств органических соединений в рамках основанных на атомных вкладах аддитивных схем. Как показал Винклер (Winkler) с соавт., этот подход может быть использован также и для предсказания некоторых видов биологической активности органических соединений [186].

Поскольку все молекулы состоят из атомов, связанных посредством химических связей, соответствующих ребрам молекулярных графов, химические связи были всегда в центре внимания при описании структур химических соединений и предсказании их свойств. Первые основанные на вкладах по связям аддитивные схемы, такие как методы Цана (Zahn) [118], Бернштейна (Bernstein) [124, 187] и Аллена (Allen) [127, 188], появились почти одновременно с первыми аддитивными схемами, основанными на атомных вкладах. В большинстве случаев они предназначены для прогнозирования термодинамических свойств, таких как теплота образования, которая непосредственно связана с энергиями химических связей. Следует, однако, отметить, что вышеупомянутые аддитивные схемы не являются основанными на вкладах изолированных связей (*separate-bond-based*), поскольку они обычно содержат перекрестные члены, которые могут быть описаны посредством молекулярных графов с 3 вершинами и 2 ребрами.

Нилаконтан (Nilakantan) с соавторами ввели понятие о *топологических торсионных углах* (*topological torsions*), которые представляют собой четверки последовательно связанных между собой неводородных атомов [189]. Таким образом, они соответствуют цепочке из 4 вершин в молекулярном графе. Каждый атом в топологическом торсионном углу описывается типом атома (который соответствует типу химического элемента), числом присоединенных неводородных атомов и числом пар π -электронов. Молекулярные дескрипторы, показывающие присутствие либо отсутствие топологических торсионных углов в химических структурах, были использованы для качественного прогнозирования биологической активности в исследованиях SAR [189]. Кирсли (Kearsley) и др. [190] осознали, что описание типа атома посредством типа химического элемента во многих случаях является чересчур специфичным и не может обеспечить достаточной гибкости, необходимой для поиска по подобию и основанного на нем широкомасштабного виртуального скрининга. В связи с этим, они предложили проводить типизацию атомов в топологических торсионных углах Нилаконтана (а также в атомных парах Кархарта, см. ниже) путем отнесения каждого из атомов к одному из семи классов: катионов, анионов, нейтральных доноров водородной связи, нейтральных акцепторов водородной связи, полярных атомов, гидрофобных атомов и др.

Четыре вышеупомянутых типа структурных фрагментов, а именно атомы, связи, перекрестные члены в основанных на связях аддитивных схемах и топологические торсионные углы, являются, с точки зрения теории графов, *цепочками* разной длины. Идея использовать число вложений цепочек разной длины в молекулярные графы в качестве дескрипторов при построении моделей «структура-свойство» была впервые предложена в 1964 г. Е.А.Смоленским [128], который показал, что энтальпия образования алканов может быть представлена как линейная комбинация чисел вложения цепочек длиной до четырех вершин (атомов) в молекулярный граф, и обосновал это с точки зрения квантовой теории. Многочисленные работы, опубликованные за последние 40 лет, свидетельствуют о том, что цепочечные структурные фрагменты являются одним из самых популярных, мощных и полезных типов фрагментов в исследова-

ниях QSPR/QSAR/SAR. И действительно, на их использовании основана работа таких компьютерных программ, предназначенных для проведения исследований QSPR/QSAR/SAR, как: CASE [156-159] Клопмана (Klopman); MULTICASE (MultiCASE, MCASE) [191, 192] Клопмана (Klopman); дескрипторный блок FRAGMENT [193] (см. разделы 5.1 и 8.3), входящий в состав программного комплекса NASAWIN [194] (см. раздел 8.2), разработанного под руководством Н.С.Зефирова на химическом факультете МГУ; программа BIBIGON [195], созданная М.И.Кумсковым с соавт.; программные комплексы TRAIL [196, 197] и ISIDA [114], разработанные В.П.Соловьевым и А.Варнеком (Varnek).

Помимо многочисленных приложений, связанных с использованием вышеупомянутых программ, цепочечные фрагменты под разными именами встречаются в ряде других исследований. В этой связи можно упомянуть *молекулярные пути (molecular pathways)* Гакха (Gakh) с соавт. [198], *молекулярные маршруты (molecular walks)* Рюкера (Rücker) [199] и др.

В отличие от цепочек, циклические и полициклические фрагменты относительно редко *в явном виде* используются в исследованиях QSAR/QSPR. Похоже, что вышеупомянутый дескрипторный блок FRAGMENT [193] является единственной программой, систематически работающей с фрагментными дескрипторами этого типа. Тем не менее, *в неявном виде* циклические фрагменты вовлечены во многочисленные исследования посредством: (а) введения специальных «циклических» и «ароматических» типов атомов и связей; (б) «сворачивания» целых циклов и даже полициклических систем в «фармакофорные» псевдоатомы; (iii) генерации циклических фрагментов как частного случая других более общих типов фрагментов, в частности, фрагментов на основе максимальных общих подграфов (см. пункт 2.2.1.5), заместителей и молекулярных остовов (см. пункт 2.2.1.7), фрагментов на основе базисных (см. пункт 2.2.1.8), «добытых» (см. пункт 2.2.1.9) и случайных (см. пункт 2.2.1.10) подграфов, а также библиотечных фрагментов (см. пункт 2.2.1.11). Кроме того, циклические фрагменты широко используются в качестве скринов при работе с химическими базами данных [200, 201].

2.2.1.2. Фрагменты WLN и SMILES

Фрагменты WLN и SMILES соответствуют подстрокам (обычно длиной в один символ) строк WLN (Wiswesser Line Notation – линейная нотация Висвессера) [202] либо SMILES (Simplified Molecular Input Line Entry System) [203, 204], которые активно используются для кодирования структур органических соединений. В структурном плане, односимвольные фрагменты WLN и SMILES представляют собой атомы либо простейшие группы, и поэтому основная отличительная особенность этого типа структурных фрагментов заключается в способах его обработки на компьютере. Поскольку простейшие операции над строками значительно более эффективны по сравнению с операциями на графах, использование дескрипторов WLN было вполне оправдано в 1970-ые годы в эпоху медленных компьютеров. В это время Адамсон (Adamson) и Бауден (Bawden) опубликовали целый ряд работ, в которых фрагментные дескрипторы WLN использовались в исследованиях QSAR и QSPR в сочетании со статистическим анализом при помощи аппарата множественной линейной регрессии [138, 140, 141, 205, 206]. Эти же авторы применили фрагментные дескрипторы WLN для проведения иерархического кластерного анализа и автоматической классификации химических структур [207]. В дальнейшем линейная нотация Висвессера была существенно усовершенствована, и на ее основе Ку (Qu) с соавт. разработал новую линейную нотацию AES (Advanced Encoding System), специально предназначенную для того, чтобы быть использованной в качестве химического языка для кодирования информации специально для методов групповых вкладов. Таким образом, хотя фрагментные дескрипторы WLN (как и сама линейная нотация Висвессера) сейчас могут казаться устаревшими, они сыграли важную историческую роль в развитии методологии исследований QSAR/QSPR, основанных на применении фрагментных дескрипторов.

Тем не менее, интерес к дескрипторам, построенных на основе линейных нотаций, полностью не исчез с приходом мощных компьютеров и со снижением интереса к нотации Висвессера как таковой. Работы, основанные на использовании фрагментных дескрипторов SMILES, все еще продолжают появляться.

В качестве характерных примеров можно привести программу SMILGP для предсказания константы распределения в системе октанол-вода $\log P$ [182] и недавно разработанную систему LINGO для расчета биофизических свойств и оценки межмолекулярного сходства на основе голографического представления канонических строк SMILES [208].

2.2.1.3. Центрированные на атомах фрагменты

Центрированные на атомах фрагменты (ЦАФ) состоят из центрального атома, окруженного одной или несколькими оболочками атомов, находящихся на одинаковом топологическом расстоянии от него. Исторически этот тип структурных фрагментов был впервые введен в практику исследований «структура-свойство» В.М.Татевским в начале 1950-ых годов [123, 209] при разработке основанных на атомах аддитивных схем для предсказания физико-химических свойств органических соединений, что потребовало разработки многоуровневой системы классификации атомов, которая эквивалентна использованию ЦАФ. Развивая эти идеи дальше, Н.Ф.Степанов с соавт. продемонстрировали на многочисленных примерах оптимальность рассмотрения соседства первого уровня для классификации атомов [210], что эквивалентно рассмотрению ЦАФ с одной оболочкой атомов вокруг центрального (т.е. ЦАФ с радиусом 1). Очень схожие идеи были также выдвинуты в конце 1950-ых годов Бенсоном (Benson) и Бассом (Buss) [126], которые в явном виде использовали ЦАФ с радиусом 1 при разработке аддитивных схем для оценки термодинамических свойств химических соединений (см. обзорную статью [211]).

ЦАФ радиуса 1 были введены в практику исследований в области хемоинформатики под названиями ЦАФ (*atom-centered fragments*) и «расширенные атомы» (*augmented atoms*) в 1971 г. Адамсоном (Adamson) [212, 213], который изучал их распределение в больших химических базах данных с целью определения преимуществ их использования в качестве скринов для подструктурного поиска. Аналогичные ЦАФ радиуса 1 были переизобретены Хоудсом (Hodes) и введены им в практику проведения исследований SAR под названием «расши-

ренные атомы» [135]. Наряду с ними, Хоудс также предложил использовать «ганглии-расширенные атомы» (*ganglia augmented atoms*), которые дополнительно учитывают связи между атомами первой и второй оболочек [214] и поэтому могут быть представлены как ЦАФ радиуса 2 с обобщенными атомами во второй оболочке. ЦАФ радиуса 1 были в дальнейшем также интегрированы в дескрипторный блок FRAGMENT [193] под названием «разветвленные фрагменты» (*branched fragments*). ЦАФ произвольного радиуса были независимо предложены и реализованы несколькими группами авторов: (а) В.П.Соловьевым и Варнеком (Varnek) в программах TRAIL [196, 197] и ISIDA [114] под именем «расширенные атомы» (*augmented atoms*); (б) Д.А.Филимоновым, В.В.Поройковым с соавт. в программе PASS [215] под именем «многоуровневые атомные окрестности» *Multilevel Neighborhoods of Atoms (MNA)* [216]; (в) Ксингом (Xing) и Гленом (Glen) под именем «структурированные по дереву отпечатки пальцев» (*tree structured fingerprints*) [217] (которые, однако, в дальнейших публикациях Бендера (Bender), Глена (Glen) и др. называются «атомными окрестностями» (*atom environments*) [218, 219] и «циркулярными отпечатками пальцев» (*circular fingerprints*) [220-222], см. Рис. 12); (г) Фолоном (Faulon) под названием «молекулярные подписи» (*molecular signatures*) [223-225].

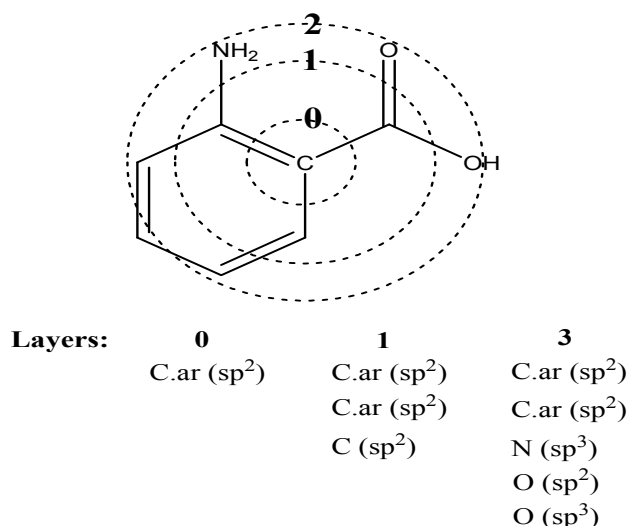


Рис. 12. «Циркулярные отпечатки пальцев» вместе с типами атомов, используемыми в mol2-файлах программного комплекса Sybyl. Индивидуальный фрагментный дескриптор вычисляется для каждого атома в молекуле с учетом атомов, отстоящих от него не больше, чем на две связи

Некоторые типы ЦАФ были первоначально разработаны для хранения в спектральных базах данных локальной (т.е. относящейся к отдельным атомам) спектральной информации, например, значений химического сдвига. Бремсер (Bremser) разработал систему подструктурных кодов, названную как «иерархически-упорядоченное сферическое окружение» (*Hierarchically Ordered Spherical Environment* (HOSE)), чтобы охарактеризовать сферическое окружение как отдельных атомов, так и целых циклических систем [226]. Эти коды генерировались автоматически из топологических представлений химических структур и служили для описания структурных контекстов для спектральных параметров (в частности, химических сдвигов). Очень близкая идея была воплощена Дюбуа (Dubois) и др. в системе DARC под именем *FREL* (Fragment Réduit à un Environment Limité – фрагмент, редуцированный до ограниченного окружения) [227, 228]. Ксяо (Xiao) с соавторами также использовали ЦАФ под названием «центрированные на атомах многоуровневые коды» (*Atom-Centered Multilayer Code* (ACMC)) для проведения структурного и подструктурного поиска в больших базах данных по химическим структурам и реакциям [229].

Одно из важных недавних приложений ЦАФ касается предсказания мишеней (*target fishing*) для данного органического соединения в хемогеномике [215, 230, 231].

2.2.1.4. Центрированные на связях фрагменты

Центрированные на связях фрагменты состоят из центральной связи, двух атомов, ею соединенных и окруженных одним или несколькими слоями атомов, находящихся на одинаковом топологическом расстоянии от ближайшего из этих двух атомов. В отличие от ЦАФ, они довольно редко используются в исследованиях SAR/QSAR/QSPR, однако они могут быть эффективно применены в качестве скринов при работе с химическими базами данных, что было продемонстрировано Адамсоном (Adamson) с соавт. [232]. Фрагменты этого типа входят в состав ключей MDL [233, 234], которые нашли применение как для

организации подструктурного поиска, кластеризации баз данных [150], так и для проведения исследований SAR [152]. Центрированные на связях фрагменты применялись также в системе DARC [227, 228].

2.2.1.5. Фрагменты на основе максимальных общих подграфов

Максимальный общий подграф (МОП) для множества графов определяется как подграф, который содержится во всех графах этого множества, но не содержится ни в одном другом МОП. В большинстве практически важных приложениях МОП определяется только для пар графов, т.е. для множеств, состоящих из двух графов. МОП могут быть найдены при помощи процедуры *пересечения* графов с использованием множества различных алгоритмов (см. обзор [235]), наиболее известный из которых состоит в поиске клик т.н. графов совместимости. Следует, однако, принять во внимание, что для пары графов может существовать несколько МОП. Основное преимущество использования МОП в качестве фрагментов в исследованиях SAR/QSAR/QSPR состоит в том, что разнообразие их строения ничем искусственно не ограничено, и поэтому с их помощью могут быть найдены ответственные за целевые свойства структурно сложные фрагменты, которые никак не могли бы попасть в поле зрения при рассмотрении только фрагментов, относящихся к какому-нибудь структурно однородному типу, такому как цепочки, циклы, ЦАФ и др..

Впервые фрагменты МОП были использованы в исследованиях SAR в начале 1980-ых годов А.Б. Розенблитом и В.Е. Голендером в рамках разработанного ими логико-комбинаторного подхода [130, 131, 236]. Поскольку в то время компьютеры были очень медленными, в практических приложениях авторам, однако, пришлось ограничиться операциями над редуцированными графами (см. обсуждение ниже), построенных на фармакофорных центрах. Следующий этап в применении фрагментов МОП в этой области относится к началу 1990-ых гг., когда их стали использовать для вычисления химического расстояния и проведения поиска по подобию [237]. В последнее время фрагменты МОП стали применять для кластеризации химических баз данных [238, 239] и,

кроме того, они снова стали использоваться для прогнозирования биологической активности органических соединений [194, 240, 241].

2.2.1.6. Атомные пары и топологические мультиплеты

Этот тип фрагментных дескрипторов был специально разработан для проведения исследований SAR для фармакологически важных свойств органических соединений. В его основе лежит понятие о *дескрипторных центрах*, под которыми подразумеваются атомы либо группы атомов, которые могли бы служить центрами межмолекулярных взаимодействий. Обычно в качестве дескрипторных центров берутся гетероатомы, ненасыщенные связи и ароматические циклы. Вторым важным элементом в спецификации этого типа дескрипторов является *расстояние* между дескрипторными центрами, под которым обычно подразумевают топологическое расстояние между атомами в химической структуре, либо кратчайшее расстояние между атомами, принадлежащими двум группам. В этом контексте, *атомная пара* определяется как пара дескрипторных центров вместе с расстоянием между ними. По аналогии с этим, *топологический мультиплет* определяется как мультиплет (обычно триплет) дескрипторных центров наряду с набором расстояний для каждой из их пар. Дескрипторы, относящиеся к этой категории, принимают обычно бинарные значения, указывающие на присутствие либо отсутствие соответствующих фрагментов в химической структуре. Таким образом, атомные пары являются частным случаем топологических мультиплетов. С позиций теории графов, атомные пары представляют собой цепочки со специфицированными типами терминальных вершин и обобщенными типами внутренних вершин и ребер в молекулярном графе. Топологические мультиплеты, однако, требуют более сложного описания с позиций теории графов.

Атомные пары впервые были введены в практику проведения исследований SAR В.В. Авидоном и названы им *фрагментарными кодами суперпозиции подструктур (ФКСП)* [131, 242]. В дальнейшем сходные дескрипторы были предложены Кархартом (Carhart) с соавт. [243], которые их использовали для

численной оценки сходства органических соединений, а также для проведения исследований SAR при помощи анализа тренд-векторов. В отличие от ФКСП, атомные пары Кархарта используют в качестве центров не только дескрипторные центры, но и все остальные неводородные атомы, которые классифицированы с учетом типа химического элемента, числа неводородных соседей и количества π -электронов. В настоящее время атомные пары Кархарта являются одним из самых распространенных фрагментных дескрипторов для проведения виртуального скрининга с целью поиска новых биологически активных соединений.

Как дальнейшее развитие дескрипторов этого типа, Хорват (Horvath) ввел топологические нечеткие биполярные фармакофорные автокоррелограммы (*Topological Fuzzy Bipolar Pharmacophore Autocorrelograms*) [244], которые можно представить как атомные пары Кархарта, в которых реальные атомы заменены на фармакофорные центры (классифицированные как гидрофобные, ароматические, акцепторы водородной связи, доноры водородной связи, катионы и анионы), тогда как топологическое расстояние между ними может принимать несколько близких значений вместо одного фиксированного. Эти дескрипторы были с успехом применены при проведении виртуального скрининга для 42 биологических мишеней с использованием поиска по подобию и нескольких четких и нечетких метрик [245], причем по эффективности использования фрагментные дескрипторы данного типа лишь очень незначительно уступают 3-мерным аналогам [244]. *Нечеткие фармакофорные триплеты* (*Fuzzy Pharmacophore Triplets*) были предложены Хорватом (Horvath) [246] как расширение топологических нечетких биполярных фармакофорных автокоррелограмм на случай трех фармакофорных центров. Важным нововведением в этом типе дескрипторов явился учет протеолитического равновесия как функции от pH среды [246]. Благодаря этой особенности, эти дескрипторы в ряде случаев оказались способными проводить эффективную дискриминацию между структурно близкими соединениями со значительно отличающимися значениями биологической активности [246].

Следует упомянуть также и другие типы топологических триплетов. В частности, *фармакофорные ключи Similog (Similog pharmacophoric keys)*, предложенные Шуффенхауэром (Schuffenhauer) и др. [247], состоят из триплетов бинарно закодированных типов атомов (фармакофорных центров) и топологических расстояний между ними. Тип атома кодируется при этом 4 битами, соответствующими следующим свойствам атома: потенциальный донор водородной связи, потенциальный акцептор водородной связи, объемность и “электроположительность” (electropositivity) (см. Рис. 13). *Топологические фармакофорные треугольники (topological pharmacophore-point triangles)*, реализованные в программном комплексе MOE [248], представляют собой триплеты атомных типов MOE, разделенные несколькими дискретными значениями топологического расстояния. Модели QSAR, полученные при помощи этих дескрипторов и аппарата «машин опорных векторов», с успехом были использованы при проведении виртуального скрининга при поиске ингибиторов циклооксигеназы-2 [249] и лиганд D₃-дофаминового рецептора [250].

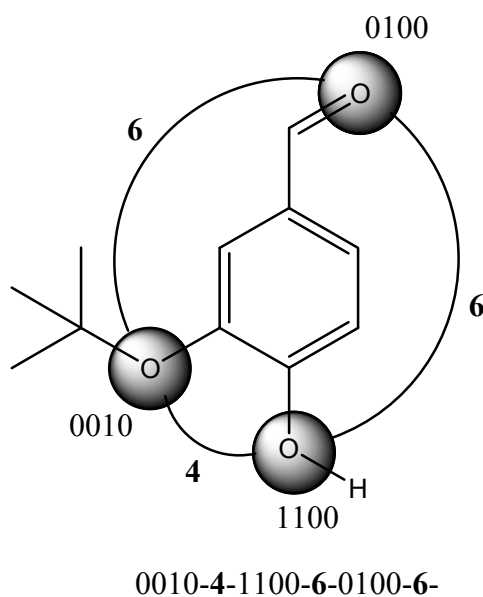


Рис. 13. Пример фармакофорных ключей Similog

2.2.1.7. Заместители и молекулярные остовы

С самого начала применения структурной теории для описания строения органических соединений декомпозиция молекул на заместители и молекулярный остов, к которому они присоединены, всегда воспринималась естествен-

ной. В историческом плане, анализ заместителей первым вошел в практику проведения исследований QSAR. Хотя на вышеупомянутом разложении основаны два классических подхода в QSAR, метод Ханча-Фуджиты (Hansch-Fujita) [251, 252] и метод Фри-Вильсона (Free-Wilson) [129], только второй из них основан на фрагментных дескрипторах, значения каждого из которых показывает наличие либо отсутствие определенного заместителя в определенном положении молекулярного остова. Пользуясь языком теории графов, подструктурные фрагменты метода Фри-Вильсона можно описать как молекулярные графы, включающие в свой состав граф заместителя и граф остова, соединенные между собой ребром. Эти бинарные дескрипторы традиционно используются в методе Фри-Вильсона в сочетании со множественным линейным регрессионным анализом, хотя последние модификации этого подхода включают использование более современных статистических методов (методов машинного обучения), таких как анализ главных компонент [253] и нейронные сети [254].

В отличие от дескрипторов, вычисляемых для заместителей, дескрипторы, описывающие строение молекулярных остовов, редко в явном виде используются в исследованиях SAR/QSAR/QSPR. Возможно, наиболее известный пример их неявного использования в исследованиях QSAR/QSPR включает использование индикаторных переменных, дискриминирующих между различными типами молекулярных остовов. Концепция молекулярных остовов и заместителей (боковых цепей) была подробно рассмотрена Бемисом (Bemis) и Мурко (Murcko) [255, 256], изучавших их распределение среди лекарств.

2.2.1.8. Фрагменты на основе базисных подграфов

Поскольку имеется огромное множество молекулярных графов, легко можно представить, что существует по крайней мере не меньшее множество подструктурных фрагментов и соответствующих фрагментных дескрипторов. Поэтому было бы очень перспективно найти такое относительно небольшое подмножество фрагментных дескрипторов, с помощью которого можно было бы аппроксимировать любое свойство. Эта идея лежит в основе концепции ба-

зисных графов, предложенной Рандичем (Randić) в 1992 г. [257], который уподобил разложение молекулярных графов по базису первичных графов (и, следовательно, разложение любого фрагментного дескриптора по базисным фрагментным дескрипторам) разложению векторов по базису векторного пространства. В цитированной работе Рандич предлагает использовать несвязанные графы, состоящие из нескольких цепочек разной длины, в качестве набора таких базисных подграфов (см. Рис. 14).

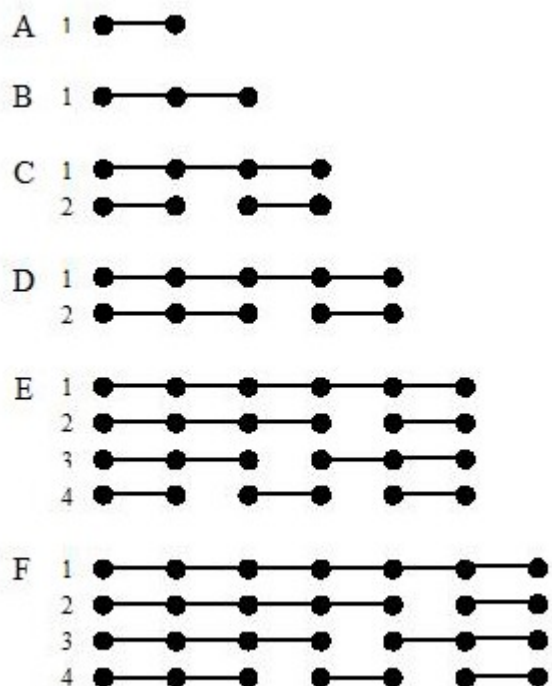


Рис. 14. Базисные подграфы Рандича для максимального числа вершин 7

Тем не менее, для случая базисных подграфов Рандича оказывается возможным найти такие примеры, когда различные структуры содержат одни и те же наборы базисных подграфов. Следовательно, такие базисные подграфы нельзя рассматривать как базисные в строгом смысле этого слова. Следует ответить, что строгое решение проблемы нахождения базисного набора инвариантов графов было найдено в 1983 г. для случая простых графов [258]. Этот результат был далее распространен на молекулярные графы И.И. Баскиным, М.И. Скворцовой с соавт. [259, 260] (см. раздел 3.2). Из этих работ, однако, следует, что полный набор базисных инвариантов графов строится на всех возможных подграфах, и поэтому невозможно ограничиться каким-либо небольшим их

подмножеством для получения дескрипторов, способных аппроксимировать все возможные свойства с любой точностью. Тем не менее, для многих задач на практике использование базисных подграфов (и соответствующих фрагментных дескрипторов) может оказаться очень полезным.

М.И. Скворцова, К.С. Федяев, И.И. Баскин и др. расширили набор базисных подграфов Рандича за счет включения как циклических фрагментов, так и составных фрагментов, состоящих из вершины, присоединенной к циклическому фрагменту [261] (этот материал не включен в данную диссертационную работу). Предложенный набор фрагментов обладает хорошей уникальностью (т.е. разные вектора дескрипторов кодируют разные структуры) и полнотой кодирования (т.е. они могут аппроксимировать большое число зависимостей структура-свойство). Базисные фрагментные дескрипторы этого типа были использованы при построении ряда QSPR-моделей [262] (см. Рис. 15).

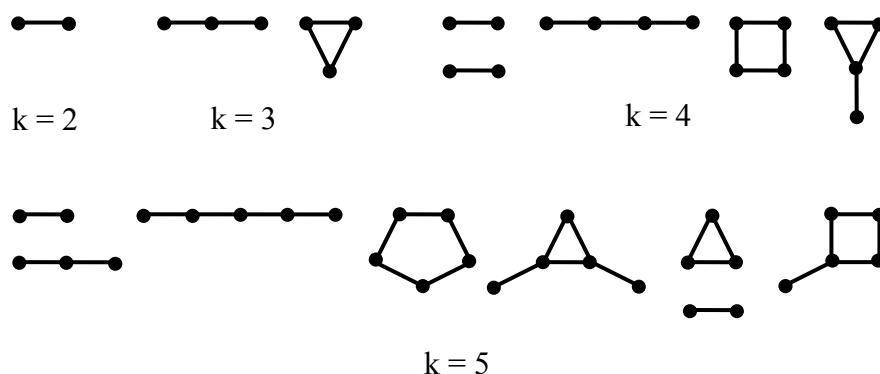


Рис. 15. Базисные подграфы Скворцовой для максимального числа вершин 5

Другим источником базисных подграфов являются результаты разложения инвариантов молекулярных графов по числам встречаемости базисных подграфов. Возможность подобного разложения следует из нескольких теоретико-графовых теорем [258, 259]. Эстрада (Estrada) развил эту методологию для *спектральных моментов* реберной матрицы смежности молекулярных графов, которые определяются как следы разных степеней такой матрицы [263-265]:

$$\mu_k = \text{tr}(E^k) \tag{61}$$

где: μ_k - это k -ый спектральный момент реберной матрицы смежности E (которая представляет собой квадратную и симметричную матрицу, элемент e_{ij} кото-

рой равен 1 только в том случае, если ребра i и j являются смежными); tr – след матрицы, т.е. сумма ее диагональных элементов. Оказывается, спектральные моменты могут быть представлены как линейные комбинации чисел встречаемости определенных связных структурных фрагментов в молекулярных графах (вышеупомянутые теоремы не гарантируют связность подграфов и простоту разложения для произвольных инвариантов молекулярных графов, и в этом, по-видимому, и заключается преимущество использования спектральных моментов в качестве таких инвариантов). Подобные линейные комбинации для простых молекулярных графов, не содержащих гетероатомов, табулированы для ациклических [263] и циклических [265] химических структур.

Для иллюстрации этого подхода рассмотрим приведенную в статье [263] корреляцию между температурой кипения алканов и спектральными моментами:

$$bp(^{\circ}C) = -76.719 + 23.992\mu_0 + 2.506\mu_2 - 2.967\mu_3 + 0.149\mu_5 \quad (62)$$

$$R = 0.9949, s = 4.21, F = 1650$$

Первые шесть спектральных моментов реберной матрицы смежности E следующим образом выражаются в виде линейных комбинаций чисел встречаемости фрагментов, приведенных на Рис. 16:

$$\mu_0 = |F_1| \quad (63)$$

$$\mu_2 = 2 \times |F_2| \quad (64)$$

$$\mu_3 = 6 \times |F_3| \quad (65)$$

$$\mu_4 = 2 \times |F_2| + 12 \times |F_3| + 24 \times |F_4| + 4 \times |F_5| \quad (66)$$

$$\mu_5 = 30 \times |F_3| + 120 \times |F_4| + 10 \times |F_6| \quad (67)$$

$$\mu_6 = 2 \times |F_2| + 60 \times |F_3| + 480 \times |F_4| + 12 \times |F_5| + 24 \times |F_6| + 6 \times |F_7| + 36 \times |F_8| + 24 \times |F_9| \quad (68)$$

где $|F_i|$ обозначает число встречаемости подграфа F_i в молекулярном графе.

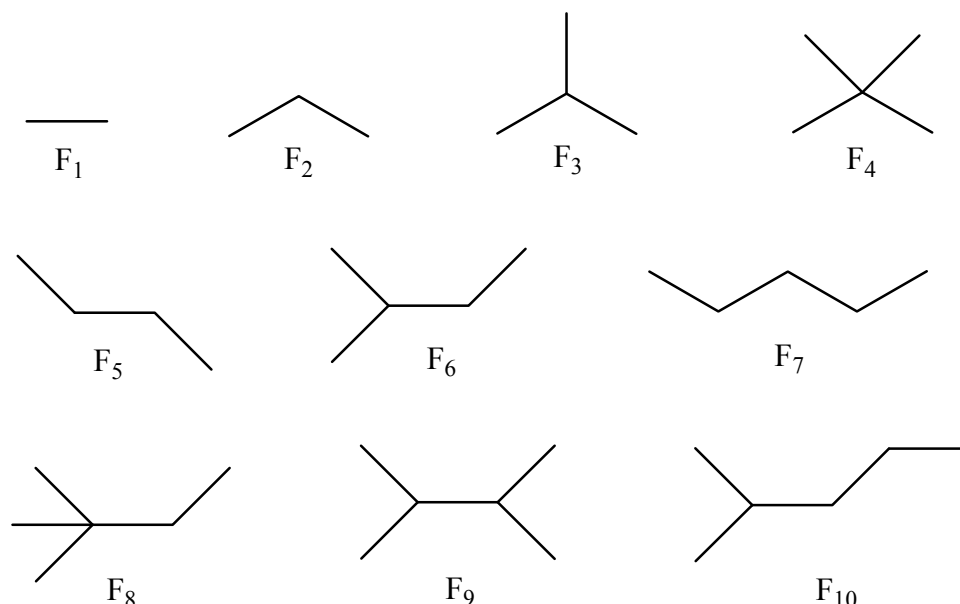


Рис. 16. Первые 10 структурных фрагментов, содержащихся в молекулярных графах алканов

Подставляя в уравнение QSPR (2.4) разложения спектральных моментов из уравнений (63)-(68), можно получить следующее уравнение QSPR, построенное на фрагментных дескрипторах:

$$bp(^{\circ}\text{C}) = -76.719 + 23.992|F_1| + 5.01|F_2| - 13.332|F_3| + 17.880|F_4| + 1.492|F_6| \quad (69)$$

В дальнейшем этот подход был распространен на молекулярные графы, содержащие гетероатомы, за счет введения весов на диагональных элементах реберной матрицы смежности [264]. В этом случае оказывается возможным оценить вклад любого фрагмента в спектральные моменты и, следовательно, в значения свойств/активности химических соединений. Эта методология легла в основу подхода *TOSS-MODE* (TOpological SubStructural MOlecular DEsign, который в дальнейшем был переименован как *TOPS-MODE* (TOPological Substructural MOlecular DEsign), т.е. *топологический подструктурный молекулярный дизайн* [266]. Этот подход был успешно применен для предсказания физико-химических свойств органических соединений (индексов удерживания в хроматографии [267], диамагнитных и магнитооптических свойств [268], дипольного момента [269], коэффициента проницаемости сквозь полиэтилен низкой плотности [270] и др.), пространственных характеристик структур [271], а также

множества различных типов биологической активности (седативной / гипнотической активности [266], противораковой активности [272], анти-ВИЧ активности [273], сенсibilизации кожи [274], гербицидной активности [275], сродства к A_1 -аденозиновому рецептору [276], ингибирования циклооксигеназы [277], антибактериальной активности [278], токсичности по отношению к *Tetrahymena pyriformis* [279], мутагенности [280-282] и др.). Во всех случаях окончательные модели были проанализированы с учетом вкладов, вносимых различными структурными фрагментами в значения свойств/активностей органических соединений.

2.2.1.9. Фрагменты на основе «добытых» (mined) подграфов

Понятие «добытых» подграфов (*mined subgraphs*) тесно связано *graph mining* (либо *subgraph mining*) – направлением в data mining (однозначного перевода «data mining» на русский язык не существует, наиболее удачный вариант – интеллектуальный анализ данных), направленным на нахождение таких графов (подграфов), которые были бы наиболее полезны для решения прикладных задач, в частности, в исследованиях SAR/QSAR/QSPR [283-288]. Преимущество этого подхода заключается в том, что в его рамках оказывается возможным осуществлять направленную генерацию только «полезных» графов (подграфов) без необходимости просмотра практически бесконечного числа всех возможных графов (подграфов). Эта методология [289, 290] обычно основывается на использовании эффективных алгоритмов генерации фрагментов, наиболее часто встречающихся в наборе графов (*frequent fragments*). В числе подобных алгоритмов упомянем: *AGM* (Apriori-based Graph Mining), разработанный Инокучи (Inokuchi) с соавт. [291]; *FSG* (Frequent Sub-Graphs), предложенный Курамоchi (Kuramochi) и Кариписом (Karypis) [292]; «алгоритм нахождения химических подструктур» (*chemical sub-structure discovery algorithm*), созданных Боргельтом (Borgelt) и Бертольдом (Berthold) [293]; *gSpan* (*graph-based Substructure pattern mining*), предложенный Яном (Yan) и Ханом (Han) [287]; *TreeMiner*, разработанный Заки (Zaki) [294]; *HybridTreeMiner*, предложенный Чи (Chi), Ян-

гом (Yang) и Мунтцем (Muntz) [295]; *CMTreeMiner* этих же авторов [296]. Первоначально этот подход использовался для классификации химических структур в рамках исследований SAR [297, 298]. Специальная модификация этого подхода с применением методики «добычи взвешенных подструктур» (*weighted substructure mining*) в сочетании со статистической процедурой linear programming boosting [299] позволяет строить количественные QSAR/QSPR регрессионные модели с использованием «добытых» фрагментных дескрипторов [288].

2.2.1.10. Фрагменты на основе случайных подграфов

Успех применения различных схем фрагментации в значительной степени зависит от начального выбора нужных типов фрагментов. Поскольку практически невозможно рассмотреть все возможные фрагменты из-за их гигантского числа, всегда приходится ограничиваться их небольшим подмножеством. К сожалению, любая попытка ограничиться каким-либо их фиксированным типом, например, только цепочками с заранее заданной максимальной длиной, чревата риском упустить из рассмотрения очень важные для решения данной задачи фрагменты. Одно из возможных решений этой проблемы состоит в использовании рассмотренных выше «добытых» (см. пункт 2.2.1.9) либо, для чисто классификационных задач, МОП-фрагментов (см. пункт 2.2.1.5). Альтернативой этому является использование стохастических процедур генерации подструктурных фрагментов.

Интересная работа в этом направлении была опубликована Грахамом (Graham) с соавт., которые получили «записи на ленту» (“tape recordings”) химических структур при помощи фрагментов атом-связь-атом, извлекаемых из молекулярных графов при помощи процедуры случайных блужданий (random walks) [300]. Для оценки структурного подобия химических соединений Батиста (Batista), Годден (Godden) и Байорат (Vajorath) разработали метод MolBlaster, основанный на генерации популяций фрагментов путем случайного удаления ребер в молекулярных графах [301]. Этот метод с успехом был использован

при проведении виртуального скрининга, основанного на поиске по подобию [302].

2.2.1.11. Библиотечные фрагменты

Во многих работах применяются фиксированные наборы фрагментов, взятых из библиотеки. Подобные библиотеки обычно содержат фрагменты, которые уже показали пользу своего использования при прогнозировании сходных свойств. Большинство аддитивных схем и методов группового вклада были разработаны на основе фиксированных наборов фрагментов. В некоторых исследованиях SAR/QSAR/QSPR также рассматриваются фиксированные наборы библиотечных фрагментов. В подобных случаях структуры фрагментов обычно задаются при помощи специального языка либо линейной нотации, специально созданных для описания списков фрагментов. В качестве характерных примеров можно привести: (а) экспертную систему DEREK, предназначенную для предсказания токсичности органических соединений, в которой используется для описания фрагментов специальный язык PATRAN [303]; (б) систему прогнозирования коэффициента распределения в системе октанол-вода $\log P$, в которой для кодирования фрагментов использован язык программирования Prolog [304]; (в) метод ALogP [180] для прогнозирования этого же свойства, основанный на использовании линейной нотации SMARTS line notation (реализованной в программном комплексе MOE (Molecular Operating Environment) [248]) для спецификации фрагментов.

2.2.2. Классификация по типам молекулярных структур

Молекулярные графы могут быть использованы для описания не только обычных молекулярных структур, но и супрамолекулярных комплексов, химических реакций, полимеров с периодической структурой и других видов химических объектов. Во всех этих случаях фрагментные дескрипторы могут быть применены для представления их структур.

Образование супрамолекулярных комплексов обычно характеризуется большим разнообразием разных типов взаимодействия между их компонентами, в частности: σ -донорно-акцепторные и π -дативные взаимодействия между металлами и лигандами, образование водородных связей, электростатические взаимодействия, образование солевых мостиков, π - π -стэкинг и π -катионные взаимодействия, ван-дер-Ваальсовы и гидрофобные взаимодействия и т.д. Некоторые из этих взаимодействий могут быть представлены специальными связями в молекулярных графах. Например, можно ввести специальную «координационную связь» для описания донорно-акцепторных и/или дативных взаимодействий между центральным атомом металла в комплексе и донорными атомами в лигандах. Также можно провести специальную «водородную связь» между атомом водорода и атомом-акцептором водородной связи, либо между донором и акцептором водородной связи в «безводородных» супрамолекулярных графах. Результирующие супрамолекулярные графы можно использовать вместо обычных молекулярных графов для вычисления фрагментных дескрипторов, и единственное отличие будет заключаться в кодировании «супрамолекулярных связей».

Концепция молекулярных графов также может быть распространена на описание химических реакций, в особенности в области органической химии (см. обзор [305]). Эта задача решается путем введения специальных типов связей, соответствующих образованию либо разрыву обычных химических связей, либо изменению их порядка. Результирующий реакционный граф содержит всю необходимую информацию, чтобы реконструировать левую и правую части соответствующего уравнения реакции. В качестве примера, рассмотрим реакцию Дильса-Альдера, изображенную на Рис. 17. Метка связи 01 в реакционном графе соответствует образованию простой связи; метка 21 соответствует превращению двойной связи в простую, тогда как метка 12 описывает превращение простой связи в двойную.

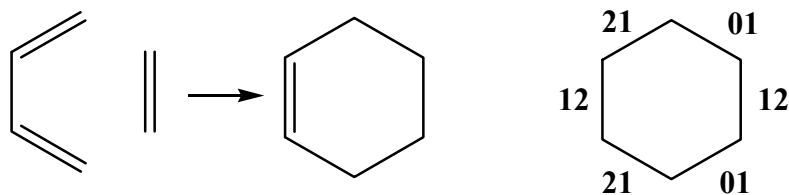


Рис. 17. Реакция Дильса-Адлера и соответствующий реакционный граф

Первые аналоги частичных реакционных графов, содержащие только связи, *подвергающиеся изменению* (т.е. разрывающиеся, образующиеся и меняющие свою кратность) в результате реакции, ранее были использованы для классификации и перечисления типов органических реакций в рамках матричного формализма Уги-Дугунджи (Ugi-Dugundji) [306] и формально-логического подхода Н.С. Зефирова и С.С. Трача [307, 308]. Владуц (Vladutz) [309] объединил структуры реагентов и продуктов реакции в единый граф, содержащий специальные метки для обозначения связей, подвергающихся изменению в результате реакции. Результирующий реакционный граф, называемый *суперпозиционным графом реакционного скелета* (*Superimposed Reaction Skeleton Graph*) может также содержать связи (вместе со смежными им атомами), не подвергающиеся изменениям в результате реакции. Сходные реакционные графы были также предложены Фуджитой (Fujita) [310, 311] и названы им *мнимыми переходными состояниями* (*imaginary transition states*), которые были использованы автором для классификации и перечисления типов органических реакций.

Следующий этап в развитии концепции реакционных графов был направлен на разработку фрагментных дескрипторов, которые могли бы быть использованы для кодирования и предсказания реакций органических соединений. Ю. Бородина с соавт. предложили расширение дескрипторов MNA, получившее название *RMNA (Reacting Multilevel Neighborhood of Atom)*, для кодирования метаболических превращений молекул [312]. В этом подходе отдельные описания субстратов и продуктов реакций при помощи MNA-дескрипторов дополнены информацией о связях, подвергающихся изменениям в результате трансформации. Дескрипторы RMNA были успешно применены для предсказания сайтов метаболического гидроксилирования с участием цитохрома-P450 [312].

Концепция реакционных графов, сходная с концепцией мнимого переходного состояния Фуджиты (Fujita) [310, 311], была недавно выдвинута Варнеком (Varnek) с соавт., предложившими использовать *конденсированные графы реакций (Condensed Graphs of Reactions)* [114]. В отличие от обычных молекулярных графов, конденсированные графы реакций содержат специальные ребра для обозначения химических связей, претерпевающих изменения в результате реакции. Конденсированные графы реакции могут быть использованы для генерации фрагментных дескрипторов точно таким же образом, как это делается для обычных молекулярных графов.

2.2.3. Классификация по типам значений дескрипторов

Обычно рассматриваются два типа значений, принимаемых фрагментными дескрипторами – бинарные и целочисленные. Бинарные значения показывают наличие (*true, yes, 1*) либо отсутствие (*false, no, 0*) данного фрагмента в химической структуре. Хотя первоначально бинарные фрагментные дескрипторы использовались главным образом в качестве скринов либо элементов «молекулярных отпечатков пальцев» (см. подробное обсуждение в разделе 2.2.4) для работы с химическими базами данных, в последнее время все чаще их стали применять для прогнозирования биологической активности, а также для проведения виртуального скрининга с использованием как поиска по подобию, так и вероятностных подходов SAR. Целочисленные значения фрагментных дескрипторов показывают, сколько раз соответствующий фрагмент встречается в химической структуре. Обычно они используются для прогнозирования физико-химических свойств (реже, биологической активности) органических соединений.

Возникает естественный вопрос: могут ли фрагментные дескрипторы принимать другие типы значений? Вероятный ответ: в принципе, да, но в этом случае они называются топологическими индексами. Например, индексы связности Кира-Холла (Kier-Hall) [313] формально могут быть представлены как фрагментные дескрипторы, значения которых равны суммам произведений оп-

ределенных атомных характеристик внутри фрагмента. Другой пример - это рассматриваемые в рамках данной диссертационной работы псевдофрагментные дескрипторы (см. раздел 5.4), которые более тесно связаны с фрагментными дескрипторами по сравнению с типичными топологическими индексами.

2.2.4, Классификация по типам дескрипторных наборов

Набор фрагментных дескрипторов, рассчитанных для химического соединения, может быть организован тремя основными способами, а именно, в виде: (а) *векторов* фиксированного размера; (б) *списков*; и (в) *хеш-таблиц*.

Чаще всего набор значений фрагментных дескрипторов, рассчитанный для химического соединения, помещают в одномерный массив фиксированного размера (т.е. вектор), каждая ячейка (элемент) которого соответствует определенному подструктурному фрагменту, а содержащееся в ней значение – значению соответствующего фрагментного дескриптора. Вектор, содержащий бинарные значения фрагментных дескрипторов, называется набором *структурных ключей* (*structural keys*), которые в контексте работы с базами данных называются также *скринами* (*screans*) (см. Рис. 18). Поскольку структурные ключи хранятся в памяти компьютера в виде битовых строк, все операции с ними осуществляются очень эффективно, и именно это обуславливает популярность их использования для работы с базами химических данных, поиска по подобию, построения моделей SAR/QSAR, а также для осуществления с их помощью виртуального скрининга.

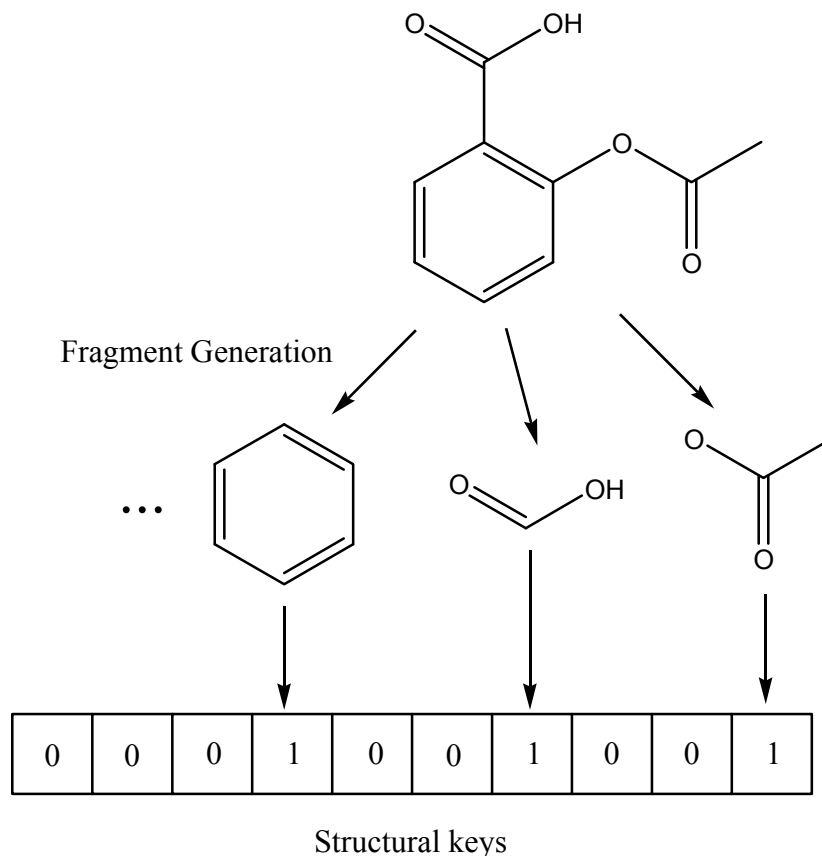


Рис. 18. Генерация структурных ключей для молекулы аспирина

Хотя структурные ключи хорошо зарекомендовали себя как эффективный инструмент исследований, однако успех их применения в значительной степени зависит от начального выбора набора фрагментов. Оказывается, что структурные ключи, построенные на фрагментах даже сравнительно небольшого размера, являются очень разреженными (т.е. они содержат лишь небольшую часть ненулевых элементов), а компьютерная обработка таких сильно несбалансированных наборов данных значительно менее эффективна по сравнению со сбалансированными. Как частичное решение этой проблемы, сгенерированные для химической структуры фрагментные дескрипторы могут быть организованы в виде списка, содержащего либо коды фрагментов (для бинарных дескрипторов) либо пары «код фрагмента – значение дескриптора» (для целочисленных фрагментных дескрипторов). Хотя такой способ представления разреженных массивов является эффективным с позиций использования памяти компьютера, однако он неэффективен с точки зрения времени, необходимого для доступа к его

элементам, что может быть очень принципиально при обработке больших баз химических данных.

Третий способ организации наборов фрагментных дескрипторов состоит в использовании *хеш-таблиц* (*hash tables*), под которыми понимается структура данных, позволяющая ассоциировать *ключи* (*keys*) с соответствующими *значениями* (*values*) [314, 315]. Хеш-таблицы позволяют эффективно осуществлять операцию *поиска по таблице* (*lookup*), которая для данного ключа (в данном случае, кода фрагмента) находит соответствующее значение (в данном случае, значение фрагментного дескриптора). Операция поиска по таблице осуществляется путем преобразования ключа (в данном случае, кода фрагмента) при помощи *хеш-функции* (*hash function*) в *хеш-код* (*hash code*), т.е. целое число, используемое как индекс в массиве, позволяющем определить местоположение *участка памяти* (*bucket*), содержащего искомое значение. Таким образом, оказывается возможным найти значение фрагментного дескриптора путем преобразования кода фрагмента (в сущности, имени фрагментного дескриптора) в хеш-код, позволяющий найти положение элемента массива, содержащего все значения дескрипторов. Однако, поскольку размер такого массива обычно значительно меньше максимального возможного значения хеш-кода, положение искомого элемента массива находят как остаток от деления хеш-кода на размер массива. Эта операция, к сожалению, может приводить к *столкновениям данных* (*collisions*), когда разные ключи указывают на один элемент массива, однако теория программирования указывает на эффективные способы *устранения столкновений данных* (*collision resolution*) за счет некоторого усложнения структур данных и алгоритмов работы с ними

Интересная модификация хеш-таблицы, предназначенная для проведения исследований QSAR, т.н. *молекулярная голограмма* (*molecular hologram*), разработана для целочисленных фрагментных дескрипторов [153]. Для получения молекулярной голограммы для химического соединения каждый найденный внутри химической структуры фрагмент кодируется при помощи линейной нотации SLN (SYBYL Line Notation) [316], потом код фрагмента переводится в 32-битный хэш-код, называемый *fragment integer ID* при помощи алгоритма

CRC [317]. (см. Рис. 19). После этого фрагмент помещается в определенную ячейку (*bin*) молекулярной голограммы, положение которой (*bin ID*) вычисляется как остаток от деления хеш-кода (*fragment integer ID*) на размер (т.е. количество ячеек) молекулярной голограммы. Каждый раз при нахождении очередного вхождения фрагмента в химическую структуру заселенность соответствующей ячейки молекулярной голограммы увеличивается на единицу. В отличие от стандартных хеш-таблиц, в молекулярных голограммах столкновения данных не устранены, и поэтому несколько разных фрагментов могут отобразиться на одну ячейку молекулярной голограммы. Следовательно, в результате анализа химической структуры общая заселенность ячейки молекулярной голограммы оказывается равной сумме целочисленных значений дескрипторов, соответствующим фрагментам, на нее отображаемым. Молекулярные голограммы легли в основу *голографического QSAR (holographic QSAR - HQSAR)* [153], в котором заселенности ячеек молекулярной голограммы выступают в качестве дескрипторов, корреляция которых с числовым значением биологической активности строится при помощи метода частичных наименьших квадратов PLS.

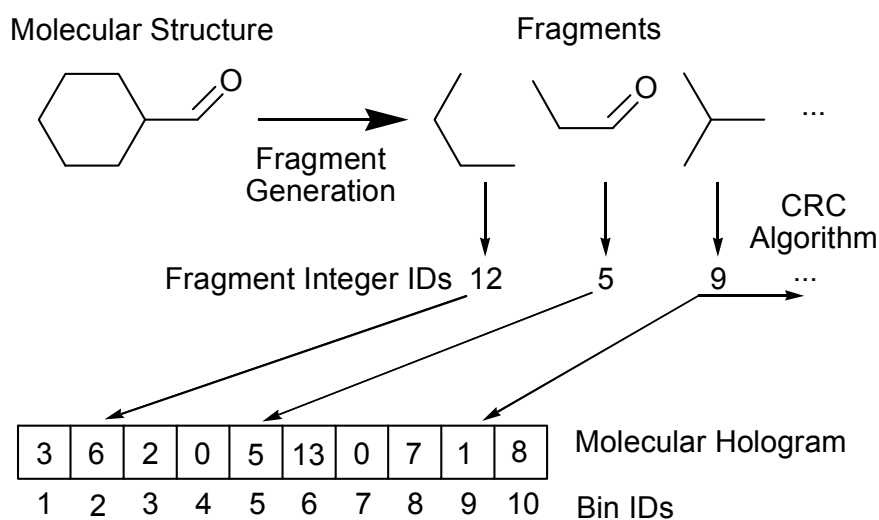


Рис. 19. Генерация молекулярной голограммы

По своей природе молекулярные голограммы очень близки к хешированным молекулярным отпечаткам пальцев (*hashed molecular fingerprints*) (или просто *молекулярным отпечаткам пальцев (molecular fingerprints)*), однако построены на основе бинарных фрагментных дескрипторов, показывающих лишь

наличие или отсутствие данного фрагмента в химической структуре. Также, в отличие от молекулярных голограмм, при построении молекулярных отпечатков пальцев каждый фрагмент может отображаться на несколько ячеек молекулярной голограммы, положения которых вычисляются при введении хеш-кода как заправки для генератора псевдослучайных чисел. Для увеличения информационной плотности (которая зависит от соотношений битов “on” и “off”), молекулярные отпечатки пальцев могут быть получены при помощи процедуры *сворачивания (folding)*, при которой каждый молекулярный отпечаток пальцев делится пополам, и две получившиеся половины комбинируются при помощи логической операции ИЛИ. Преимущество хешированных молекулярных отпечатков пальцев заключается в возможности использовать большое число дескрипторов для описания химической структуры. Недостаток же их связан с тем, что в них столкновения данных не устраняются (см. обсуждение выше). Тем не менее, в некоторых случаях этот недостаток может быть частично устранен путем подбора оптимальной длины хеш-буфера, при котором исключены столкновения наиболее важных фрагментных дескрипторов (см. Рис. 20).

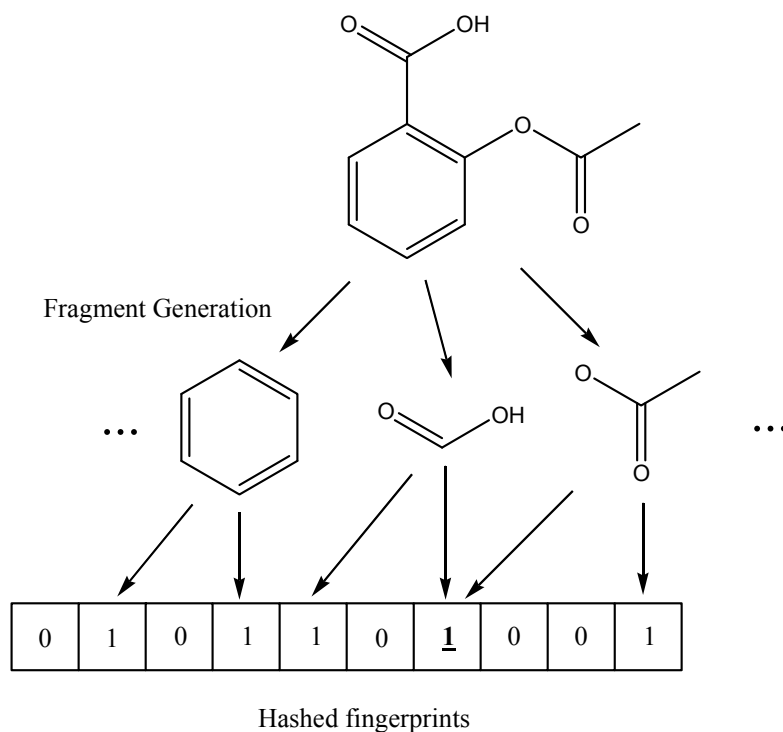


Рис. 20. Генерация хешированных молекулярных отпечатков пальцев. Каждый фрагмент приводит к установке нескольких битов. Бит, на котором произошла коллизия, отмечен жирным шрифтом и подчеркнут

2.2.5. Классификация по связности фрагментов

Фрагменты, используемые во фрагментных дескрипторах, могут быть связными (*connected*) и несвязными (*disconnected*). В абсолютном большинстве работ используются связные фрагменты. Оказывается, значения дескрипторов, построенных на несвязных фрагментах, всегда может быть выражено через значения дескрипторов, построенных на базе их компонент связности [259]. Следовательно, дескрипторы, построенные на основе несвязных фрагментов, являются избыточными, поскольку не содержат дополнительной информации по сравнению дескрипторами, построенными на основе связных фрагментов.

Тем не менее, дескрипторы построенные на несвязных фрагментах, могут в ряде случаев оказаться полезными, поскольку уравнения SAR/QSAR/QSPR с их участием могут оказаться более простыми. В частности, нелинейные QSAR/QSPR-модели с целочисленными дескрипторами на основе связных фрагментов могут быть заменены линейными моделями на основе несвязных фрагментов, поскольку числа встречаемости несвязных подграфов в молекулярном графе нелинейно выражаются через числа встречаемости связных подграфов. Таким образом, использование несвязных фрагментов можно рассматривать как неявный способ введения нелинейности в модели QSAR/QSPR, построенных на основе целочисленных фрагментных дескрипторов. То же самое касается бинарных фрагментных дескрипторов, но только в случае с ними нелинейные выражения заменяются логическими операциями конъюнкции. Таким образом, в случае бинарных дескрипторов несвязные фрагменты в неявном виде вводят логическую операцию конъюнкции в модели SAR.

Идея применения дескрипторов, основанных на несвязных фрагментах, лежит в основе концепции *компаунд-дескрипторов* (определяемых как комбинации несвязанных между собой фрагментов в молекулярной структуре), была недавно введена В.А. Тарасовым с соавт. [318]. В цитируемой работе было показано, что компаунд-дескрипторы существенно улучшают качество SAR модели, позволяющей прогнозировать мутагенность на основе Байесовского вероятностного подхода. Кроме того, дескрипторы на основе несвязных фрагментов

использовались в неявном виде (в форме конъюнкций бинарных дескрипторов на основе связанных фрагментов) в ряде работ, основанных на вероятностных методах прогнозирования (см. книгу [319] и ссылки в ней).

2.2.6. Классификация по уровням детализации молекулярных графов

В отличие от исследований QSPR, практически целиком основанных на рассмотрении молекулярных графов, вершины которых соответствуют всем атомам (по крайней мере, неводородным) в молекуле, при работе с биологической активностью, и особенно на качественном уровне, часто требуется более высокий уровень абстракции. В последнем случае бывает удобно описывать химические структуры при помощи специальных *редуцированных (reduced)* графов, вершины которых, иногда называемые дескрипторными либо фармакофорными центрами, представляют атом или группу атомов, способные взаимодействовать с биологической мишенью, тогда как ребра описывают удаленность дескрипторных центров друг от друга, например, по числу химических связей между ними (т.н. топологическое расстояние). Подобное биологически-ориентированное представление химических структур было впервые предложено в 1982 г. В.В. Авидоном с соавт. под именем *графа связности дескрипторных центров (ГСДЦ)* [131] как обобщение предложенных ранее дескрипторов ФКСП (см. пункт 2.2.1.6).

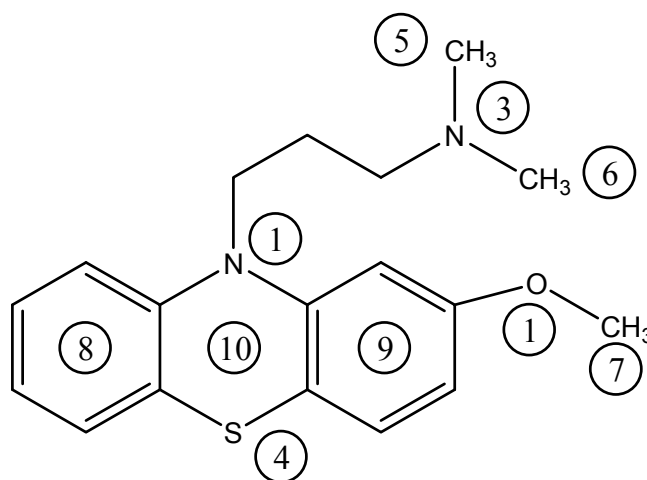


Рис. 21. Структура фенотиазина с отмеченными на ней дескрипторными центрами

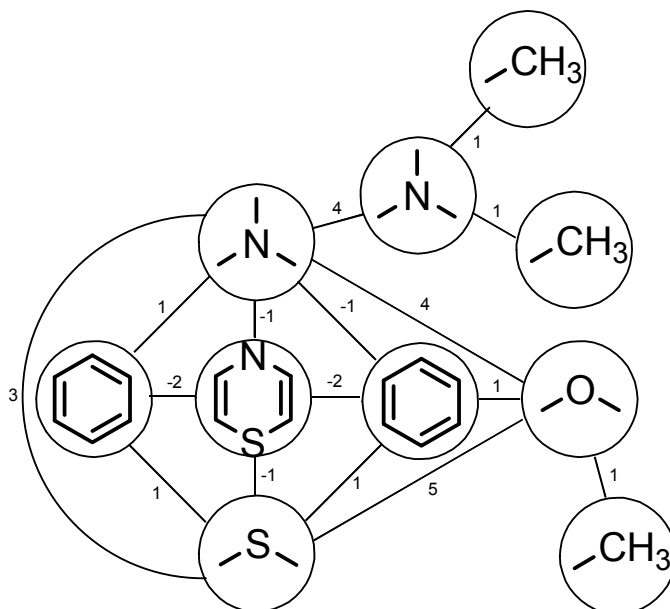


Рис. 22. Граф связности дескрипторных центров

На Рис. 22 приведен ГСДЦ для молекулы фенотиазина. В этом случае редуцированный граф состоит из 10 вершин, соответствующих дескрипторным центрам, показанным на Рис. 21, и 16 ребер. Набор дескрипторных центров включает: (а) четыре гетероатома (см. нумерацию на Рис. 21), которые могут принимать участие в донорно-акцепторных взаимодействиях и образовании водородных связей с биомолекулами; (б) три метильные группы 5, 6, 7, которые могут участвовать в гидрофобных взаимодействиях с биологическими молекулами; (в) два бензольных кольца 8, 9 и один гетероцикл 10, которые могут принимать участие в π - π и π -катионных взаимодействиях с биологическими молекулами. Одиннадцать ребер в ГСДЦ помечены положительными числами, показывающими топологическое расстояние (по числу связей) между дескрипторными центрами, тогда как отрицательные числа обозначают пересечения дескрипторных центров, когда они содержат один либо несколько общих атомов. ГСДЦ оказались полезными не только как источник биологически-ориентированных фрагментных дескрипторов (например, дескрипторы ФКСП можно рассматривать как «атомные пары», рассчитанные при использовании ГСДЦ вместо молекулярных графов), но также и при поиске фармакофоров,

В дальнейшем, редуцированные графы и основанные на них фрагментные дескрипторы неоднократно вводились разными группами авторов. Так, пред-

ложенные в 1985 г. атомные пары Кархарта (Carhart) [243] оказались близкими к вышеупомянутым дескрипторам ФКСП, и, следовательно, их тоже можно рассматривать как дескрипторы, основанные на двухвершинных связных подграфах специальных редуцированных графов, в которых ребра соответствуют путям между атомами. Предложенный в 1996 г. Кирсли (Kearsley) модифицированный вариант атомных пар [190], в котором классификация атомов основана на их физико-химических свойствах, еще выше поднял уровень абстракции этого типа дескрипторов. В 2003 г. Жиллет (Gillet), Виллвет (Willett) и Брэдшоу (Bradshaw) предложили новый тип редуцированных графов (в дальнейшем мы их будем называть GWB-редуцированными графами) и продемонстрировали их высокую эффективность в осуществлении поиска по подобию [320]. На Рис. 23 показан GWB-редуцированный граф, состоящий из 6 вершин и 5 ребер, наряду с несколькими химическими структурами, отображаемыми в него. Три его вершины с меткой *R* соответствуют кольцам (Rings), две вершины с меткой labeled *L* – линкерам (Linkers), а одна вершина с меткой *F* соответствует структурным особенностям (Features) – в данном случае это атом кислорода, способный образовывать водородные связи. В отличие от вышеупомянутых редуцированных графов ГСДЦ, ребра GWB-редуцированных графов специальным образом не помечены и соответствуют обычным химическим связям.

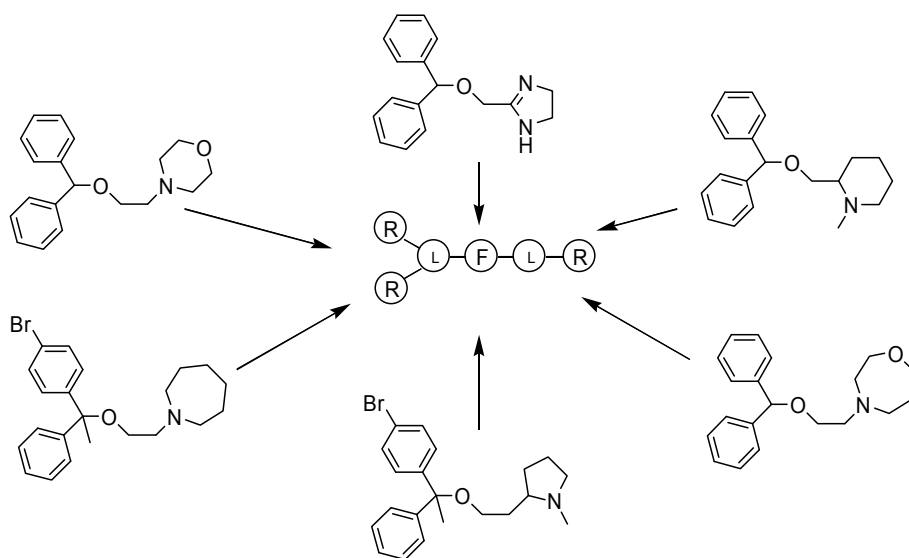


Рис. 23. Примеры химических структур, соответствующих одному GWB-редуцированному графу (показан в центре)

Еще одна отличительная черта GWB-редуцированных графов заключается в иерархической организации меток вершин. Например, метка Ar_n (ароматический цикл, не образующий водородных связей) более конкретна по сравнению с меткой Ar (любой ароматический цикл), которая, в свою очередь, является более конкретной по сравнению с R (любое кольцо). Благодаря этой особенности, GWB-редуцированные графы также могут быть организованы иерархически, причем уровень их абстрактности может быть контролируем (см. Рис. 24). Все это приводит к более высокой гибкости в их использовании. Кроме поиска по подобию, фрагментные дескрипторы на основе GWB-редуцированных графов с успехом были применены при построении классификационных моделей SAR с использованием деревьев принятия решений [321].

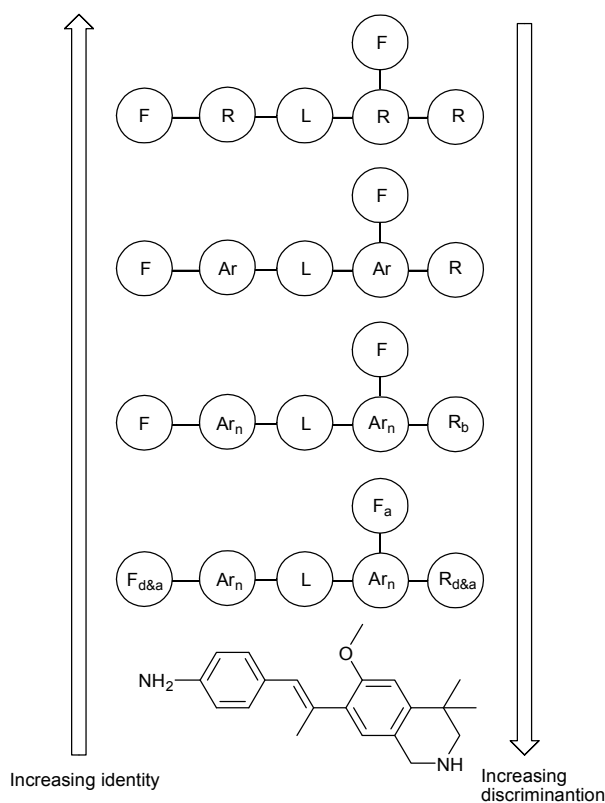


Рис. 24. Иерархия GWB-редуцированных графов

2.2.7. Фрагментные дескрипторы с выделенными атомами

Фрагментные дескрипторы с отмеченными атомами рассмотрены в разделе 5.3.

2.3. Ограничения фрагментных дескрипторов

Несмотря на успешное применение и большую популярность фрагментных дескрипторов, они все-таки не лишены определенных ограничений. В литературе упоминается о трех основных проблемах, связанных с ними: (1) проблема «редких» либо «отсутствующих» фрагментов; (2) проблема адекватного представления стереохимической информации; (3) отсутствие физической интерпретации.

Проблема «редких» и «отсутствующих» фрагментов [322] является, по-видимому, наиболее серьезной из упомянутых трех. Действительно, число фрагментов (и, следовательно, количество фрагментных дескрипторов) практически неограниченно: оно значительно превышает число возможных химических структур. В результате этого любая химическая структура содержит такие фрагменты, которые отсутствуют (либо присутствуют в слишком малом количестве) в обучающей выборке, использованной для построения моделей SAR/QSAR/QSPR, необходимых для прогнозирования нужного свойства. Поскольку для фрагментных дескрипторов, соответствующих отсутствующим либо редким фрагментам, нельзя сколько-нибудь надежно оценить значение соответствующего регрессионного коэффициента (и, следовательно, оценить насколько он важен для прогнозирования определенного свойства), то в том случае, если он все-таки важен для прогнозирования данного свойства, оно не будет надежно предсказано. Отсюда возникает следующий парадокс, который, очевидно, противоречит всей практике применения фрагментных дескрипторов: нельзя надежно прогнозировать свойства органических соединений, отсутствующих в обучающей выборке.

Одно из возможных практических рекомендаций, вытекающих из анализа проблемы «отсутствующих фрагментов», заключается в необходимости введения ограничений на классы фрагментов, вводимых в статистический анализ, в результате чего становится возможным определить область применимости моделей QSAR/QSAR/QSPR как множество молекулярных графов, не содержащих «проблематичных» (т.е. отсутствующих, редких либо принимающих постоян-

ное значение на обучающей выборке) фрагментных дескрипторов. На решение этой проблемы были отчасти также направлены разработанные в рамках данной диссертационной работы псевдофрагментный подход (см. раздел 5.4) и не-векторный QSAR/QSPR-анализ (см. раздел 4.5).

Альтернативным подходом к решению проблемы «отсутствующих» фрагментов является оценка “*ab initio*” значений регрессионных коэффициентов при отсутствующих фрагментных дескрипторов, что и было реализовано в рамках программы CLOGP для прогнозирования липофильности органических соединений [323]. В процитированной статье утверждается, что подобная операция позволяет прогнозировать липофильность органических соединений с отсутствующими фрагментами с ошибкой меньше 0.5 log единиц, но это утверждение было подвергнуто критике в статье [322].

Вторая из проблем, связанных с использованием фрагментных дескрипторов, связана с необходимостью учета стереохимической информации, без чего модели QSAR/QSPR должны давать идентичный прогноз для всех диастереомеров и цис-транс-изомеров. К сожалению, полностью корректное решение этой проблемы невозможно осуществить в рамках представления структур органических молекул в виде графов: оно требует явное рассмотрение гиперграфов. Тем не менее, в большинстве практически важных случаев достаточно вводить специальные метки, специфицирующие стереохимическую конфигурацию хиральных центров либо конфигурацию при двойных связях, и их использовать при спецификации фрагментов, как это было, например, сделано для голографических фрагментных дескрипторов [324], а также специфицировано в специализированном языке описания фрагментов PARTAN [325].

Что же касается третьей проблемы, т.е. невозможности предоставить *физическую* интерпретацию построенным с участием фрагментных дескрипторов моделям, то многими исследователями это не считается недостатком фрагментных дескрипторов, поскольку интерпретация с точки зрения *физики* не является приоритетной задачей в области хемоинформатики. С этой целью можно обратиться к другим областям вычислительной химии, в частности к квантовой химии и молекулярному моделированию.

ГЛАВА 3. МАТЕМАТИЧЕСКОЕ ОБОСНОВАНИЕ ВЫБРАННОГО ПОДХОДА

3.1. Химическая значимость поиска базиса инвариантов помеченных графов

Поиск соотношений «структура-свойство» является важнейшей проблемой современной химии, и методы описания молекул играют существенную роль в таких исследованиях. Один из наиболее популярных подходов к решению этой проблемы основан на представлении молекулярной структуры в виде взвешенного (помеченного) молекулярного графа и использовании инвариантов графов (т.е. числовых характеристик, не зависящих от нумерации вершин графа) для его описания. Такими инвариантами графов являются как молекулярные дескрипторы (см. [105]), описывающие химические структуры (но не отдельные их конформации!), так и любые функции, аппроксимирующие свойства соответствующих химических соединений. Заметим, что ряд топологических (т.е. вычисляемых без учета явного пространственного строения молекул) молекулярных дескрипторов, вычисляемых в результате формальных математических операций на графах, называют по историческим причинам топологическими индексами [326-331]. Фрагментные дескрипторы также являются топологическими молекулярными дескрипторами, но их не принято называть топологическими индексами.

Возникает вопрос: существует ли конечный набор базисных инвариантов графов, такой чтобы любой инвариант графа мог бы быть однозначно представлен в виде линейной комбинации базисных инвариантов? Если подобный набор существует, то его элементы образуют конечный базис алгебры инвариантов графов (множество инвариантов графов в совокупности с операциями сложения, умножения и умножения на действительное число образуют алгебру инвариантов графов). В этом случае можно было бы выбирать молекулярные дескрипторы из этого базисного набора и рассматривать только линейные зависимости в поиске количественных соотношений «структура-свойство».

Проблема нахождения базисных подграфов была рассмотрена Рандичем в 1992 г. [257]. В случае ее решения стало бы возможно с их помощью однознач-

но представлять химические структуры. Рандичем было предложено использовать в качестве базисных подграфов графы-пути, а в качестве значений базисных инвариантов – числа вложений базисных подграфов в молекулярный граф. Тем не менее, на нескольких примерах было показано, что разные химические структуры могут содержать одинаковые наборы подграфов-путей, и поэтому предложенные «базисные» подграфы таковыми, строго говоря, не являются.

Тем не менее, анализируя математическую литературу, мы обнаружили, что строгое решение вышеупомянутой проблемы было найдено еще в 1983 г. для случая простых графов [332], однако, будучи опубликовано на русском языке в малодоступном для зарубежных специалистов издании, оно оставалось практически неизвестным. Суть предложенного решения заключается в следующем. Пусть $\Gamma^{(n)}$ обозначает множество всех простых (т.е. с невзвешенными вершинами и ребрами), как связных так и несвязных графов. Показано методами коммутативной алгебры [332], что любой инвариант $f(G)$ графа $G \in \Gamma^{(n)}$ может быть однозначно представлен в виде:

$$f(G) = \sum_j c_j g_j(G) \quad (70)$$

где c_j обозначает некоторые константы, независимые от G , $g_j(G)$ – число вложений графа $G_j \in \Gamma^{(n)}$ в G (т.е. количество различных подграфов G , изоморфных G_j), а суммирование идет по всем графам $G_j \in \Gamma^{(n)}$. Это означает, что множество $\{g_j\}$ образует базис алгебры инвариантов графов из $\Gamma^{(n)}$. Кроме того, любой инвариант графа $G \in \Gamma^{(n)}$ задается числом его подграфов, получаемых удалением из G ребер всеми возможными неэквивалентными способами.

Между тем, для решения большинства задач в области химии представляют наибольший интерес не простые графы, а те, которые несут веса на своих вершинах и ребрах. Эти веса определяются типами соответствующих атомов и связей. Вследствие этого взвешенный граф значительно точнее описывают молекулярную структуру химического соединения, чем простой граф.

Кроме взвешенных графов, в математике также рассматриваются помеченные графы, которые получаются при отнесении вершин и ребер к определенным классам путем приписывания им соответствующих меток. Если же в

качестве меток использовать действительные числа, то от помеченных графов можно перейти ко взвешенным. Таким образом, взвешенные графы в определенной мере можно рассматривать частным случаем помеченных графов. Следовательно, решение задачи нахождения базиса алгебры помеченных графов позволило бы распространить рассмотренные выше математические результаты на предсказания свойств реальных химических соединений.

3.2. Две основные теоремы о базисе инвариантов графов

Построим множество помеченных графов. Рассмотрим сначала множество простых графов $\Gamma^{(n)}$ и два конечных множества произвольных меток (символов), $V = \{v_1, \dots, v_{p_1}\}$, $E = \{e_1, \dots, e_{p_2}\}$, $v_i \neq v_j$, $e_i \neq e_j$, $i \neq j$. Поместим метки на вершины (из V) и ребра (из E) графов из $\Gamma^{(n)}$ всеми неэквивалентными способами. Обозначим через $H_{V,E}^{(n)}$ множество построенных таким образом помеченных по вершинам и ребрам графов, а через N – число элементов в множестве $H_{V,E}^{(n)}$. Возможно также, что в графах из $\Gamma^{(n)}$ метятся только вершины ($E = \emptyset$ – пустое множество) или только ребра ($V = \emptyset$ – пустое множество). Обозначим получаемые таким образом множества графов соответственно через $H_V^{(n)}$ и $H_E^{(n)}$.

Рассмотрим метки как переменные, принимающие вещественные числовые значения. Тогда любой граф $H \in H_{V,E}^{(n)}$ может быть представлен как симметричная матрица $A = \|a_{ij}\|$, в которой диагональный элемент a_{ii} соответствует метке вершины i , а недиагональный элемент a_{ij} ($i \neq j$) соответствует метке ребра, соединяющего вершины i и j , тогда как для несмежных вершин i и j он равен нулю.

Определение. Инвариантом помеченного графа $H \in H_{V,E}^{(n)}$ называется скалярная функция от матричных элементов a_{ij} , значения которой не зависят от нумерации вершин графа.

Теорема 1. Любой инвариант $f(H)$ помеченного графа $H \in H_{V,E}^{(n)}$ может единственным образом быть представлен в виде:

$$f(H) = \sum_{j=1}^N c_j g_j(H) \quad (71)$$

где: c_j – это некоторые константы, не зависящие от H и зависящие от f ; $g_j(H)$ – это число вложений графа $H_j \in H_{V,E}^{(n)}$ в граф H (т.е. количество различных подграфов графа H , изоморфных H_j). Таким образом, множество g_j образует базис в алгебре инвариантов графов из $H_{V,E}^{(n)}$. Кроме того, величина любого инварианта $f(H)$ для графа H определяется числом подграфов в H , получаемых из H путем удаления ребер всеми неэквивалентными способами.

Доказательство. Упорядочим графы из $H_{V,E}^{(n)}$ следующим способом. Сначала пронумеруем произвольным образом все графы с $n(n-1)/2$ ребрами, потом все графы с $[n(n-1)/2]-1$ ребром и т.д., пока не будут пронумерованы графы, состоящие из изолированных вершин. Обозначим через B квадратную матрицу с элементами $b_{ij} = g_j(H_i)$, $(i, j = \overline{1, N})$. Очевидно, что: 1) если графы H_i и H_j имеют одинаковое количество ребер, то $b_{ij} = g_j(H_i) = b_{ji} = g_i(H_j) = 0$ и $b_{jj} = g_j(H_j) = 1$; и 2) если графы H_i и H_j имеют разное количество ребер и $j < i$, то $b_{ij} = g_j(H_i) = 0$. Таким образом, матрица B является треугольной, на ее диагонали находятся только единицы, а все элементы под диагональю равны нулю. Следовательно, существует обратная матрица B^{-1} . Запишем систему уравнений:

$$f(H_i) = \sum_{j=1}^N c_j g_j(H_i) = \sum_{j=1}^N b_{ij} c_j \quad (i = \overline{1, N}) \quad (72)$$

или в матричной форме $\bar{f} = B\bar{c}$, где $\bar{f} = (f(H_1), \dots, f(H_N))$, $\bar{c} = (c_1, \dots, c_N)$ – вектора-колонки. Система уравнений (2) всегда имеет единственное решение: $\bar{c} = B^{-1}\bar{f}$. Следовательно, существует единственное разложение (71) инварианта $f(H)$ для заданной нумерации графов H_j .

Покажем, что разложение (71) не зависит от нумерации графов H_j . Предположим, что некоторая нумерация приводит к векторам \bar{f}' , \bar{c}' и матрице B' (не обязательно треугольной). Переход от первой нумерации ко второй можно осуществить при помощи подстановки π : $j \rightarrow \pi(j)$ ($i = \overline{1, N}$) либо соответствующей матрицы подстановки X размера $N \times N$, причем $\det X \neq 0$. Очевидно, что $X\bar{f} = \bar{f}'$, $X\bar{c} = \bar{c}'$ и $XBX^{-1} = B'$. Как было показано выше, по крайней мере для од-

ной нумерации графов в разложении (71) справедливо $\bar{f} = B\bar{c}$. Умножая обе части этого уравнения на X , имеем: $\bar{f}' = X\bar{f} = (XBX^{-1})(X\bar{c}) = B'\bar{c}'$. Следовательно, разложение (71) верно при любой нумерации графов H_j .

Теорема 1 доказана. ■

Теорема 2. Любой инвариант $f(H)$ помеченного графа $H \in H_{V,E}^{(n)}$ может быть представлен при помощи полинома от переменных, равных числам встречаемости некоторых связных подграфов в H . Количество вершин в таких подграфах и степень полинома меньше либо равно n .

Доказательство. Прежде всего покажем, что число встречаемости любого несвязанного подграфа C в графе H может быть выражено через числа встречаемости некоторых связных подграфов в H . Предположим, что C состоит из k компонент связности, т.е. $C = \bigcup_{i=1}^k C_i$, где $\{C_i\}$ – связанные подграфы, причем $C_i \cap C_j = \emptyset$, $i \neq j$. В общем случае возможно, что некоторые подграфы из $\{C_i\}$ изоморфны друг другу. Разобьем множество $\{C_i\}$ на p групп Ω_i ($i = \overline{1, p}$) таким образом, чтобы подграфы в каждой из групп были изоморфны друг другу, а подграфы из разных групп, наоборот, друг другу неизоморфны. Пусть m_i – число элементов в Ω_i , $m_i \geq 1$, $\sum_{i=1}^p m_i = k$ и $i = \overline{1, p}$. Пронумеруем подграфы из $\{C_i\}$ следующим образом: сначала пусть идут подграфы из $\{C_i\}$, относящиеся к группе Ω_1 , потом относящиеся к группе Ω_2 и т.д. Пусть M_i – множество всех подграфов графа H , изоморфных подграфам из группы Ω_i , а l_i – число элементов в M_i ($i = \overline{1, p}$). Очевидно, что $l_i \geq m_i$.

Построим новые подграфы графа H , выбирая всеми возможными способами m_i разных элементов из M_i одновременно для всех $i = \overline{1, p}$. Число таких подграфов равно $\prod_{i=1}^p C_{l_i}^{m_i}$, $C_{l_i}^{m_i} = l_i! / [m_i!(l_i - m_i)!]$. Полученные из M_i подграфы можно отнести к двум типам. В первом случае исходные подграфы из M_i не пересекаются, во втором – пересекаются. Обозначим через t_1 и t_2 число подграфов первого и второго типа, соответственно. Очевидно, что $t_1 + t_2 = \prod_{i=1}^p C_{l_i}^{m_i}$. Заметим, что t_1 равно числу встречаемости подграфа C в H и совпадает, согласно определению, с числом подграфов в H , изоморфных C . Подграфы же второго

типа имеют меньше k компонент связности, и сумма $t_1 + t_2 = \prod_{i=1}^p C_{l_i}^{m_i}$ является полиномом степени $k = \sum_{i=1}^p m_i$ от переменных l_i ($i = \overline{1, p}$).

Таким образом, число встречаемости t_1 несвязного подграфа C с k компонентами связности можно выразить через числа встречаемости связных компонент и некоторых подграфов с меньшим чем k числом компонент связности. Применяя многократно этот результат ко всем несвязным подграфам, можно прийти к формулировке теоремы 2.

Теорема 2 доказана. ■

3.3. Теоретические основы сочетания искусственных нейронных сетей и фрагментных дескрипторов

Традиционно принято считать, что теоретическую основу использования многослойных нейронных сетей составляет нейросетевая интерпретация теоремы Колмогорова о представлении непрерывных функций нескольких переменных в виде суперпозиций непрерывных функций одного переменного и сложения [333], которая в исходном виде была сформулирована следующим образом.

Теорема. При любом целом $n \geq 2$ существуют такие определенные на единичном отрезке $E^1 = [0; 1]$ непрерывные действительные функции $\psi^{pq}(x)$, что каждая определенная на n -мерном единичном кубе E^n непрерывная действительная функция $f(x_1, \dots, x_n)$ представима в виде

$$f(x_1, \dots, x_n) = \sum_{q=1}^{2n+1} \chi_q \left[\sum_{h=1}^n \psi^{pq}(x_p) \right], \quad (73)$$

где функции $\chi_q(y)$ действительны и непрерывны.

Эта теорема, появившаяся в 1957 году в результате научной полемики между академиками А. Н. Колмогоровым и В. И. Арнольдом, первоначально не имела никакого отношения к нейронным сетям и только в 1987 году была переложена в термины теории нейронных сетей в работе Р. Хехт-Нильсена (R. Hecht-Nielsen) [334]. В этой своей новой формулировке теорема доказывает представимость функции многих переменных достаточно общего вида $R^n \rightarrow R^m$ с помощью двухслойной (трехслойной с формальным учетом входного слоя)

нейронной сети с прямыми полными связями с n компонентами входного сигнала, $2n+1$ компонентой первого (скрытого) слоя с заранее известными ограниченными функциями активации (например, сигмоидными) и m компонентами второго слоя с неизвестными функциями активации. Теорема, таким образом, в *неконструктивной форме* доказывает решаемость задачи представления функции достаточно произвольного вида на нейронной сети и указывает для каждой задачи минимальные значения числа нейронов сети, необходимых для ее решения. Решаемость задачи представления функции при помощи теоремы Колмогорова в *конструктивной форме* было найдено несколько позже в работах Д. А. Шпрехера (D. A. Sprecher) [335, 336], в которых приведен численный алгоритм нахождения неизвестных функций активации второго слоя.

Поскольку теорема Колмогорова в интерпретации Хехт-Нильсена описывает нейронную сеть, в которой один из слоев содержит нейроны с нефиксированной функцией активации, она не может быть непосредственно применена к наиболее популярным архитектурам нейронных сетей, например, ко многослойному персептрону либо к нейросетям радиальной базисной функции, в которых все нейроны имеют функции активации фиксированного вида. Эта проблема была, однако, решена в 1992 г. в работе Куркова (Kůrková) [337], в которой, основываясь на теореме Колмогорова, доказывается способность многослойной нейронной сети с двумя слоями скрытых нейронов с фиксированными функциями активации (например, сигмоидными) аппроксимировать любую непрерывную функцию многих переменных, а также оценить, исходя из свойств аппроксимируемой функции, необходимое для этого число скрытых нейронов и точность аппроксимации.

С другой стороны, как показано было нами выше (теорема 2 из раздела 3.2), *любой инвариант помеченного графа может быть представлен при помощи полинома от переменных, равных числам встречаемости некоторых связанных подграфов в этом графе*, и принимая во внимание, что

- 1) полином является непрерывной функцией,
- 2) молекулярное свойство является инвариантом молекулярного графа, не зависящим от нумерации его вершин,

3) число встречаемости некоторого подграфа в графе является значением соответствующего фрагментного дескриптора для этого графа, мы сразу приходим к формулировке центрального положения данной диссертационной работы: любая сколь угодно сложная зависимость между структурой органического соединения и его свойством может быть аппроксимирована при помощи многослойной нейронной сети персептронного типа с двумя скрытыми слоями нейронов и набора фрагментных дескрипторов. Следует, однако, отметить, что в большинстве случаев для аппроксимации зависимости «структура-свойство», как показывает опыт, достаточно и одного слоя скрытых нейронов.

ГЛАВА 4. РАЗРАБОТКА НЕЙРОСЕТЕВЫХ ПОДХОДОВ

Данная глава содержит описание предложенных нами подходов к решению перечисленных в разделе 1.4 проблем, связанных с применением искусственных нейронных сетей для решения прикладных задач, в частности, для поиска количественных корреляций «структура-свойство».

4.1. Подход к решению проблемы «переучивания» нейронных сетей

Одной из основных проблем, с которой мы столкнулись в начале 1990-ых годов уже в ходе самых первых работ по применению аппарата искусственных нейронных сетей для прогнозирования свойств органических соединений была связана с эффектом «переучивания» и необходимостью поиска эффективных методов его предотвращения.

4.1.1. Суть эффекта «переучивания» нейросетей

Эффект «переучивания» (overtraining) нейросетей был, по-видимому, впервые описан в математической литературе в 1990 г (см. [338]). Он наблюдается при обучении многослойных нейронных сетей с обратным распространением ошибки (т.е. многослойных персептронов) в том случае, когда число примеров в обучающей выборке невелико по сравнению с числом настраиваемых параметров нейросети (т.е. синаптических весов и порогов активации нейронов). В настоящее время его принято считать особым проявлением эффекта «переподгонки данных» (overfitting), наблюдаемого во многих методах машинного обучения (о сходстве и различии понятий «переучивания» и «переподгонки» см. в статье [339]). Суть эффекта «переучивания» заключается в следующем: процесс обучения нейросети может быть условно разделен на две последовательные фазы – «обобщения» (generalization) и «запоминания» (memorization). Для химических соединений, содержащихся в обучающей выборке, среднеквадратичная ошибка прогнозирования их свойств постоянно уменьшается

по ходу обучения в обеих фазах. В то же время, для соединений, отсутствующих в обучающей выборке, среднеквадратичная ошибка прогнозирования сначала уменьшается по ходу обучения в фазе «обобщения», но потом начинает расти в последующей фазе «запоминания». В результате этого «переобученная» нейросеть хорошо воспроизводит свойства соединений из обучающей выборки, но плохо прогнозирует свойства любых других соединений, например, содержащихся в контрольных выборках. Эффект «переучивания» схематически показан на Рис. 25.

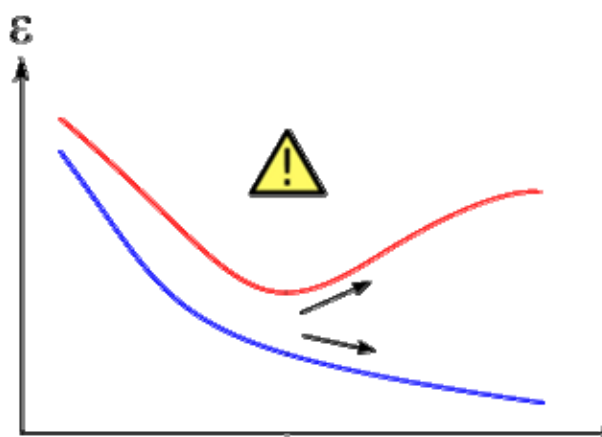


Рис. 25. Эффект "переучивания" нейросети. Нижняя кривая показывает ход изменения (при обучении нейросети) ошибки прогнозирования для соединений, входящих в обучающую выборку, а верхняя – в контрольную выборку. Восклицательным знаком отмечена точка перехода из фазы «обобщения» в фазу «запоминания».

Природу эффекта «переучивания» обычно связывают с постепенным увеличением эффективного числа дескрипторов (а вместе с этим и сложности модели) по мере обучения нейросети (см. [18, 338]). Настраиваемые параметры нейросети, каковыми являются значения всех синаптических весов и порогов активации, перед началом обучения инициализируются обычно случайными числами, близкими к нулю. В этом случае во всех нейронах функция активации срабатывает при значениях аргумента, близких к нулю. Поскольку в окрестностях нуля любая нелинейная непрерывная функция приближается к линейной, то и нейросеть в самом начале обучения формирует на выходе сигналы, связанные со входными сигналами зависимостями, близкими к линейным. Таким образом, на начальном этапе обучения выходные сигналы представляют собой

линейные комбинации входных. В этом случае эффективное число дескрипторов равно числу линейно независимых дескрипторов в базе, и это число не может превышать числа входных нейронов. По мере обучения нейросети значения настраиваемых параметров растут по абсолютной величине, и в разложении в ряд Тейлора-Маклорена функции активации все большую роль начинают играть члены со второй, третьей и более высокими степенями. В результате этого нейросеть постепенно переходит к моделированию квадратичных, кубических и более сложных зависимостей, которые описываются все более возрастающим числом параметров. Таким образом, по ходу обучения нейросети эффективное число параметров постоянно возрастает, пока не достигает определенного максимального числа, которое равно числу настраиваемых параметров нейросети (т.е. суммарного числа синаптических весов и порогов активации), деленному на порядок группы автоморфизмов помеченного графа, соответствующего нейросети. Параллельно с эффективным числом дескрипторов при обучении нейросети растет и емкость класса моделируемых функций, которая в теории статистического обучения выражается размерностью Вапника-Червоненкиса. Упрощенно можно сказать, что в тот момент времени, когда емкость этого класса начнет превышать объем используемых для обучения данных, и наступает «переучивание».

4.1.2. Методы предотвращения «переучивания» нейросетей

В литературе описано несколько методов предотвращения «переучивания» [338]. Наиболее простым из них является уменьшение общего числа настраиваемых параметров нейросети за счет уменьшения числа входных и скрытых нейронов. В исследованиях, проведенных в рамках настоящей диссертационной работы, уменьшение числа входных нейронов осуществлялось за счет предварительного отбора дескрипторов при помощи линейно-регрессионного метода БПМЛР (см. подраздел 4.1.5), а числа скрытых нейронов – за счет варьирования их числа и определения из них оптимального. Тем не менее, этот метод предотвращения «переучивания» не является панацеей – его недостатком

являются слишком упрощенные модели, получаемые на небольших выборках (т.н. «недоподгонка данных», т.е. *underfitting*).

Второй способ предотвращения «переучивания», в соответствии с общими положениями теории статистического обучения и основанного на ней принципа минимизации структурного риска, состоит во введении регуляризационного члена в минимизируемую в процессе обучения нейросети функцию риска. Частным случаем такого введения регуляризаторов в нейросеть является обучение «с забыванием», имеющее очевидные нейрофизиологические аналогии. В рамках нейросетевого программного комплекса NASAWIN нами был реализован и этот метод предотвращения переучивания за счет введения четырех разных регуляризаторов. Тем не менее, этот способ обладает существенным недостатком – для нахождения оптимального значения относительного веса регуляризатора в функции риска требуется многократно проводить полное обучение нейросети для разных его значений, что делает метод малопривлекательным с вычислительной точки зрения.

Наконец, самым эффективным методом предотвращения «переучивания» является остановка обучения при достижении наименьшей среднеквадратичной ошибки прогнозирования на контрольной выборке. Показано, что подобная остановка обучения является одной из форм регуляризации [340]. Получаемые при этом модели сравнимы по прогнозирующей способности с моделями, при построении которых явным образом используются регуляризаторы, но при этом тратится вычислительных ресурсов значительно меньше, так как для построения модели требуется всего лишь однократное (и к тому же неполное) обучение нейросети. Именно эта схема предотвращения «переучивания» и является основной в программном комплексе NASAWIN, разработанном в рамках данной диссертационной работы.

Тем не менее, в ходе практического применения вышеупомянутого метода остановки обучения обнаружилась проблема, суть которой состоит в том, что поскольку контрольная выборка используется для остановки обучения, т.е. для отбора модели, содержащаяся в ней информация в неявном виде частично попадает в отобранную модель, и поэтому такая выборка уже не может счи-

таться внешней по отношению к этой модели, а ошибка прогнозирования на ней – для объективной оценки прогнозирующей способности этой модели. Иными словами, если критерий минимума средней ошибки на контрольной выборке используется для выбора статистической модели, то само это значение является искаженным в оптимистическую сторону оценкой прогнозирующей способности отобранной модели. Ниже изложено предложенное нами в 1995 г. эффективное решение этой проблемы [341].

4.1.3. Трехвыборочный подход

Для решения вышеизложенной проблемы, связанной с некорректностью использования одной и той же контрольной выборки для отбора модели и оценки ее прогнозирующей способности, предлагается использовать трехвыборочный подход, согласно которому производится деление всего набора данных на 3 выборки: обучающую (training set), внутреннюю контрольную (validation set) и внешнюю контрольную (prediction set). По обучающей выборке производится построение последовательности моделей с возрастающей сложностью (емкостью класса моделей). В случае линейно-регрессионных моделей, формируемых путем наращивания числа отбираемых дескрипторов, в качестве такого критерия сложности может выступать число отобранных дескрипторов, а при обучении нейросети – номер шага (эпохи) обучения. Для определения оптимальной сложности модели (и тем самым отбора модели с оптимальной сложностью) используется критерий минимума среднеквадратичной ошибки прогнозирования, вычисляемой для внутренней контрольной выборки. Поскольку информация из внешней контрольной выборки никаким образом не участвует ни в построении, ни в отборе моделей, то среднеквадратичная ошибка прогнозирования на ней может быть использована для оценки прогнозирующей способности отобранной модели. Разбивку набора данных на три выборки можно осуществлять либо случайным образом, либо систематично в рамках процедуры скользящего контроля.

Трехвыборочный метод был нами впервые представлен в 1995 г. в рамках приглашенного пленарного доклада на конференции по интеллектуальной обработке данных (г. Оберн, штат Алабама, США) и был положительно воспринят сообществом математиков, специализирующихся в области нейросетей. Почти одновременно с нами и независимо от нас сходные идеи были также опубликованы И.Тетко с соавт. [339] и впоследствии легли в основу разработанного им позже метода ассоциативных нейронных сетей [342]. С тех пор трехвыборочный метод превратился в обязательный атрибут нейросетевых исследований в данной области.

Трехвыборочный метод, в сочетании с идеями ансамблевого подхода к построению QSAR/QSPR-моделей, лег в основу как более ранней методики, изложенной в подразделе 6.3.1 (т.н. одноуровневого комбинаторного подхода), так и более поздней разработки – процедуры двойного скользящего контроля, примененной в целом ряде разделов данной диссертационной работы.

4.1.4. Процедура двойного скользящего контроля

Для построения и объективной оценки прогнозирующей способности линейно-регрессионных и нейросетевых моделей нами была предложена процедура $N \times (N-1)$ - кратного двойного скользящего контроля [343]. В этом подходе исходная база данных систематически разбивается на 3 части: обучающую, внутреннюю контрольную и внешнюю контрольную выборки в соотношении $(N-2):1:1$. Информация из внутренней контрольной выборки используется для отбора моделей с наилучшей прогнозирующей способностью. Информация из внешней контрольной выборки никаким образом не используется при построении и отборе моделей, и поэтому ошибка прогнозирования на ней (как среднеквадратичная, так и средняя абсолютная) может быть использована для оценки реальной прогнозирующей способности моделей. При таких разбиениях каждое соединение из исходной базы данных попадает в обучающую выборку N^2-3N+2 раза, во внутреннюю контрольную выборку - $N-1$ раз и во внешнюю контрольную выборку - также $N-1$ раз.

Предсказанное значение свойства для каждого соединения вычисляется как среднее из предсказанных значений при всех $N-1$ разбиениях, при которых оно попадает во внешнюю контрольную выборку, тогда как дисперсия предсказанных значений может быть использована для оценки точности прогноза для данного соединения. На Рис. 26 представлена диаграмма разбиения исследуемых баз данных для $N = 5$.

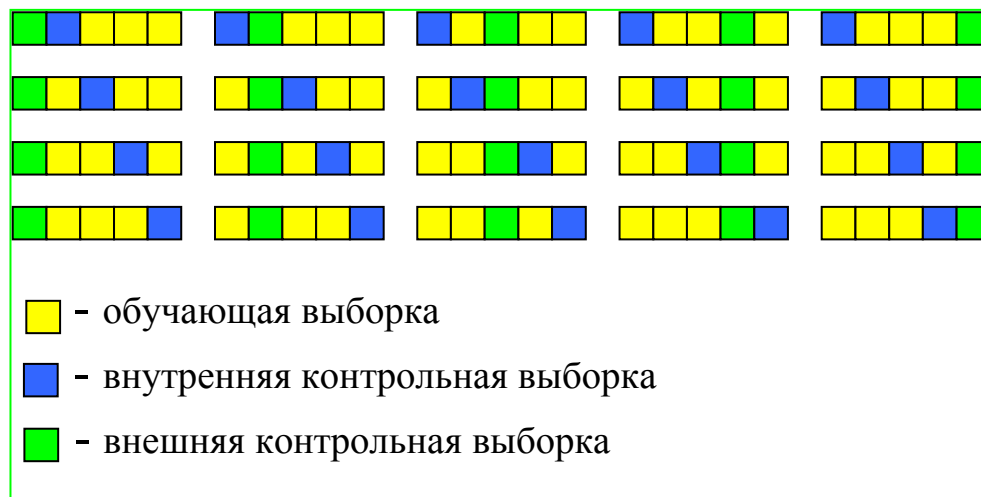


Рис. 26. Схема 5x4-кратного двойного скользящего контроля

В результате на основе усреднения $N \times (N-1)$ частных моделей, выводимых при разных разбиениях исходной базы данных, получаются соответствующие комбинированные модели. Вычисляемые статистические характеристики включают: (1) Q^2_{DCV} - параметр Q^2 ($Q^2 = (SS - PSS) / SS$, где PSS сумма квадратов ошибок прогноза свойства, SS - сумма квадратов отклонения свойства от среднего значения) для усредненных спрогнозированных значений, (2) $RMSE_{DCV}$ - среднеквадратичная ошибка прогнозирования, (3) MAE_{DCV} - средняя абсолютная ошибка прогнозирования.

Метод двойного скользящего контроля обеспечивает объективную оценку реальной прогнозирующей способности моделей, процедура отбора которых предполагает использование контрольной выборки либо процедуры скользящего контроля. Он не только позволяет эффективно предотвращать «переучивание» нейросетей (благодаря трехвыборочному подходу), но и обращает стохастические свойства нейросетевых моделей из кажущегося недостатка в преимущество, поскольку благодаря этому позволяет оценивать

ожидаемую ошибку прогноза.

Из описанных в математической литературе метод двойного скользящего контроля больше всего похож на процедуру вложенного скользящего контроля (nested cross-validation), однако между ними имеются принципиальные отличия в критериях отбора моделей, не позволяющие использовать последнюю для аналогичной работы с нейросетями. Подчеркнем также, что то, что иногда в литературе называется «процедурой двойного скользящего контроля» (double cross-validation), на деле является обычной процедурой двукратного скользящего контроля.

4.1.5. Быстрая пошаговая множественная линейная регрессия

Трехвыборочный подход применен нами также и в рамках метода быстрой пошаговой множественной линейной регрессии (БПМЛР) – специального линейно-регрессионного метода, разработанного нами для предварительного отбора дескрипторов для нейросетей. В данном случае внутренняя контрольная выборка используется для определения оптимального числа включаемых в модель дескрипторов. В рамках метода БПМЛР текущий вектор ошибок (невязок) инициализируется экспериментальными значениями свойств соединений из обучающей выборки. На каждой итерации дескриптор, наилучшим образом коррелирующий с текущим вектором ошибок на обучающей выборке, добавляется к текущему набору отобранных дескрипторов, а соответствующая регрессионная модель, построенная на этом дескрипторе, используется для пересчета текущего вектора ошибок, который уже используется на следующей итерации для отбора следующего дескриптора и т.д. Интересной и нетривиальной особенностью этого приема является то, что каждый дескриптор может быть включен в модель несколько раз на разных итерациях. При добавлении очередного дескриптора регрессионный коэффициент при свободном члене из построенного на нем регрессионного уравнения суммируется с текущим коэффициентом при свободном члене в многомерной (т.е. включающей множество дескрипторов) модели. Что касается регрессионного коэффициента при самом деск-

рипторе, то он переносится в многомерную модель, если дескриптор включается в нее в первый раз, либо суммируется с уже имеющимся значением при последующем включении его в модель. Процесс пошагового отбора дескрипторов и построения результирующей модели останавливается по достижению наименьшей ошибки прогнозирования на внутренней контрольной выборке, тогда как ошибка прогнозирования на внешней контрольной выборке, информация из которой никаким образом не используется в проводимом статистическом анализе, используется для оценки прогнозирующей способности результирующей многомерной линейной регрессионной модели.

Хотя метод БПМЛР первоначально был предназначен только для предварительного отбора дескрипторов для построения нейросетевых моделей, однако за время эксплуатации он успел себя зарекомендовать как самостоятельный мощный метод статистического анализа, обладающий очень высокой производительностью и позволяющий даже на персональном компьютере эффективно обрабатывать выборки огромного размера как по числу дескрипторов (миллионы) так и соединений. Последнее свойство очень важно при работе с фрагментными дескрипторами ввиду их очень большого числа. Из существующих методов регрессионного анализа самый близкий к БПМЛР подход – это аддитивная регрессия, однако между ними есть существенные различия.

4.2. Подход к интерпретации нейросетевых моделей

Одной из основных проблем, возникающих при применении нейросетей для выявления количественных соотношений «структура-свойство» и «структура-активность», обычно считалась неинтерпретируемость нейросетевых моделей. Нейросеть обычно рассматривалась как «черный ящик», способный осуществлять прогноз, но не предоставляющий никакой возможности понять, как он это делает (см., например, [344]). Именно это и считалось основным недостатком применения нейросетевой методологии в химических исследованиях, поскольку для обоснованного использования построенных моделей часто

требуется понимание лежащих в их основе физико-химических и биологических явлений.

Действительно, наборы весовых коэффициентов нейросетей не могут быть непосредственно использованы для интерпретации нейросетевых моделей, поскольку их числовые значения, как правило, меняются при перестроении последних и сильно зависят от особенностей архитектуры нейросетей, например, от числа скрытых нейронов. Все это препятствует их непосредственному использованию для описания моделей «структура-свойство» и «структура-активность» на содержательном уровне.

Следует отметить, что задача интерпретации нейросетевых моделей осознана специалистами в области искусственного интеллекта и частично решена для случая бинарных нейросетей (в которых сигналы на входах и выходах принимают только бинарные значения 0 и 1), для которых разработаны специальные методики извлечения явных правил (типа если..., то...) из нейросетевых моделей [16, 345-348], а также технология «вербализации», позволяющая в автоматическом режиме давать словесное описание таким моделям [349]. Тем не менее, проблема все еще оставалась неразрешенной для нейросетей с непрерывными выходами, а именно такие нейросети используются для построения количественных моделей «структура-свойство» и «структура-активность». Единственным из применимых для этого случая подходов является т.н. анализ «чувствительности» (sensitivity analysis), позволяющий определять относительную важность входов нейросетей путем сравнения ошибок прогнозирования исходной нейросети с ошибками прогнозирования обученных на этих же данных других нейросетей, получаемых из исходной путем удаления по одному каждого из входных нейронов [350]. В этом случае величина возрастания ошибки при удалении входного нейрона определяет его важность (следовательно, и важность соответствующего дескриптора при построении нейросетевых моделей «структура-свойство» и «структура-активность»). Хотя такая характеристика действительно очень важна, однако ее информативность явно уступает тому, что дают методы статистического анализа (например, множественная линейная регрессия, метод частичных наименьших квадратов и др.).

Для решения этой проблемы мы предложили использовать специальный набор описывающих нейросетевые модели статистических характеристик, значения которых, в отличие от значений весовых коэффициентов нейросетей, почти не меняются при перестроении моделей, слабо зависят от числа скрытых нейронов и вполне могут быть использованы для интерпретации нейросетевых моделей. Более того, с их помощью можно анализировать даже такие характеристики соотношений «структура-свойство» и «структура-активность», которые обычно невозможно извлечь при помощи стандартных статистических подходов и которые, как будет показано ниже, могут быть важны для понимания соответствующих физико-химических и биологических процессов. Но сначала, для лучшего понимания сущности предлагаемого подхода, рассмотрим, как может быть интерпретируемо уравнение множественной линейной регрессии.

Пусть функция f линейна по переменным x и y :

$$f(x, y) = a \cdot x + b \cdot y + c \quad (74)$$

Значения коэффициентов a , b и c такой функции могут быть найдены по методу множественной линейной регрессии исходя из известных значений x , y и f для набора описываемых ими объектов (точек). Влияние x на f описывается при помощи коэффициента a , представляющего собой значение частной производной функции f по отношению к переменной x , причем оно одинаково для всех объектов (пронумеруем все объекты от 1 до N):

$$a = \left. \frac{\partial f(x, y)}{\partial x} \right|_{x=x_1, y=y_1} = \dots = \left. \frac{\partial f(x, y)}{\partial x} \right|_{x=x_N, y=y_N} \quad (75)$$

По аналогии, влияние y на f выражается посредством коэффициента b , равного одинаковым значениям частной производной функции f по отношению к переменной y на всех N объектах выборки:

$$b = \left. \frac{\partial f(x, y)}{\partial y} \right|_{x=x_1, y=y_1} = \dots = \left. \frac{\partial f(x, y)}{\partial y} \right|_{x=x_N, y=y_N} \quad (76)$$

Таким образом, уравнение линейной регрессии может быть интерпретировано при помощи регрессионных коэффициентов a и b , выражающих влияние соответствующих переменных на значение функции. Заметим, однако, что

выражение (74) можно рассматривать как начало разложения Тэйлора-Маклорена функции $f(x,y)$ в окрестности точки $(0,0)$:

$$f(x, y) = f(0,0) + \left. \frac{\partial f(x, y)}{\partial x} \right|_{x=0, y=0} \cdot x + \left. \frac{\partial f(x, y)}{\partial y} \right|_{x=0, y=0} \cdot y + \frac{1}{2} \left. \frac{\partial^2 f(x, y)}{\partial x^2} \right|_{x=0, y=0} \cdot x^2 +$$

$$+ \frac{1}{2} \left. \frac{\partial^2 f(x, y)}{\partial y^2} \right|_{x=0, y=0} \cdot y^2 + \frac{1}{2} \left. \frac{\partial^2 f(x, y)}{\partial x \partial y} \right|_{x=0, y=0} \cdot xy + \dots \quad (77)$$

Основная идея разработанного нами подхода состоит в использовании статистических характеристик, основанных на коэффициентах в разложении функции по Тэйлору-Маклорену, для интерпретации нейросетевых моделей.

Рассмотрим теперь, как извлечь из набора данных ту же информацию о влиянии x и y на f при помощи нейросетей. В этом случае, если при построении нейросетевой модели отождествить x и y со входами нейросети, а f с ее выходом, то влияние x на f может быть выражено при помощи среднего значения частной производной функции f по отношению к переменной x , усредненного по всей выборке:

$$a \sim M_x = \frac{1}{N} \sum_{i=1}^N \left. \frac{\partial f(x, y)}{\partial x} \right|_{x=x_i, y=y_i} \quad (78)$$

Здесь основное отличие от рассмотренного выше случая множественной линейной регрессии состоит в том, что значения частной производной может несколько отличаться на разных точках вследствие нелинейности функции f , что и обуславливает необходимость усреднения. Аналогично, влияние другой переменной y на функцию f может быть выражено при помощи усредненного значения частной производной по отношению к ней:

$$b \sim M_y = \frac{1}{N} \sum_{i=1}^N \left. \frac{\partial f(x, y)}{\partial y} \right|_{x=x_i, y=y_i} \quad (79)$$

Итак, предлагаются следующие статистические характеристики для интерпретации нейросетевых моделей:

- M_x – среднее значение первой частной производной по выборке:

$$M_x = \frac{1}{N} \sum_{i=1}^N \left. \frac{\partial f(x, \dots)}{\partial x} \right|_{x=x_i, \dots} \quad (80)$$

- D_x – среднее значение дисперсии первой частной производной по выборке:

$$D_x = \frac{1}{N} \sum_{i=1}^N \left(\left. \frac{\partial f(x, \dots)}{\partial x} \right|_{x=x_i, \dots} - M_x \right)^2 \quad (81)$$

- M_{xx} – среднее значение второй частной производной по выборке:

$$M_{xx} = \frac{1}{N} \sum_{i=1}^N \left. \frac{\partial^2 f(x, \dots)}{\partial x^2} \right|_{x=x_i, \dots} \quad (82)$$

- M_{xy} – среднее значение второй смешанной частной производной по отношению к двум переменным:

$$M_{xy} = \frac{1}{N} \sum_{i=1}^N \left. \frac{\partial^2 f(x, \dots)}{\partial x \partial y} \right|_{x=x_i, y=y_i, \dots} \quad (83)$$

Еще одна статистическая характеристика I_x (сумма квадратов значений первой частной производной) может быть использована (и реально используется в программном комплексе NASAWIN) для определения относительной важности переменных:

$$I_x = \sum_{i=1}^N \left(\left. \frac{\partial f(x, \dots)}{\partial x} \right|_{x=x_i, \dots} \right)^2 \quad (84)$$

Для многослойной нейросети с обратным распространением ошибки значения первых частных производных $\left. \frac{\partial f}{\partial x} \right|_{x=x_i, \dots}$ могут быть легко получены из значений величин δ на входных нейронах, тогда как значения вторых частных производных $\left. \frac{\partial^2 f}{\partial x^2} \right|_{x=x_i, \dots}$ и $\left. \frac{\partial^2 f}{\partial x \partial y} \right|_{x=x_i, y=y_i, \dots}$ можно вычислить по методу конечных разностей. Значение M_x можно рассматривать как аналог коэффициента в уравнении линейной регрессии для переменной x , D_x выражает степень нелинейности функции по отношению к переменной x , а M_{xx} описывает взаимодействие между переменными x и y . Остановимся подробнее на использовании этих статистических характеристик для выявления типов нелинейного характера зависимости.

Пусть функция f линейна по своим аргументам – переменным x и y :

$$f(x, y) = ax + by \quad (85)$$

В этом случае только значения M_x и M_y будут ненулевыми ($M_x = a$, $M_y = b$), тогда как значения других статистических характеристик будут равно нулю ($D_x = 0$, $D_y = 0$, $M_{xx} = 0$, $M_{xy} = 0$, $M_{yy} = 0$). Параболическая зависимость

$$f(x, y) = ax^2 + by^2 \quad (86)$$

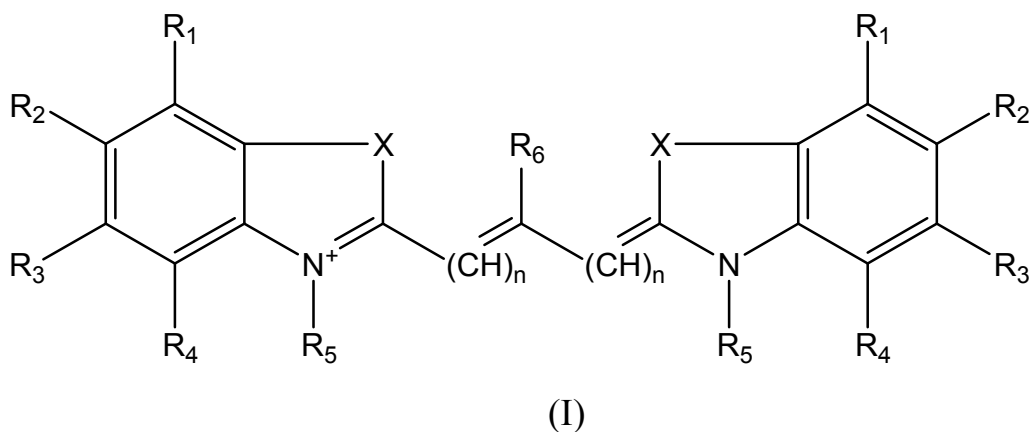
может быть выявлена по ненулевым значениям статистических характеристик M_{xx} и M_{yy} ($M_{xx} = a$, $M_{yy} = b$) и по нулевому значению M_{xy} . Гиперболический характер зависимости

$$f(x, y) = axy \quad (87)$$

может быть определен по ненулевому значению M_{xy} ($M_{xy} = a$) и нулевым значениям M_{xx} и M_{yy} .

Следует также отметить, что рассматриваемые статистические характеристики вполне могут быть использованы при дискретных и даже при булевых (индикаторных) значениях переменных, хотя характер интерпретации в последнем случае несколько иной: значения M_x тогда обозначают вклад наличия определенного признака у объекта в значение функции (например, вклад фрагмента X химической структуры в значение какого-либо свойства химического соединения), а значения M_{xy} – либо (если Y – непрерывная переменная, а X – булева) влияние признака X на M_y , т.е. на характер зависимости функции от ее аргумента y (например, влияние наличие фрагмента X внутри химической структуры на зависимость какого-либо свойства химического соединения от значения дескриптора Y) либо (если X и Y – булевы переменные) вклад конъюнкции признаков X и Y в значение функции (например, вклад факта одновременного присутствия фрагментов X и Y в химическом соединении в значение его свойства).

Рассмотрим теперь, каким образом введенные выше статистические характеристики могут быть использованы для интерпретации нейросетевой модели «структура-свойство», на примере предсказания положения длинноволновой полосы поглощения цианиновых красителей (I) в этаноле.



Подробно построение нейросетевой модели для этого случая рассмотрено в разделе 7.1.1 данной диссертационной работы, поэтому здесь мы остановимся лишь на возможности дать ей содержательную интерпретацию при помощи рассматриваемых статистических характеристик. Для целей интерпретации была отобрана модель, построенная при помощи трехслойной нейросети с 10 скрытыми нейронами, показавшая наилучшую прогнозирующую способность на контрольной выборке. В качестве дескрипторов использованы энергии НОМО и LUMO, рассчитанные при помощи полуэмпирического квантово-химического метода PM3, длина (число ацетиленовых фрагментов) полиметиновой цепочки N , а также индикаторные переменные X : X_S (для $\mathbf{X} = \text{S}$, см. структурную формулу), X_N (для $\mathbf{X} = \text{N}$), X_O (для $\mathbf{X} = \text{O}$), X_{CC} ($\mathbf{X} = -\text{CH}=\text{CH}-$), X_{CCC} ($\mathbf{X} = \text{C}(\text{CH}_3)_2$).

Табл. 1. Значения статистических характеристик

Дескриптор X	M_x	D_x	M_{xx}	M_{xy}
$E_{\text{НОМО}}$	97.8	37.6	0.293	-1.043 ($Y = X_N$)
N	94.4	37.4	0.408	-0.692 ($Y = X_N$)
E_{LUMO}	-39.3	17.2	0.767	0.521 ($Y = E_{\text{НОМО}}$)
X_N	-26.8	9.1	0.522	-1.043 ($Y = E_{\text{НОМО}}$)
X_O	-23.1	9.0	0.082	-0.278 ($Y = E_{\text{НОМО}}$)
X_S	-20.3	7.7	0.031	0.277 ($Y = X_N$)
X_{CCC}	-10.6	4.1	0.019	-0.153 ($Y = E_{\text{НОМО}}$)
X_{CC}	3.7	2.1	-0.041	-0.425 ($Y = E_{\text{НОМО}}$)

В Табл. 1 приведены названия дескрипторов вместе со значениями рассматриваемых статистических характеристик для каждого из них. Для удобства рассмотрения значения M_x и D_x приведены в первоначальной форме, тогда как значения M_{xx} , D_{xx} и M_{xy} шкалированы таким образом, чтобы разброс значений всех дескрипторов и свойств был одинаков. Дескрипторы в таблице отсортированы в порядке возрастания абсолютного значения M_x . В результате анализа приведенных в таблице данных можно прийти к выводу, что нейронная сеть четко отделила влияние размера энергетической щели между граничными молекулярными орбиталями НОМО и LUMO от влияния описывающего электронную корреляцию конфигурационного взаимодействия на положение длинноволновой полосы поглощения красителя. Согласно значению статистической характеристики M_x , длина полиметиновой цепочки N является одним из наиболее важных параметров, влияющих на положение этой полосы поглощения, причем это влияние не связано напрямую с величиной энергетической щели между граничными орбиталями. Возможное объяснение этого эффекта состоит в том, что при удлинении полиметиновой цепочки увеличивается плотность одноэлектронных уровней вблизи граничных орбиталей, что приводит к усилению взаимодействия электронных конфигураций, получаемых при электронных переходах между этими уровнями, что, в свою очередь, приводит к уменьшению энергетической щели между основным и первым возбужденным электронными состояниями, и, значит, к батохромному сдвигу длинноволновой полосы поглощения.

Следующими по важности двумя дескрипторами являются $E_{НОМО}$ и E_{LUMO} . Для них значения статистической характеристики M_x можно интерпретировать следующим образом: основной вклад в длинноволновую полосу поглощения вносит переход электрона с НОМО на LUMO. Действительно, длина волны поглощаемого света, вызывающего этот электронный переход, должна быть обратно пропорциональна разнице между этими энергетическими уровнями:

$$\lambda \propto \frac{1}{E_{LUMO} - E_{НОМО}} \quad (88)$$

В соответствии с выражением (88), значения частной производной λ по отношению к E_{LUMO} должны быть отрицательными во всех точках, тогда как соответствующие значения частной производной λ по отношению к E_{HOMO} – положительными. Это точно соответствует знакам приведенных в Табл. 1 значений M_x , а также тому, что значения D_x существенно меньше, чем M_x . Таким образом, данные из Табл. 1 согласуются с формулой (88) и, следовательно, с вышеизложенной интерпретацией.

Следующими по важности являются индикаторные переменные, определяющие тип гетероциклов. Отрицательные (и меньшие по абсолютной величине по сравнению с N , E_{HOMO} и E_{LUMO}) значения M_x для X_N , X_O и X_S можно объяснить исходя из того, что введение атомов азота, кислорода и серы в соответствующее положение в цианиновых красителях приводит к понижению плотности одноэлектронных энергетических уровней вблизи граничных орбиталей, что приводит к уменьшению взаимодействия соответствующих электронных конфигураций (см. рассуждение выше) и, как следствие, к гипсохромному сдвигу длинноволновой полосы поглощения света. Это предположение отчасти подтверждается существенно меньшими по величине значениями M_x для X_{CC} и X_{CCC} .

Рассмотрим теперь значения статистических характеристик M_{xx} и M_{xy} , описывающих нелинейный характер нейросетевой модели. Данные из Табл. 1 свидетельствуют о том, что зависимость λ от E_{LUMO} является наиболее «параболической» - она напоминает отрицательную ветвь (поскольку значение M_x отрицательно, а M_{xx} положительно) параболы $y = x^2$. В принципе, это не противоречит выражению (88), поскольку определяемая этой формулой часть гиперболы действительно локально близка по форме к отрицательной ветви параболы. Тем не менее, относительно небольшое значение M_{xx} в сочетании с относительно большим (в сравнении с M_x) значением D_{xx} для E_{HOMO} указывает на более сложный характер нелинейной зависимости λ от E_{HOMO} . Можно предположить, что причиной этого является сильное взаимодействие между E_{HOMO} и X_N , о чем свидетельствует большое отрицательное значение перекрестного члена M_{xy} между ними. Это взаимодействие может быть объяснено большим вкладом атом-

ных орбиталей азота в высшую занятую молекулярную орбиталь НОМО. Подобным же образом может быть объяснено отрицательное значение M_{xy} для $E_{НОМО}$ и X_O , причем его меньшее значение (по сравнению с X_N) может быть объяснено разницей в электроотрицательности между азотом и кислородом, ведущей к меньшей вовлеченности атомных орбиталей кислорода в систему π -электронную систему сопряжения в молекуле красителя. По аналогии, большее значение M_{xy} для $E_{НОМО}$ и X_{CC} может быть объяснено прямым вовлечением двойной связи $C=C$ в сопряжение, тогда как меньшее значение M_{xy} для $E_{НОМО}$ и X_{CCC} можно объяснить отсутствием сопряжения между группой $C(CH_3)_2$ и π -электронной системой в молекуле красителя.

Из приведенного выше рассмотрения можно сделать вывод, что получаемая при анализе нейросетевой модели интерпретация вполне согласуется с основными физическими принципами теории поглощения света.

Таким образом, при использовании статистических характеристик, описывающих распределение получаемых нейронной сетью значений частных производных переменных, соответствующих свойствам/активностям химических соединений, по отношению к переменным, соответствующим значениям дескрипторов, возможно извлечь из набора данных не только такую же информацию (например, о влиянии дескрипторов на свойства/активности), что и «традиционные» статистические подходы (например, методы линейного регрессионного анализа), но и получить дополнительную ценную информацию о нелинейном характере зависимостей структура-свойство и структура-активность. Подобная информация является уникальной, поскольку ее крайне затруднительно извлечь из набора данных как с использованием параметрических методов статистики (в которых тип нелинейности «зашит» в используемые параметрические уравнения), так и при помощи непараметрических подходов (которые довольно плохо аппроксимируют распределения частных производных). В то же время, ценность подобной информации несомненна.

Таким образом, искусственные нейронные сети больше не являются «черными ящиками», не поддающимися интерпретации. Напротив, при исполь-

зовании рассмотренных выше статистических характеристик, они являются ценными инструментами анализа химических данных.

4.3. Концепция обучаемой симметрии

Формулировка проблемы. Классический подход к выявлению количественной зависимости структура-активность (QSAR) для узкой серии соединений, обладающих одинаковым скелетом, предполагает использование в качестве дескрипторов параметров заместителей (например, константы Гамметта σ , константы Тафта E_s , константы липофильности π , параметров STERIMOL: L , B_1 , B_2 и др.). В этом случае может возникнуть проблема тогда, когда несколько положений заместителей топологически эквивалентны (например на Рис. 27а положения 2 и 6, а также 3 и 5, топологически эквивалентны).

Рассмотрим эту проблему на следующем простом примере. Пусть база данных содержит монозамещенные пиридины **A-D**, приведенные на Рис. 27b. Соединения **A** и **B** несут заместители в положении 2, тогда как соединения **C** и **D** – в положении 6, которое топологически эквивалентно положению 2. Возникает вопрос: как можно построить единое уравнение QSAR для всей базы с использованием констант заместителей, находящихся в этих положениях? Наиболее очевидным решением для этого случая было бы введение канонической нумерации для положений замещения, при которой заместители во всех структурах оказались бы присоединенными к одному положению, например, к положению 2 (как в структурах **G-H** на Рис. 27b). Такая база данных могла бы быть использована для разработки уравнений QSAR с использованием параметров заместителей для положения 2 в качестве дескрипторов.

Дополним теперь базу данных пиридинами **I** и **J** с двумя заместителями в положениях 2 и 6 (см. Рис. 27c). Решение проблемы для этого случая уже не столь очевидно, как в предыдущем случае, поскольку различные критерии канонизации могут привести к различной нумерации положений замещения и, следовательно, к разным уравнениям QSAR. На Рис. 27c показана нумерация положений замещения для двух критериев канонизации: (i) заместители с лек-

сикографически «младшими» именами присоединяются в положения с меньшими номерами (структуры **K** и **M**), и (ii) заместители с меньшим числом электронов присоединяются к положениям с меньшими номерами (структуры **L** и **N**). Очевидно, что применение подобных произвольных критериев не может составлять серьезную основу для исследований в области QSAR, поскольку построенные подобным образом уравнения QSAR зависят от факторов, не релевантных природе зависимости структура-активность.

Альтернативный подход к решению этой проблемы заключается в использовании *симметрических функций* параметров заместителей в качестве независимых переменных в уравнениях QSAR (см. Рис. 28). Простейшим примером подобных функций является сумма σ -констант всех заместителей (см., например, работу [351]). На практике, однако, в подобных случаях успех QSAR/QSPR-анализа в значительной степени зависит от способности исследователя предложить по интуиции форму симметрических функций. В ряде случаев подобный подход действительно работает. Например, если величина какого-либо биологического эффекта строго аддитивна по всем топологически эквивалентным положениям замещения, то использование суммы констант заместителей, находящихся в этих положениях, в качестве симметрической функции может привести к построению хорошей QSAR-модели. Тем не менее, основанные на интуиции подходы не всегда приводят к наилучшему решению.

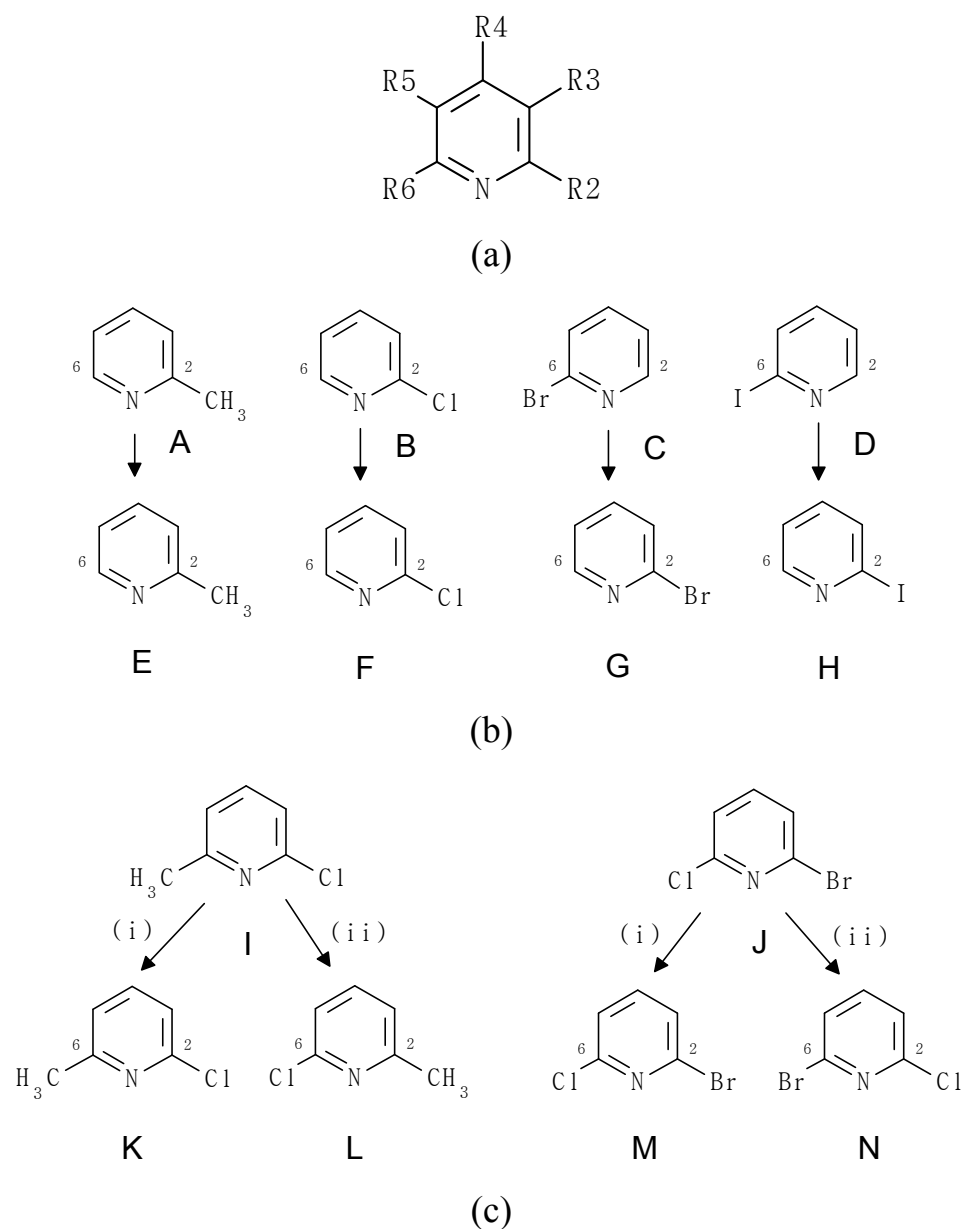
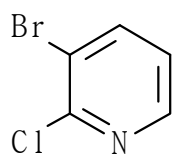


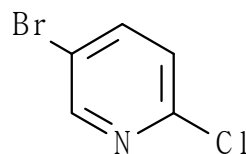
Рис. 27. (a) Существуют две пары топологически эквивалентных положения заместителей в пиридинах: два α -положения, несущие заместители R_2 и R_6 , и два β -положения для заместителей R_3 и R_5 . (b) Выборка α -замещенных пиридинов до канонизации нумерации положений замещений для критериев (i) и (ii), обсуждаемых в тексте.

Еще одним недостатком этого подхода является то, что простые наборы симметрических функций не всегда могут различить разные соединения и, следовательно, предсказать для них разные величины активности (см. пример на Рис. 28).

$f_1 = \sigma_2 + \sigma_6$
$f_2 = \pi_2 + \pi_6$
$f_3 = \sigma_3 + \sigma_5$
$f_4 = \pi_3 + \pi_5$
$f_5 = \pi_2 + \pi_3 + \pi_4 + \pi_5 + \pi_6$



O



P

Рис. 28. Примеры симметрических функций, которые могли бы быть использованы для построения QSAR-моделей. Значения этих функций не меняются при перестановках заместителей по топологически эквивалентных положениям: $2 \Leftrightarrow 6$ и $3 \Leftrightarrow 5$. Очевидно, что значения всех пяти симметрических функций одинаковы для соединений O и P, хотя их биологическая активность может отличаться.

Следует отметить, что эта проблема для случая 2,6-замещенных пиридинов и одного параметра заместителей может быть решена путем использования двух симметрических функций (а именно *суммы* и *произведения* параметров заместителей) в качестве независимых переменных при построении зависимости структура-активность. Действительно, из основной теоремы о симметрических многочленах [352] следует, что любая симметрическая функция от двух аргументов может быть представлена как многочлен от суммы и произведения этих аргументов:

$$f(x, y) \equiv f(y, x) \Leftrightarrow f(x, y) = P(x + y, x \cdot y) \quad (89)$$

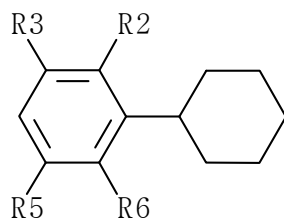
где P – произвольный многочлен от двух переменных.

В рассматриваемом примере мы принимаем, что переменная x – это σ -константа заместителя в положении 2, а переменная y – это σ -константа заместителя в положении 6. Следовательно, для решения проблемы необходимо: (а) создать две симметрические функции (*сумму* и *произведение* σ -констант) и использовать их в качестве независимых переменных при построении QSAR-моделей, (б) применить статистический метод, способный выявлять нелинейные зависимости между переменными (поскольку многочлен в общем случае является нелинейной функцией) для анализа количественной зависимости «структура-активность» произвольной функциональной сложности. Функция f в уравнении (89) является *инвариантной* относительно перестановки переменных x и y , тогда как функции $x + y$ и $x \cdot y$ являются *базисным набором инвариантов*, поскольку через них может быть выражен любой инвариант. В принципе, любой базисный набор инвариантов относительно группы автоморфизмов графа, представляющего собой наибольший общий фрагмент (скелет) набора химических соединений, и действующей на множестве положений заместителей на этом скелете, может быть использован для построения симметрических функций. Тем не менее, необходимые базисные наборы инвариантов известны лишь для простейших групп подстановок. Если мы добавим в нашу выборку химических соединений пиридины, замещенные по другим положениям, либо, в дополнение к σ -константам, будем использовать еще и другие параметры заместителей, уравнение (89) уже не сможет быть применено, и нам придется либо формулировать и доказывать математические теоремы для каждого конкретного случая, либо искать альтернативные подходы к решению проблемы.

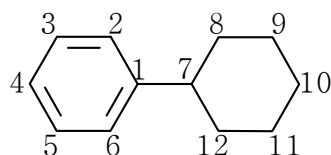
Следует подчеркнуть, что существует одно неперемutable условие, которому должна удовлетворять любая функция, инвариантная относительно перестановки своих аргументов: ее общий вид должен быть нелинейным относительно этих аргументов. Докажем это математически. Произвольная линейная функция от двух аргументов x и y может быть выражена как

$f(x, y) = a \cdot x + b \cdot y + c$, где a , b и c – произвольные коэффициенты. Очевидно, что в этом случае $f(x, y) \equiv f(y, x)$ тогда и только тогда, когда $a = b$. Следовательно, для общего случая $a \neq b$ симметрическая функция f не может быть линейной. Аналогичным образом можно доказать, что общий вид симметрической функции от трех аргументов не может быть ни линейным ни квадратичным. Это означает, что такие традиционные статистические подходы, как линейный и квадратичный регрессионный анализ, метод частичных наименьших квадратов (PLS) и др., не могут быть применены для нахождения произвольной симметрической функциональной зависимости в наборе экспериментальных данных. Следовательно, только методы (например, искусственные нейронные сети), способные аппроксимировать нелинейные функции произвольного вида, могут быть использованы для этой цели.

Решение проблемы. Для решения этой проблемы мы предлагаем: (а) расширить обучающую выборку соединений в N раз (где N – порядок группы подстановок, действующей на множестве позиций присоединения заместителей к общей подструктуре и индуцированной действующей на ней группой автоморфизмов, см. Рис. 29) путем добавления копий соединений с той же активностью, но различающихся перестановкой топологически эквивалентных позиций присоединения заместителей (см. Рис. 30), и (б) использовать искусственные нейронные сети для выявления количественной зависимости «структура-активность».



Общая формула соединений выборки



Общая подструктура для выборки

$$A = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 2 & 6 & 2 & 6 \\ 3 & 5 & 3 & 5 \\ 4 & 4 & 4 & 4 \\ 5 & 3 & 5 & 3 \\ 6 & 2 & 6 & 2 \\ 7 & 7 & 7 & 7 \\ 8 & 8 & 12 & 12 \\ 9 & 9 & 11 & 11 \\ 10 & 10 & 10 & 10 \\ 11 & 11 & 9 & 9 \\ 12 & 12 & 8 & 8 \end{pmatrix} \Rightarrow B = \begin{pmatrix} 2 & 6 \\ 3 & 5 \\ 5 & 3 \\ 6 & 2 \end{pmatrix}$$

Рис. 29. Группа автоморфизмов **A** общей подструктуры (подграфа) для набора соединений индуцирует группу **B**, действующую на множестве четырех положений замещения 2, 3, 5, 6. Группа **B** определяет каким образом исходная выборка должна быть расширена за счет добавления копий соединений с переставленными заместителями.

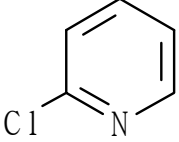
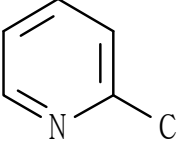
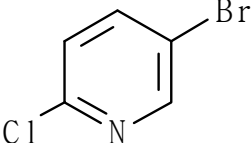
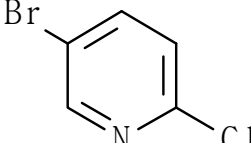
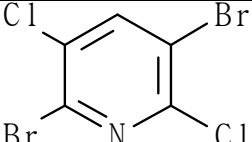
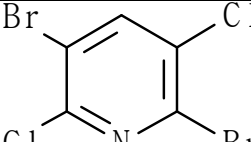
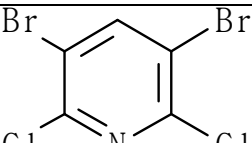
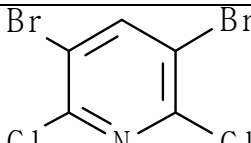
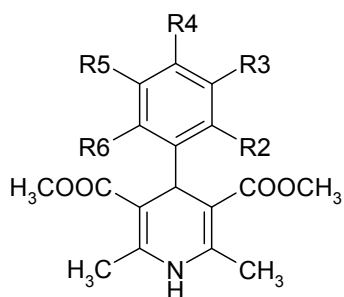
	
	
	
	

Рис. 30. Структуры из левой колонки таблицы должны быть дополнены структурами из правой колонки для того, чтобы нейронная сеть могла обучиться необходимым свойствам симметрии. В некоторых случаях, когда общая подструктура замещена симметрично (как в четвертой строке), это означает, что подобные структуры должны быть дублированы (либо им приписан весовой фактор 2).

В этом случае нейронные сети обучаются строить нелинейные зависимости «структура-активность» с необходимыми свойствами симметрии. В начале обучения нейронные сети дают разные предсказания активности для соединений с перестановкой заместителей в эквивалентных положениях (например, для 2- и 6-хлорпиридинов, которые, очевидно, эквивалентны), поскольку подгоночные параметры нейросети инициализируются случайными числами, однако в процессе обучения эта разность становится незначительной. Следует отметить, что при построении реальных количественных зависимостей «структура-активность» на выборках небольшого размера эта разность исчезает не полностью, поскольку максимально приемлемое число скрытых нейронов, не приводящее к сильному «переучиванию», обычно меньше минимального числа скрытых нейронов, необходимого для полного ее исчезновения. На практике, одна-

ко, эта разница всегда оказывается значительно меньшей, чем погрешность нейросетевой модели. Следовательно, при осуществлении прогноза по такой нейросетевой модели для нового соединения необходимо сделать прогнозы для всех копий соединения и результирующие значения усреднить.

Пример 1. Блокаторы кальциевых каналов L-типа. В этом примере рассматривается применение концепции обучаемой симметрии к изучению количественных соотношений «структура-активность» для принадлежащих к 1,4-дигидропиридиновому ряду блокаторов кальциевых каналов L-типа (II).



II

Данные по биологической активности были взяты со статьи [353]. Ранее эти данные уже были обработаны с использованием констант заместителей в качестве дескрипторов и линейного регрессионного анализа для получения статистической модели [354], однако оценки прогнозирующей способности построенной модели в этой работе сделано не было. В работе [353] в дополнение к константам заместителей в качестве дескрипторов были использованы еще и топологические индексы, а для построения количественной модели «структура-активность» была применена искусственная нейронная сеть в комбинации с факторным анализом для предобработки дескрипторов. Прогнозирующая способность полученной нейросетевой модели оказалась в этой работе не очень высокой (хотя и значительно лучшей по сравнению с линейно-регрессионной моделью, построенной на том же наборе дескрипторов): наилучшее значение коэффициента корреляции 0.733, а наименьшая среднеквадратичная ошибка прогноза 1.019 логарифмических единиц.

В нашем исследовании исходная выборка, состоящая из 46 соединений, была расширена в 2 раза за счет добавления копий соединений с переставлен-

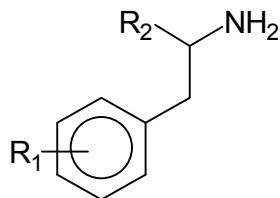
ными позициями присоединения заместителей, а полученная выборка из 92 соединений была случайным образом разбита на обучающую выборку из 84 соединений и контрольную выборку из 8 соединений. Было использовано 5 дескрипторов, описывающих заместители: π -константы для *para*- (R_4) и двух *meta*-положений (R_3 и R_5) и E_s -константы двух *ortho*-положений (R_2 и R_6). Эффективная концентрация, вызывающая 50% блокирование кальциевых каналов ($\log(1/EC_{50})$), была взята в качестве биологической активности, которая была нами скоррелирована со значениями этих пяти дескрипторов при помощи многослойной нейросети с обратным распространением ошибок, включающей два скрытых нейрона (при большем числе скрытых нейронов наблюдалась худшая предсказательная способность нейросетевых моделей). Нейросеть была обучена по стандартному алгоритму «обобщенного дельта-правила». В процессе обучения не было «переучивания» (среднеквадратичная ошибка на контрольной выборке все время уменьшалась), и оно было остановлено после 200.000 итераций, когда изменение среднеквадратичной ошибки на обучающей выборке на 100 последовательных итерациях стало меньше 0.001 логарифмических единиц. В результате обучения нейросети среднеквадратичная ошибка на обучающей выборке составила 0.79 логарифмических единиц (коэффициент корреляции 0.832), а на контрольной выборке – 0.71 логарифмическая единица.

Для сравнения мы провели аналогичное исследование с тем же самым набором дескрипторов при той же самой разбивке базы на обучающую и контрольную выборки, но без дублирования соединений. В этом случае среднеквадратичная ошибка на обучающей выборке оказалась 0.70 логарифмических единиц (коэффициент корреляции 0.87), а на контрольной выборке – 1.59 логарифмических единиц. Таким образом, расширение базы за счет добавления копий соединений с переставленными эквивалентными позициями присоединения заместителей обеспечило значительное повышение прогнозирующей способности нейросетевой модели (среднеквадратичная ошибка на контрольной выборке упала с 1.59 до 0.71 логарифмической единицы). Далее, мы применили построенную на исходной (нерасширенной) выборке нейросетевую модель для прогнозирования активности всех клонов (т.е. тех же самых соединений, но с

переставленными эквивалентными позициями присоединения заместителей). Среднеквадратичная ошибка прогноза в этом случае оказалась 1.57 логарифмических единиц. Поскольку клоны являются теми же самыми соединениями с той же самой биологической активностью, то можно сделать вывод, что нейросеть, обученная на исходном нерасширенном наборе данных, неспособна корректно воспроизвести свойства симметрии в количественных зависимостях «структура-активность».

Таким образом, можно сделать вывод, что нейросеть, обученная на расширенном наборе данных делает более корректные предсказания по сравнению с нейросетью, обученной на исходном наборе данных. Эти эксперименты были повторены нами для различных разбинок набора соединений на обучающую и контрольную выборки, и во всех случаях общая картина оставалась неизменной.

Пример 2. Галлюциногенная активность фенилалкиламинов. Целью данного исследования явилось изучение применимости концепции обучаемой симметрии на примере галлюциногенной активности фенилалкиламинов (III).



III

Данные по галлюциногенной активности этой группы соединений взяты из работы [355]. Поскольку в исходном наборе химических структур имелось несколько двухпозиционных «мостиковых» заместителей для R_1 , для которых не определены константы заместителей, мы их преобразовали путем «разрезания» в однопозиционные (например, «мостиковые» заместители 4,5-(OCH_2O) и 4,5-($\text{OCH}_2\text{CH}_2\text{O}$) были преобразованы в однопозиционные заместители 4- OCH_3 и 5- OCH_3). Для поиска количественных зависимостей «структура-активность» мы использовали набор из 7 дескрипторов: σ -константы для двух *орто*-положений и π -константы для двух *мета*- и одного *пара*-положения в R_1 , а также индикаторную переменную, указывающую на присутствие алкильного

заместителя в R_2 . Исходная выборка, включающая 35 соединений, была, как и в предыдущем примере, удвоена, и получившиеся 70 соединений были случайным образом разбиты на обучающую и контрольную выборки в соотношении 10:1. Как и в предыдущем примере, была использована многослойная ИНС с обратным распространением ошибок с двумя скрытыми нейронами. При обучении не наблюдался эффект «переучивания», что, как и в предыдущем примере, сделало ненужным использование третьей выборки для объективной оценки прогнозирующей способности нейросетевой модели. В результате обучения среднеквадратичная ошибка составила 0.55 логарифмических единиц на обучающей выборке (коэффициент корреляции 0.932) и 0.47 логарифмических единиц на контрольной выборке. Как и в предыдущем примере, мы повторили построение модели с использованием исходного (нерасширенного) набора данных. В этом случае уже наблюдался сильный эффект «переучивания» вследствие неблагоприятного соотношения между числом соединений и числом подстроечных параметров в нейросети. Среднеквадратичная ошибка нейросетевой модели, взятой при прохождении среднеквадратичной ошибки на контрольной выборке через минимум (т.е. до начала «переучивания»), составила 0.89 логарифмических единиц на обучающей выборке (коэффициент корреляции 0.82) и 0.54 логарифмические единицы на контрольной выборке, тогда как «переученная» нейросеть показала ошибку в 0.49 логарифмических единиц на обучающей выборке (коэффициент корреляции 0.95) и 0.98 логарифмических единиц на контрольной выборке. Обе эти модели дали близкие среднеквадратичные ошибки при прогнозировании галлюциногенной активности «клонов» исходных соединений (1.19 и 1.15 логарифмических единиц). Таким образом, расширение исходной выборки соединений за счет их «клонов» (получаемых путем перестановок эквивалентных позиций присоединения заместителей) позволило улучшить соотношение между числом соединений в выборке и числом подстроечных параметров нейросети (70:17 против 35:17), что, в свою очередь, привело к улучшению качества нейросетевой модели.

Следует отметить, что построенные нами количественные модели «структура-активность» существенно лучше опубликованных (обзор известных моде-

лей приведен в работе [355]): все опубликованные модели построены только на небольших подмножествах использованного в нашей работе набора соединений (коэффициенты корреляции варьируются от 0.79 для выборки из 26 соединений до 0.97 для выборки из 10 соединений), и ни в одной из работ не оценивалась прогнозирующая способность моделей на контрольной выборке.

Как и в предыдущем случае, все вычислительные эксперименты были повторены для разных разбинок исходных соединений на обучающую и контрольные выборки, и во всех случаях качественные результаты совпали.

Выводы. Нами предложен подход (концепция обучаемой симметрии), позволяющий осуществлять построение количественных моделей «структура-активность» в рамках основанного на параметрах заместителей «классического» подхода для однородных наборов химических соединений с симметричным общим скелетом, позволяющий обходиться без произвольных симметрических функций от констант заместителей. Нейронная сеть в этом случае обучается не только воспроизводить зависимость биологической активности от значений дескрипторов, но и воспроизводить необходимые свойства симметрии в количественных соотношениях «структура-активность». Следует также отметить, что разработанная методология применима не только к «классическому» подходу, основанному на использовании констант заместителей в качестве дескрипторов: она применима к любому исследованию, в котором требуется аппроксимировать количественную зависимость «структура-свойство» или «структура-активность» для симметрично построенных химических систем (при небольшом порядке группы симметрии). Таким образом, концепция обучаемой симметрии позволяет улучшать прогнозирующую способность количественных нейросетевых моделей «структура-активность» и «структура-свойство» за счет использования дополнительной информации о свойствах симметрии этих соотношений.

ГЛАВА 5. РАЗРАБОТКА ФРАГМЕНТНЫХ ПОДХОДОВ

Данная глава содержит описание разработанных нами концепций, методов, программ и алгоритмов, нацеленных на то, чтобы превратить фрагментный подход в мощный инструмент максимально точного моделирования широкого разнообразия свойств органических соединений. В главе не только приводятся способы преодоления перечисленных в разделе 2.3 ограничений фрагментных дескрипторов, но и предлагаются методики, направленные на значительное расширение сферы применения фрагментного подхода.

5.1. Принципы построения и генерации фрагментных дескрипторов

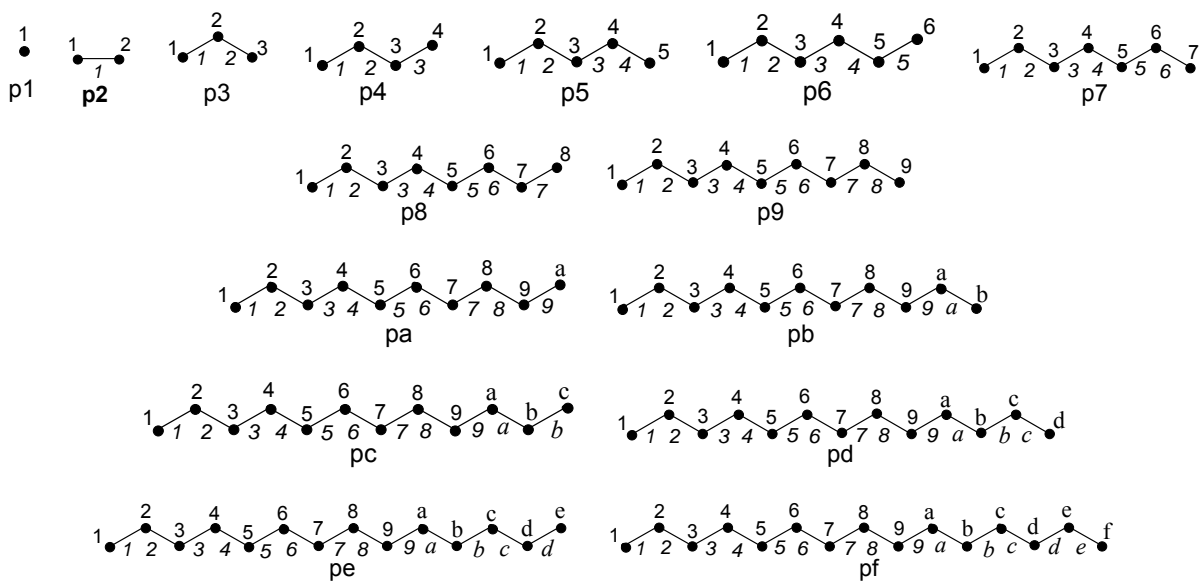
Основными отличительными особенностями разработанного нами варианта фрагментных дескрипторов является чрезвычайная гибкость (и, как следствие, универсальность их применения для моделирования самых разнообразных свойств органических соединений), а также очень высокая производительность их генерации. Гибкость достигается наличием большого числа типов генерируемых фрагментов (см. подраздел 5.1.1) в сочетании с развитой четырехуровневой классификацией типов атомов (см. подраздел 5.1.2), наличием механизма их автоматического обобщения (см. подраздел 5.1.3) и нескольких стратегий комбинирования разных уровней классификации атомов внутри фрагментов (см. подраздел 5.1.4). Эффективность достигается за счет совершенного алгоритма, генерирующего все типы фрагментов за два просмотра структуры, использования оригинального трехуровневого иерархического списка кодов генерируемых фрагментов с очень быстрым доступом к его элементам, а также поддержанием динамически меняющегося списка групп статистически эквивалентных дескрипторов (см. подраздел 5.1.5). Следует также отметить использование оригинальной методики поиска ароматических циклов, а также алгоритмов поиска изоморфных вложений графов и определения их групп автоморфизмов. Важными особенностями также является возможность работы с «вы-

деленными» атомами (см. раздел 5.3), полимерными структурами (см. раздел 5.4) и стереохимической информацией.

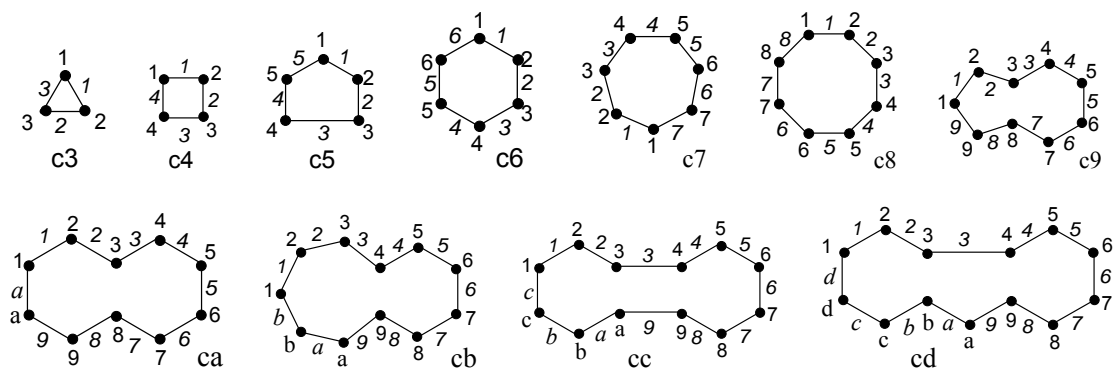
5.1.1. Типы фрагментов

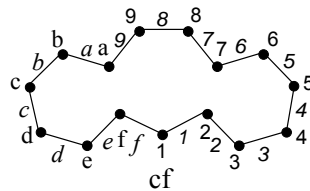
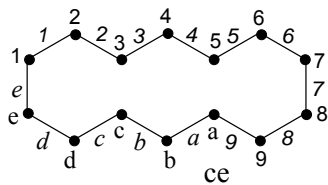
Наш рабочий набор фрагментов, генерируемый программой Fragment (см. раздел 8.3), включает: цепочечные фрагменты длиной от 1 (*p1*) до 15 (*pf*) атомов, циклические: длиной от 3 (*c3*) до 15 (*cf*) атомов, три разветвленных фрагмента (*s4*, *s5*, *s6*), бициклические: длиной от 6 (*b0*) до 15 (*ba*) атомов в различных сочетаниях, трициклические: длиной от 12 (*t0*) до 15 (*ta*) атомов в различных сочетаниях (Рис. 31).

Цепочки

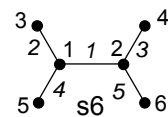
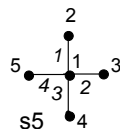
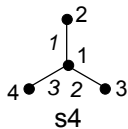


Циклы

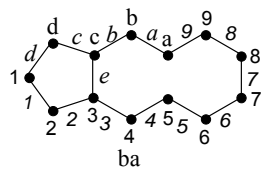
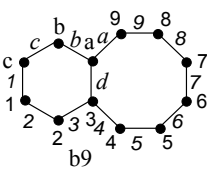
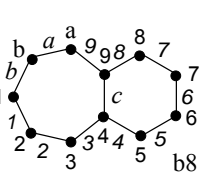
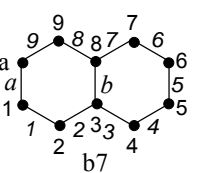
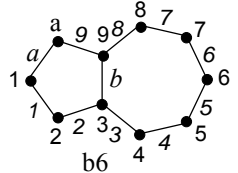
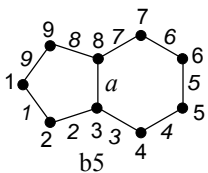
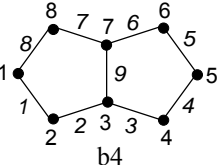
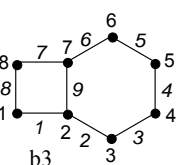
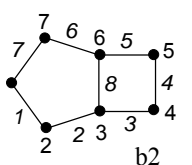
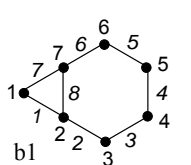
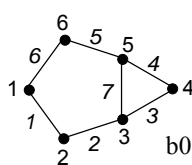




Разветвления



Бициклы



Трициклы

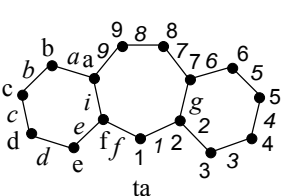
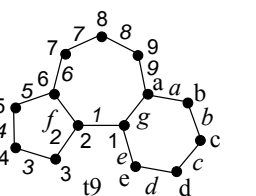
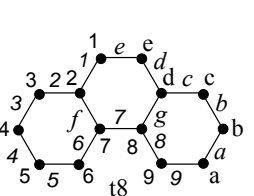
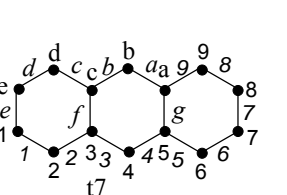
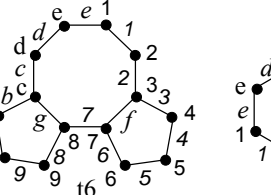
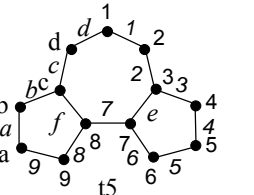
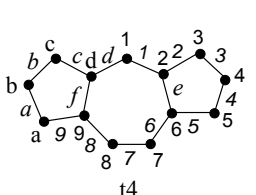
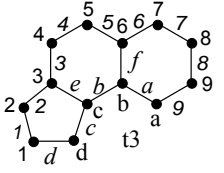
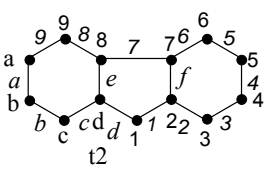
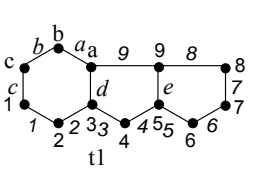
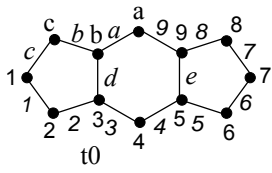


Рис. 31. Типы фрагментов



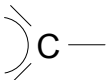


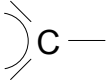
5.1.2. Иерархическая классификация атомов во фрагментах

В рамках нашего подхода была разработана специальная схема иерархической классификации типов атомов, в рамках которой каждому из них соответствует код, состоящий из 3 символов. Классификация основана на следующем принципе: каждый последующий символ конкретизирует предыдущий. Исключения составляют строчные символы, входящие в состав двухбуквенных кодов химических элементов, например, Cl, Br, Si и т.д. Первый символ (или первые два символа для двухбуквенных кодов химических элементов) отражает тип химического элемента, тогда как последующие символы служат для уточнения ближайшего окружения атома. Если для определения типа атома достаточно двух символов, то в качестве третьего используется символ «_». Таким образом, наш подход предусматривает возможность введения нескольких уровней обобщения типов атомов. Кроме того, имеется возможность создания специальных типов атомов, объединяющих по тем или иным принципам несколько типов атомов.

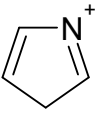
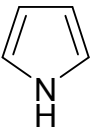
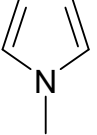
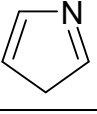
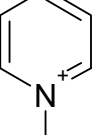
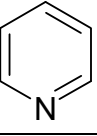

Первый уровень классификации соответствует «обобщенному» типу атомов, типы химических элементов учитываются на втором уровне, тогда как третий и, возможно, четвертый уровни классификации дополнительно учитывают тип гибридизации, связевое окружение атома, его формальный заряд, а также число водородных соседей. Полный набор типов атомов, соответствующих элементам-органогенам (C, N, O, S, Se, P, As, Si, F, Cl, Br, I) приведен в Табл. 2. Следует отметить, что для атомов Se классификация типов атомов аналогична S, для атомов As аналогична P, а для атомов Br и I – аналогична Cl. Щелочные и щелочноземельные металлы обозначаются MXX, где XX - тип химического элемента. Например, атом натрия обозначается MNa, атом калия МК_ и т.д. Все остальные элементы кодируются как XX_, например, La_, Be_ и т.д.

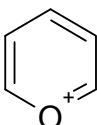
Табл. 2. Классификация типов атомов

Код атома	Условное изображение	Расшифровка кода атома
-----------	----------------------	------------------------

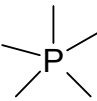
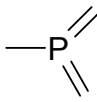
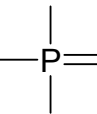
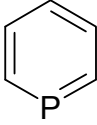
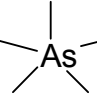
	•	Произвольный атом
C_	C	Атом углерода
CA_	-C _{sp³}	Атом углерода в sp ³ -гибридизации
CA0	CH ₄	Атом углерода в sp ³ -гибридизации, связанный с 4 атомами водорода (атом углерода в составе метана)
CA1	-CH ₃	Атом углерода в sp ³ -гибридизации, связанный с 3 атомами водорода (первичный атом углерода, атом углерода в составе метильной группы)
CA2	>CH ₂	Атом углерода в sp ³ -гибридизации, связанный с 2 атомами водорода (вторичный атом углерода, атом углерода в составе метиленовой группы)
CA3	>CH-	Атом углерода в sp ³ -гибридизации, связанный с 1 атомом водорода (третичный атом углерода, атом углерода в составе метиновой группы)
CA4	>C<	Атом углерода в sp ³ -гибридизации, не связанный с атомами водорода (четвертичный атом углерода)
CD_	=C	Атом углерода при двойной неароматической связи
CD1	=CH ₂	Атом углерода при двойной неароматической связи, имеющий 2 водородных соседей
CD2	=CH-	Атом углерода при двойной неароматической связи, имеющий одного водородного соседа
CD3	=C<	Атом углерода при двойной неароматической связи, имеющий 3 неводородных соседей
CD4	=C=	Атом углерода при двух двойных неароматических связях (алленовый атом углерода)
CB_		Атом углерода, входящий в состав 6-членного ароматического цикла
CB1		Атом углерода, входящий в состав 6-членного ароматического цикла и связанный с 1 атомом водорода
CB2		Атом углерода, входящий в состав 6-членного ароматического цикла и не связанный с атомом водорода
CH_		Атом углерода, входящий в состав 5-членного ароматического цикла
CH1		Атом углерода, входящий в состав 5-членного ароматического цикла и связанный с 1 атомом водорода
CH2		Атом углерода, входящий в состав 5-членного ароматического цикла и не связанный с атомами водорода
CT_	≡C	Атом углерода при тройной связи
CT1	≡CH	Атом углерода при тройной связи, имеющий одного

		водородного соседа
CT2	$\equiv\text{C}-$	Атом углерода при тройной связи, не имеющий водородных соседей
CTN	$\equiv\text{C}^-$	Атом углерода при тройной связи, несущий формальный отрицательный заряд
N_	N	Атом азота
NA_	-N	Атом азота в sp^3 -гибридизации
NA0	NH_3	Формально незаряженный атом азота в sp^3 -гибридизации, связанный с 3 атомами водорода (атом азота в молекуле аммиака)
NA1	$-\text{NH}_2$	Формально незаряженный атом азота в sp^3 -гибридизации, связанной с 2 атомами водорода (атом азота в аминогруппе)
NA2	$>\text{NH}$	Формально незаряженный атом азота в sp^3 -гибридизации, связанный с 1 атомом водорода
NA3	$>\text{N}-$	Формально незаряженный атом азота в sp^3 -гибридизации, не связанный с атомами водорода
ND_	$=\text{N}$	Формально незаряженный атом азота при двойной неароматической связи
ND1	$=\text{NH}$	Формально незаряженный атом азота при двойной неароматической связи, имеющий водородного соседа
ND2	$=\text{N}-$	Формально незаряженный атом азота при двойной неароматической связи, не имеющий водородных соседей
NDN	$=\text{N}^-$	Формально отрицательно заряженный атом азота при двойной связи
NT_	$\equiv\text{N}$	Атом азота при тройной связи (в составе нитрильной группы)
NC_	$>\text{N}^+<$	Формально положительно заряженный атом азота в sp^3 -гибридизации
NC0	NH_4^+	Формально положительно заряженный атом азота в sp^3 -гибридизации, связанный с 4 атомами водорода (атом азота в составе катиона аммония)
NC1	$-\text{NH}_3^+$	Формально положительно заряженный атом азота в sp^3 -гибридизации, связанный с 3 атомами водорода (атом азота в составе протонированной аминогруппы)
NC2	$>\text{NH}_2^+$	Формально положительно заряженный атом азота в sp^3 -гибридизации, связанный с 2 атомами водорода
NC3	$>\text{NH}^+<$	Формально положительно заряженный атом азота в sp^3 -гибридизации, связанный с 1 атомом водорода
NC4	$>\text{N}^+<$	Формально положительно заряженный атом азота в sp^3 -гибридизации, не связанный с атомами водорода
NE_		Формально положительно заряженный атом азота


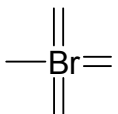
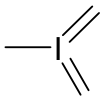
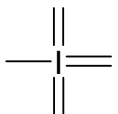
		при кратной связи
NE1	$=\text{NH}^+$	Формально положительно заряженный атом азота при двойной связи, имеющий водородного соседа
NE2	$=\text{N}^+=$	Формально положительно заряженный атом азота при двух двойных связях
NE3	$=\text{N}^+<$	Формально положительно заряженный атом азота при двойной связи, не имеющий водородных соседей
NE4	$\equiv\text{N}^+—$	Формально положительно заряженный атом азота при тройной связи
NH_		Атом азота в пятичленном ароматическом гетероцикле
NHC		Формально положительно заряженный атом азота в пятичленном ароматическом гетероцикле
NH1		Формально незаряженный атом азота в пятичленном ароматическом гетероцикле, связанный с атомом водорода
NH2		Формально незаряженный атом азота в пятичленном ароматическом гетероцикле, не связанный с атомами водорода
NHD		Формально незаряженный атом азота в пятичленном ароматическом гетероцикле, имеющий двух соседей
NB_		Атом азота в шестичленном ароматическом гетероцикле
NBC		Формально положительно заряженный атом азота в шестичленном ароматическом гетероцикле
NBD		Формально незаряженный атом азота в шестичленном ароматическом гетероцикле
N5_		Формально пятивалентный атом азота с тремя соседями, например, в составе нитро-группы
O_	O	Атом кислорода
OA_	O	Формально незаряженный атом кислорода, не входящий в ароматический гетероцикл
OA0	H ₂ O	Атом кислорода в составе молекулы воды
OA1	-OH	Атом кислорода в составе гидроксильной группы
OA2	-O-	Формально незаряженный атом кислорода, не связанный с атомами водорода и не входящий в

		ароматический гетероцикл
ON_	$-O^-$	Атом кислорода, несущий формальный отрицательный заряд
OC_	$>O^+$	Атом кислорода, несущий формальный положительный заряд и не входящий в ароматический гетероцикл
OD_	$=O$	Формально незаряженный атом кислорода при неароматической двойной связи (атом кислорода в составе карбонильной группы)
OH_		Атом кислорода в составе пятичленного ароматического гетероцикла
OB_		Атом кислорода в составе шестичленного ароматического гетероцикла
S_	S	Атом серы
SA_		Формально незаряженный двухвалентный атом серы, не образующий кратных связей
SA0	H_2S	Атом серы в составе молекулы сероводорода
SA1	$-SH$	Атом серы, связанный с атомом водорода (атом серы в составе сульфгидрильной группы)
SA2	$-S-$	Формально незаряженный двухвалентный атом серы, связанный с двумя неводородными атомами и не входящий в состав ароматического гетероцикла (атом серы в составе тиоэфирной группы)
SD_	$=S$	Формально незаряженный атом серы при двойной связи, не входящий в ароматический гетероцикл
SD1	$=S$	Формально незаряженный двухвалентный атом серы, ковалентно связанный с одним неводородным атомом
SD2	$=S=$	Формально незаряженный четырехвалентный атом серы, ковалентно связанный с двумя неводородными атомами
SD3	$>S=$	Формально незаряженный четырехвалентный атом серы, ковалентно связанный с тремя неводородными атомами
SD4		Формально незаряженный шестивалентный атом серы, ковалентно связанный с четырьмя атомами
SN_	$-S^-$	Атом серы, несущий формальный отрицательный заряд и ковалентно связанный с одним атомом
SC_	$>S^+$	Атом серы, несущий формальный положительный заряд и ковалентно связанный с тремя атомами

S6_		Атом серы, ковалентно связанный с шестью атомами
SH_		Атом серы в составе пятичленного ароматического гетероцикла
SB_		Атом серы в составе шестичленного ароматического гетероцикла
E_	Se	Атом селена
EA_	Se	Формально незаряженный двухвалентный атом селена, не образующий кратных связей
EA0	H ₂ Se	Атом селена в составе молекулы селеноводорода
EA1	-she	Атом селена, связанный с атомом водорода
EA2	-Se-	Формально незаряженный двухвалентный атом селена, связанный с двумя неводородными атомами и не входящий в состав ароматического гетероцикла
ED_	=Se	Формально незаряженный атом селена при двойной связи, не входящий в ароматический гетероцикл
ED1	=Se	Формально незаряженный двухвалентный атом селена, ковалентно связанный с одним неводородным атомом
ED2	=Se=	Формально незаряженный четырехвалентный атом селена, ковалентно связанный с двумя неводородными атомами
ED3	>Se=	Формально незаряженный четырехвалентный атом селена, ковалентно связанный с тремя неводородными атомами
ED4		Формально незаряженный шестивалентный атом селена, ковалентно связанный с четырьмя атомами
EN_	-Se ⁻	Атом селена, несущий формальный отрицательный заряд и ковалентно связанный с одним атомом
EC_	>Se ⁺	Атом селена, несущий формальный положительный заряд и ковалентно связанный с тремя атомами
EH_		Атом селена в составе пятичленного ароматического гетероцикла
EB_		Атом селена в составе шестичленного ароматического гетероцикла
P_	P	Атом фосфора
PA_		Атом фосфора, не образующий двойных связей ^a
PA0	PH ₃	Атом фосфора в молекуле PH ₃

PA1	-PH ₂	Атом фосфора, ковалентно связанный с двумя атомами водорода
PA2	>PH	Атом фосфора, ковалентно связанный с одним атомом водорода
PA3	>P-	Атом фосфора, ковалентно связанный с тремя неводородными атомами и не образующий двойных связей ^a
PAC	>P ⁺ <	Атом фосфора, ковалентно связанный с четырьмя неводородными атомами и несущий формальный положительный заряд
PA5		Атом фосфора, ковалентно связанный с пятью неводородными атомами
PD ₋		Атом фосфора, образующий одну или несколько двойных связей ^a
PD2	-P=	Атом фосфора, ковалентно связанный с двумя неводородными атомами, не входящий в 6-членный ароматический гетероцикл и образующий одну двойную связь ^a
PD3		Атом фосфора, ковалентно связанный с тремя неводородными атомами и образующий две двойные связи ^a
PD4		Атом фосфора, ковалентно связанный с четырьмя неводородными атомами и образующий двойную связь ^a
PDB		Атом фосфора в составе шестичленного ароматического гетероцикла
A ₋	As	Атом мышьяка
AA ₋		Атом мышьяка, не образующий двойных связей ^a
AA0	AsH ₃	Атом мышьяка в молекуле AsH ₃
AA1	-AsH ₂	Атом мышьяка, ковалентно связанный с двумя атомами водорода
AA2	>AsH	Атом мышьяка, ковалентно связанный с одним атомом водорода
AA3	>As-	Атом мышьяка, ковалентно связанный с тремя неводородными атомами и не образующий двойных связей ^a
AAC	>As ⁺ <	Атом мышьяка, ковалентно связанный с четырьмя неводородными атомами и несущий формальный положительный заряд
AA5		Атом мышьяка, ковалентно связанный с пятью неводородными атомами

AD_		Атом мышьяка, образующий одну или несколько двойных связей ^a
AD2	-As=	Атом мышьяка, ковалентно связанный с двумя неводородными атомами, не входящий в 6-членный ароматический гетероцикл и образующий одну двойную связь ^a
AD3		Атом мышьяка, ковалентно связанный с тремя неводородными атомами и образующий две двойные связи ^a
AD4		Атом мышьяка, ковалентно связанный с четырьмя неводородными атомами и образующий двойную связь ^a
ADB		Атом мышьяка в составе шестичленного ароматического гетероцикла
Si_	Si	Атом кремния
Si0	SiH ₄	Атом кремния в молекуле SiH ₄
Si1	-SiH ₃	Атом кремния, ковалентно связанный с тремя атомами водорода
Si2	>SiH ₂	Атом кремния, ковалентно связанный с двумя атомами водорода
Si3	>SiH-	Атом кремния, ковалентно связанный с одним атомом водорода
Si4	>Si<, =Si<, >Si<	Атом кремния, не связанный ковалентно с атомами водорода
F_	F	Атом фтора
F_0	F ⁻	Фторид-анион
F_1	-F	Ковалентно связанный атом фтора
Cl_	Cl	Атом хлора
Cl0	Cl ⁻	Хлорид-анион
Cl1	-Cl	Атом хлора, образующий одну ковалентную связь
Cl2	-Cl=	Формально незаряженный атом хлора, образующий две ковалентные связи
ClC	-Cl ⁺ -	Формально положительно заряженный атом хлора, образующий две ковалентные связи
Cl3		Атом хлора, образующий три ковалентные связи (атом хлора в степени окисления +5)
Cl4		Атом хлора, образующий четыре ковалентные связи (атом хлора в степени окисления +7)
Br_	Br	Атом брома
Br0	Br ⁻	Бромид-анион
Br1	-Br	Атом брома, образующий одну ковалентную связь

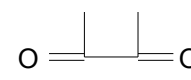
Br2	-Br=	Формально незаряженный атом брома, образующий две ковалентные связи
BrC	-Br ⁺ -	Формально положительно заряженный атом брома, образующий две ковалентные связи
Br3		Атом брома, образующий три ковалентные связи (атом брома в степени окисления +5)
Br4		Атом брома, образующий четыре ковалентные связи (атом брома в степени окисления +7)
I	I	Атом иода
I_0	I ⁻	Иодид-анион
I_1	-I	Атом йода, образующий одну ковалентную связь
I_2	-I=	Формально незаряженный атом йода, образующий две ковалентные связи
I_C	-I ⁺ -	Формально положительно заряженный атом йода, образующий две ковалентные связи
I_3		Атом йода, образующий три ковалентные связи (атом йода в степени окисления +5)
I_4		Атом йода, образующий четыре ковалентные связи (атом йода в степени окисления +7)

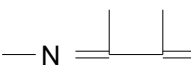
(а) Семиполярные связи также считаются двойными

5.1.3. Построение фрагментного дескриптора

Значением фрагментного дескриптора является число вхождений соответствующего подграфа в молекулярный граф, т.е. сколько раз соответствующая подструктура содержится в химической структуре. Код фрагмента содержит 3 поля (для фрагментов *p1* два), разделенные точками: код типа фрагмента, коды атомов и коды связей (1 – простая связь, 2 – двойная связь, 3 – тройная связь, 4 – ароматическая связь). Код типа фрагмента содержит два символа, первый из которых является одной из следующих букв: *p* – для цепочек, *c* – для циклических фрагментов, *s* – для разветвленных фрагментов, *b* – для бициклических фрагментов, *t* – для трициклических фрагментов. Второй символ соответствует числу вершин (в шестнадцатеричной системе исчисления) в соответствующем подграфе. Полный перечень типов фрагментов приведен в подразделе

ле 5.1.1. В тех случаях, когда возможно несколько вариантов кодирования, из них выбирается лексикографически минимальный.

Пример 1. Фрагмент  имеет код p4.OD_CD3CD3OD_.212. Здесь код типа фрагмента p4, тип атомов кислорода OD_, тип атомов углерода CD3, типы связей - 2, 1 и 2.

Пример 2. Фрагмент  может быть закодирован двумя способами: p4.ND2CD3CD3OD_.212 и p4.OD_CD3CD3ND2.212. Из этих двух вариантов выбирается лексикографически наименьший код: p4.ND2CD3CD3OD_.212

Для задания маски, соответствующей целой группе фрагментов, можно также пользоваться подстановочным символом '*', который может соответствовать любому символу в коде фрагмента.

Пример 3. Рассмотрим две маски: p1.*** и p2.*****.*. В этом случае программа Fragment сгенерирует все возможные фрагменты с одним и двумя неводородными атомами.

5.1.4. Генерация кодов фрагментов с обобщенными типами атомов

Программа Fragment позволяет автоматически добавлять к каждому коду фрагмента ряд его вариантов с различными уровнями обобщения типов атомов. Эти обобщенные варианты рассматриваются как самостоятельные дескрипторы. В настоящей версии программы предусмотрено 4 способа обобщения: 1) генерация кодов фрагментов с учетом только максимально подробного уровня классификации (*none*); 2) генерация кодов фрагментов, при которой изменение уровня классификации для всех атомов происходит одинаковым образом (*level1*); 3) генерация кодов фрагментов с учетом разных уровней классификации для атомов (*level2*); 4) генерация кодов фрагментов с учетом всех возможных уровней классификации (*full*). Например, если программа находит фрагмент NH₂CH=O с кодом

p3.NA1CD2OD_.12,

то в дальнейшем будут сгенерированы в соответствующих режимах обобщения следующие коды фрагментов:

1) *none*: p3.NA1CD2OD_.12

2) *level1*: p3.NA1CD2OD_.12

p3.NA_CD_OD_.12

p3.N__C__O__.12

p3._____.12

3) *level2*:

p3.NA1CD2OD_.12

P3.NA1CD2O__.12

P3.NA_CD2OD_.12

p3.NA_CD2O__.12

p3.N__CD2OD_.12

p3.N__CD2O__.12

p3.NA1CD_OD_.12

p3.NA1CD_O__.12

p3.NA_CD_OD_.12

p3.NA_CD_O__.12

p3.N__CD_OD_.12

p3.N__CD_O__.12

p3.NA1C__OD_.12

p3.NA1C__O__.12

p3.NA_C__OD_.12

p3.NA_C__O__.12

p3.N__C__OD_.12

p3.N__C__O__.12

p3._____.12

4) *full*:

p3.NA1CD2OD_.12

p3.NA1CD2O__.12

p3.NA1CD2____.12

p3.NA_CD2OD_.12

p3.NA_CD2O__.12

p3.NA_CD2____.12

p3.N__CD2OD_.12

p3.N__CD2O__.12

p3.N__CD2____.12

p3.____CD2____.12

p3.NA1CD_OD_.12

p3.NA1CD_O__.12

p3.NA1CD____.12

p3._____.12

p3.NA_CD_OD_.12

p3.NA_CD_O__.12

p3.NA_CD____.12

p3.N__CD_OD_.12

p3.N__CD_O__.12

p3.N__CD____.12

p3.____CD____.12

p3.NA1C__OD_.12

p3.NA1C__O__.12

p3.NA1C____.12

p3.NA_C__OD_.12

p3.NA_C__O__.12

p3.NA_C____.12

p3.OD_CD2____.21

p3.N__C__OD_.12

p3.N__C__O__.12

p3.N__C____.12

p3.____C____.12

p3.NA1__OD_.12

p3.NA1__O__.12

p3.NA1____.12

p3.NA____OD_.12

p3.NA____O__.12

p3.NA____.12

p3.N____OD_.12

p3.N____O__.12

p3.N____.12

p3.O__CD2____.21

p3.OD_CD____.21	p3.O__CD____.21	p3.OD_C____.21
p3.O__C____.21	p3.OD____.21	p3.O____.21

5.1.5. Алгоритм генерации фрагментных дескрипторов

Нами разработан и реализован в программе Fragment эффективный алгоритм нахождения/генерации фрагментов. Данный алгоритм включает два прохода по базе данных химических соединений. Во время первого прохода осуществляется поиск необходимых фрагментов и определяется число появлений каждого из них в каждой из химических структур исследуемой базы данных, а при втором проходе формируется матрица, содержащая числа вхождений каждого из найденных фрагментов в каждой химической структуре из базы данных.

При первом проходе из базы данных считывается каждая из имеющихся структур и приводится к «стандартному» виду (явно заданные атомы водорода преобразуются в неявные, меняются резонансные формы некоторых функциональных групп, например семиполярная связь в нитро-группе заменяется на двойную и т.д.). Далее производится поиск ароматических циклов и полициклических систем. После этого все содержащиеся в текущей химической структуре атомы классифицируются с помощью рассмотренной выше кодировки из трех символов. Далее каждая структура анализируется в три этапа. На первом этапе ищутся все фрагменты типов *p1, p3, c3, p5, c5, s5, p7, c7, b1, b2, p9, c9, b5, pb, cb, b8, pd, cd, bb, bc, t3, t4, t5, t6, t7, t8, t9, pf, cf* и *te* с применением специальной процедуры поиска, состоящей из 16 вложенных циклов и множества специальных условий проверки для прореживания поискового дерева на как можно более ранней стадии. На втором этапе ищутся все фрагменты типов *p2, p4, c4, s4, p6, c6, b0, s6, p8, c8, b3, b4, pa, ca, b6, b7, pc, cc, b9, ba, t0, t1, t2, pe, ce, bd, ta, tb, tc, td* с использованием аналогичной процедуры поиска. Наконец, на третьем этапе, все указанные пользователем нестандартные фрагменты ищутся с использованием рекурсивной процедуры нахождения подграфов в графе.

После нахождения первоначального набора фрагментных дескрипторов, содержащих коды атомов в наиболее подробной классификации, программа генерирует фрагментные дескрипторы с различными уровнями обобщения классификации атомов (в соответствии с выбранной схемой, см. выше) и формирует канонические кодирующие строки для каждого из них. При этом просматриваются все возможные перестановки из группы автоморфизмов соответствующего фрагмента, и осуществляется выбор лексикографически наименьшей строки. Каждая каноническая строка сравнивается сначала с указанным пользователем либо сформированным вызывающей программой (NASAWIN, NETPROGNOSIS, и т.д.) списком масок (кодов) фрагментов, а затем она ищется в иерархически сформированном списке уже найденных фрагментов. Если такая строка соответствует какой-либо из масок и содержится в этом списке, то число вложений соответствующего фрагмента увеличивается на единицу, в противном случае, если строка соответствует какой-либо маске, но отсутствует в списке, то соответствующий фрагмент добавляется к списку найденных фрагментов с числом вложений, равным единице. Для нестандартных фрагментов число вложений определяется путем деления числа изоморфных вложений соответствующего подграфа в молекулярный граф на предварительно найденный порядок группы автоморфизмов этого подграфа. Кроме того, программа хранит в памяти список фрагментов, содержащих указатели на группы статистически идентичных дескрипторов (значения которых пропорциональны друг другу для всех уже пройденных химических структур), тогда как сам список и все группы реорганизуются после завершения анализа каждой из химических структур.

После завершения первого прохода подсчитывается число появлений во всей базе данных для каждого из фрагментов, накопленных в иерархическом списке, и те фрагменты, которые содержатся в слишком малом числе соединений, и, соответственно, не удовлетворяют пороговому условию, заданному пользователем, удаляются из списка. Кроме того, из каждой группы статистически идентичных дескрипторов в списке оставляется только один. На втором проходе формируется файл с именами оставшихся дескрипторов и файл, со-

держащий матрицу значений дескрипторов (т.е. числа вложений каждого из фрагментов в каждую из структур).

5.2. Примеры прогнозирования физико-химических свойств органических соединений с использованием фрагментных дескрипторов и линейно-регрессионных моделей

Описанные выше фрагментные дескрипторы впервые были нами предложены в 1990-1991 г. [356, 357] и запрограммированы в виде дескрипторного блока FRAGMENT (см. раздел 8.3), который вошел в состав программных комплексов EMMA (см. раздел 8.1) и NASAWIN (см. раздел 8.2). В наших работах фрагментные дескрипторы себя проявили как очень эффективные инструменты для построения моделей, позволяющих прогнозировать разнообразные свойства органических соединений. В частности, как показано ниже на примере прогнозирования поляризуемости химических соединений (см. подраздел 5.2.1) и энтальпии образования алифатических полинитросоединений (см. подраздел 5.2.2), они, в сочетании с аппаратом множественной линейной регрессии, являются очень удобным средством автоматического создания аддитивных схем расчета физико-химических свойств. В подразделах от 5.2.3 до 5.2.7 приведен цикл работ (сделанных в соавторстве с Н.И.Жоховой), в которых предложенные фрагментные дескрипторы, в сочетании с аппаратом множественной линейной регрессии, успешно использованы для прогнозирования нескольких видов физико-химических свойств органических соединений, которые лишь с очень большим трудом либо вообще не поддаются расчету при помощи методов квантовой химии и молекулярного моделирования. Отметим, что во всех случаях нами были построены модели, превышающие по качеству все, опубликованные ранее в литературе. В следующей главе диссертации будет также показано, что замена множественной линейной регрессии на аппарат искусственных нейронных сетей ведет к дальнейшему улучшению прогнозирующей способности полученных моделей.

5.2.1. Прогнозирование поляризуемости органических соединений

Поляризуемость, α , - одно из наиболее важных электрических свойств молекул, характеризующее способность электронной системы деформироваться под действием внешнего электрического поля [358]. $\mu_{\text{индуц.}} = \alpha \cdot \epsilon$. Поскольку поляризуемость является тензорной величиной, то, строго говоря, моделируемой величиной является инвариант этого тензора – среднее значение его диагональных членов (одна треть от следа матрицы тензора).

Молекулярная рефракция является аддитивным свойством, и детальные аддитивные схемы расчета этой величины были предложены как на основе атомных инкрементов, так и инкрементов связей [358]. Известно много методов расчета поляризуемости разной степени сложности и точности [359-362]. В силу вышесказанного, очевидно, что принцип аддитивности можно было бы распространить и на поляризуемость. И действительно, тензор поляризуемости был приписан отдельным связям и функциональным группам в соответствии с гипотезой, что покомпонентное сложение групповых тензоров приведет к тензору молекулярной поляризуемости [360]. На этой основе был разработан полуэмпирический метод расчета компонент молекулярной поляризуемости, который для базы данных, включающей 120 структур (База 1) дал точность с величиной стандартного отклонения 3.5% [361, 362]. Однако наиболее неожиданными оказались данные работы [363], где было показано, что простой набор только атомных поляризуемостей для десяти элементов позволяет хорошо вычислять молекулярную поляризуемость без учета каких-либо других структурных параметров [363]. Интересно, что применение квантово-химических методов, таких как MINDO, AM1, PM3 и даже DFT, дает худшие результаты, чем аддитивный подход.

Целью нашей работы была проверка способности разработанных нами фрагментных дескрипторов выступать, в сочетании с линейным регрессионным анализом, в качестве средства автоматического создания аддитивных схем физико-химических свойств химических соединений на примере прогнозирования поляризуемости. Работа проводилась с использованием наших QSAR/QSPR

программ EMMA (см. раздел 8.1) и NASAWIN (см. раздел 8.2). Для данного исследования было создано несколько баз данных. Во-первых, по данным работ [361, 362] была сформирована База 1, состоящая из 293 структур, в которой представлены разнообразные классы органических соединений и некоторых неорганических веществ (как, например, H₂, O₂, N₂, N₂O, CO, SO₂, H₂S, H₂O, NH₃, Cl₂ и др.). Во-вторых, по данным работы [363] была создана База 2, содержащая 426 соединений, включающих C, H, O, N, S, P, F, Cl, Br, I (циклические и ациклические неароматические углеводороды; ароматические углеводороды, галогенированные и перфторированные производные, спирты, фенолы, простые и сложные эфиры, альдегиды, кетоны, карбоновые кислоты; амины, нитрилы, нитропроизводные, амиды, серу- и фосфорсодержащие соединения). Кроме того, была сформирована комбинированная база данных, База 3, которая объединяла две предыдущих базы и, после исключения дубликатов, состояла из 613 соединений.

На основе комбинированной Базы 3 нами был построен ряд моделей, наилучшая из которых следующая:

$$\alpha_{\text{calc}} = 0.04 + 1.08 f_1 + 0.38 f_2 + 0.92 f_3 + 0.61 f_4 + 3.04 f_5 + 2.18 f_6 + 0.44 f_7 + 2.34 f_8 + 3.35 f_9 + 5.49 f_{10} + 0.38 f_{11} + 0.15 f_{12} + 0.34 f_{13} + 0.36 f_{14} \quad (2)$$

$$n = 552, r^2 = 0.9967, s = 0.38 \text{ \AA}^3, F = 10931$$

где $f_1, f_2, f_3, f_4, f_5, f_6, f_7, f_8, f_9, f_{10}$ - число атомов C, H, N, O, S, P, F, Cl, Br, I, соответственно; f_{11} - количество тройных связей в молекуле, $\bullet \equiv \bullet$; f_{12} - количество двойных связей в молекуле, $\bullet = \bullet$; f_{13} - количество ароматических связей, $\bullet \div \bullet$; f_{14} - количество атомов в сочленениях циклов в ароматической системе, C_{Ar}(C_{Ar})₃. Следует отметить, что ранее опубликованные аддитивные схемы показали на Базе 3 существенно более высокое значение стандартного отклонения ($s = 0.61 \text{ \AA}^3$).

Таким образом, в рамках рассмотренной методологии нами построены уравнения, позволяющие прогнозировать поляризуемость органических соединений различных классов с высокой точностью исходя из их элементного состава и числа фрагментов, отражающих наличие кратных и ароматических свя-

зей, а также учитывающих конденсированные ароматические системы. Этим примером продемонстрировано, что предложенные фрагментные дескрипторы в сочетании со статистическим аппаратом множественной линейной регрессии являются мощным инструментом для разработки аддитивных схем прогнозирования физико-химических свойств органических соединений.

5.2.2. Прогнозирование энтальпий образования алифатических полинитросоединений

Алифатические полинитросоединения находят практическое применение главным образом благодаря своей высокой энергетической емкости [364]. Именно поэтому из физико-химических свойств этой группы соединений наиболее хорошо экспериментально изучены термодинамические свойства, в частности теплоты образования [364]. Цель настоящей работы – анализ пригодности автоматического метода создания аддитивных схем на основе использования фрагментных дескрипторов и сравнение точности прогноза с результатами популярных методов молекулярно-механического и полуэмпирических квантово-химических расчетов для прогнозирования энтальпий образования алифатических полинитросоединений. В данной работе мы использовали экспериментальные данные по теплотам образования 31 алифатического полинитросоединения [364].

Построенную нами в результате выполнения работы при помощи программного комплекса ЕММА (см. раздел 8.1) с использованием блока Fragment (см. раздел 8.3) эмпирическую схему расчета энтальпий образования алифатических полинитросоединений можно представить при помощи уравнения (в ккал/моль):

$$\Delta H_f^0 = -13.2 - 6.29f_1 - 3.81f_2 - 4.59f_3 + 3.13f_4 + 3.65f_5 + 6.47f_6,$$

$$R = 0.9922; s = 2.65; F = 253.5,$$

где f_1 – число атомов углерода; f_2 – число связей между вторичным и четвертичным атомами углерода; f_3 – число связей между первичным и четвертичным атомами углерода; f_4 – число пар первичных атомов углерода, присоединенных

к одному четвертичному атому углерода; f_5 – количество комбинаций четвертичных атомов углерода и нитро-групп, присоединенных к одному четвертичному атому углерода; f_6 – число пар нитро-групп, присоединенных к одному атому углерода. Дескриптор f_1 отражает атомные вклады в теплоту образования, дескрипторы f_2 и f_3 – вклады связей, а дескрипторы f_4 , f_5 и f_6 – поправки на невыгодное взаимодействие определенных групп, присоединенных к одному атому.

Для сравнения были проведены расчеты энтальпий образования этих же соединений при помощи молекулярно-механического метода ММХ (при помощи программы PCMODEL) и эмпирических квантово-химических методов AM1 и PM3 (при помощи программы Hyperchem). В Табл. 3 приведены значения среднеквадратических ошибок прогноза для каждого из методов. Соответствующие диаграммы разброса приведены на Рис. 32. Для построенной нами аддитивной схеме она оказалась в несколько раз ниже, чем для ММХ, AM1 и PM3.

Таким образом, из всех рассмотренных методов наилучшие результаты дает применение подструктурных аддитивных схем. Однако область применения подобных схем ограничена теми классами соединений, на которых они были построены. Поэтому проводить прогноз по построенной аддитивной схеме можно только для алифатических полинитросоединений, в остальных же случаях из рассмотренных методов можно рекомендовать только PM3.

Табл. 3. Среднеквадратические ошибки прогноза энтальпии образования алифатических полинитросоединений

Метод	ММХ	AM1	PM3	Аддитивная схема
Среднеквадратическая ошибка в ккал/моль	21.8	33.4	11.6	2.3

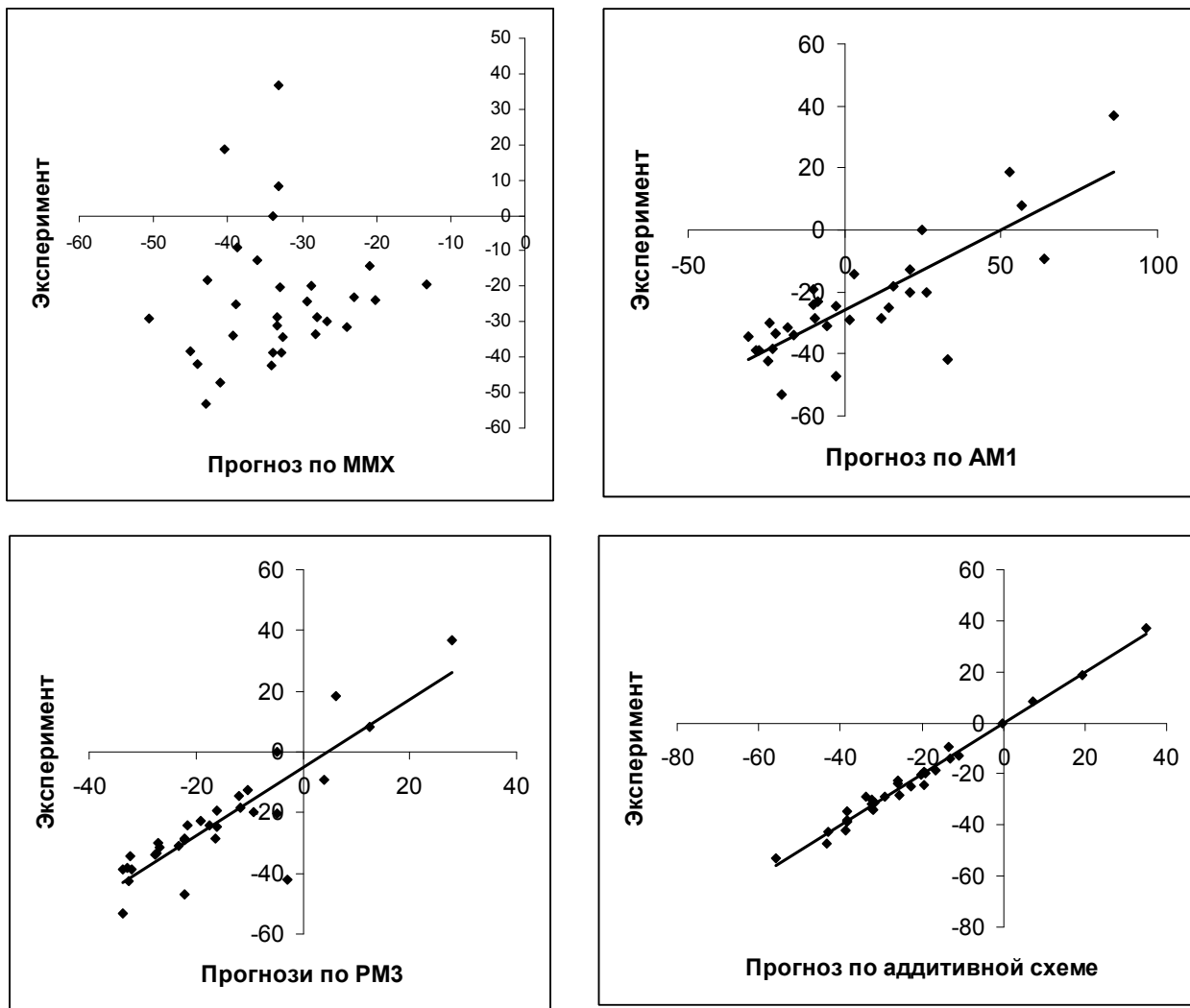


Рис. 32 Диаграммы разброса для прогнозирования энтальпии образования алифатических полинитросоединений разными методами (в ккал/моль)

5.2.3. Прогнозирование магнитной восприимчивости органических соединений

Магнитная восприимчивость – величина, характеризующая связь намагниченности вещества с магнитным полем в этом веществе. Намагниченность вещества, \mathbf{J} , прямо пропорциональна напряженности поля, \mathbf{H} , вызывающего намагничивание: $\mathbf{J}=\chi\mathbf{H}$, где χ - величина, называемая магнитной восприимчивостью вещества [365]. Поскольку магнитная восприимчивость является тензорной величиной, то, строго говоря, моделируемой величиной является инвариант этого тензора – среднее значение его диагональных членов (одна треть от следа матрицы тензора).

Одним из наиболее простых и распространенных методов расчета магнитной восприимчивости является аддитивный способ, основанный на сум-

мировании атомных и связевых вкладов (инкременты Паскаля) [366], а также схемы, учитывающие связевые взаимодействия [367]. Описаны также квантово-химические подходы, как относительно простые [368], так и метод DFT [369].

Поскольку QSAR/QSPR метод успешно применялся для моделирования большого числа физико-химических свойств, то было бы также интересно использовать его для расчета магнитной восприимчивости. В литературе имеются данные по расчету магнитной восприимчивости с помощью основанного на теории графов подхода [367, 370], спектральных моментов топологической матрицы связей [268], а также для элементоорганических галоидопроизводных элементов четвертой группы на основе топологических индексов [370].

В данной работе мы исследовали применение фрагментных дескрипторов, генерируемых блоком FRAGMENT (см. раздел 8.3), для прогнозирования магнитной восприимчивости диамагнетиков.

Составление баз данных. В качестве модельной базы экспериментальных данных по магнитной восприимчивости (База 1) были выбраны данные работы [268]. Они включают обучающую выборку из 233 алифатических и 85 ароматических соединений и контрольную выборку из 20 алифатических и 20 ароматических соединений (всего 358 структур). Хотя часть данных по ароматическим структурам (28 соединений) в работе [268] была вынесена в отдельную таблицу, мы решили использовать эти данные совместно с данными Базы 1 и включили их в выборку, в результате чего была сформирована База 2, которая после исключения дубликатов составила 378 структур. Далее мы сформировали Базу 3 за счет дополнения Базы 2 некоторыми литературными данными по магнитной восприимчивости. Во-первых, мы использовали данные работы [370] по магнитной восприимчивости органических галогенпроизводных с целью увеличения набора уже имеющихся в базе структур такого типа. Именно эти соединения не очень хорошо моделировались в работе [268]. Во-вторых, мы дополнили базу данными по гетероциклическим соединениям, взятыми из источников [371-373]. Наконец, в базу были добавлены два примера циклопропановых структур, чтобы убедиться в способности модели работать с напряженными структурами.

Построение QSPR-моделей. QSPR-моделирование проводили с использованием QSAR программ EMMA (см. раздел 8.1) и NASAWIN (см. раздел 8.2). При работе с программой EMMA сначала рассчитывали все фрагментные (максимальный размер фрагментов до 6 атомов) и два дополнительных дескриптора (см. ниже), далее формировали обучающую и контрольную выборки, и на основе пошаговой регрессии и предварительного отбора из групп взаимно скоррелированных ($R > 0.9$) дескрипторов тех, которые наилучшим образом коррелируют с моделируемым свойством, строили QSPR-модели.

Прежде всего, мы решили повторить результаты работы [268], но с использованием фрагментных дескрипторов. Полученные данные приведены в Табл. 4 (стр. 165). Прежде чем перейти к обсуждению и сравнению литературных и полученных нами данных, отметим, что авторы работы [268] использовали в качестве дескрипторов спектральные моменты топологической матрицы связей и, самое главное, рассматривали алифатические и ароматические структуры по отдельности. При этом QSPR-модель [268] для алифатических структур имела следующие статистические характеристики: R^2 (коэффициент детерминации) = 0.960, s (стандартное отклонение) = 6.06 (10^{-6} единиц), среднеквадратичная ошибка на прогнозе 8.49 (10^{-6} единиц).

Табл. 4. Статистические характеристики QSPR-моделей для магнитной восприимчивости (в 10^{-6} единиц)

Модель	База	Обучающая выборка			Контрольная выборка	
		$N_{дескр}$	R^2	s	$R^2_{прогн}$	$MAE_{прогн}$
1	1	4	0.937	7.63	0.949	7.50
2	1	4	0.943	7.30	0.984	4.17
3	1	4	0.982	4.14	0.985	4.56
4	2	3	0.871	6.79	0.948	6.28
5	2	6	0.989	1.99	0.934	6.58
6	2	6	0.987	6.48	0.937	8.41
7	2	8	0.991	5.44	0.931	7.87
8	3	7	0.985	4.99	0.934	7.02

Для построения модели 1 (Табл. 4) на основе фрагментных дескрипторов мы использовали обучающую и контрольную выборки алифатических струк-

тур, идентичные работе [268]. Из Табл. 4 видно, что статистические характеристики модели 1 немного уступают вышеприведенным литературным данным. Тем не менее, эта модель имеет неплохую прогнозирующую способность: так, средняя ошибка на прогнозе для модели 1, построенной с использованием 4 дескрипторов, составляет даже $7.5 (10^{-6})$ единиц).

Далее мы исследовали смешанные модели с единичным включением дескрипторов другого типа, обратив особое внимание на простоту вычисления таких добавочных дескрипторов. Оказалось, что добавление в модель такого простого дескриптора, как молекулярная масса, позволяет несколько улучшить качество QSPR-модели (Табл. 4, модель 2). Этот дескриптор включается в QSPR-модель, построенную с помощью пошаговой регрессии, вторым, что приводит к улучшению качества прогноза (средняя ошибка на прогнозе достигает 7-6.3).

Однако существенное улучшение качества модели было достигнуто при включении в уравнение, полученное на основе фрагментных дескрипторов, дескриптора V_x [374]. Этот дескриптор был введен для описания молекулярного объема при учете сольватационных эффектов. Использование этого дескриптора приводит к резкому улучшению даже однопараметровой модели (Табл. 4 на стр. 165, модель 3). Модель, включающая 5 дескрипторов, имеет превосходные статистические характеристики уменьшает среднюю ошибку на прогнозе до 4.8. Ниже приведено уравнение этой модели:

$$-\chi_m \times 10^6 = -2.91 + 0.82 V_x + 3.42 fr_1 + 6.40 fr_2 - 4.88 fr_3 - 2.99 fr_4 \quad (2)$$

$n = 355$, $R^2 = 0.9856$, $s = 3.7 (10^{-6})$ единиц, $F = 3104$, средняя ошибка (по модулю) на прогнозе 4.82, где fr_i равно числу следующих фрагментов в молекулах: $fr_1 - Br$, $fr_2 - Hal$, $fr_3 - \bullet = \bullet$, (\bullet – произвольный атом), $fr_4 - C(Hal)_2$.

Рассмотрим теперь ароматические соединения. Литературная QSPR-модель для ароматических структур [268] (85 соединений в обучающей и 20 соединений в контрольной выборках, 5 дескрипторов) имела следующие статистические характеристики: $R^2 = 0.9604$, $s = 3.82 (10^{-6})$ единиц, средняя ошибка при скользящем контроле $4.12 (10^{-6})$ единиц, среднеквадратичная ошибка на прогнозе 4.00. Модель 4 (Табл. 4, стр. 233) построена на тех же данных, что и в

работе [268], но с применением фрагментных дескрипторов. Как и в случае алифатических соединений, для ароматической выборки (обучающая - 85 соединений, контрольная - 20 соединений) статистические параметры модели 4 (Табл. 4, стр. 165), построенной на фрагментных дескрипторах, немного уступают литературным данным. Тем не менее, ее прогнозирующая способность выше.

Включение в модель дескриптора молекулярной массы существенно не улучшает ни статистических показателей модели, ни ее прогнозирующей способности. Напротив, использование дескриптора V_x приводит к резкому улучшению QSPR- модели (Табл. 4 на стр. 233, модель 5). Модель, содержащая 6 дескрипторов, имеет превосходные статистические характеристики ($s = 1.99 \times 10^{-6}$ единиц) и уменьшает среднюю ошибку на прогнозе до 6.6 (10^{-6} единиц).

Для построения QSPR-модели ароматических соединений была использована обучающая и контрольная выборка ароматических структур, составленные по данным работы [268]. Как видно из данных Табл. 4 на стр. 165, модель 6, построенная только на фрагментных дескрипторах, имеет достаточно хорошие статистические характеристики и обладает хорошей предсказательной силой.

Применение дескриптора V_x также приводит к резкому улучшению QSPR-модели (Табл. 4 на стр. 233, модель 7). Модель, включающая 8 дескрипторов, имеет превосходные статистические характеристики ($s = 5.44 \times 10^{-6}$ единиц) и уменьшает среднюю ошибку на прогнозе до 7.8×10^{-6} единиц. Уравнение для этой модели приведено ниже:

$$-\chi_M \times 10^6 = -4.87 + 0.823 V_x - 6.64 fr_1 + 11.8 fr_2 - 8.05 fr_3 - 6.09 fr_4 - 2.20 fr_5 + 1.08 fr_6 + 9.85 fr_7 \quad (3)$$

$n = 378$, $R^2 = 0.9908$, $s = 5.44$ ($\times 10^{-6}$ единиц), средняя ошибка (по модулю) на прогнозе 7.87, где fr_i равно числу следующих фрагментов в молекулах: $fr_1 - Cl$, $fr_2 - Hal$, $fr_3 - N-O$, $fr_4 - C=O$, $fr_5 - \bullet-\bullet\div\bullet-\bullet=\bullet$, (\div - ароматическая связь), $fr_6 - \bullet=\bullet-\bullet\div\bullet\div\bullet\div\bullet$, $fr_7 - RC_{Ar}\div C_{Ar}(C_{Ar}H)_2$.

В задачи следующего этапа нашей работы входило исследование применимости фрагментного подхода на примере расширенной выборки органиче-

ских соединений, содержащей в том числе галоидпроизводные и гетероциклические структуры ароматической природы – Базы 3. Полученные модели имеют достаточно высокие статистические показатели. Наилучшей прогнозирующей способностью обладает модель 8, построенная на семи дескрипторах:

$$-\chi_M \times 10^6 = -3.91 + 3.93 fr_1 + 6.41 fr_2 - 5.90 fr_3 - 2.93 fr_4 + 0.728 fr_5 + 9.77 fr_6 + 0.823 V_x \quad (4)$$

$n = 420$, $R^2 = 0.9846$, $s = 5.0$ ($\times 10^{-6}$ единиц), средняя ошибка (по модулю) на прогнозе 7.02 ($\times 10^{-6}$ единиц).

где fr_i равно числу следующих фрагментов в молекулах: fr_1 – Br, fr_2 – Hal, fr_3 – =O, fr_4 – C(Hal)₂, fr_5 – ●=●-●÷●÷●, fr_6 – RC_{Ar}÷C_{Ar}(C_{Ar}H)₂.

На Рис. 33 приведены диаграммы разброса экспериментальных и расчетных значений магнитной восприимчивости для обучающей и контрольной выборок согласно вышеприведенной модели.

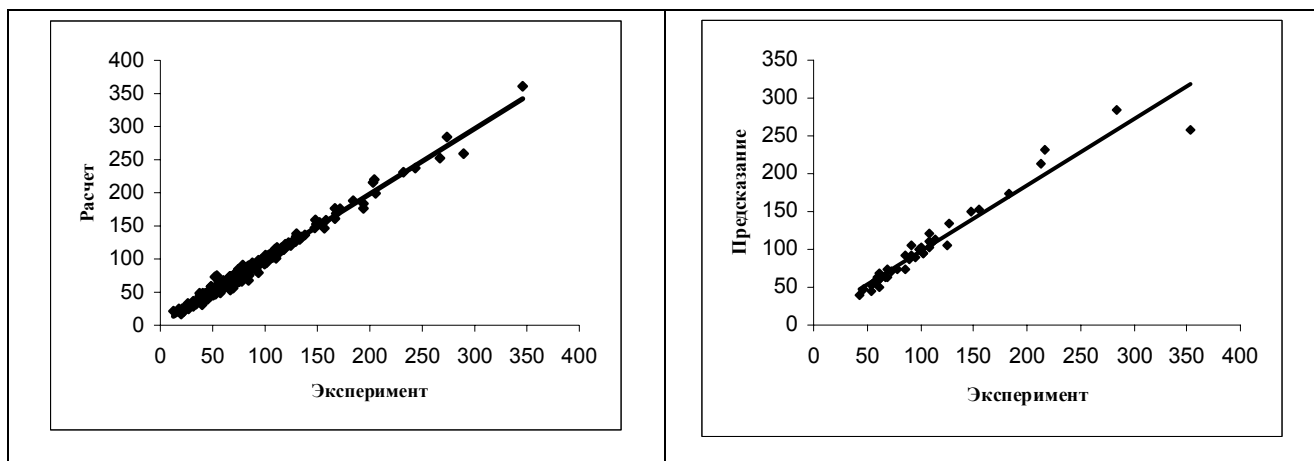


Рис. 33. Диаграмма разброса экспериментальных и расчетных значений магнитной восприимчивости для обучающей (слева) и контрольной (справа) выборок соединений (База 3) согласно модели 9.

Таким образом, нами продемонстрирована применимость фрагментного подхода в рамках методологии QSPR для расчета магнитной восприимчивости органических соединений различных классов. Предложенные модели по статистическим характеристикам превосходят описанные в литературе. Этим примером продемонстрировано, что предложенные фрагментные дескрипторы в сочетании со статистическим аппаратом множественной линейной регрессии яв-

ляются удобным инструментом для прогнозирования таких физических свойств органических соединений, которые лишь с очень большим трудом поддаются оценке при помощи строгих квантово-механических методов расчета.

5.2.4. Прогнозирование энтальпии парообразования органических соединений

Данная работа была стимулирована появлением публикации Е. В. Сагдеева и В. П. Барабанова [375], в которой авторы делают попытку найти зависимость энтальпии парообразования, $\Delta H_{\text{пар}}$, от температуры кипения в соответствии с литературными данными. Авторы установили полиномиальный характер такой зависимости, но для каждого класса органических соединений эти зависимости имеют собственные параметры и, таким образом, универсальное уравнение не было получено [375]. Более того, температура кипения является не расчетным, а экспериментально определяемым параметром, что затрудняет использование полученных закономерностей для прогноза величин $\Delta H_{\text{пар}}$ для других, и, особенно, неизвестных соединений.

В связи с этим нам представлялось интересным попытаться применить QSPR-методологию для получения универсального и прогностического QSPR-уравнения на экспериментальном материале по величинам $\Delta H_{\text{пар}}$, взятым из работы [375]. Отметим, что в литературе имеются примеры применения методов QSPR для расчета $\Delta H_{\text{пар}}$ с использованием физико-химических, топологических и структурных дескрипторов [376-379].

В настоящей работе мы исследовали применение фрагментных дескрипторов для QSPR-рассмотрения энтальпии парообразования, $\Delta H_{\text{пар}}$. В качестве модельной базы были взяты экспериментальные данные по $\Delta H_{\text{пар}}$, отнесенные к стандартным условиям (25°C), для 52 соединений из работы [375]. Подчеркнем, что этот набор достаточно репрезентативен и включает органические соединения тринадцати различных классов, такие как алканы, циклоалканы, олефины, ацетилены, спирты, карбонильные соединения, карбоновые кислоты, амины. База была разделена на обучающую (39 соединений) и контрольную (13 соединений, по одному соединению из каждого класса. Для оценки предсказательной

способности модели мы использовали независимую контрольную выборку, в которую были включены данные по $\Delta H_{\text{пар}}$ для тринадцати соединений [380], каждое из которых представляло один из классов модельной базы и которые не участвовали в построении модели.

QSPR-моделирование проводили с использованием программы NASAWIN (см. раздел 8.2) и дескрипторного блока FRAGMENT (см. раздел 8.3). Построение QSPR-модели методом пошаговой регрессии осуществляли на основе предварительного расчета фрагментных дескрипторов и последующего отбора из группы взаимно скоррелированных ($R > 0.9$) дескрипторов тех из них, которые наилучшим образом коррелируют с моделируемым свойством. Рассчитывали фрагменты с максимальным размером от 1- до 6-атомных.

На первом этапе работы мы получили единое линейно-регрессионное QSPR-уравнение для соединений базы с использованием обучающей и контрольной выборок (в ккал/моль):

$$\Delta H_{\text{пар(расч.)}} = 3.7272 + 5.2361fr1 + 7.9110fr2 + 5.6798fr3 + 23.9276fr4 + 4.7953fr5 \quad (1)$$

Уравнение построено на пяти одноатомных дескрипторах и имеет следующие параметры: число соединений в обучающей выборке - 38, число соединений в контрольной выборке - 13, квадрат коэффициента корреляции для обучающей выборки, $R^2 = 0.993$, квадрат коэффициента корреляции для контрольной выборки, $R^2_{\text{контр.}} = 0.982$, стандартное отклонение, $s = 1.785$ ккал/моль, критерий Фишера, $F = 908.19$, среднеквадратичная ошибка на обучающей выборке, $RMSE_{\text{обуч.}} = 1.64$ ккал/моль. В уравнении (1) $fr1$ равно числу следующих фрагментов в молекулах: $fr1$ - Cl, $fr2$ - NH_2 , $fr3$ - $=\text{O}$, $fr4$ - OH, $fr5$ - общее число неводородных атомов в молекуле.

Прогнозирующие свойства фрагментной модели оценивали с помощью независимой выборки, составленной по данным, приведенным в [380] и включающей 13 соединений: $R^2_{\text{прогн.}} = 0.988$, $RMS_{\text{прогн.}} = 1.57$ ккал/моль. Диаграммы разброса расчетных и экспериментальных значений энтальпии парообразования для обучающей выборки (слева) и независимой выборки для прогноза (справа), для этой модели представлены на Рис. 34 (стр. 171). В отличие от уравнений, предложенных в работе [375] и представляющих собой частные случаи для рас-

чета этого свойства для каждой группы из тринадцати классов, включающей по четыре соединения базы, полученная линейно-регрессионная QSPR-модель является единым уравнением для расчета энтальпии парообразования исследованных соединений. Модель позволяет избежать использования таких экспериментальных параметров, как температура кипения, и ограничиться только знанием структурной формулы соединения.

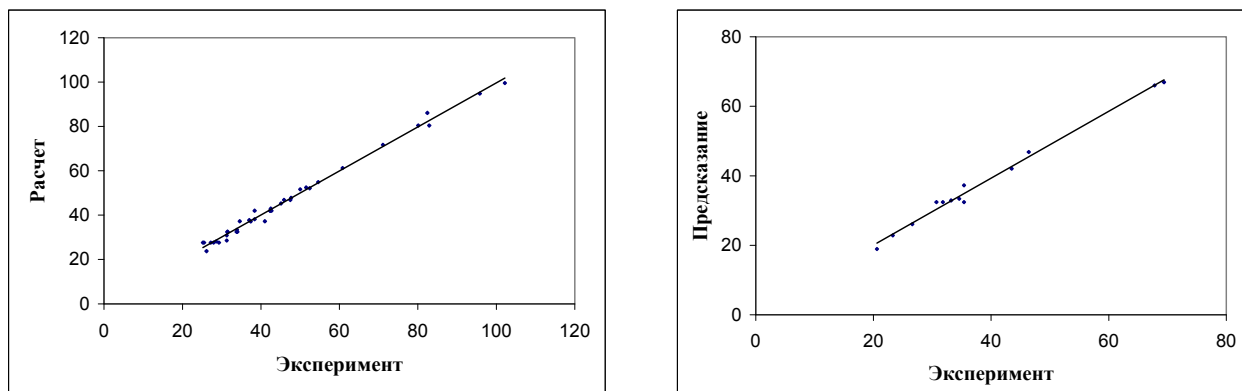


Рис. 34. Диаграмма разброса экспериментальных и расчетных значений энтальпии парообразования для обучающей выборки (слева) и выборки для независимого прогноза (справа). Единица измерения – ккал/моль.

Таким образом, применение метода QSPR/QSAR позволяет получить общую модель для расчета и прогноза энтальпии парообразования исследованных органических соединений различных классов только на основе знания структурной формулы соединения. Фрагментная модель является альтернативой набору уравнений зависимости энтальпии парообразования от температуры кипения, предложенному для расчета энтальпии парообразования органических соединений в работе [375].

5.2.5. Прогнозирование энтальпии сублимации органических соединений

Энтальпия сублимации, $\Delta_{\text{sub}}H$, - энтальпия перехода вещества из твердого состояния непосредственно (без плавления) в газообразное [381]. Это свойство представляет определенный практический интерес в химии кристаллического состояния и, в частности, для проблем диспергирования красителей, выцветания материалов, а также таких экологических проблем, как перенос органиче-

ских загрязнителей в атмосфере и т.д. [381-384]. Экспериментальное определение энтальпий сублимации, $\Delta_{\text{sub}}H$, как и других термодинамических величин, требует дорогих и длительных процедур. Поэтому, в литературе уделялось значительное внимание как расчетным теоретическим, так и эмпирическим QSPR методам. Так, например, отметим, что значения энтальпий сублимации были получены на основе расчета кристаллических упаковок [381-383]. Для QSPR использовались методы регрессионного анализа [381], нейронные сети [381], а также 3D-QSAR (CoMFA) [384]. В случае линейного регрессионного анализа с обучающей выборкой из 62 соединений (контрольная выборка состояла из 10 соединений) было получено трехпараметровое уравнение, в котором в качестве дескрипторов использовались число атомов углерода, а также число доноров и акцепторов водородной связи [381] (подробнее см. ниже).

В настоящей работе фрагментные дескрипторы применены нами для QSPR-исследования энтальпии сублимации. В качестве модельной базы экспериментальных данных по энтальпиям сублимации (База 1) были выбраны данные работы [381]: обучающая выборка из 62 соединений и контрольная выборка - 10 структур (соединения 63-72). Полная выборка включала молекулы с известной кристаллической структурой, содержащие атомы С, Н, О, N, в том числе алифатические и ароматические углеводороды, их оксо- и аза-производные, карбоновые кислоты, амиды и аминокислоты, цианиды, хиноны, гетероциклы. Преимуществом данной выборки соединений является наличие для нее расчета энтальпий сублимации тремя способами: (1) теоретическим расчетом кристаллических упаковок (со следующими статистическими параметрами: $n = 62$, $r^2 = 0.971$, $s = 0.939$ ккал/моль, максимальная ошибка = 3.5 ккал/моль), (2) регрессионным анализом (со следующими статистическими параметрами: три дескриптора, $n = 62$, $r^2 = 0.92$, $s = 1.6$ ккал/моль, максимальная ошибка = 8.9 ккал/моль, средняя ошибка на прогнозе = 2.8 ккал/моль), и (3) с помощью нейронной сети (со следующими параметрами для лучшей модели: семь скрытых нейронов, $n = 62$, $r^2 = 0.865$, $s = 2.2$ ккал/моль, максимальная ошибка = 10.1 ккал/моль, средняя ошибка на прогнозе = 3.6 ккал/моль). Это дает хорошую основу для сравнения, хотя сама выборка и не очень велика.

База 2 (88 соединений) была создана путем добавления в Базу 1 экспериментальных данных работы [382] и исключения дубликатов, а База 3 – путем добавления в Базу 2 экспериментальных данных работы [384] по хлорированным дифенилам (15 структур) и после исключения дубликатов База 3 в результате включала 104 соединения. Включение хлорированных дифенилов обусловлено как важностью данного типа соединений, находящихся широкое применение в качестве изоляционных материалов и замедлителей горения, так и желанием расширить структурное разнообразие выборки на хлорсодержащие соединения.

QSPR моделирование проводилась с использованием наших QSAR программ EMMA (см. раздел 8.1) и NASAWIN (см. раздел 8.2). Фрагментные дескрипторы вычислялись блоком FRAGMENT (см. раздел 8.3), на работу которого налагались следующие ограничения: длина цепочек составляла 1-6, отбор фрагментных дескрипторов осуществляли как в автоматическом режиме, так и вручную, при отборе из группы скоррелированных друг с другом дескрипторов выбирались наиболее коррелирующие с активностью.

Рассмотрим теперь сравнительные QSPR результаты. В Табл. 5 представлены характеристики моделей, полученных на основе фрагментных дескрипторов. Прежде всего, мы построили QSPR-модель (Модель 1, Табл. 5), используя ту же выборку, что и в работе [381], то есть взяли 62 соединения в качестве обучающей выборки и 10 соединений для прогноза (База 1). Из Табл. 5 видно, что на 3 фрагментных дескрипторах, получается удовлетворительная статистика, сравнимая с данными работы [381] и дающая разумный прогноз (Модель 1). Интересно, что первый дескриптор (число неводородных атомов) моделирует первый дескриптор работы [381], а два последующих фрагментных дескриптора непрямым образом моделируют число центров, образующих водородные связи (как и в работе [381]).

Табл. 5. Статистические характеристики QSPR-моделей для энтальпии сублимации (в ккал/моль)

Модель	База	Обучающая выборка			Контрольная выборка	
		$N_{\text{дескр}}$	R^2	s	$R^2_{\text{прогн}}$	$MAE_{\text{прогн}}$
1	1	3	0.924	2.38	0.769	2.7
2	2	3	0.852	2.92	0.752	2.41
3	3	2	0.845	2.97	0.816	2.16

Естественно, что, имея в распоряжении Базу 3, было интересным получить QSPR-модель, используя расширенную обучающую выборку. С этой целью в качестве обучающей выборки использовались соединения 1-62 (База 1), 73-88 (База 2) и 13 соединений ряда хлорированных дифенилов (см. выше). Контрольная выборка включала в соответствии с работой [381] те же 10 соединений (63-72), но была дополнена тремя соединениями ряда хлорированных дифенилов: мы взяли два соединения, использованных для прогноза в работе [384]. Таким образом, обучающая выборка состояла из 91, а контрольная выборка из 13 соединений. Построенная QSPR-модель (Модель 2, Табл. 5 на стр. 174) отличается хорошей предсказательной способностью, превосходящей показатели Модели 1, и позволяет прогнозировать исследуемое свойство для соединений ряда хлорированных дифенилов. Средняя ошибка на прогнозе 2.4 ккал/моль (модель на основе трех дескрипторов).

Наконец, рассмотрение структур, выпадающих из корреляции, привело к идее изменить обучающую и контрольную выборки следующим образом: мы перенесли две структуры адамантан и диметилглиоксим, из контрольной в обучающую выборку, а муравьиную кислоту, наоборот, из обучающей в контрольную выборку. Таким образом, обучающая выборка состояла из 92, а контрольная выборка из 12 соединений. Построенная QSPR-модель (Модель 3, Табл. 5 на стр. 233) имеет лучшую прогнозирующую способность по сравнению с моделями 1 и 2. Столь резкое понижение ошибки прогноза на контрольной выборке при столь небольшой модификации разбивки данных на обучающую и контрольную выборку, однако, свидетельствует о наличии проблемы

«редких фрагментов», когда отдельные соединения из контрольной выборки содержат фрагменты, плохо представленные в обучающей выборке.

Уравнение, соответствующее модели, полученной на основе трех фрагментных дескрипторов (Модель 3, Табл. 5 на стр. 174), которая характеризуется высокой прогнозирующей способностью, приведено ниже:

$$\Delta_{\text{sub}} H_{\text{расч.}} = +5.57 + 1.23 \text{ fr1} + 6.92 \text{ fr2} + 6.95 \text{ fr3} \quad (1)$$

$n = 104$, $r^2 = 0.8450$, $s = 2.97$ ккал/моль, $F = 160$, средняя ошибка (по модулю) на прогнозе 2.16 ккал/моль, где fr1 – число любых неводородных атомов, \bullet ; fr2 – количество фрагментов вида $=\text{CR-OH}$; fr3 – количество фрагментов вида $\text{N}_{\text{sp}^3}\text{-C=O}$.

На рисунках Рис. 35 приведен разброс экспериментальных и расчетных значений энтальпии сублимации, соответствующий данной модели.

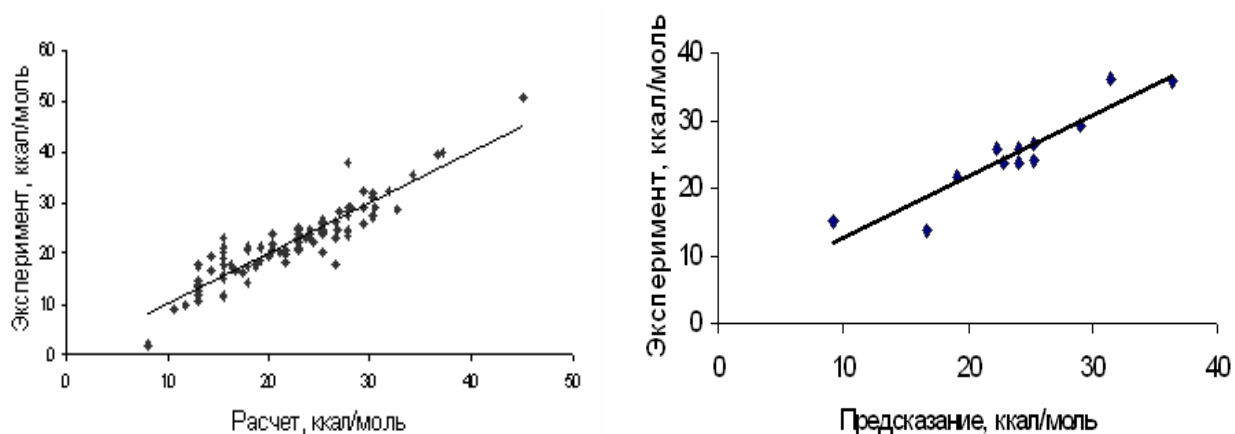


Рис. 35. Диаграмма разброса расчетных и экспериментальных значений энтальпии сублимации для обучающей (слева) и контрольной (справа) выборок из Базы 3 согласно линейно-регрессионной модели (уравнение 1)

Таким образом, впервые исследованы энтальпии сублимации органических соединений различных классов в рамках фрагментного подхода на основе метода QSPR. Показано, что данная методология позволяет получить модели расчета энтальпии сублимации с параметрами, сравнимыми, а в ряде случаев превосходящими характеристики регрессионных уравнений, предложенных в литературе. Иными словами, предложена модель, позволяющая прогнозировать

энтальпию сублимации соединений исходя из дескрипторов, учитывающих фрагментный состав молекулы.

5.2.6. Прогнозирование температуры вспышки органических соединений

Температура вспышки (T_f) – одна из важных характеристик горючих свойств органических веществ [385-387]. Она определяется как нижняя граница температуры, при которой смесь паров данного вещества с воздухом может быть подожжена при иницировании [385-387].

Величины T_f известны для многих соединений [387]; однако они не всегда публикуются даже для промышленно важных соединений. Более того, во многих случаях экспериментальное определение этой величины для токсичных, летучих, взрывчатых и радиоактивных веществ затруднительно. Все это диктует необходимость разработки теоретических методов оценки температуры вспышки. Для T_f были предложены различные схемы расчета, в том числе основанные на QSPR-исследованиях [386, 387].

В настоящей работе мы рассмотрели возможности применения структурных дескрипторов для QSPR-исследования температуры вспышки. Для проведения исследований мы использовали программные комплексы EMMA (см. раздел 8.1) и NASAWIN (см. раздел 8.2) в сочетании со входящим в оба комплекса дескрипторным блоком FRAGMENT (см. раздел 8.3).

Составление баз данных. По данным работы [386] была сформирована База 1, состоящая из 400 структурно-разнородных органических соединений. Кроме того, по данным из статьи [387] была создана База 2, содержащая 271 соединение. Она также включает в себя разнообразные классы органических соединений.

Результаты и обсуждение. Прежде всего, следует рассмотреть данные QSPR-исследования, приведенные в работах [386, 387], что необходимо для сравнения с результатами, полученными нами. В работе [386] были получены модели для расчета температуры вспышки с использованием PLS (метода частичных наименьших квадратов) и нейронной сети. В последнем случае при ис-

пользовании 25 дескрипторов, характеризующих вклады функциональных групп и атомов различных типов, для 135 соединений обучающей выборки, 133 контрольной и 132 выборки для прогноза авторы получили величины s (стандартного отклонения) 10.8°C, 14.1°C и 14.3°C, соответственно. Однако для метода PLS результаты были значительно менее удовлетворительны: величина s для каждого из этих случаев составляла 21°C, 25°C и 23°C, соответственно, что может свидетельствовать о ее нелинейном характере моделируемой зависимости. В работе [387] проведено моделирование температуры вспышки с использованием программы CODESSA. Авторы получили трехпараметровое уравнение со следующими статистическими характеристиками: R^2 (коэффициент детерминации) = 0.9020, R^2_{cv} (квадрат коэффициента корреляции при скользящем контроле) = 0.8985, s (стандартное отклонение) = 16.1 °C.

Табл. 6. Статистические характеристики QSPR-моделей для температуры вспышки

Модель	База	Обучающая выборка			Контрольная выборка	
		$N_{дескр}$	R^2	$s, ^\circ\text{C}$	$R^2_{прогн}$	$MAE_{прогн}, ^\circ\text{C}$
1	1	9	0.872	18.8	0.833	15.2
2	1A	9	0.871	18.9	0.829	15.3
3	2	9	0.932	13.7		
4	2A	9	0.935	13.3		
5	2	9	0.920	14.8	0.931	9.9

На первом этапе работы мы решили повторить результаты работы [386] (исследуя обучающую и контрольную выборки, идентичные приведенным в работе), но используя фрагментные дескрипторы. Данные, полученные на основе линейно-регрессионного анализа для Баз 1 и 1A, приведены в Табл. 6 на стр. 177 (Модели 1 и 2, соответственно). При построении моделей использовали процедуру пошагового включения рассчитанных дескрипторов в модель. Модель 1, построенная с использованием 9 фрагментных дескрипторов, имеет статистические параметры, превосходящие показатели PLS модели (средняя абсо-

лютная ошибка для обучающей выборки 20.6 °С, для контрольной выборки - 23.3 °С):

$$T_{f \text{ расч.}} = - 0.826 + 0.285 fr1 + 0.497 fr2 + 0.151 fr3 - 6.718 fr4 + 0.208 fr5 + 0.130 fr6 - 1.87 fr7 + 4.50 fr8 + 0.369 fr9$$

$n = 398$, $R^2 = 0.8724$, $s = 18.8$ °С, средняя ошибка (по модулю) на прогнозе 15.2 °С, где fr_i равно числу следующих фрагментов в молекулах: $fr1$ - N, $fr2$ - OH, $fr3$ - • (произвольный атом), $fr4$ - CH₃, $fr5$ - C-S, $fr6$ - C-C=O, $Fr7$ - •-•-• (цепочка из трех произвольных атомов), $fr8$ - C_{Ar}H÷C_{Ar}H÷C_{Ar}R÷C_{Ar}H (÷ - ароматическая связь), $fr9$ - C-C-C-Hal

Расширение числа используемых фрагментных дескрипторов до 25 позволяет улучшить качество линейно-регрессионной модели практически до качества нейросетевой [386]. В их число входят дескрипторы, характеризующие количество в молекуле атомов галогенов, N, O, S; а также двух- и трехатомных фрагментов с различными типами связей (двойной, тройной, ароматической: $fr1$ - I, $fr2$ - F, $fr3$ - Br, $fr4$ - S, $fr5$ - N, $fr6$ - OH, $Fr7$ - •, $fr8$ - C=O, $fr9$ - CH₃NR₂, $fr10$ - CH₂Hal, $fr11$ - =CR-NHR, $fr12$ - =CR-OH, $fr13$ - CH₃-C_{sp}³, $fr14$ - HC_{Ar}÷C_{Ar}R÷C_{Ar}, $fr15$ - C-C=O, $fr16$ - =CR-C_{sp}³-Cl, $fr17$ - CH₂-CH₂-C≡, $fr18$ - C-C_{sp}³-Cl, $fr19$ - =C-C_{Ar}÷C_{Ar}-OH, $fr20$ - C-C-C-N, $fr21$ - C_{Ar}÷C_{Ar}÷C_{Ar}÷C_{Ar}-N, $fr22$ - C-C-C-S-C, $fr23$ - C-C-C-C-C-O, $fr24$ - CH₃- C_{Ar}(÷C_{Ar}H)₂, $fr25$ - Hal-C(-C)₂. На Рис. 36 (стр. 179) представлена диаграмма разброса расчетных и экспериментальных значений температуры вспышки для обучающей и контрольной выборок соединений Базы 1 согласно модели, построенной на 25 фрагментных дескрипторах ($R^2 = 0.9557$, $s = 11.4$ °С, средняя абсолютная ошибка прогноза = 11.8, среднеквадратичная ошибка для обучающей выборки, $RMS_{обуч.} = 10.87$ °С, среднеквадратичная ошибка прогноза $RMS_{прог.} = 15.75$ °С).

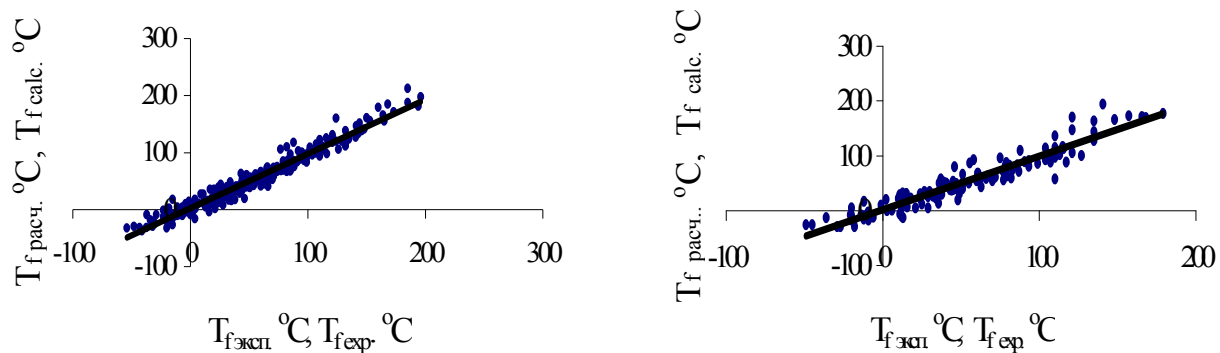


Рис. 36. Диаграмма разброса расчетных и экспериментальных значений температуры вспышки для обучающей (слева) и контрольной (справа) выборок Базы 1 согласно линейно-регрессионной модели, построенной на 25 фрагментных дескрипторах

Уменьшение количества соединений в Базе 1 за счет исключения 12 структур приводит к незначительному ухудшению качества моделей для Базы 1А (ср. Модели 1 и 2, Табл. 6 на стр. 177), при этом природа используемых в модели дескрипторов остается в целом неизменной, кроме замены фрагмента ($C_{Ar}H \div C_{Ar}H \div C_{Ar}R \div C_{Ar}H$) на фрагмент ($-O-CR=O$).

Далее мы использовали фрагментные дескрипторы для построения моделей для Базы 2 и “уменьшенной” Базы 2А (Табл. 6 на стр. 177, Модели 3 и 4). Как это было сделано в работе [387], для обучающей выборки, куда были включены все соединения, представленные в Базе 2, мы получили модели, по качеству не уступающие моделям 1 и 2 и превосходящие по статистическим показателям модель (см. выше), приведенную в работе [387]. Например, модель, построенная для Базы 2 на 25 дескрипторах, имеет следующие статистические показатели: $R^2 = 0.9566$, $s = 11.2$ °C, $RMSE_{обуч.} = 10.67$ °C.

Предсказательную способность QSPR-модели для Базы 2 мы оценили, используя ее разбивку на обучающую (179 соединений) и контрольную (89 соединений) выборки. Модель, построенная на 9 фрагментных дескрипторах, имеет весьма высокие прогнозирующие свойства ($R^2_{прогн.} = 0.9315$, средняя ошибка (по модулю) прогноза = 9.9 °C (Табл. 6 на стр. 177, Модель 5).

Таким образом, нами построены на основе фрагментных дескрипторов линейно-регрессионные модели, позволяющие прогнозировать температуру

вспышки с точностью, в ряде случаев, приближающейся к точности ее экспериментального определения.

5.2.7. Прогнозирование сродства азо- и антрахиноновых красителей к целлюлозному волокну

Взаимодействие красителей различной природы с хлопчатобумажным волокном представляет многостадийный физико-химический процесс, определяемый специфическими особенностями структуры текстильного полимера и природой молекулы красителя. Одной из основных характеристик, описывающей взаимодействие красителя с волокном, является химическое сродство красителя к волокну (аффинность), экспериментально определяемое разностью химических потенциалов красителя в волокне и в растворе в стандартных условиях, $-\Delta\mu^0$ (кДж·моль⁻¹). Этот параметр зависит от множества физико-химических факторов, оказывающих влияние на взаимодействие красителя с волокном (электростатические и ван дер Ваальсовы взаимодействия, образование водородных связей, гидрофобность и др.) [388]. Поэтому для исследования аффинности широко используются методы QSAR и 3D-QSAR. Так, методом CoMFA было показано, что на сродство анионных и нейтральных азо- [389, 390], гетероциклических моноазо- [391], симметричных биазо- [392] и антрахиноновых [393] красителей к целлюлозному волокну доминирующее влияние оказывают электростатические взаимодействия.

В задачу данной работы входило исследование сродства краситель-целлюлоза в рамках фрагментного подхода с использованием методологии QSPR. Исследование проводили с помощью программного QSAR/QSPR-комплекса NASAWIN (см. раздел 8.2) с использованием дескрипторного блока FRAGMENT (см. раздел 8.3). В работе исследовали 3 выборки соединений, включающие 30 серосодержащих азо-красителей [390] (База 1); 49 антрахиноновых красителей [394] (База 2); и комбинированную выборку, содержащую оба набора структур (База 3).

На первом этапе работы для исследуемых выборок с помощью программного комплекса NASAWIN были построены линейно-регрессионные модели с использованием дескрипторов, характеризующих фрагменты с максимальной длиной цепочек 6, 10 и 15 атомов и внешней контрольной выборки, включающей каждое пятое соединение базы. Эти модели (Табл. 7, модели 1-12) обладают хорошими описательными и прогнозирующими свойствами. Наилучшее качество прогноза достигнуто при включении в модель фрагментов длиной до 15 атомов.

Табл. 7. Статистические параметры QSPR моделей на основе фрагментных дескрипторов для сродства азо- и антрахиноновых красителей к целлюлозному волокну

	База	$R^2_{\text{обуч.}}$	$S, \text{ кДж} \cdot \text{моль}^{-1}$	$R^2_{\text{прог.}}$	F	Число фрагментных дескрипторов / максимальное число атомов во фрагменте
1	База 1 (азо-)	0,949	0,87	0,896	88,6	4 / 6
2		0,957	0,81	0,850	81,7	5 / 6
3		0,958	0,83	0,839	64,9	6 / 6
4		0,949	0,87	0,900	88,6	4 / 10
5		0,957	0,81	0,850	81,7	5 / 10
6		0,971	0,95	0,908	161,7	4 / 15
7	База 2 (антрахиноны)	0,918	0,56	0,860	51,1	7 / 15
8		0,924	0,55	0,866	47,5	8 / 15
9		0,931	0,53	0,866	44,9	9 / 15
10	База 3 (азо- и антрахиноны)	0,955	0,79	0,854	136,8	9 / 15
11		0,960	0,75	0,832	136,5	10 / 15
12		0,968	0,68	0,807	154,0	11 / 15

Для сравнительной оценки качества фрагментных и литературных моделей, полученных для тех же выборок методами сравнительного анализа молекулярного поля (CoMFA) и множественной линейной регрессии (MRL) с использованием квантово-химических дескрипторов (азо- соединения) [390] и методом сравнительного анализа молекулярной поверхности (CoMSA, азо- и антрахиноновые соединения) [394], мы построили серию моделей при использо-

ванием скользящего контроля с исключением по одному соединению. По своим статистическим показателям полученные модели сопоставимы с цитируемыми в литературе, а в ряде случаев их превосходят. Так, регрессионная модель, построенная для Базы 1 на 4 фрагментных дескрипторах (длина цепочки во фрагменте 15 атомов) имеет показатели ($R^2_{\text{обуч.}} = 0,967$ кДж·моль⁻¹; $F = 181,6$; квадрат коэффициента корреляции при скользящем контроле, $Q^2 = 0,949$; стандартное отклонение, $s = 0,66$ кДж·моль⁻¹; среднеквадратичная ошибка, $RMSE_{\text{ск}} = 0,74$ кДж·моль⁻¹; стандартное отклонение при скользящем контроле, $s_{\text{ск}} = 0,80$ кДж·моль⁻¹), превосходящие параметры лучшей регрессионной модели, полученной для этой же выборки на основе использования в качестве дескрипторов энергий высшей занятой и низшей свободной молекулярных орбиталей $E_{\text{НОМО}}$, $E_{\text{ЛУМО}}$ и среднего арифметического между ними, (3 дескриптора, $R^2_{\text{обуч.}} = 0,92$; стандартное отклонение, $s = 1,02$ кДж·моль⁻¹; $F = 95,0$; $Q^2 = 0,89$; стандартная ошибка при скользящем контроле 1,19) [390]. Для PLS CoMFA [390] и CoMSA моделей [394] значения Q^2 лежат в пределах 0,63-0,75 и 0,829-0,970, соответственно. В полученную модель входят дескрипторы, описывающие следующие фрагменты молекул азо-соединений: $=\text{RC}-\text{C}$, $\text{RC}_{\text{Ar}}\div\text{C}_{\text{Ar}}-\text{NH}_2$, $\bullet=\bullet-\bullet\div\bullet\div\bullet\div\bullet$ (\bullet - произвольный атом) и $\text{C}-\text{C}-\text{C}\div(\text{C}\div)_2\text{C}-\text{N}=\text{N}-\text{C}\div(\text{C}\div)_4\text{C}-\text{N}$.

Фрагментные модели, построенные для выборки антрахиноновых красителей (База 2), по прогнозирующим свойствам также не уступают литературным моделям [394]. Квадрат коэффициента корреляции при скользящем контроле модели, включающей 8 дескрипторов (цепочки из 15 атомов), ($R^2_{\text{обуч.}} = 0,942$; $s = 0,46$; $F = 81,5$; $Q^2 = 0,915$; $RMSE_{\text{ск}} = 0,50$ кДж·моль⁻¹; $s_{\text{ск}} = 0,55$ кДж·моль⁻¹), превышает максимальное значение Q^2 (0,88) CoMSA модели [394]. Наиболее значителен вклад фрагментов: RC_{Ar} , $\text{C}_{\text{Ar}}-\text{N}_{\text{sp}}^3-\text{C}$, $\text{HC}_{\text{Ar}}\div\text{C}_{\text{Ar}}-\text{NHR}$ и $\text{C}_{\text{sp}}^3-\text{O}-(\text{C}_{\text{Ar}}\div)_7\text{C}_{\text{Ar}}-\text{N}_{\text{sp}}^3$.

На основе слияния узких выборок азо- и антрахиноновых красителей (База 3) мы получили более универсальную модель для описания сродства краситель-целлюлоза и оценили ее прогнозирующую способность при помощи скользящего контроля. Модель включает 10 фрагментных дескрипторов (цепочки из 15 атомов) и имеет следующие характеристики: $R^2_{\text{обуч.}} = 0,954$; $Q^2 =$

0,935; $s = 0,76$ кДж·моль⁻¹; $F = 139,5$; $s_{\text{ск}} = 0,89$ кДж·моль⁻¹; $RMSE_{\text{ск}} = 0,83$ кДж·моль⁻¹:

$$-\Delta\mu^0 = -0.49 + 2.19 \text{ Fr1} - 1.03 \text{ Fr2} - 1.01 \text{ Fr3} - 0.56 \text{ Fr4} + 3.13 \text{ Fr5} + 0.21 \text{ Fr6} + 0.11 \text{ Fr7} + 0.85 \text{ Fr8} + 0.45 \text{ Fr9} + 1.10 \text{ Fr10} \quad (1)$$

Где $R^2_{\text{обуч.}} = 0,954$; $Q^2 = 0,937$; $s = 0,70$ кДж·моль⁻¹; $F = 139,5$; $s_{\text{ск}} = 0,82$ кДж·моль⁻¹; $RMSE_{\text{ск}} = 0,76$ кДж·моль⁻¹;

Fr1/ $C_{\text{Ar}}-N$, Fr2/ $C_{\text{sp}^3}-N_{\text{sp}^3}-C$, Fr3/ $HC_{\text{Ar}}\div C_{\text{Ar}}-NH_2$, Fr4/ $RC_{\text{Ar}}\div C_{\text{Ar}}-N$,

Fr5/ $N_{\text{sp}^3}-C_{\text{Ar}}\div C_{\text{Ar}}-N=$, Fr6/ $\bullet-\bullet\div\bullet\div\bullet-\bullet$, Fr7/ $C-(C\div)_6C$,

Fr8/ $C_{\text{sp}^3}-O-(C_{\text{Ar}}\div)_7C_{\text{Ar}}-N_{\text{sp}^3}$,

Fr9/ $C_{\text{Ar}}\div(C_{\text{Ar}}\div)_3C_{\text{Ar}}-N=N-C_{\text{Ar}}\div(C_{\text{Ar}}\div)_2C_{\text{Ar}}-C=C$,

Fr10/ $N_{\text{sp}^3}-C_{\text{Ar}}\div(C_{\text{Ar}}\div)_4C_{\text{Ar}}-N=N-C_{\text{Ar}}\div(C_{\text{Ar}}\div)_2C_{\text{Ar}}-N_{\text{sp}^3}$

Наибольший вклад в модель вносят фрагментные дескрипторы $RC_{\text{Ar}}-N$, и $N_{\text{sp}^3}-C_{\text{Ar}}\div C_{\text{Ar}}-N=$. Таким образом на основе дескрипторов, учитывающих фрагментный состав молекулы, предложены линейно-регрессионные QSPR-модели, позволяющие прогнозировать сродство азо- и антрахиноновых красителей к целлюлозному волокну. Этим примером продемонстрировано, что предложенные фрагментные дескрипторы в сочетании со статистическим аппаратом множественной линейной регрессии являются мощным инструментом для прогнозирования сложных промышленно-важных свойств органических соединений.

5.3. Фрагментные дескрипторы с «выделенными» атомами

Мы предлагаем подход, который позволяет значительно расширить круг свойств, для прогнозирования которых можно применять фрагментные дескрипторы за счет указания специальных «выделенных» атомов, играющих специфическую роль в природе моделируемого свойства. Например, при моделировании константы основности аминов логично отметить тот самый атом азота внутри химической структуры, который участвует в рассматриваемом кислотно-основном равновесии. Суть предлагаемого метода заключается в том, что: 1) такие «выделенные» атомы помечаются определенными метками в соответст-

вии с тем, по каким причинам этот атом выделен; 2) при генерации фрагментных дескрипторов каждая такая метка рассматривается как отдельный псевдоатом с именем, соответствующем символу метки; 3) при построении уравнений «структура-свойство» должна иметься возможность включать в модели только те дескрипторы, которые содержат такой псевдоатом.

Мы предлагаем использовать фрагментные дескрипторы с «выделенными» атомами для моделирования широкого круга свойств: (1) при расчете локальных характеристик молекул, таких, например, как химические сдвиги в спектрах ЯМР либо кислотно-основные свойства определенных атомов в молекулах; (2) при прогнозировании биологической активности для однородных выборок соединений, содержащих общий фрагмент с анкерными атомами, к которым присоединены заместители; (3) для прогнозирования кинетических параметров химических реакций одного типа; (4) при прогнозировании физических свойств полимеров (за счет добавления специальных меток к атомам, принадлежащим основной цепи полимера); (5) для прогнозирования свойств, обусловленных образованием супрамолекулярных комплексов (за счет добавления специфических меток, указывающих на роль атомов в супрамолекулярном взаимодействии); (6) для учета стереохимической информации (путем добавления меток S и R либо D и L к стереохимическим центрам, а также E и Z к атомам, связанным двойной связью). В каждом случае предлагаемый прием обеспечивает использование в построении моделей наиболее важных по смыслу фрагментных дескрипторов. Таким образом, использование фрагментных дескрипторов с «выделенными» атомами позволяет значительно расширить сферу применения фрагментного подхода в поиске количественных соотношений «структура-свойство», а также снять некоторые ограничения, которые ранее были свойственны фрагментным дескрипторам.

Применение таких дескрипторов нами проиллюстрировано на примерах моделирования: (1) химических сдвигов в ^{31}P ЯМР спектрах производных монофосфинов, (2) способности аналогов 1-[(2-гидроксиэтокси)-метил]-6(фенилтио)тимина (НЕРТ) к ингибировать обратную транскриптазу вируса ВИЧ-1 и (3) констант скорости гидролиза эфиров карбоновых кислот. Еще один

пример использования такого вида фрагментных дескрипторов для прогнозирования констант ионизации рассмотрен в подразделе 7.1.2.

Расчет фрагментных дескрипторов с “выделенными” атомами и построение QSAR/QSPR-моделей методами быстрой пошаговой множественной линейной регрессии (БПМЛР) и трехслойной нейросети обратного распространения (ИНС) осуществляли с помощью программного комплекса NASAWIN (см. раздел 8.2).

5.3.1. Прогнозирование химических сдвигов в ^{31}P ЯМР спектрах замещенных монофосфинов

Для построения QSPR-моделей химических сдвигов в ^{31}P ЯМР спектрах замещенных монофосфинов мы использовали базу данных, включающую 291 фосфинов $\text{RN}_{3-n}\text{R}_n$, в том числе 29 первичных, 38 вторичных и 224 третичных с различными заместителями [395]. Разброс в экспериментальных значениях прогнозируемого параметра составил от -183 до +61 ppm. Известно, что величины химических сдвигов зависят от степени экранирования ядер атомов электронным облаком, плотность которого зависит от характера присоединенных к этим атомам заместителей. Поэтому представлялось целесообразным использование дескрипторов, описывающих электронное и пространственное влияние этих заместителей. В качестве таковых были выбраны дескрипторы, основанные на числе вхождения в структуру фрагментов, содержащих от 4 до 10 неводородных атомов и включающих атом P, маркированный меткой “a”. Лучшая из серии полученных нами БПМЛР и ИНС комбинированных моделей модель БПМЛР имеет следующие характеристики прогнозирующей способности: $Q^2_{\text{DCV}} = 0.8298$, $RMSE_{\text{DCV}} = 0.5679$ ppm, $MAE_{\text{DCV}} = 6.1$ ppm. Диаграмма разброса для нее приведена на Рис. 37.

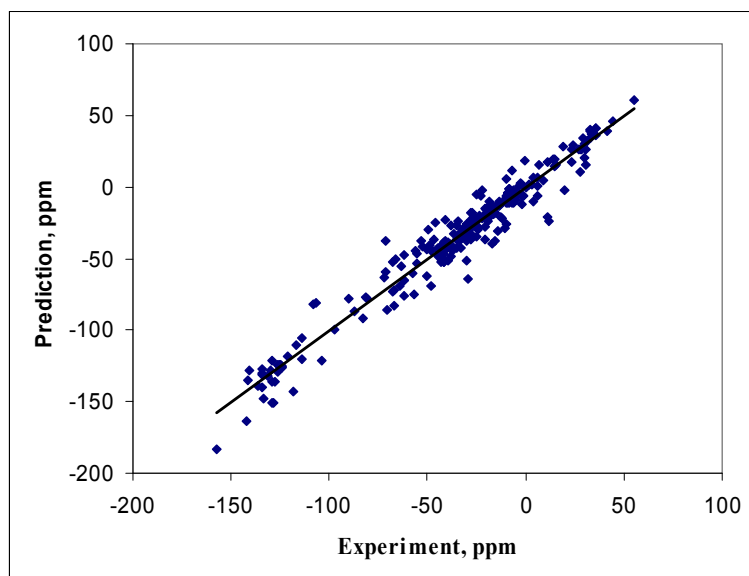


Рис. 37. Диаграмма разброса при прогнозировании химических сдвигов в ^{31}P ЯМР спектрах замещенных монофосфинов

Наиболее значимыми для описания исследуемого свойства являются приведенные на Рис. 38 фрагменты с “выделенным” атомом P^a . Первые три фрагмента отражают σ -индукционное влияние алкильных заместителей на атом фосфора, четвертый – эффект сопряжения с ароматическим ядром, пятый – влияние расположенного в орто-положении атома фтора.

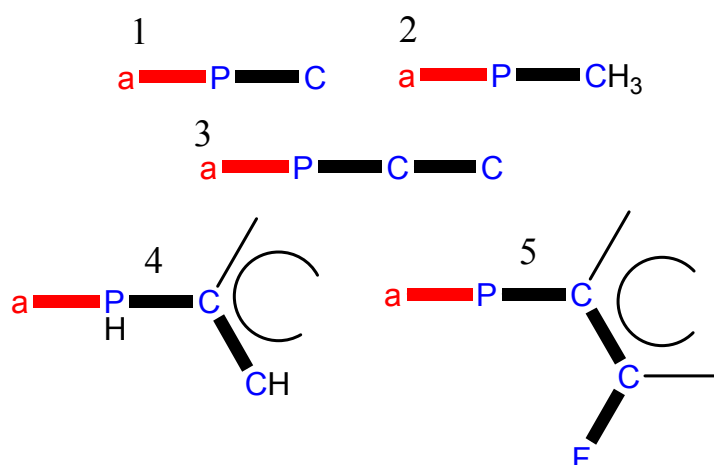


Рис. 38. Наиболее важные фрагменты для химического сдвига в ^{31}P ЯМР спектрах замещенных монофосфинов.

Данный пример иллюстрирует возможность использовать фрагментные дескрипторы с «выделенными» атомами для прогнозирования локальных свойств химических соединений, которые можно приписать определенным атомам или группам атомов внутри молекулы. В этом случае использование це-

почечных фрагментов с терминальными «выделенными» атомами позволяет получать легко интерпретируемые модели, наглядно показывающие пути влияния отдельных атомов или групп внутри молекулы на изучаемое свойство.

5.3.2. Прогнозирование способности аналогов 1-[(2-гидроксиэтокси)-метил]-6(фенилтио)тимина (НЕРТ) ингибировать обратную транскриптазу вируса ВИЧ-1

Ингибирующую активность в отношении обратной транскриптазы вируса ВИЧ-1, представленную эффективной концентрацией соединений, необходимой для достижения 50% защиты клеток линии МТ-4 от цитотоксического действия вируса ($\log 1/EC_{50}$), мы исследовали для однородной выборки производных НЕРТ [396]. На Рис. 39 приведены общий структурный элемент соединений выборки и фрагменты заместителей R_1 , R_2 и R_3 , которые соответственно связаны с анкерными атомами общего фрагмента, маркированными метками “b”, ”c” и ”d”, и которые вносят наибольший вклад в лучшую комбинированную модель:

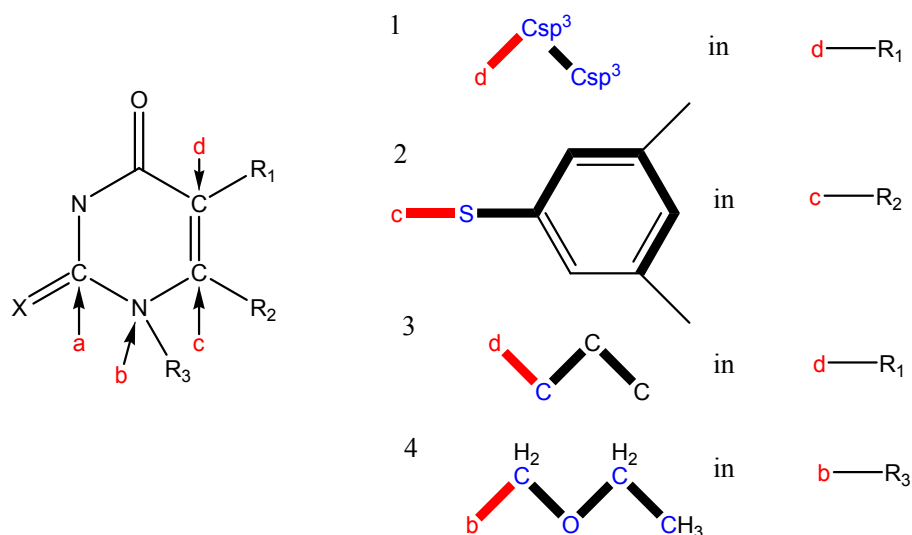


Рис. 39. Наиболее важные фрагменты для ингибирования обратной транскриптазы ВИЧ-1 производными НЕРТ

Модель получена с помощью метода ИНС и имеет следующие параметры прогнозирующей способности: $Q^2_{DCV} = 0.8561$, $RMSE_{DCV} = 0.520$ и $MAE_{DCV} = 0.41$. Диаграмма разброса для нее представлена на Рис. 40.

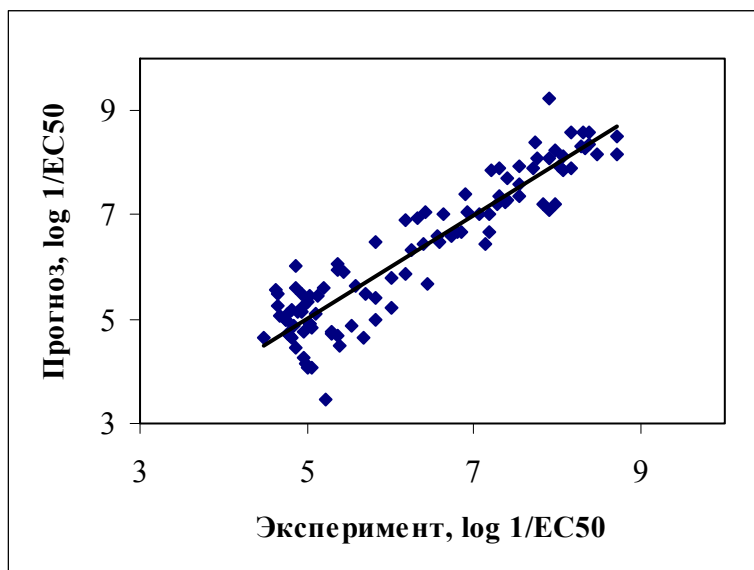


Рис. 40. Диаграмма разброса для ингибирования обратной транскриптазы вируса ВИЧ-1 производными НЕРТ

Рассматриваемый пример иллюстрирует возможность применения фрагментных дескрипторов с «выделенными» атомами для количественного прогнозирования биологической активности органических соединений внутри рядов соединений с одинаковым общим фрагментом (скелетом). Следует отметить, что обычно фрагментные дескрипторы редко используются для этой цели, поскольку аппроксимируемый с их помощью вклад конкретной группировки атомов в общее свойство оказывается независимым от того, где именно внутри химической структуры она находится. Поскольку это плохо соотносится с природой биологической активности, которая связана с точным пространственно-электронным распознаванием молекул, то это часто приводит к плохой прогнозирующей способности построенных QSAR-моделей и к невозможности их интерпретации с целью выявления факторов, влияющих на биологическую активность.

Предлагаемые фрагментные дескрипторы с «выделенными» атомами полностью решают эту проблему, поскольку позволяют позиционировать все рассматриваемые фрагменты относительно заранее заданных внутри химиче-

ской структуры «реперных точек». На приведенной на Рис. 40 общей структуре для рассматриваемого ряда соединений такими «реперными точками» являются места подсоединений заместителей к общему скелету, которые мы «выделили» путем приписывания им меток *a*, *b*, *c* и *d*. Благодаря этому аппроксимируемый при помощи фрагментных дескрипторов с «выделенными» таким образом атомами вклад группировки атомов в общую биологическую активность оказывается зависимым от ее положения внутри химической структуры. Это приводит не только к существенному росту прогнозирующей способности получающихся QSAR-моделей, но и делает их легко интерпретируемой со структурно-химической точки зрения, поскольку значения регрессионных коэффициентов в линейных моделях и введенной нами характеристики M_x (см. раздел 4.2) для нейросетевых моделей четко показывают, какая группировка атомов в каком положении вносит какой вклад в биологическую активность, и, следовательно, какие изменения нужно внести для ее оптимизации. Более того, рассмотрение характеристик M_{xy} (см. раздел 4.2) позволяет выявить синергию и диссинергию во влиянии различных группировок атомов на биологическую активность. В определенном смысле предлагаемый подход можно считать дальнейшим развитием классического метода Фри-Вильсона [129].

5.3.3. Прогнозирование констант скорости гидролиза эфиров карбоновых кислот

База данных, содержащая сведения по константам скорости гидролиза, измеренным в диапазоне температур от 0 до 154°C в бинарных системах вода:растворитель (концентрация неводного компонента 0-98%), для 2092 эфиров карбоновых кислот, была использована для прогнозирования константы скорости реакции, $\lg k$ [397, 398]. В зависимости от природы заместителей у атомов С и О кислотного остатка эфиров экспериментальные значения $\lg k$ изменялись от -7.53 до -0.17. QSPR-модели строили с помощью метода ИНС с использованием в качестве дескрипторов температуры, концентрации органических растворителей, параметров, характеризующих их свойства [398], а также фрагментов, со-

державших “выделенные” атомы, которые, в соответствии с основными концепциями механизма реакции [399], входят в состав реакционных центров на какой-либо из ее стадий. Каждый из таких фрагментов описывает влияние ближайших к реакционным центрам групп атомов на скорость реакции. Лучшая комбинированная модель для этой выборки получена с помощью метода ИНС и имеет: $Q^2_{DCV} = 0.9162$, $RMSE_{DCV} = 0.31$ и $MAE_{DCV} = 0.19$. Диаграмма разброса для полученной модели приведена на Рис. 41.

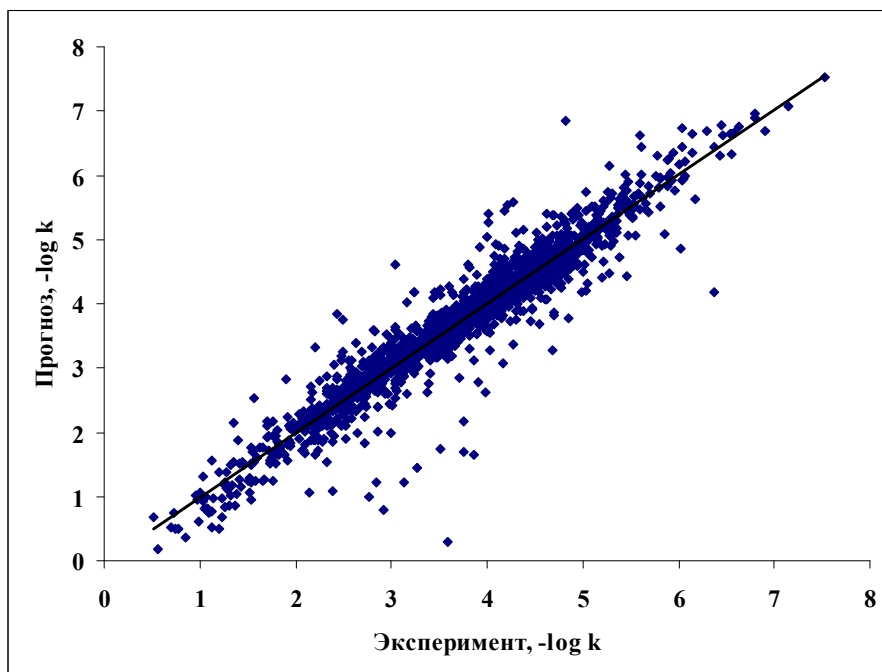


Рис. 41. Диаграмма разброса для констант скорости гидролиза сложных эфиров

На Рис. 42 схематически приведены три фрагмента, наличие которых в структуре наиболее сильно отражается на величине константы скорости гидролиза.

Первый фрагмент описывает стерическое влияние заместителей при α -углеродном атоме карбоновой кислоты, второй – электронное влияние расположенного в уходящей группе атома кислорода, несущего неподеленные электронные пары, третий – влияние фенильной группы при карбоксиле.

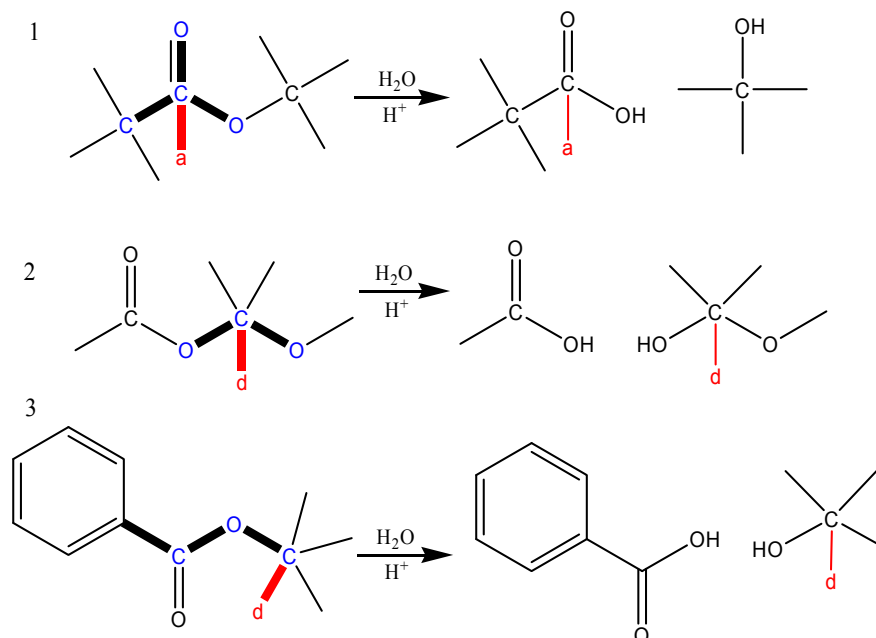


Рис. 42. Наиболее важные фрагменты для прогнозирования констант скоростей гидролиза сложных эфиров

Таким образом, данный пример иллюстрирует возможность применения фрагментных дескрипторов с «выделенными» атомами для количественного прогнозирования кинетических констант органических реакций, а также для автоматизированного извлечения из огромной массы экспериментальных данных основных факторов, влияющих на протекание органических реакций. Можно надеяться, что в будущем подобного рода анализ займет достойное место в широком арсенале средств теоретической органической химии.

5.4. Псевдофрагментные подходы. FRAGPROP. Прогнозирование физических свойств полимеров

Одним из недостатков фрагментных дескрипторов, является проблема редких фрагментов, которые могут отсутствовать в обучающей выборке, но присутствовать в соединениях, для которых осуществляется прогноз. Поскольку величины вкладов редких фрагментов не могут быть определены по обучающей выборке, то можно ожидать значительных ошибок прогнозирования для соединений, содержащих такие фрагменты. Мы предлагаем решать эту проблему путем введения дополнительных дескрипторов, значения которых в

какой-то мере были бы связаны с величинами вкладов фрагментов в прогнозируемое свойство, Мы также предлагаем использовать для этого особую категорию фрагментных дескрипторов, значения которых вычисляются путем комбинирования свойств присутствующих в этих фрагментах атомов. Дескрипторы такого рода мы будем называть псевдофрагментными дескрипторами, чтобы их отличать от «настоящих» фрагментных дескрипторов, имеющих в качестве значения числа встречаемости либо индикаторы наличия тех или иных фрагментов в структурах химических соединений. В качестве свойств атомов для прогнозирования физико-химических свойств органических молекул можно, например, использовать атомную массу, число электронов, ковалентный радиус, электроотрицательность, потенциал ионизации и т.д., поскольку предполагается, что от них зависят величины вкладов фрагментных дескрипторов в прогнозируемое свойство. Важно также, чтобы используемые комбинации свойств имели ясный физический смысл, поскольку в этом случае возрастают шансы наличия корреляции их значений с величинами вкладов фрагментов, При такой корреляции небольшое число псевдофрагментных дескрипторов начинает входить в статистические модели вместо многочисленных «настоящих» фрагментных дескрипторов, в том числе и потенциально редких, выступая тем самым в качестве сжатого обобщения последних. Это в значительной степени и решает проблему редких фрагментов, если псевдофрагментные дескрипторы строятся на основе часто встречающихся фрагментов, состоящих из отдельных атомов или небольших цепочек из произвольных атомов, которые присутствуют практически во всех молекулах.

В качестве первого примера псевдофрагментного дескриптора рассмотрим конструкцию $p1_AR3 = \frac{1}{N_a} \sum_{i=1}^{N_a} R_i^3$. В качестве атомного свойства здесь выступает атомный радиус. Очевидно, что куб атомного радиуса пропорционален «объему» атома. Поскольку суммирование идет по атомам, то они и выступают в качестве базового фрагмента для вычисления дескриптора. Физический смысл всего дескриптора – средний объем атома. Можно предположить, что он будет играть существенную роль при прогнозировании волюметрических свойств

веществ, например, плотности. Если даже будет требоваться осуществить прогноз подобного свойства для химического соединения, содержащего редкий элемент, отсутствующий в обучающей выборке, то все равно будет дана разумная аппроксимация его вклада в прогнозируемое свойство.

Рассматриваемые псевдофрагментные дескрипторы могут быть использованы при построении статистических моделей в сочетании с «настоящими» фрагментными дескрипторами. Эффективность отдельных комбинаций дескрипторов этого типа с фрагментными дескрипторами была также показана нами в работах [400, 401].

В настоящей работе мы исследовали дескрипторы на основе комбинаций атомов во фрагментах при прогнозировании трех ключевых физических характеристик полимеров: показателя преломления (n , 298K), температуры стеклования (T_g , K) и плотности в аморфном состоянии (ρ , г/см³, 298K.). Ранее эти свойства моделировались с использованием метода групповых вкладов Ван Кревелена [402] и схем Аскадского [403]. Эти методы не являются по своей сути статистическими, и поэтому для них не оцениваются статистические характеристики моделей. QSPR-модели для расчета свойств полимеров описаны в работе Бицерано [404], однако, для этих моделей не определена прогнозирующая способность с помощью скользящего контроля или независимой внешней выборки, что делает невозможным прямое сопоставление их статистических характеристик.

Рабочие выборки, включающие сведения об экспериментальных значениях показателя преломления, температуры стеклования и плотности в аморфном состоянии формировали на основе монографии [404].

Расчет фрагментных дескрипторов и построение количественных моделей структура-свойство осуществляли методами быстрой пошаговой множественной линейной регрессии (БПМЛР, см. подраздел 4.1.5) и трехслойной искусственной нейронной сети (нейросети обратного распространения, см. подраздел 1.2.4) с помощью программного комплекса NASAWIN (см. раздел 8.2). Генерировали наборы фрагментов, включающих от 1 до 5 неводородных атомов с учетом кратных связей, гетероатомов, функциональных групп и т.д. при

помощи дескрипторного блока FRAGMENT (см. раздел 8.3). Для расчета комбинаций свойств атомов во фрагментах использовали дескрипторный блок FRAGPROP (см. раздел 8.4). Этот дескрипторный блок позволяет вычислять 50 комбинаций свойств атомов (или дескрипторы FRAGPROP, fragmental properties) для фрагментов размерами от 1 до 5 неводородных атомов. Полный набор дескрипторов, вычисляемый блоком FRAGPROP, приведен в разделе 8.4 данной диссертационной работы. Для оценки прогнозирующей способности QSPR-моделей была применена процедура 5x4-кратного двойного скользящего контроля (см. подраздел 4.1.4). Вычисляемые статистические характеристики включают: (1) Q^2_{DCV} - параметр Q^2 ($Q^2=(SS-PSS)/SS$, где PSS - сумма квадратов ошибок прогноза свойства, SS - сумма квадратов отклонения свойства от среднего значения) для усредненных спрогнозированных значений, (2) $RMSE_{DCV}$ - среднеквадратическая ошибка прогнозирования, (3) MAE_{DCV} - средняя абсолютная ошибка прогнозирования.

Как показали расчеты, качество QSPR моделей, как линейно-регрессионных, так и нейросетевых, полученных для всех трех исследованных характеристик полимеров – показателя преломления, плотности в аморфном состоянии и температуры стеклования, значительно улучшается при включении в модели наряду с фрагментными дескрипторами, дескрипторов, описывающих комбинации свойств атомов во фрагментах. Это наблюдается для всего исследованного диапазона размеров фрагментов - от 1 до 5 неводородных атомов. Так, лучшая QSPR модель для показателя преломления, была получена методом БПМЛР на основе фрагментных дескрипторов, содержащих от 1 до 4 неводородных атомов, и имела следующие статистические характеристики: Q^2_{DCV} 0.7822, $RMSE_{DCV}$ 0.033, MAE_{DCV} 0.021. При включении в эту модель дескрипторов, описывающих свойства атомов во фрагментах (см. Табл. 8 на стр. 195), эти показатели улучшаются, соответственно, до 0.872, 0.026 и 0.015. В случае температуры стеклования добавление дескрипторов FRAGPROP в лучшую БПМЛР модель, построенную с использованием фрагментных дескрипторов, включающих от 1 до 5 неводородных атомов, также позволяет улучшить ее статистические показатели: от 0.849 до 0.864 (Q^2_{DCV}), от 45.0 до 42.7 ($RMSE_{DCV}$) и от 32.0

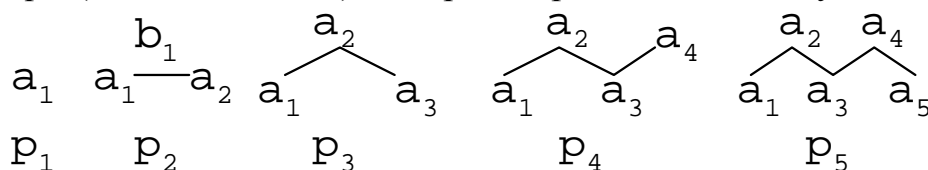
до 28.0 (MAE_{DCV}). Повышение прогнозирующей способности в наибольшей степени наблюдается в случае QSPR моделей, построенных для расчета плотности полимеров в аморфном состоянии. Например, статистические показатели лучшей из БПМЛР моделей, построенной с использованием фрагментов с размерами от 1 до 2 неводородных атомов (Q^2_{DCV} 0.474, $RMSE_{DCV}$ 0.159 и MAE_{DCV} 0.959), при комбинировании фрагментных дескрипторов с дескрипторами FRAGPROP, становятся, соответственно 0.910, 0.066 и 0.043. Комбинации свойств атомов во фрагментах, имеющие наибольшую значимость для описания исследованных свойств, приведены в Табл. 8.

Табл. 8. Формулы для расчета комбинаций свойств атомов во фрагментах и названия дескрипторов, наиболее часто встречающихся в QSPR-моделях, полученных для прогнозирования свойств полимеров (дескрипторы приведены по степени убывания частоты встречаемости в частных моделях для соответствующего свойства).

N	Название дескриптора	Формула
Плотность в аморфном состоянии		
1	Отношение числа электронов к числу атомов в молекуле или среднее количество электронов в атоме	$p1_ANe = N_e / N_a$
2	Среднее значение произведения электроотрицательностей атомов для всех связей в молекуле.	$p2_APE = \frac{1}{N_b} \sum_{p2} \chi(a_1) \cdot \chi(a_2)$
3	Максимальное значение произведения модуля разности электроотрицательностей для всех связей в молекуле на порядок соответствующей связи	$p2_HDE = \max_{p2} (\chi(a_1) - \chi(a_2) \cdot n_b)$
4	Сумма произведений разностей электроотрицательности атомов в положениях 1-2 и 5-4 для всех 5-атомных цепочек	$p5_SPDE = \sum_{p5} (\chi(a_1) - \chi(a_2)) \cdot (\chi(a_5) - \chi(a_4))$
5	Отношение суммы кубов атомных радиусов к числу атомов в молекуле	$p1_AR3 = \frac{1}{N_a} \sum_{i=1}^{N_a} R_i^3$
6	Среднее значение произве-	

	дений разностей электротрицательности атомов в положениях 1-2 и 5-4 для всех 5-атомных цепочек	$p5_APDE = \frac{1}{N_{p5}} \sum_{p5} (\chi(a_1) - \chi(a_2)) \cdot (\chi(a_5) - \chi(a_4))$
Температура стеклования		
7	Среднее значение произведений атомных радиусов в положениях 1-4 по всем 4-атомным цепочкам.	$p4_APR = \frac{1}{N_{p4}} \sum_{p4} R(a_1) \cdot R(a_2)$
8	Число π -электронов в молекуле	$p1_Npi = N_\pi$
9	Сумма модулей разностей электроотрицательностей для всех связей X-H в молекуле, где X-гетероатом	$p2_SDEHnc = \sum_{p2 a_1 \neq C} \chi(a_1) - \chi(H) $
Поляризуемость		
10	Средний атомный потенциал ионизации в молекуле.	$p1_AIP = \frac{1}{N_a} \sum_{i=1}^{N_a} I_i$
11	См. дескриптор 7	
12	Минимальная электроотрицательность атома в молекуле	$p1_LE = \min(\chi_i)$
13	См. дескриптор 9	
14	Среднее значение произведений разностей электротрицательности атомов в положениях 1-2 и 3-2 для всех трех-атомных связанных фрагментов без учета связей с атомами водорода	$p3_APDEnh = \frac{1}{N_{p3}} \sum_{p3} (\chi(a_1) - \chi(a_2)) \cdot (\chi(a_3) - \chi(a_2))$
30	Сумма произведений разностей электротрицательности атомов в положениях 1-2 и 4-3 для всех 4-атомных цепочек	$p4_SPDE = \sum_{p4} (\chi(a_1) - \chi(a_2)) \cdot (\chi(a_4) - \chi(a_3))$

где χ - электроотрицательность, ra – ковалентный атомный радиус, an (атомы), bn (связи) и pn (цепочки атомов), которые определяются следующим образом:



Таким образом, псевдофрагментные дескрипторы позволяют в существенной мере улучшать качество моделей, использующих фрагментные дескрипторы, и мы предполагаем, что это происходит за счет решения проблемы редких фрагментов. Следует отметить, что хотя псевдофрагментные дескрипторы могут и сами по себе участвовать в построении моделей «структурасвойство», наилучшие модели всегда получаются только в сочетании с «настоящими» фрагментными дескрипторами. Поэтому их применение следует рассматривать как способ улучшения моделей, построенных на базе фрагментных дескрипторов.

Кроме рассмотренного выше прогнозирования некоторых физических свойств полимеров, преимущество использования псевдофрагментных дескрипторов в качестве добавки к фрагментным дескрипторам продемонстрировано нами для прогнозирования температуры плавления ионных жидкостей (см. раздел 6.4) и констант связывания циклодекстрина с органическими молекулами [400]. Кроме того, псевдофрагментные дескрипторы в сочетании с дескрипторами, описывающими распределение зарядов в молекуле, хорошо себя зарекомендовали при прогнозировании эмбриотоксичности синтетических аналогов природных аминов [405].

ГЛАВА 6. СОЧЕТАНИЕ ИСКУССТВЕННЫХ НЕЙРОННЫХ СЕТЕЙ И ФРАГМЕНТНЫХ ДЕСКРИПТОРОВ

Данная глава посвящена изучению эффекта от совместного использования искусственных нейронных сетей и фрагментных дескрипторов. На большом числе примеров проводится сравнение с линейными моделями, построенными на тех же базах данных с применением тех же самых дескрипторов.

6.1. Первые свидетельства эффективности совместного использования искусственных нейронных сетей и фрагментных дескрипторов

В 1993 г. мы опубликовали статью, в которой искусственные нейронные сети и пошаговая множественная линейная регрессия были систематически применены при построении количественных корреляций «структура-свойства» (QSPR-моделей) для разнообразных физико-химических свойств углеводородов (главным образом, алканов) [406]. В частности, были построены модели для прогнозирования: 1) температуры алканов (выборка, насчитывающая 177 соединений, была взята из справочника [407]); 2) температуры плавления алканов (выборка, насчитывающая 90 соединений, была взята из справочника [407]); 3) октанового числа алканов, алкенов и циклоалкенов (выборка, насчитывающая 153 соединения, была взята из работы [408]); 4) одновременно шести свойств (молярного объема, молярной рефракции, теплоты испарения, критической температуры, критического давления и поверхностного натяжения) алканов (выборка, насчитывающая 69 соединений, была взята с работы [409]).

В ходе исследования два альтернативных набора дескрипторов были использованы для описания химических структур: топологические индексы (ТИ) [326] и фрагментные дескрипторы (ФД) [356]. Набор топологических индексов включал индексы молекулярной связности ${}^0\chi$, ${}^1\chi$, ${}^2\chi$, ${}^3\chi$, ${}^3\chi_c$, ${}^4\chi$, ${}^4\chi_c$, индекс Винера W и индексы молекулярной формы 0k , 1k , 2k , 3k . Топологические индексы рассчитывались при помощи разработанных нами дескрипторных блоков CONNECT, BALABAN и KARPA. В качестве структурных фрагментов брались

цепочки длиной до двух атомов. Основанные на них фрагментные дескрипторы рассчитывались при помощи разработанного нами дескрипторного блока FRAGMENT (см. разделы 5.1 и 8.3).

В N -м компьютерном эксперименте выборка, взятая из соответствующего литературного источника, была разбита на обучающую выборку с N_t соединениями и контрольную выборку с N_v соединениями. Для обеих выборок были рассчитаны молекулярные дескрипторы. Для построения нейросетевой модели использовалась искусственная нейронная сеть с обратным распространением ошибок, содержащая n_i входных, n_o выходных, n_h скрытых и 2 псевдонеурона смещения (bias). Каждый входной нейрон соответствовал одному из рассчитанных молекулярных дескрипторов, каждый выходной – прогнозируемому свойству, а число скрытых нейронов бралось таким, чтобы максимально уменьшить «переучивание» при сохранении точности прогноза. Обучение велось при помощи алгоритма «обобщенного дельта-правила», скорость обучения была взята $\eta = 0.8$, момент $\mu = 0.9$, а критерием завершения обучения являлось уменьшение изменения шкалированной суммарной среднеквадратичной ошибки для обучающей выборки после очередной эпохи ниже порогового значения 0.0001. Качество работы искусственной нейронной сети определялось по среднеквадратичной ошибке прогнозирования значений свойства на обучающей выборке s_t , по коэффициенту корреляции между прогнозируемыми и экспериментальными значениями свойства на обучающей выборке R и среднеквадратичной ошибке прогноза на контрольной выборке s_v . При проведении данного исследования была использована первая версия разработанной нами программы-эмулятора искусственных нейронных сетей NASA (см. раздел 8.1).

Для проведения сравнения нейросетевых моделей с линейно-регрессионными, те же самые выборки при тех же наборах рассчитанных молекулярных дескрипторов и тех же разбивках выборок на обучающие и контрольные были обработаны на программном комплексе «ЭММА» (см. раздел 8.1), предназначенном для проведения QSPR/QSAR-исследований при помощи пошаговой процедуры множественного линейно-регрессионного анализа количественных зависимостей между свойствами химических соединений и описы-

вающими химические структуры молекулярными дескрипторами. В процессе проведения исследований из множества регрессионных моделей отбиралась одна, дающая наилучший прогноз на обучающей и контрольной выборках, и для нее вычислялись среднеквадратичная ошибка прогноза на обучающей выборке s_t , коэффициент корреляции между прогнозируемыми и экспериментальными значениями свойства на обучающей выборке R и среднеквадратичная ошибка прогноза на контрольной выборке s_v .

Результаты компьютерных экспериментов приведены в Табл. 9. В компьютерных экспериментах 1-6 прогнозировалось по одному свойству (один выходной нейрон), тогда как в компьютерных экспериментах 7 и 8 одновременно прогнозировались шесть различных свойств (шесть выходных нейронов).

Табл. 9. Результаты нейросетевого и линейно-регрессионного моделирования физико-химических свойств углеводородов (обозначения см. в тексте)

№	Выборка			МД	Архитектура нейронной сети			Статистические показатели нейросетевых моделей			Статистические показатели линейно-регрессионных моделей		
	Свойство	N_t	N_v		n_i	n_h	n_o	s_t	R	s_v	s_t	R	s_v
1	Температура кипения алканов, 1 атм., °С	159	18	ТИ	12	2	1	4.08	0.999	2.33	9.44	0.996	10.9
2	-- // --	159	16	ФД	18	2	1	4.74	0.999	2.18	23.0	0.979	22.5
3	Температура плавления алканов, °С	81	9	ТИ	12	2	1	16.2	0.976	13.8	29.4	0.924	28.5
4	-- // --	81	9	ФД	18	2	1	16.0	0.977	16.8	32.9	0.902	31.8
5	Октановое число алканов, алкенов, циклоалканов	138	15	ТИ	12	2	1	10.9	0.841	12.1	13.2	0.761	17.0
6	-- // --	138	15	ФД	34	2	1	5.97	0.954	4.37	10.6	0.858	10.4
7	Молярный объем алканов, см ³ /моль	63	6	ТИ	12	5	6	0.84	0.999	0.89	0.45	1.000	0.64
7	Молярная рефракция	63	6	ТИ	12	5	6	0.15	1.000	0.18	0.04	1.000	0.09

	алканов, см ³ /моль												
7	Теплота испарения алканов, кДж/моль	63	6	ТИ	12	5	6	0.44	0.994	0.51	0.27	0.999	0.21
7	Критиче- ская темпе- ратура ал- канов, °С	63	6	ТИ	12	5	6	3.80	0.994	3.94	5.25	0.996	2.82
7	Критиче- ское дав- ление ал- канов, атм.	63	6	ТИ	12	5	6	0.46	0.984	0.39	0.68	0.988	0.39
7	Поверхно- стное на- тяжение алканов, дин/см	63	6	ТИ	12	5	6	0.18	0.996	0.28	0.28	0.990	0.29
8	Молярный объем ал- канов, см ³ /моль	63	6	ФД	14	6	6	0.88	0.999	1.10	0.62	1.000	0.42
8	Молярная рефракция алканов, см ³ /моль	63	6	ФД	14	6	6	0.20	0.999	0.18	0.04	1.000	0.09
8	Теплота испарения алканов, кДж/моль	63	6	ФД	14	6	6	0.44	0.996	0.56	0.18	1.000	0.07
8	Критиче- ская тем- пература алканов, °С	63	6	ФД	14	6	6	3.37	0.995	3.58	7.52	0.993	4.96
8	Критиче- ское дав- ление ал- канов, атм.	63	6	ФД	14	6	6	0.44	0.986	0.23	0.79	0.986	0.40
8	Поверхно- стное на- тяжение алканов, дин/см	63	6	ФД	14	6	6	0.17	0.996	0.17	0.31	0.989	0.23

Из сравнительного анализа данных в таблице можно сделать следующие выводы.

1) Для углеводородов температура кипения, плавления, октановое число, критическая температура и поверхностное натяжение прогнозируются существ-

венно лучше при использовании искусственных нейронных сетей по сравнению с линейным регрессионным анализом, что свидетельствует о нелинейном характере зависимости этих свойств от рассматриваемых дескрипторов.

2) При прогнозировании молярного объема, молярной рефракции и теплоты испарения акланов предпочтительно использовать линейный регрессионный анализ по сравнению с искусственными нейронными сетями, что свидетельствует о практически строгой линейной зависимости этих свойств от рассматриваемых дескрипторов.

3) В большинстве случаев (для 7 свойств из 9) использование фрагментных дескрипторов приводит к построению моделей с лучшей прогнозирующей способностью по сравнению с применением топологических индексов.

4) Сочетание искусственных нейронных сетей с фрагментными дескрипторами чаще всего приводит к построению моделей с наилучшей прогнозирующей способностью.

Именно этот последний вывод (в то время совершенно неожиданный и противоречащий бытовавшим тогда убеждениям о преимуществах использования топологических индексов и аппарата множественной линейной регрессии) и послужил отправным толчком для проведения большой серии разноплановых исследований, которые и легли в основу данной диссертационной работы.

Сейчас, оценивая рассмотренную выше работу, можно сказать, что она во многих отношениях явилась пионерной.

1) Она явилась одной из первых работ, в которых аппарат искусственных нейронных сетей был применен для прогнозирования физико-химических свойств органических соединений, и однозначно первой работой, где это было сделано систематически. В настоящее время аппарат нейросетей является неоспоримым лидером в прогнозировании многочисленных свойств органических соединений.

2) В ней впервые применено сочетание аппарата искусственных нейронных сетей и фрагментных дескрипторов для прогнозирования свойств органических соединений. К настоящему времени это сочетание продемонстрировало

свое преимущество в прогнозировании многочисленных свойств органических соединений.

3) В ней впервые было успешно применено многозадачное обучение, позволяющее одновременно осуществлять прогноз нескольких свойств в рамках одной модели. Следует заметить, что вообще в теории машинного обучения первые работы по многозадачному обучению, предвосхитившие появление ныне популярного целого направления в вычислительной математике, были опубликованы в том же 1993 г., т.е. не раньше данной работы. В настоящее время многозадачное обучение является одним из перспективных направления развития работ по прогнозированию свойств органических соединений (см. подраздел 7.4.2).

6.2. Прогнозирование физико-химических свойств органических соединений с использованием фрагментных дескрипторов и нейросетевых моделей

В разделе 5.2 мы привели ряд примеров прогнозирования физико-химических свойств органических соединений с использованием фрагментных дескрипторов и стандартного аппарата пошаговой множественной линейной регрессии. В задачу следующего этапа рассмотренных там исследований входило установление того, приводит ли замена линейно-регрессионного анализа на нейросетевой при том же наборе дескрипторов и разбивке базы данных к повышению прогнозирующей способности полученных моделей. Для построения нейросетевой модели была использована трехслойная однонаправленная нейронная сеть, реализованная в рамках программы NASAWIN (см. раздел 8.2). Число нейронов входного слоя соответствовало числу дескрипторов, а внутренний слой было помещено 2 нейрона, а выходной слой состоял из одного нейрона, соответствующего прогнозируемому свойству. В качестве алгоритма обучения было взято обобщенное “дельта-правило” (см. пункт 1.2.4.4), параметр скорости обучения 0.25, значение параметра “момента” обучения 0.9. Процесс обучения был остановлен по достижению наименьшей ошибки прогноза на контрольной выборке. Трехвыборочный подход не был применен, по-

сколькx явление «переучивания» было выражено очень слабо либо вообще не наблюдалось. В Табл. 10 приведено сравнение точности прогноза для построенных линейно-регрессионных моделей.

Табл. 10. Точность прогноза для линейно-регрессионных и нейросетевых моделей

Свойство	Подраздел	MAE _{пред} или RMSE _{пред} * для линейно-регрессионной модели	MAE _{пред} или RMSE _{пред} * для нейросетевой модели
Магнитная восприимчивость. $\times 10^{-6}$ единиц	5.2.3	7.02	<u>6.25</u>
Энтальпия парообразования, ккал/моль	5.2.4	<u>1.57</u>	1.77
Энтальпия сублимации, ккал/моль	5.2.5	2.16	<u>1.66</u>
Температура вспышки, °C	5.2.6	15.8*	<u>14.6*</u>

Как видно из Табл. 10, для трех из четырех свойств (т.е. для магнитной восприимчивости, энтальпии сублимации и температуры вспышки) применение нейронных сетей приводит к уменьшению ошибок прогноза. Что же касается энтальпии парообразования, то можно предположить, что более высокая прогнозирующая способность линейно-регрессионной модели обусловлена строгим аддитивным характером этого свойства. Таким образом, в большинстве случаев применение нейронных сетей вместо аппарата множественной линейной регрессии приводит к повышению прогнозирующей способности количественных моделей «структура-свойство».

6.3. Моделирование физических свойств органических жидкостей в рамках процедуры трехвыборочного скользящего контроля

6.3.1. Общая методология моделирования

Для демонстрации эффективности использования фрагментных дескрипторов в сочетании с аппаратом искусственных нейронных сетей при прогнозировании физических свойств самых разнообразных органических соединений было проведено как линейно-регрессионное, так и нейросетевое моделирование вязкости, плотности (для жидких веществ), давления насыщенных паров и температуры кипения на основе единой методики, которую можно назвать процедурой трехвыборочного скользящего контроля. Его разработка явилась дальнейшим развитием трехвыборочного подхода (см. подраздел 4.1.3). Основная идея метода – использование процедуры скользящего контроля и ансамбля нейросетевых моделей вместо единичной модели для того, чтобы сделать прогноз и оценку его качества более обоснованным и независимым от конкретной разбивки базы на три выборки – обучающую, внутреннюю и внешнюю контрольные. Эта процедура была нами применена только в данном цикле работ и в дальнейшем была заменена на более эффективную (вследствие генерации большего разнообразия нейросетевых моделей) процедуру двойного скользящего контроля (см. подраздел 4.1.4).

Во всех случаях исследования в рамках этого подхода проводилось по следующей схеме. На первом этапе для всех соединений из базы данных, включающей информацию о структурах химических соединений и их свойствах, проводился расчет фрагментных дескрипторов (чисел вхождений структурных фрагментов в химическую структуру), причем максимальный размер фрагментов варьировался от 1 до 10 атомов. При расчете исключались фрагменты, встречающиеся в выборке менее, чем в 1 % соединений, а также статистически идентичные. Далее для каждого дескриптора были рассчитаны нелинейные модификации (квадрат (D_i^2), квадратный корень ($D_i^{1/2}$), десятичный логарифм ($\lg(D_i)$), отношение значения дескриптора к числу неводородных атомов в молекуле (D_i/n_a)).

Следует отметить, что использование, наряду с фрагментными дескрипторами, их нелинейных модификаций вполне оправдано. Для исследования этого

вопроса нами предварительно был проведен сравнительный анализ как линейно-регрессионных так и нейросетевых моделей (методика их построения рассмотрена ниже) для нескольких наборов дескрипторов, различающихся максимальным числом атомов во фрагментах (1 и 2) и наличием/отсутствием нелинейных модификаций дескрипторов. Анализ полученных результатов показал, что статистические характеристики построенных моделей с дескрипторами и их нелинейными модификациями заметно лучше аналогичных характеристик для моделей, построенных без включения нелинейных модификаций дескрипторов. Этот результат кажется вполне логичным для линейно-регрессионных моделей, поскольку подобные модификации в определенной мере позволяют учесть нелинейности зависимости «структура-свойство», но может показаться непонятным в случае искусственных нейронных сетей, которые сами по себе способны моделировать нелинейные зависимости. Одной из возможных причин этого явления может служить тот факт, что для предварительного отбора дескрипторов используется пошаговая процедура построения линейно-регрессионных зависимостей, и привнесение в нее нелинейности при помощи приведенных модификаций дескрипторов делает отбор дескрипторов для нелинейного метода, каковым являются искусственные нейронные сети, более обоснованным. Интересно отметить, что в литературе отсутствует описание этого явления, и потому оно заслуживает дальнейшего исследования.

Далее после проведения нелинейных модификаций часть дескрипторов отбрасывалась таким образом, чтобы все парные коэффициенты корреляции r между оставшимися дескрипторами не превышали 0.97. После этого база данных разбивалась на три выборки – обучающую (80% соединений), внутреннюю контрольную (10% соединений) и внешнюю контрольную (10% соединений). Разбивка проводилась 10 разными способами таким образом, чтобы каждое соединение из базы данных присутствовало по одному разу в каждой из двух контрольных выборок. Затем для каждого первоначального набора дескрипторов (различающихся максимальным размером фрагментов) и каждой разбивки базы данных проводился отбор дескрипторов при помощи процедуры БПМЛР (см. подраздел 4.1.5). После этого из 10 первоначальных наборов дескрипторов от-

бирался оптимальный в соответствии со средней ошибкой прогноза на внутренних контрольных выборках и отобранные из него наборы дескрипторов были далее использованы в исследовании при помощи многослойных нейронных сетей с обратным распространением ошибок.

На следующем этапе для каждой разбивки базы данных строилось по 5 нейросетевых моделей для каждого числа скрытых нейронов, которое варьировалось от 2 до 8. Обучение проводилось при помощи «обобщенного дельта-правила» (параметр скорости 0,25, момент 0,9) до достижения минимальной среднеквадратичной ошибки на внутренней контрольной выборке. После этого определялось оптимальное число скрытых нейронов, обеспечивающее наименьшие ошибки на внутренних контрольных выборках, и результаты прогнозирования полученных моделей для всех соединений усреднялись. В результате для каждого соединения были получены результаты прогноза ансамблевой модели, для оценки качества которой вычислялись следующие статистические показатели: множественный коэффициент корреляции R , а также среднеквадратичные значения ошибок для обучающей ($RMSE_t$), внутренней контрольной ($RMSE_v$) и внешней контрольной ($RMSE_p$) выборок. Для оценки эффекта перехода к ансамблевому моделированию проводился также расчет средних значений этих показателей, вычисленных для каждой из моделей до усреднения.

6.3.2. Моделирование вязкости органических соединений

При моделировании вязкости органических соединений была использована база данных, взятая из работы [410]. Из выборки, приведенной в работе [410], были исключены два соединения (266 и 267), для которых авторами ошибочно приведены одинаковые названия, но разные значения вязкости. Моделируемое свойство для данной базы представлено в виде десятичного логарифма от значения вязкости органического соединения, измеренного в единицах Па·с. При построении моделей вся база данных, состоящая из 367 органических соединений различных классов ((367 структур – линейные, разветвленные и циклические (моно- и бициклические) алканы, алкены и алкины, арены, спирты,

простые и сложные эфиры, кетоны, альдегиды, карбоновые кислоты, нитрилы, имины, амины, амиды, галоген- и серосодержащие соединения, нитро-соединения)), разбивалась 10-ю разными способами на три выборки: обучающую (293 соединения), контрольную (37 соединения) и выборку для оценки прогнозирующей способности (37 соединения). Согласно описанной выше схеме, с помощью процедуры БПМЛР из рассчитанного множества дескрипторов проводился их отбор для 10 различных вариантов разбивки базы данных. В процессе построения каждой линейной регрессионной модели проводилось последовательное включение дескрипторов до достижения наименьшей средне-квадратической ошибки на внутренней контрольной выборке.

Табл. 11. Усредненные статистические характеристики линейно-регрессионных моделей при варьировании максимального размера дескрипторов

Количество атомов	Общее количество дескрипторов	Среднее количество отобранных дескрипторов	МЛР			
			$R_{обуч}$	$RMS_{обу}$ <i>с</i>	$RMS_{конт}$ <i>р</i>	$RMS_{пре}$ <i>д</i>
1	146	38±20	0,9204	0,2172	0,2366	0,2407
2	531	53±12	0,9740	0,1260	0,1857	0,1853
3	1757	46±16	0,9794	0,1113	0,1950	0,2119
4	1974	42±22	0,9593	0,1336	0,2079	0,2341
5	2183	34±21	0,9531	0,1470	0,2113	0,2330
6	2413	36±21	0,9681	0,1307	0,1960	0,2207
7	2566	33±19	0,9662	0,1302	0,2088	0,2392
8	2649	35±22	0,9656	0,1337	0,2075	0,2305
9	2703	33±20	0,9652	0,1348	0,2077	0,2322
10	2732	35±22	0,9658	0,1330	0,2081	0,2316
11	2945	35±22	0,9657	0,1331	0,2044	0,2297
12	2759	35±22	0,9657	0,1331	0,2044	0,2297
13	2770	35±22	0,9657	0,1331	0,2044	0,2297

МЛР – множественная линейная регрессия; $R_{ср}$ – коэффициент корреляции; $RMS_{обуч}$, $RMS_{контр}$, $RMS_{предск}$ - среднеквадратичная ошибка на обучающей, контрольной выборках и на выборке для оценки предсказательной способности, соответственно.

Результаты полученных линейно-регрессионных моделей для 13 наборов дескрипторов с различным максимальным размером фрагментов (130 моделей) представлены в Табл. 11 и на Рис. 43. Как видно из Рис. 43, минимумы для обучающей и контрольной выборок, а также для выборки для оценки прогнозирующей способности приходится на множество дескрипторов с максимальным числом атомов, равным 2, 3 и 6, соответственно. Однако, при построении нейросетевых моделей наилучшие статистические характеристики были получены для множества дескрипторов с максимальным размером фрагментов, равным трем. Выбор оптимального набора дескрипторов проводился по значению среднеквадратичной ошибки для внутренней контрольной выборки, поскольку некорректно ориентироваться как на минимум для обучающей выборки (во избежание построения переопределенных моделей), так и на внешнюю контрольную выборку (поскольку данные для этой выборки следует использовать только для оценки предсказательной способности, а не для построения и отбора моделей).

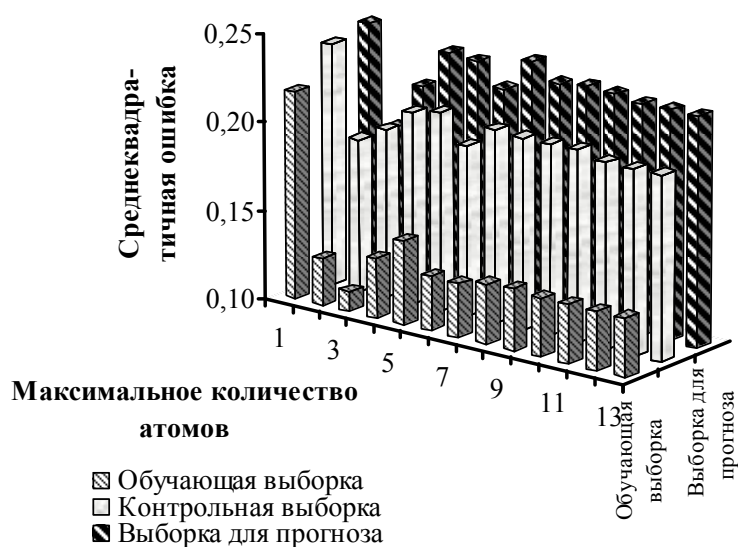


Рис. 43. Гистограмма зависимости среднеквадратичной ошибки от максимального размера фрагментных дескрипторов

Само по себе наличие оптимального значения максимального размера, обеспечивающего наилучшую прогнозирующую способность моделей, для генерируемых фрагментов не является очевидным, и поэтому заслуживает от-

дельного рассмотрения. Связано это, очевидно, с тем, что при увеличении размеров фрагментов число их типов, а, следовательно, и число фрагментных дескрипторов резко возрастает. В то же время, при прочих равных условиях (т.е. при одинаковой ошибке на обучающей выборке и одинаковом числе отобранных дескрипторов), как следует из целого ряда математических теорий (см. ниже), прогнозирующая способность статистической модели ухудшается с увеличением первоначального числа дескрипторов, из которого производится отбор. Действительно, согласно статистической теории прогнозирования Вапника-Червоненкиса [411], минимальный размер выборки соединений, необходимый для достижения заданного качества прогнозирования зависит как от числа отобранных дескрипторов, так и от первоначального числа дескрипторов, причем в последнем случае для бинарных дескрипторов (т.н. признаков) показан логарифмический характер зависимости минимального размера выборки от логарифма числа первоначальных дескрипторов. Следовательно, при фиксированном размере выборки качество модели ухудшается при увеличении первоначального числа дескрипторов. Таким образом, эффективное число дескрипторов в статистической модели (т.н. размерность Вапника-Червоненкиса) в общем случае не равно числу отобранных дескрипторов и зависит также от первоначального числа дескрипторов, из которого производился их отбор. К аналогичным выводам приходит и теория индуктивных выводов [412, 413]. Согласно Риссанену, ожидаемая ошибка статистической модели на данных, не входящих в обучающую выборку, определяется степенью сжатия информации с помощью этой модели. Чем меньше суммарная длина описания данных с помощью модели и описания самой модели, тем ниже ошибка предсказаний при помощи этой модели. Длина описания модели M равна количеству информации, необходимой для выбора этой модели из множества с априорным распределением вероятностей $P(M)$, что равно величине $-\log P(M)$. Ясно, что чем из большего первоначального числа отбираются дескрипторы, тем меньше априорная вероятность получаемой модели, и, следовательно, тем больше длина описания модели и, следовательно, ожидаемая ошибка прогноза.

При анализе дескрипторов, участвующих в построении всех 350 моделей, оказалось, что наиболее важными являются: общее число неводородных атомов в молекуле (n_a), отношение количество метильных групп, связанных с углеродным атомом, к числу неводородных атомов ($n(\text{CH}_3\text{-C})/n_a$), а также отношение числа пропильных групп к числу неводородных атомов ($n(\text{CH}_3\text{-CH}_2\text{-CH}_2)/n_a$). Кроме того, следует отметить значимость таких дескрипторов, как количество аминогрупп ($n(-\text{NH}_2)$), атомов азота при двойной связи ($n(=\text{N})/n_a$), цепочек, содержащих гидроксильные группы ($n(\text{C}_{\text{sp}^3}\text{-O-C}_{\text{sp}^3}\text{-OH})$ и $n(\text{C}_{\text{sp}^3}\text{-C}_{\text{sp}^3}\text{-C}_{\text{sp}^3}\text{-OH})$), атомов галогенов ($n(\text{F-})$, $n(\text{C-I})$), количество амидных групп ($n(\text{N-C=O})$). Можно предположить, что первые три дескриптора описывают ван-дер-ваальсово взаимодействие между молекулами, а остальные – электростатическое (включая образование водородных связей).

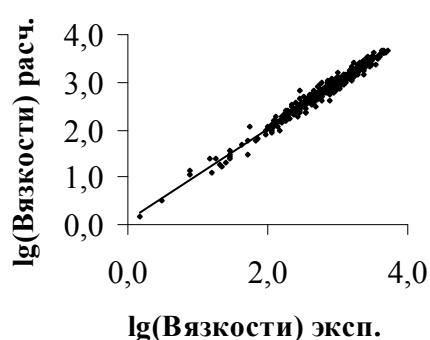
После построения ряда нейросетевых моделей (350 моделей) с варьированием числа нейронов в скрытом слое от 2 до 8 было выбрано оптимальное число скрытых нейронов, равное семи (Табл. 12), хотя практически при любом количестве скрытых нейронов статистические параметры модели были приблизительно одинаковыми. В Табл. 13 представлены полученные статистические параметры моделирования. Корреляция усредненных по всему массиву моделей расчетных данных для всех выборок с экспериментальными значениями представлена на Рис. 44.

Табл. 12. Зависимость значения RMSE от числа нейронов в скрытом слое

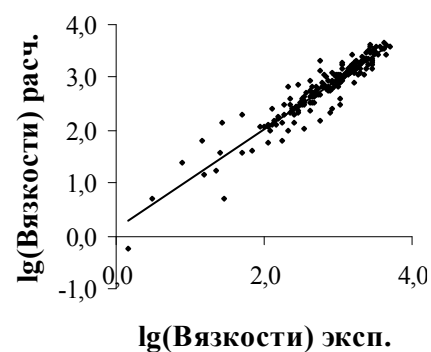
Количество нейронов в скрытом слое	$RMSE_{\text{обуч}}$	$RMSE_{\text{контр}}$	$RMSE_{\text{предск}}$
2	0,110	0,193	0,226
3	0,106	0,191	0,222
4	0,108	0,192	0,220
5	0,106	0,192	0,219
6	0,105	0,191	0,219
7	0,105	0,189	0,219
8	0,105	0,191	0,220

Табл. 13. Статистические показатели полученных моделей для вязкости органических соединений

Статистические показатели моделей Название этапа исследования	R	$RMSE_t$	$RMSE_v$	$RMSE_p$
Линейно-регрессионные модели	0,9794	0,111	0,195	0,212
Средние значения показателей по всем индивидуальным нейросетевым моделям	0,9815	0,105	0,189	0,219
Показатели ансамблевой модели, усредняющей прогнозы индивидуальных нейросетевых моделей	0,9904	0,078	0,177	0,208



(а)



(б)

Рис. 44. Результаты нейросетевого моделирования вязкости: (а) корреляция экспериментальных значений с результатами прогноза, полученными путем усреднения по всем моделям, при построении которых данные соединения входили в обучающие выборки; (б) корреляция экспериментальных значений с результатами прогноза, полученными путем усреднения по всем моделям, при построении которых данные соединения входили во внешние контрольные выборки

Из Табл. 13 видно, что прогнозирующая способность нейросетевых моделей (которую наиболее корректно оценивать по среднеквадратичным ошибкам для внешних контрольных выборок, превосходит аналогичные показатели линейных (они являются линейными по отношению к регрессионным коэффициентам, но нелинейными по отношению к значениям дескрипторов) регрессионных моделей. Кроме того, построенные в ходе данной работы модели для предсказания вязкости жидких органических соединений существенно превосходят по всем показателям наилучшие из ранее опубликованных моделей (см. [410,

414]). Следует также обратить внимание на заметное различие средних значений статистических показателей по ансамблю нейросетевых моделей и статистических показателей ансамблевой модели, усредняющей прогнозы, даваемые этими моделями. То, что вторые существенно лучше первых, свидетельствует о больших преимуществах использования ансамблей нейросетевых моделей по сравнению с индивидуальными моделями.

6.3.3. Моделирование плотности жидких органических соединений

В качестве источника для формирования использованной в данной работе базы был взят электронный каталог органических соединений фирмы Fluka [415], содержащий 16793 записи. База данных была автоматически из него отобрана путем задания следующих условий: 1) наличие в каталоге значения плотности для соединения; 2) чистота образца 98% и выше; 3) наличие значения показателя преломления (что означает, что данные приведены для жидкости). Сформированная таким образом база данных содержала 803 соединения, относящиеся ко следующим классам: алканы, алкины, арены, аллены, спирты, простые и сложные эфиры, нитро-соединения, альдегиды, карбоновые кислоты, кетоны, нитрилы, амины, имины, амиды, гетероциклические соединения, моно-, би- и трициклические структуры.

При обработке базы данных была применена рассмотренная выше (см. подраздел 6.3.1) методика. Каждый раз база разбивалась на обучающую выборку (641 соединение), контрольную выборку (81 соединение) и выборку для оценки предсказательной способности (81 соединение). Из четырех указанных выше модификаций дескрипторов было использовано три: 1) квадрат значения дескриптора; 2) квадратный корень из значения дескриптора; и 3) отношение значения дескриптора к числу неводородных атомов в молекуле.

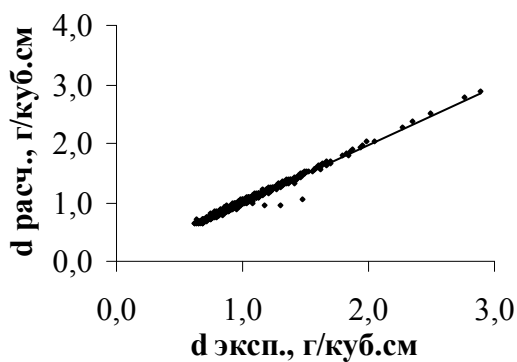
Для определения оптимального размера фрагментов нами было сгенерировано 11 наборов фрагментных дескрипторов при варьировании максимального размера фрагмента от 1 до 11 атомов. Для каждого из этих наборов дескрипторов было построено по методу БПМЛР по одной (линейной по регрессион-

ным коэффициентам, но нелинейных по дескрипторам) модели для каждой из 10 разбивок базы на три выборки. Из сравнения усредненных по разбивкам статистических показателей полученных моделей было найдено, что наименьшие ошибки на внутренних контрольных выборках получаются при использовании наборов фрагментных дескрипторов, сгенерированных при задании величины максимального размера фрагмента от 3 до 5 атомов. Именно эти 3 набора дескрипторов и были использованы в ходе дальнейшего моделирования.

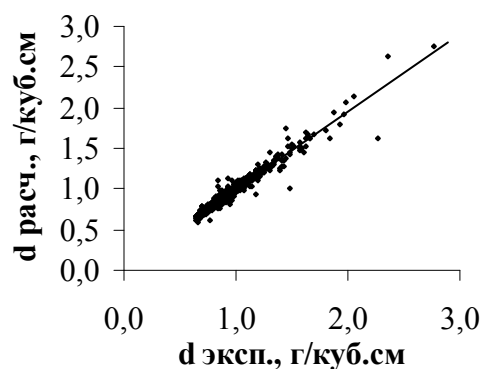
На следующем этапе было построено по 350 нейросетевых моделей (по 5 моделей для каждого количества скрытых нейронов, которое варьировалось от 2 до 8) для каждого из этих 3 наборов дескрипторов. При сравнении статистических показателей (по критерию наименьших среднеквадратичных ошибок на внутренних контрольных выборках) выявилось, что наилучшими являются модели, максимальный размер фрагментных дескрипторов в которых равен 4 атомам. Из моделей, построенных с этим набором дескрипторов, была отобрана группа из 50 моделей (5 моделей для каждой из 10 разбивок базы) с оптимальным числом скрытых нейронов, равным четырем. Следует отметить, что оптимальное число скрытых нейронов для трех типов выборок (т.е. для обучающих, внутренних и внешних контрольных выборок) различалось, поэтому этот параметр выбирался по внутренним контрольным выборкам.

При анализе наборов отобранных фрагментных дескрипторов выяснилось, что наиболее важными (степень важности определялась по количеству содержащих их моделей) являются относительное число sp^3 - и sp^2 -гибридизованных атомов углерода ($n(C_{sp^3})$ и $n(H_2C=)/n_a$), а также относительное количество различных гетероатомов (в частности, галогенов, кислорода, азота, кремния, серы и т.д.), что можно объяснить различием масс, ковалентных и ван-дер-ваальсовых радиусов у этих элементов. Разнообразные поправки описываются такими дескрипторами как количество тройных углерод-углеродных связей, и дескрипторами, характеризующими разветвленность.

Диаграммы разброса усредненных по всему массиву моделей расчетных данных для плотности жидких органических соединений по всем выборкам с экспериментальными значениями представлена на Рис. 45.



(а)



(б)

Рис. 45. Результаты моделирования плотности: (а) обучающая выборка; (б) выборка для оценки предсказательной способности

Статистические показатели полученных моделей представлены в Табл. 14. Из их сравнения легко видеть, что прогнозирующая способность нейросетевых моделей (которую можно оценить по значению среднеквадратичной ошибки для внешней контрольной выборки, $RMSE_p$) превосходит таковую для линейных регрессионных моделей (даже построенных на основе нелинейных модификаций дескрипторов). Статистические показатели наших моделей для прогнозирования плотности жидкостей для разнородных органических соединений оказались близки к наилучшей из опубликованных моделей (см. [416]), однако наши модели построены по значительно более представительной выборке.

Табл. 14. Статистические показатели полученных моделей для плотностей жидких органических соединений (в г/см^3)

Статистические показатели моделей Название этапа исследования	R	$RMSE_t$	$RMSE_c$	$RMSE_p$
Линейно-регрессионные модели	0,9897	0,036	0,055	0,067
Средние значения показателей по всем индивидуальным нейросетевым моделям	0,9911	0,034	0,052	0,061
Показатели ансамблевой модели, усредняющей прогнозы индивидуальных нейросетевых моделей	0,9943	0,018	0,036	0,043

Данные таблицы также свидетельствуют о преимуществах использования ансамблей нейросетевых моделей по сравнению с индивидуальными моделями.

6.3.4. Моделирование давления насыщенных паров

Моделирование давления насыщенных паров велось по созданной на основе опубликованных данных [417] выборке из 352 соединений (углеводороды и галогенуглеводороды), которая в процессе работы разбивалась 10 разными способами на три выборки: обучающую (279 соединений), контрольную (36 соединений) и выборку для оценки предсказательной способности (36 соединений). На первом этапе по методу БПМЛР производился отбор дескрипторов, причем, как оказалось, наилучшим моделям соответствует множество фрагментных дескрипторов с максимальным числом атомов во фрагменте, равным 6.

При моделировании давления паров среди наиболее значимых дескрипторов, присутствующих практически во всех моделях, оказались: квадрат числа углеродных атомов ($n^2(C)$); логарифм общего числа неводородных атомов ($\lg n_a$); количество атомов галогенов, связанных с углеродным атомом, входящим в состав шестичленных ароматических циклов ($n[C_{Ar}-Hal]$); количество метиленовых групп, связанных с углеродным атомом, входящим в состав шестичленных ароматических циклов ($n[C_{Ar}-CH_2]$); квадратный корень от количества атомов фтора ($\sqrt{n[F]}$); количество простых углерод-углеродных связей ($n(C-C)/n_a$); количество двухатомных углерод-углеродных фрагментов ароматических систем ($n[C_{Ar}\dot{-}C_{Ar}]$) и др. Подобный набор наиболее важных дескрипторов, по-видимому, обусловлен доминирующей ролью ван-дер-ваальсовых взаимодействий.

При построении нейросетевых моделей с различным числом скрытых нейронов (от 2 до 8) был проведен анализ зависимости статистических показателей моделей от числа скрытых нейронов. Оптимальным количеством скрытых нейронов для данной выборки оказалось три. Сводные данные, содержащие основные статистические показатели построенных моделей, приведены в

Табл. 15. Корреляция усредненных по всему ансамблю моделей расчетных данных для давления насыщенных паров с экспериментальными значениями представлена на Рис. 46.

Табл. 15. Статистические показатели полученных моделей для давления насыщенных паров органических соединений (в $\lg(\text{Па})$)

Статистические показатели моделей Название этапа исследования	R	$RMSE_t$	$RMSE_c$	$RMSE_p$
Линейно-регрессионные модели	0,9902	0,198	0,248	0,258
Средние значения показателей по всем индивидуальным нейросетевым моделям	0,9969	0,118	0,143	0,161
Показатели ансамблевой модели, усредняющей прогнозы индивидуальных нейросетевых моделей	0,9979	0,095	0,140	0,158

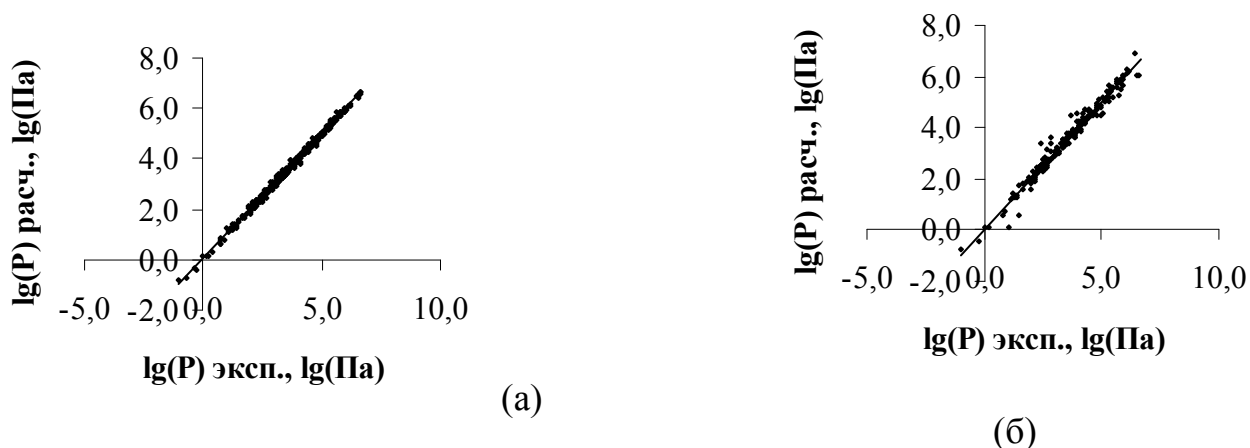


Рис. 46. Результаты моделирования давления насыщенных паров: (а) обучающая выборка; (б) внешняя контрольная выборка

Из Табл. 15 видно, что прогнозирующая способность нейросетевых моделей (которую корректно оценивать по значению $RMSE_{пред}$, т.е. по среднеквадратичной ошибке на внешней контрольной выборке, превосходит аналогичные показатели линейных регрессионных моделей (даже содержащих нелинейные модификации дескрипторов). Точность предсказания давления насыщенных паров в построенных моделях оказалась сравнимой с моделью Голла-Джурса [417] и существенно выше других опубликованных моделей (см. [416]).

6.3.5. Моделирование температуры кипения разнородных органических соединений

Температура кипения моделировалась по выборке, содержащей разнородные органические соединения. В качестве источника данных был взят электронный каталог органических соединений фирмы Fluka [415], содержащий 16 793 записей. База данных «структура-свойство» создавалась путем автоматизированного отбора записей из электронного каталога со следующими условиями: 1) наличие в каталоге значения температуры кипения для данного соединения при атмосферном давлении; 2) чистота образца 99% и выше.

В процессе построения моделей вся база данных разбивалась 10-ю разными способами на три выборки: 1) обучающую (409 соединений); 2) контрольную (50 соединений); и 3) выборку для оценки прогнозирующей способности (50 соединений). Согласно описанной выше схеме (см. подраздел 6.3.1), для базы данных был проведен расчет фрагментных дескрипторов с варьированием максимального размера фрагментов от 1 до 10 атомов. Далее для каждого из полученных дескрипторов были рассчитаны 4 нелинейные модификации. После этого, для 10 различных вариантов разбивки базы данных из первоначального набора с помощью процедуры быстрой пошаговой множественной линейной регрессии (БПМЛР) был проведен отбор дескрипторов. Усредненные результаты полученных линейно-регрессионных моделей (с нелинейными модификациями дескрипторов) для 10 наборов дескрипторов с переменным максимальным размером фрагментов (всего 100 моделей) представлены в Табл. 16.

Табл. 16. Усредненные статистические показатели линейно-регрессионных моделей для прогнозирования температуры кипения органических соединений при варьировании максимального размера фрагментов

Максимальное количество атомов во фрагментах	Общее количество фрагментных дескрипторов	Среднее число отобранных фрагментных дескрипторов	БПМЛР			
			R_t	$RMSE_t$	$RMSE_c$	$RMSE_p$
1	138	33±19	0,9642	17,7	18,6	20,5
2	555	46±13	0,9814	12,9	16,7	18,6
3	1744	46±20	0,9821	12,2	17,9	20,4
4	2104	44±17	0,9838	11,8	18,2	19,5
5	2327	42±14	0,9835	11,9	18,3	19,9
6	2561	35±15	0,9801	13,1	18,3	20,4
7	2706	37±16	0,9812	12,7	18,4	20,0
8	2781	38±16	0,9827	12,1	17,3	20,0
9	2821	37±17	0,9811	12,7	18,5	19,6
10	2851	37±17	0,9811	12,7	18,6	19,6

БПМЛР – быстрая пошаговая множественная линейная регрессия; $R_{обуч}$ – множественный коэффициент корреляции (квадратный корень от коэффициента детерминации) на обучающей выборке; $RMSE_{обуч}$, $RMSE_{контр}$, $RMSE_{пред}$ – среднеквадратичная ошибка в °С на обучающей, контрольной выборке и для выборки для оценки предсказательной способности, соответственно.

Как видно из Табл. 16, минимальные значения среднеквадратичных ошибок для обучающей и двух контрольной выборок приходятся на наборы фрагментных дескрипторов с максимальным числом атомов, равным 2, 4 и 5, соответственно. В дальнейшем по ходу данной работы для построения нейросетевых моделей для прогнозирования температуры кипения органических соединений использовался только набор фрагментных дескрипторов с максимальным числом атомов, равным двум, поскольку при этом предсказательная способность модели, оцененная по среднеквадратичной ошибке на внутренней контрольной выборке, оказывается наилучшей.

По частотам вхождения в отбираемые при построении линейно-регрессионных моделей дескрипторов можно сделать вывод об их относительной значимости. В соответствии с этим критерием, при моделировании температуры кипения разнородных органических соединений наиболее весомыми являются вклады: метильных групп, связанных с любыми неводородными атомами ($n[\text{H}_3\text{C}-\bullet]/n_a$ и $n[\text{H}_3\text{C}-\bullet]$); sp^2 -гибридизованных атомов углерода ($n[\text{C}_{sp2}]/n_a$); фрагментов ароматических систем ($n^2[\text{C}_{Ar}\div\text{C}_{Ar}]$); произвольных неводородных атомов ($\log\{n[\bullet]\}$ и $n^2[\bullet]$). Значительный вклад также вносят группы, содержащие полярные атомы и связи, в частности: sp -, sp^2 - и sp^3 -гибридизованные атомы азота ($n(\text{N})$, $\sqrt{n[=\text{C}-\text{N}]}$, $n[=\text{N}-]$ и $n[\text{C}_{sp2}-\text{N}]/n_a$); гидроксильные группы, связанные с атомом углерода ($n[\text{C}-\text{OH}]$, $n^2[\text{HC}_{Heterocycle}-\text{OH}]$); атомы кислорода при двойной связи ($n[\text{O}=\bullet]/n_a$); атомы галогенов в различном структурном контексте ($n[\text{C}_{sp3}-\text{I}]/n_a$, $n[\text{H}_2\text{C}-\text{Hal}]$, $\sqrt{n[\text{C}-\text{F}]}$, $n[\text{Br}]$); атомы бора, кремния и серы ($n^2[\text{B}-\bullet]$, $n[\text{C}_{sp2}-\text{N}]/n_a$, $n[\text{Hal}-\text{Si}]$ и $n[\text{C}-\text{S}]/n_a$).

После построения ряда нейросетевых моделей (350 моделей) с варьированием числа скрытых нейронов было выбрано оптимальное число скрытых нейронов, равное двум (как обеспечивающее наименьшие ошибки на внутренних контрольных выборках). В Табл. 17 приведены статистические показатели построенных моделей.

Табл. 17. Статистические показатели полученных моделей для температуры кипения разнородных органических соединений (ошибки приведены в °C)

Статистические показатели моделей	R	$RMSE_t$	$RMSE_c$	$RMSE_p$
Название этапа исследования				
Линейно-регрессионные модели	0,9814	12,9	16,7	18,6
Средние значения показателей по всем индивидуальным нейросетевым моделям	0,9869	11,0	16,1	17,2
Показатели ансамблевой модели, усредняющей прогнозы индивидуальных нейросетевых моделей	0,9911	9,1	16,1	16,9

Как видно из Табл. 16 и Табл. 17, прогнозирующая способность построенных нейросетевых моделей заметно выше линейно-регрессионных. Кроме того, следует обратить внимание на тот факт, что (как и во всех других случаях, см. подразделы 6.3.2, 6.3.3 и 6.3.4) статистические показатели ансамблевой модели, усредняющей прогнозы по нейросетевому ансамблю, всегда заметно средних статистических показателей индивидуальных нейросетевых моделей в ансамбле. Это еще раз подтверждает известное из теории и практики машинного обучения утверждение о существенных преимуществах использования ансамблей нейросетевых моделей по сравнению с индивидуальными моделями. По-видимому, два основных фактора вносят вклад в это явление. Во-первых, усреднение по моделям, получаемым при разных разбиениях базы данных позволяет эффективно использовать для обучения информацию из внутренних контрольных выборок, что эквивалентно увеличению эффективного размера обучающих выборок. Во-вторых, уменьшается вклад дисперсии в среднеквадратичную ошибку прогнозирования, поскольку дисперсия среднего нескольких случайных независимых переменных всегда ниже средней дисперсии каждой из этих переменных (т.е. происходит подавление «шума» при усреднении).

Как известно, статистические показатели отдельно взятой модели при небольшом размере базы данных не может служить основой для вывода о качестве методики моделирования и иметь какую-либо статистическую значимость при отсутствии корректного скользящего контроля. Так, например, одна из полученных для данной базы данных статистических моделей характеризовалась следующими статистическими показателями: среднеквадратичная ошибка для обучающей выборки $RMSE_t$ равна 5.6°C , для внутренней контрольной выборки $RMSE_v = 4.4^{\circ}\text{C}$, а для внешней контрольной выборки $RMSE_p = 5.0^{\circ}\text{C}$, что в несколько раз ниже усредненных показателей. Статистические показатели подобных индивидуальных моделей могут не характеризовать их истинную прогнозирующую способность, особенно когда в процессе их построения производится отбор дескрипторов. Хотя в отдельных публикациях, как, например [418], встречаются подобные результаты, ориентироваться на них нецелесообразно. Поэтому усредненные по множеству моделей результаты являются статистиче-

ски более достоверными, чем показатели индивидуальных моделей. При этом важно, чтобы усреднение проводилось таким образом, чтобы информация об экспериментальном значении прогнозируемого свойства для каждого из химических соединений никаким образом не участвовала ни в построении, ни в отборе моделей, по которым проводился для него прогноз.

Корреляция между спрогнозированной по ансамблю моделей (в ансамбль моделей для прогнозирования свойства данного соединения включались только те модели, при построении которых оно не участвовало в составе обучающей либо контрольной выборки) температурой кипения органических соединений и экспериментальным значением этого свойства показана в виде диаграммы разброса точек на Рис. 47.

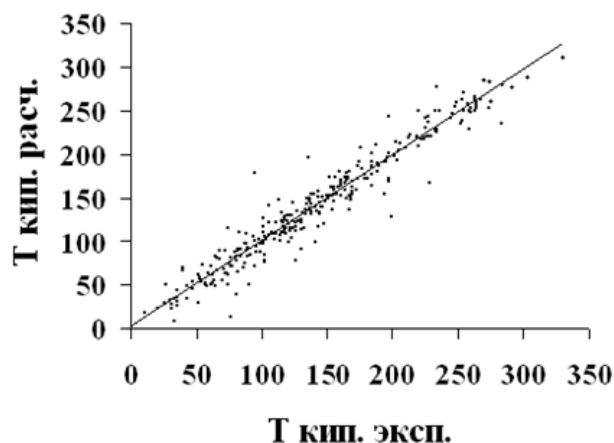


Рис. 47. Корреляция расчетных и экспериментальных данных по температурам кипения (в °С)

После получения вышеизложенных результатов нами было проведено их сравнение с литературными данными. При этом было выделено два типа работ: 1) работы, в которых база данных включала не более 100 соединений, и 2) работы, представляющие результаты обработки более представительных выборок соединений. О разнообразии выборок в первом случае говорить не приходится из-за малого количества соединений, и моделирование в таких случаях проводилось лишь в узких сериях соединений. Работ, в которых исследования проводились с большими базами данных, оказалось всего несколько: статьи Игольфа (Egolf) и др. [419, 420] (в данных работах нейросеть с обратным распростране-

нием ошибки применяли в комбинации с физико-химическими дескрипторами), работы Холла (Hall) и др. [418, 421] (в данном случае нейросеть с обратным распространением ошибки применяли в комбинации с электротопологическими индексами), а также работа Тетте (Tatteh) и др. [386] (нейросеть функции радиального базиса сочетали с топологическими индексами). В работах [418-421] база данных содержала 298 органических соединений со специально отобранными высокоточными экспериментальными данными, однако в качестве результатов были приведены характеристики лишь одной «лучшей» модели ($RMSE_t = 5,4^{\circ}\text{C}$ и $RMSE_v = 5,9^{\circ}\text{C}$), что ставит под вопрос статистическую достоверность результатов. Как оказалось, сравнить полученные нами результаты можно лишь с данными работы [386]. Для меньшей по размеру и менее разнообразной выборки Тетте и др. были получены следующие статистические показатели: $RMSE_t = 11,4^{\circ}\text{C}$, $RMSE_v = 15,1^{\circ}\text{C}$ и $RMSE_p = 19,4^{\circ}\text{C}$, что хуже результирующих данных по моделированию, полученных нами.

6.4. Прогнозирование температуры плавления ионных жидкостей

Температура плавления является одним из наиболее сложных для прогнозирования свойств химических соединений, далеко не полный список причин чего включает: плохая воспроизводимость экспериментальных данных, зависимость от типа кристаллической упаковки, сосуществование нескольких аллотропных модификаций кристаллов, зависимость от наличия микропримесей, существование эвтектик, возможность затвердевания в аморфное либо жидкокристаллическое состояния и др. В то же время, температура плавления является важнейшей технической характеристикой, которая определяет сферу применения ионных жидкостей – материалов, широко используемых в качестве экологически-безопасных растворителей в химической промышленности. Именно поэтому это «тяжелое» свойство является удобным объектом для сравнения различных методик построения QSPR-моделей.

Мы приняли участие в совместном исследовании, проведенном несколькими группами авторов, в ходе которого широкий набор современных методов

машинного обучения (ассоциативные нейронные сети, машины опорных векторов, метод ближайших соседей, метод частичных наименьших квадратов, нейронные сети обратного распространения и множественная линейная регрессия), реализованные в нескольких программных комплексах (VCCLAB, ISIDA и NASAWIN [см. раздел 8.2]), в сочетании с разнообразными типами дескрипторов (несколько типов фрагментных дескрипторов, псевдофрагментные дескрипторы типа FRAGPROP [см. разделы 5.4 и 8.4], дескрипторы на основе электроно-топологических состояний атомов, а также все виды дескрипторов, генерируемых программой DRAGON) были впервые применены для обработки больших и структурно разнородных баз по температурам плавления ионных жидкостей [401].

В данной работе были построены QSPR-модели для четырех выборок (см. Рис. 48), включающих: (1) 126 бромидов производных пиридинов (PYR, IV и V); (2) 384 бромидов производных имидазолов и бензимидазолов (IMZ, VI и VII); (3) 207 бромидов четвертичных аммониев (QUAT, VIII); (4) 717 соединений, входящих во все вышеупомянутые наборы (FULL).

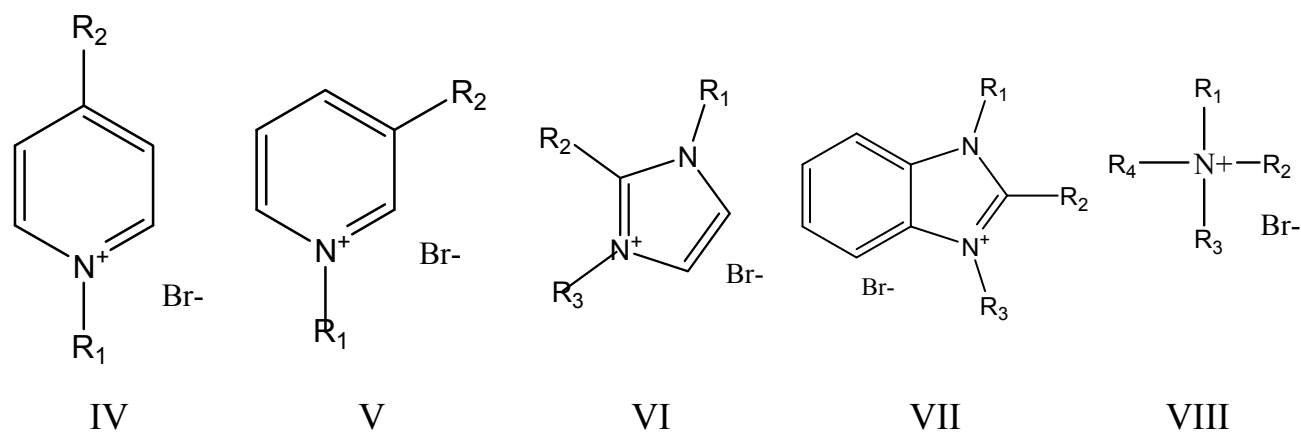


Рис. 48. Структуры ионных жидкостей

Оценка прогнозирующей способности построенных моделей проводилась при помощи процедуры 5-кратного *внешнего* (т.е. при котором информация из контрольных выборок никак не может участвовать в отборе лучших моделей) скользящего контроля по трем показателям (Q^2 , $RMSE$, MAE) для четырех выборок. В нашей части этой большой совместной работы в качестве методов машинного обучения мы использовали реализованные в программном комплексе

NASAWIN (см. раздел 8.2) нейросети обратного распространения (BPNN, см. подраздел 1.2.4), метод БПМЛР (FSMLR, см. подраздел 4.1.5) и метод частичных наименьших квадратов (PLS), а в качестве дескрипторов – набор фрагментных дескрипторов, вычисляемых блоком FRAGMENT (см. раздел 8.3), к которым был применен набор псевдофрагментных дескрипторов (см. раздел 5.4), вычисляемых блоком FRAGPROP (см. раздел 8.4). Использование псевдофрагментных дескрипторов было обусловлено тем, что как показали предварительные вычислительные эксперименты, они в данном случае значительно повышают прогнозирующую способность построенных моделей. Кроме того, следует отметить, что обучение нейросети велось на полном наборе дескрипторов (попытки использовать процедуру БПМЛР для их предварительного отбора заканчивались значительным падением прогнозирующей способности модели). Вследствие этого всякий раз проводился визуальный контроль синапсов нейросети и в случае «паралича» процедура обучения вручную останавливалась и перезапускалась заново. В Табл. 18 представлены значения средней абсолютной ошибки прогнозирования (*MAE*), вычисленной при 5-кратном внешнем скользящем контроле. Отметим, что нейросеть обратного распространения приводит к построению лучших моделей по сравнению с БПМЛР и методом частичных наименьших квадратов PLS.

Табл. 18. Значения средней абсолютной ошибки прогнозирования температуры плавления ионных жидкостей (в градусах)

	PYR	IMZ	QUAT	FULL
BPNN	<u>26.2</u>	32.4	<u>30.3</u>	<u>31.5</u>
FSMLR	34.8	36.2	36.1	33.7
PLS	32.5	<u>31.9</u>	31.8	31.9

Для сравнения QSPR-моделей, получаемых разными методами, каждой комбинации выборки и статистического показателя оценивалось среднее значение этих показателей, и каждой модели присваивался ранг “0”, если по всем трем показателям она оказывалась лучше средней, и “1” если хотя бы по одному показателю она уступала среднему. Далее ранги полученных моделей скла-

дывались, и результирующее число было использовано для сравнения методик построения QSPR-моделей. В Табл. 19 представлены полученные значения рангов для различных методов построения QSPR-моделей.

Табл. 19. Сравнение различных методов построения QSPR-моделей для прогнозирования температуры плавления ионных жидкостей

Метод	Программа	Q^2	<i>RMSE</i>	<i>MAE</i>	Итого
BPNN	NASAWIN	0	0	0	0
ASNN/E-counts	VCCLAB	0	0	0	0
ASNN/Dragon	VCCLAB	0	0	0	0
SVM/Dragon	VCCLAB	0	0	0	0
SVM/E-state	VCCLAB	0	0	1	1
SVM/E-counts	VCCLAB	0	1	1	2
ASNN/E-state	VCCLAB	1	1	1	3
PLSM	NASAWIN	2	1	2	5
MLR/Dragon	VCCLAB	2	2	2	6
MLR-CM/SMF	ISIDA	3	3	2	8
kNN/Dragon	VCCLAB	3	3	3	9
FSMLR	NASAWIN	4	4	4	12
kNN/E-state	VCCLAB	3	3	4	10
MLR/E-state	VCCLAB	3	4	4	11
kNN/E-counts	VCCLAB	4	4	4	12
MLR/E-counts	VCCLAB	4	4	4	12

Как видно из Табл. 19, нейросеть обратного распространения BPNN, реализованная в рамках программного комплекса NASAWIN, занимает первые два места наряду с ASNN/E-counts. Если учесть, что ASNN построена на основе нейросетей обратного распространения, а дескрипторы E-counts являются фрагментными, то можно сделать вывод, что именно комбинация нейросетей обратного распространения с фрагментными дескрипторами приводит к построению наилучших моделей для прогнозирования температуры плавления ионных жидкостей.

ГЛАВА 7. РАЗРАБОТКА ИНТЕГРИРОВАННЫХ ПОДХОДОВ

Данная глава посвящена рассмотрению предложенных нами подходов, которые включают разного рода интеграцию нейросетей: (а) с методами молекулярного моделирования; (б) с комбинацией дескрипторных описаний одно- и многокомпонентных химических систем и внешних условий; а также (в) между собой. Все это ведет к значительному расширению круга свойств химических соединений, поддающихся надежному прогнозированию с использованием разрабатываемых нами методов.

7.1. Совместное применение методологии искусственных нейронных сетей и методов молекулярного моделирования

На современном этапе научно-технического развития определяющую роль играет практическое использование сложных молекулярных и супрамолекулярных систем, что со всей актуальностью ставит задачу прогнозирования всего комплекса их практически значимых свойств и на основе этого проведение их целенаправленного дизайна. Не менее актуальной является также задача компьютерной обработки накопленных экспериментальных данных, относящихся к подобным системам, и извлечения информации, необходимой для эффективного конструирования новых систем.

В настоящее время для решения вышеуказанных задач все большее значение приобретают методы молекулярного моделирования, в основе которых лежат разнообразные методы молекулярно-механического и квантово-химического расчета модельных молекулярных систем. Несмотря на успехи в области молекулярного моделирования, следует признать, что ни одна даже самая совершенная молекулярная модель неспособна охватить всего комплекса взаимодействий, в которые вовлечена реальная молекулярная система, равно как и учесть эти взаимодействия с достаточной точностью. Это служит серьезным препятствием к практическому применению многих теоретических моделей.

В связи с этим особую актуальность приобретает проблема соотнесения теоретически рассчитываемых характеристик молекулярных систем с проявляемыми ими в эксперименте свойствами. Трудность решения этой проблемы обусловлена тем, что общий вид зависимости неучтенных в модели факторов от учитываемых молекулярных характеристик всегда является неизвестным, что является препятствием к применению стандартного аппарата математической статистики. Вследствие этого прогнозирование с достаточной точностью большинства практически важных свойств на основе теоретических моделей возможно в лучшем случае только внутри очень узкой группы молекулярных систем при помощи упрощенной линейной модели, параметризованной по имеющимся экспериментальным данным. Это делает невозможным применение полученной модели для большинства практически важных систем.

Генеральным направлением в решении указанной проблемы нам видится использование математического аппарата обработки данных, позволяющего выявлять любые сколь угодно сложные нелинейные зависимости неизвестного вида между теоретически рассчитываемыми молекулярными характеристиками и экспериментальными данными. Наиболее подходящими для этого являются искусственные нейронные сети, которые позволяют проводить “интеллектуальную обработку” экспериментальных данных, содержащихся в химических базах данных, выявляя существующие (заранее неизвестные) корреляции между имеющимися экспериментальными и структурными данными, с одной стороны, и прогнозируемыми молекулярными характеристиками, с другой стороны.

В данном разделе будет показано на нескольких примерах, что искусственные нейронные сети представляют собой мощный статистический аппарат обработки данных, который, в сочетании с элементами молекулярного моделирования, способен обеспечить надежный прогноз разнообразных свойств сложных молекулярных систем. Преимущество применения нейросетей заключается в их уникальной способности извлекать из эксперимента и обобщать зависимости, которые крайне трудно вывести из теоретических соображений. Поэтому аппарат нейросетей является необходимым дополнением к методам молекуляр-

ного моделирования, способным резко повысить их прогнозирующую способность и, следовательно, решать задачи дизайна сложных молекулярных систем.

Возникает вопрос: если нейросети в комбинации с фрагментными дескрипторами могут аппроксимировать любое свойство, то зачем их надо комбинировать с методами молекулярного моделирования? Все зависит от объема имеющихся экспериментальных данных (см. Табл. 20). Если данных достаточно много, то сочетания нейросетей с фрагментными дескрипторами действительно достаточно для моделирования любого свойства. Если данных очень мало либо они вообще отсутствуют, то нейросети не могут быть обучены, и поэтому для прогнозирования остаются только методы молекулярного моделирования. В промежуточной же ситуации, когда имеется определенный объем экспериментальных данных, но его недостаточно для построения нейросетевой модели на одних фрагментных дескрипторах, наилучший эффект дает интеграция молекулярного и нейросетевого моделирования. Это может быть достигнуто, например, путем использования определенных величин, вычисляемых при помощи методов молекулярного моделирования в качестве дескрипторов при построении моделей «структура-свойство». Чем больше экспериментальных данных, тем более простые методы молекулярного моделирования могут быть для этого применены.

Табл. 20. Выбор метода моделирования в зависимости от объема эданных

Объем экспериментальных данных	Предпочтительный метод моделирования
Мало либо отсутствуют	Молекулярное моделирование
Промежуточный объем данных	Сочетание молекулярного и нейросетевого моделирования
Достаточно много	Нейросетевое моделирование

7.1.1. Предсказание положения длинноволновой полосы поглощения симметричных цианиновых красителей.

Целью данной работы является иллюстрация эффективности применения искусственных нейронных сетей для предсказания практически важных

свойств сложных молекулярных систем на примере прогнозирования положения длинноволновой полосы поглощения симметричных цианиновых красителей в спиртовом растворе.

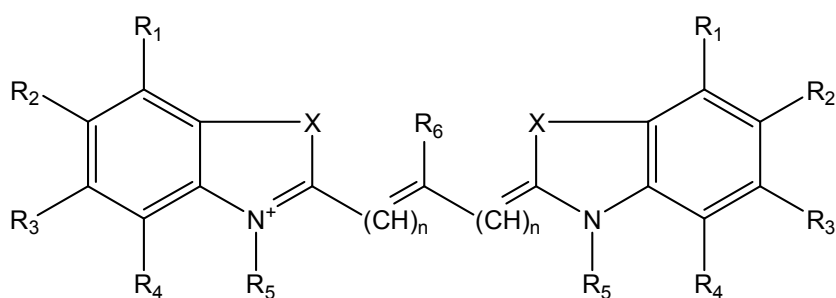
Основными областями применения цианинов (красителей, содержащих цепочку атомов $N^+=(CH-CH)_n=CH-N$) является их использование в качестве спектральных сенсбилизаторов и лазерных красителей. Ввиду чрезвычайной важности практического применения за последние 30 лет было проведено множество работ по выявлению зависимости физико-химических свойств цианиновых красителей от их строения на количественном уровне (см. обзорную статью [422] и монографии [423, 424]). В большинстве публикаций рассматривалось применение методов корреляционного анализа (уравнение Гаммета) для прогнозирования кислотности и потенциалов окисления и восстановления в очень узких рядах симметричных цианиновых красителей с одним варьируемым заместителем. Что касается предсказания положения длинноволновой полосы поглощения, то во всех опубликованных работах выявленные зависимости носили качественный либо полуколичественный характер. В качестве примеров можно привести сдвиг на ~ 100 нм при удлинении полиметиновой цепочки красителей на одно виниленовое звено [424, 425] и правило Ферстера-Дьюара-Нотта [425-427], описывающее влияние заместителей на окраску цианинов в нечетных положениях полиметиновой цепочки. В то же время, все попытки найти хотя бы полуколичественную зависимость окраски красителей от параметров заместителей не привели к желаемому результату.

Альтернативным методом предсказания окраски цианиновых красителей является использование квантово-химических расчетов. Уже простейшие подходы на основе теории возмущений молекулярных орбиталей позволили описать на качественном уровне изменение окраски красителей при варьировании ряда структурных параметров [423-427]. Особенно плодотворным оказалось использование выведенного на основе простого метода Хюккеля параметра эффективной длины концевых групп [428], который позволил на полуколичественном уровне описать зависимость окраски цианиновых красителей от строения гетероциклов. Тем не менее, непосредственное применение метода Хюккеля

ля и даже более совершенного метода Парра-Паризера-Попла с учетом конфигурационного взаимодействия дает очень большие ошибки, в ряде случаев превышающие 100 нм.

Наши предварительные эксперименты показали, что при использовании существенно более совершенного метода ZINDO/S с учетом конфигурационного взаимодействия возможно достичь точности прогноза 20-30 нм внутри групп красителей с одинаковой длиной цепочки и одинаковыми типами гетероциклов, если осуществлять подбор с учетом экспериментальных данных подстроечных параметров этого метода (факторов взвешивания для интегралов σ - σ - и π - π -перекрываний) внутри каждой из этих групп. Очевидными недостатками этого подхода являются как недостаточная точность прогноза (для практических целей желательно не больше 5-10 нм), так и наличие большого числа групп красителей, требующих отдельной параметризации, что не дает возможности осуществить такой прогноз для большинства красителей ввиду отсутствия экспериментальных данных, необходимых для параметризации.

В настоящей работе при помощи искусственных нейронных сетей (многослойных персептронов) нами обработана выборка из 398 симметричных цианиновых красителей, описываемых общей формулой (I). Этой формулой охватывается большинство используемых в промышленности цианиновых красителей.



где $n=0-6$, $X=O, S, NR, CH=CH, C(CH_3)_2$. Выборка была случайным образом разделена на две части: обучающую выборку, состоящую из 359 красителей, и контрольную, насчитывающую 39 соединений. Кроме этого, из данной выборки была получена подвыборка, включающая красители с незамещенным мезоположением ($R_6=H$), которая тоже была случайным образом разделена на обучающую (157 красителей) и контрольную (17 красителей).

На первом этапе работы для всех 398 красителей было определено геометрическое строение молекул при помощи процедуры автоматического молекулярного моделирования, включающей проведение в автоматическом режиме построения 3D-моделей молекул с последующим уточнением моделей путем поочередного применения методов молекулярной механики, молекулярной динамики (для вывода молекул из ложных локальных минимумов) и, наконец, полумпирического квантово-химического метода РМЗ. Технические детали разработанного нами и использованного в этом исследовании клиент/серверного программного комплекса описаны в работе [429].

На втором этапе работы проводилась нейросетевая обработка баз данных при помощи многослойных персептронов с использованием компьютерной программы NASAWIN. В качестве дескрипторов использовались рассчитанные на первом этапе по методу РМЗ энергии высшей занятой молекулярной орбитали (ВЗМО) $E_{ВЗМО}$ и низшей свободной молекулярной орбитали (НСМО) $E_{НСМО}$, длина полиметиновой цепочки n , индикатор наличия заместителя в мезоположении полиметиновой цепочки, а также индикаторы типа X в формуле (I): X_O , X_N , X_S , $X_{CH=CH}$, $X_{C(CH_3)_2}$.

Обучение нейросети проводилось по обучающей выборке с использованием алгоритма обобщенного дельта-правила (см. [42]) при начальном значении параметра скорости обучения 0.25 с последующим снижением до 0.01. Прогнозирующая способность нейросети оценивалась при помощи независимого прогноза на контрольной выборке. Для нахождения оптимальной архитектуры сети обучение проводилось при разном числе нейронов во внутреннем слое. В Табл. 21 приведены значения коэффициентов корреляции R и среднеквадратичных ошибок на обучающей выборке s_t и среднеквадратичных ошибок прогноза на контрольной выборке s_v для разного числа внутренних нейронов n_h для соединений из полной выборки, а в Табл. 22 дана та же информация для красителей с $R_6=H$.

Из информации, содержащейся в Табл. 21 и Табл. 22 можно сделать вывод, что для обеспечения наилучшей прогнозирующей способности следует брать нейросеть с 8-10 внутренними нейронами для произвольного симметрич-

ного цианинового красителя и с 7-8 внутренними нейронами для незамещенных в цепочке симметричных цианиновых красителей, при этом качество прогноза (среднеквадратичная ошибка 7-11 нм для общего случая и 3-5 нм для красителей с $R_6=H$) значительно превосходит все то, что было достигнуто ранее (см. обсуждение выше) и обеспечивает достаточную точность для решения практических задач дизайна красителей с заданным положением длинноволновой полосы поглощения.

Табл. 21. Результаты обучения ИНС для полной выборки (обозначения см. в тексте)

n_h	R	$RMSE_t$ (в нм)	$RMSE_v$ (в нм)
2	0.9884	13.5	10.0
3	0.9910	11.9	9.3
4	0.9912	11.8	8.3
5	0.9924	10.9	7.5
6	0.9924	10.9	8.0
7	0.9931	10.4	9.0
8	0.9923	11.0	7.4
9	0.9923	11.0	7.4
10	0.9928	10.6	7.0
11	0.9921	11.1	8.5
12	0.9930	10.6	9.4

Табл. 22. Результаты обучения ИНС для выборки с $R_6=H$ (обозначения см. в тексте)

n_h	R	$RMSE_t$ (в нм)	$RMSE_v$ (в нм)
2	0.9959	7.5	4.9
3	0.9958	7.6	4.7
4	0.9966	6.8	4.5
5	0.9954	8.0	5.5
6	0.9961	7.4	5.2
7	0.9986	4.4	3.4
8	0.9976	5.7	4.5
9	0.9955	7.8	6.9
10	0.9980	5.3	4.8

7.1.2. Оценка значений констант ионизации для различных классов органических соединений

Как известно, константа ионизации, K_a (или обратный логарифм $-\log K_a = pK_a$) является одной из важнейших характеристик органических соеди-

нений, отражающей их кислотно-основные свойства. К настоящему времени опубликован ряд работ, посвященных предсказанию констант ионизации для различных классов органических соединений с использованием различных подходов. Так, например, для оценки констант ионизации фенолов, карбоновых кислот и азотсодержащих соединений были построены линейно-регрессионные модели с использованием квантово-химических дескрипторов [430-432]. В работах [433, 434] оценка pK_a фенолов и карбоновых кислот осуществлялась с помощью термодинамических вычислений с предварительным расчетом атомных зарядов и оптимизацией геометрии молекул. Также для прогнозирования значений pK_a восемнадцати азотсодержащих соединений было предпринято построение CoMFA моделей [435]. Результаты, полученные в работах [430-435], показывают, что стандартная ошибка предсказания (s) значений pK_a , получаемая с помощью регрессионных моделей, а также с помощью термодинамических расчетов составляет 0.3 – 0.6 ккал/моль, тогда как значение s , полученное с помощью CoMFA-модели составило 0.193 ккал/моль. Однако следует отметить, что несмотря на то, что метод CoMFA дает наименьшую ошибку предсказания, использование данного метода представляется эффективным только для оценки значений pK_a очень узкой однородной выборки соединений, обладающих общим скелетом, что накладывает серьезные ограничения на применимость метода CoMFA для предсказания свойств широких разнородных выборок.

В данной работе предпринята попытка использования фрагментного и квантово-химического подходов для моделирования значений pK_a различных классов органических соединений, а так же попытка построения общей QSPR-модели для всех рассматриваемых классов соединений. Для этого, с помощью программы MOLED [436] были созданы четыре базы данных (БД): (1) фенолы (170 соединений); (2) карбоновые кислоты (238); 3) азотсодержащие соединения (268); 4) общая база (676). Далее каждая БД была случайным образом разделена на обучающую (90% соединений) и контрольную (10% соединений) выборки.

Значения констант ионизации для различных соединений были взяты из работ [430, 431]. Построение моделей “структура-свойство” было осуществлено с помощью нейросетевого программного комплекса NASAWIN [194] (см. раздел 8.2). Для моделирования локальных свойств нами были использованы фрагментные дескрипторы с выделенными атомами.

Известно, что константа ионизации достаточно хорошо коррелирует с квантово-химическими дескрипторами [430-432]. В данной работе нами были при помощи специальной утилиты, входящей в программный комплекс NASAWIN, рассчитаны значения двенадцати дескрипторов, описывающих внутримолекулярные электронные свойства молекул, такие как: энергия высшей занятой и низшей незанятой молекулярных орбиталей, заряд на меченом атоме, максимальный отрицательный заряд на атоме, максимальный заряд на атоме водорода, дипольный момент молекулы, электрофильная и нуклеофильная граничная электронная плотность, электрофильная, нуклеофильная и радикальная суперделокализация, атомная самополяризуемость. При этом, с помощью метода БПМЛР (см. подраздел 4.1.5) проводился отбор наиболее значимых дескрипторов для каждой QSPR-модели. Прогнозирующая способность полученных моделей для каждой БД была проверена предсказанием значений pK_a соединений контрольной выборки.

На первом этапе методами БПМЛР и ИНС были построены частные QSPR-модели для фенолов, карбоновых кислот и азотсодержащих соединений с использованием как фрагментных дескрипторов рассчитанных для соединений с метками, а так же с использованием фрагментных и квантово-химических дескрипторов. Статистические параметры для этих моделей представлены в Табл. 23. Анализируя полученные QSPR-модели, можно сделать вывод, что во всех случаях модели “структура-свойство”, построенные методом ИНС, характеризуются несколько более высокими значениями коэффициента корреляции и дают меньшую ошибку как на обучающих, так и на контрольных выборках, по сравнению с моделями, построенными с использованием статистического аппарата множественной линейной регрессии. Следует так же отметить, что использование квантово-химических дескрипторов также способствует улучшению

статистических параметров моделей по сравнению с результатами, получаемыми при использовании только фрагментных дескрипторов.

Табл. 23. Статистические показатели моделей, построенных для фенолов, карбоновых кислот и азотсодержащих соединений

Класс соединений	Параметры моделей, построенных с использованием только ФД	Параметры моделей построенных с использованием ФД и квантово-химических дескрипторов
Фенолы	МЛР: $R^2 = 0.9746$, $s = 0.40$, $RMSE_t = 0.38$, $RMSE_v = 0.57$	МЛР: $R^2 = 0.9794$, $s = 0.36$, $RMSE_t = 0.33$, $RMSE_v = 0.41$
	ИНС: $R^2 = 0.9815$, $RMSE_t = 0.32$, $RMSE_v = 0.53$	ИНС: $R^2 = 0.9831$, $RMSE_t = 0.30$, $RMSE_v = 0.42$
Карбоновые кислоты	МЛР: $R^2 = 0.8966$, $s = 0.33$, $RMSE_t = 0.31$, $RMSE_v = 0.51$	МЛР: $R^2 = 0.9122$, $s = 0.31$, $RMSE_t = 0.28$, $RMSE_v = 0.34$
	ИНС: $R^2 = 0.9115$, $RMSE_t = 0.28$, $RMSE_v = 0.48$	ИНС: $R^2 = 0.9534$, $RMSE_t = 0.21$, $RMSE_v = 0.27$
Азотсодержащие соединения	МЛР: $R^2 = 0.9302$, $s = 0.99$, $RMSE_t = 0.93$, $RMSE_v = 1.14$	МЛР: $R^2 = 0.9611$, $s = 0.75$, $RMSE_t = 0.69$, $RMSE_v = 0.94$
	ИНС: $R^2 = 0.9306$, $RMSE_t = 0.93$, $RMSE_v = 1.13$	ИНС: $R^2 = 0.9692$, $RMSE_t = 0.62$, $RMSE_v = 0.60$

где: R^2 - коэффициент детерминации, $RMSE_t$, $RMSE_v$ – среднеквадратичная ошибка на обучающей и контрольной выборке, s – стандартное отклонение.

Как видно из данных, приведенных в Табл. 23, лучшие статистические параметры были получены методом ИНС для фенолов. Однако в этом случае статистические параметры моделей, построенных с использованием только 20 фрагментных дескрипторов ($R^2 = 0.9815$, $RMSE_t = 0.32$, $RMSE_v = 0.53$) и дополнительных 7 квантово-химических дескрипторов ($R^2 = 0.9831$, $RMSE_t = 0.30$, $RMSE_v = 0.41$) оказались достаточно близки. В отличие от этого, улучшение статистических параметров при использовании квантово-химических дескрипторов в комбинации с фрагментными дескрипторами для моделирования pK_a карбоновых кислот и азотсодержащих соединений достаточно очевидно.

Так как предложенный подход оказался эффективным при моделировании трех отдельных баз данных, нами была предпринята попытка построить общую модель “структура-свойство” для общей базы данных. С целью проверки прогностической способности QSPR-моделей база данных также была раз-

делена на обучающую выборку (609 соединений) и контрольную выборку (67 соединений). При моделировании общей базы данных было использовано различное количество дескрипторов. Первоначально, методом БПМЛР было отобрано 226 фрагментных и 7 квантово-химических дескрипторов, а в дальнейшем осуществлялось последовательное уменьшение их числа (Табл. 24). Такой подход представляется эффективным для выявления оптимального количества дескрипторов, что, в свою очередь, дает возможность получения более устойчивой предсказательной QSPR-модели.

Табл. 24. Статистические параметры QSPR-моделей построенных с использованием фрагментных и квантово-химических дескрипторов

$D_{\text{фр}}/D_{\text{кк}}$	R^2	$RMSE_{\text{обуч}}$	$RMSE_{\text{контр}}$
226/7	0.9938	0.34	0.40
194/6	0.9931	0.36	0.55
96/4	0.9862	0.36	0.46
46/4	0.9832	0.56	0.64
13/2	0.9658	0.78	0.92

Анализ полученных результатов показал что модель “структура – свойство”, построенная с использованием 96 фрагментных и 4 квантово-химических дескрипторов (Рис. 49), представляется наиболее оптимальной. Как видно из Табл. 24, переход от модели, построенной на двухстах дескрипторах, к модели на ста приводит лишь к незначительному уменьшению коэффициента корреляции при сохранении значения $RMSE$ для обучающей выборки и улучшению $RMSE$ для контрольной выборки. При этом дальнейшее уменьшение числа дескрипторов до 50 хотя и позволяет сохранить близкий коэффициент корреляции, но приводит к резкому увеличению значений $RMSE$ как для контрольной, так и для обучающей выборки.

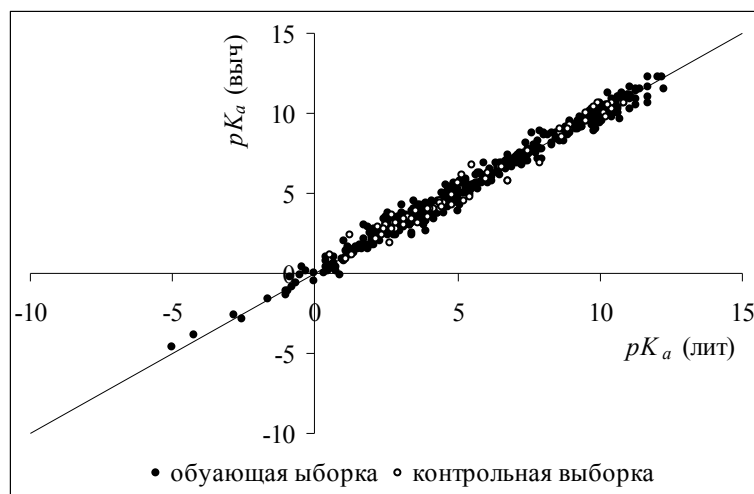


Рис. 49. График разброса для QSPR-модели, построенной с использованием ста дескрипторов для общей базы данных.

Таким образом, в результате данной работы была построена количественная модель зависимости “структура-свойство” для констант ионизации органических соединений, принадлежащих различным классам. Полученные результаты показали хорошую применимость фрагментного подхода и искусственных нейронных сетей к моделированию данного свойства и возможность использования полученных моделей “структура-свойство” для предсказания констант ионизации новых соединений, принадлежащих к классам, использованным нами для построения моделей.

7.1.3. Моделирование мутагенной активности замещенных полициклических нитросоединений с помощью искусственных нейронных сетей

Экспериментальные исследования, проведенные в течение последних трех десятилетий, показали, что немалое число химических соединений, используемых в промышленности, сельском хозяйстве, медицине и в быту, обладают мутагенной активностью и представляют собой генетическую опасность для человека, сопоставимую с опасностью радиации. В настоящее время на основании полученных данных предпринимаются попытки идентифицировать фрагменты структуры, квантовые параметры и физико-химические свойства химического соединения, которые могут определять его мутагенную актив-

ность. Основной целью этих работ является разработка компьютерного подхода для предварительного (внеэкспериментального) отбора безопасных в генетическом отношении соединений среди вновь синтезированных веществ, представляющих потенциальную ценность в качестве пестицидов, фармакологических и косметических средств, пищевых добавок, красителей и т.д.

В настоящее время апробирован ряд компьютерных методов прогноза мутагенной активности соединений. Наиболее известным из них является «CASE» (Computer Automated Structure Evaluation), основанный на поиске с помощью линейно-регрессионного анализа фрагментов структуры – биофоров, которые вносят наибольший вклад в биологическую активность соединения [437]. Известны также попытки описания мутагенной активности с помощью квантово-химических дескрипторов. Например, для аминокислотных производных бифенила, бензидина, стибена была показана корреляция активности с гидрофобностью – коэффициентом распределения октанол-вода ($\log P$), энергией низшей незанятой орбитали (E_{LUMO}), значениями констант σ^+ Гаммета [438, 439].

Мутагенная активность нитро- и амино- замещенных флуоренонов, бифенилов [440, 441] и гетероциклических аналогов пирена и фенантрена уже изучалась ранее [442]. Полученные в этих работах экспериментальные данные были использованы ранее использованы нами для построения линейно-регрессионных уравнений количественной зависимости мутагенной активности этих соединений от ФД и квантово-химических дескрипторов с помощью программного комплекса ЕММА. Для выборки замещенных бифенилов (21 соединение) были получены 2 линейно-регрессионных уравнения, включающие как фрагменты структуры, так и квантово-химические дескрипторы (минимальный квадрат коэффициента вклада атомной орбитали углерода в низшую свободную молекулярную орбиталь; минимальный квадрат коэффициента вклада атомной орбитали азота в низшую свободную молекулярную орбиталь; максимальный индекс свободной валентности для атомов углерода; среднее значение индекса свободной валентности для атомов кислорода) [443]. Для гетероциклических аналогов пирена и фенантрена (22 соединения) лучшим из серии линейно-регрессионных уравнений было одно, включающее в себя только квантово-

химические дескрипторы (минимальный индекс свободной валентности; максимальный π -заряд на атоме азота; средний квадрат коэффициента вклада атомной орбитали кислорода в высшую занятую молекулярную орбиталь) и гидрофобность $\log P$ [444]. Полученные статистические модели хорошо прогнозировали мутагенную активность химических соединений, входящих в исследуемую выборку, однако отобранные дескрипторы не были информативными с точки зрения представлений о механизмах действия этих соединений.

Поэтому для нейросетевого моделирования были использованы те же экспериментальные данные, но модели строились на основе дескрипторов, отобранных экспертным путем в соответствии с гипотезами о механизме действия нитроароматических соединений и эмпирическими заключениями о влиянии элементов структуры на мутагенную активность.

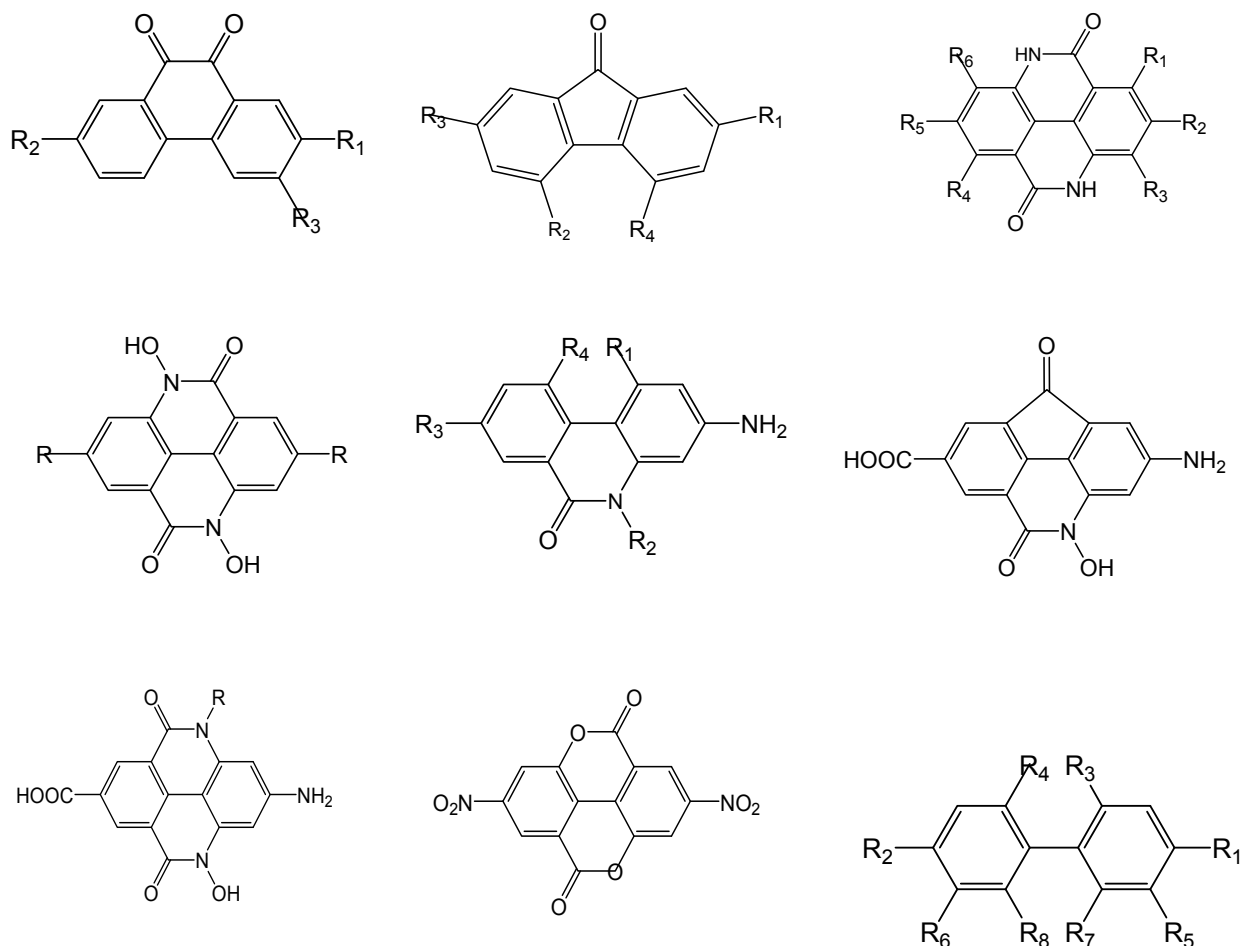


Рис. 50. Структуры мутагенных полициклических нитросоединений. В качестве заместителей выступают в различных комбинациях группы NO_2 , COOH , CONH_2

На Рис. 50 приводятся структуры соединений, использованных в настоящем исследовании [445]. Были использованы экспериментальные данные по мутагенной активности в штамме *Salmonella typhimurium* TA 1538 (*hisD3052*, *rfa*, *uvr*), регистрирующем мутации сдвига рамки считывания, без метаболической активации фракцией S9 печени млекопитающих [440-446]. Исходная выборка включала в себя 54 соединения; исследуемая активность выражалась как логарифм числа his^+ -ревертантов при дозе, относящейся к середине линейного участка кривой доза-эффект.

Вначале мы рассмотрели основные факторы, определяющие или влияющие на мутагенную активность нитроароматических соединений, чтобы определить набор дескрипторов для включения в математические модели. Известно, что основным путем биотрансформации нитроаренов, приводящим к образованию мутагенных, канцерогенных и токсичных метаболитов, является восстановление нитрогруппы нитроредуктазами клетки [447]. Способность к восстановлению нитроаренов коррелирует с таким параметром как энергия низшей незанятой молекулярной орбитали E_{LUMO} (дескриптор d_1) [447]. Кроме того, в модель были включены 2 квантово-химических дескриптора, которые характеризуют состояние атомов азота и кислорода в молекулах: максимальный заряд на атоме азота (дескриптор d_2) и максимальный заряд на атоме кислорода (дескриптор d_3). В качестве дескриптора d_4 в модель был включен коэффициент распределения октанол-вода $\log P$ (гидрофобность), характеризующий способность молекулы достигать сайтов взаимодействия в живом организме. Квантовые расчеты проводились по методу AM1, расчет $\log P$ – по методу Реккера.

Наибольшую активность в экспериментах показали соединения с *пара*-положением нитрогруппы, гетероциклические аналоги пирена с *пара*-положением аминогруппы, тогда как наличие заместителей в *орто*- и *мета*-положениях снижало активность. Поэтому в качестве подструктурных дескрипторов в модель были введены следующие дескрипторы: наличие нитрогруппы в *пара*-положении - d_5 ; наличие аминогруппы в *пара*-положении - d_6 ; наличие *мета*- и *орто*-заместителей - d_7 .

Нейросетевое моделирование проводилось с использованием программного комплекса NASAWIN (см. раздел 8.2) и EMMA (см. раздел 8.1). Каждая исследуемая выборка разбивалась случайным образом на обучающую и контрольную подвыборки. Были изучены построенные на одинаковых наборах дескрипторов линейно-регрессионные и нейросетевые модели. Обучение ИНС проводилось по методу обобщенного дельта-правила со скоростью обучения, равной 0,25 и моментом 0,9. Обучение прерывалось в момент наступления переучивания.

Предварительно нами был проведен отбор наиболее значимых дескрипторов с помощью множественной линейной регрессии. Для общей выборки такими дескрипторами оказались 3 дескриптора – E_{LUMO} (d_1), $\log P$ (d_4) и дескриптор, характеризующий наличие нитрогруппы в *para*-положении (d_5). Результаты нейросетевого моделирования для общей выборки, содержащей 49 обучающих и 5 контрольных соединений, приведены в Табл. 25.

С целью улучшения параметров модели в рамках заданных дескрипторов общая выборка из 54 соединений была разбита на 2 подвыборки структурно-родственных соединений. Для первой подвыборки, состоящей из 33 гетероциклических аналогов пирена и фенантрена и замещенных флуоренонов, значимыми оказались все дескрипторы, за исключением липофильности (d_4). Для второй подвыборки, содержащей замещенные бифенилы, значимыми оказались дескрипторы E_{LUMO} (d_1) и $\log P$ (d_4).

Сравнивая между собой результаты отбора дескрипторов для различных выборок, легко заметить, что дескриптор E_{LUMO} выступает в роли основного, способного характеризовать мутагенную активность как молекул с конденсированными бензольными кольцами, так и производных бифенила. Чем энергия низшей свободной орбитали ниже, тем стабильнее соответствующие активные метаболиты, в частности первый в цепи восстановления нитрогруппы – анион-радикал, тем больше мутагенная активность. Вторым по значению дескриптором, характеризующим положение нитрогрупп, является d_5 , прямо пропорциональный мутагенной активности. Влияние *мета*- и *орто*- заместителей, уменьшающее планарность молекулы и ее активность, оказалось более важным для

молекул с конденсированными бензольными кольцами, чем для бифенилов. Следует отметить, что планарность молекулы для мутагенной активности нитроароматических соединений имеет особое значение [448]. Во-первых, планарные молекулы обладают большей способностью интеркалировать в ДНК, чем непланарные, и, во-вторых, полагают, что они имеют повышенное сродство к нитроредуктазам, чем другие соединения [449]. Липофильность оказалась существенным параметром при описании активности производных с конденсированными бензольными кольцами. Модели, описанные в литературе для этих производных, также как и для бифенилов, как включают этот параметр [444], так и не включают его [443]. Этот факт можно объяснить тем, что липофильность играет второстепенную роль в определении мутагенной активности рассматриваемых соединений.

Табл. 25. Статистические характеристики нейросетевых моделей

Выборка соединений	Дескрипторы входного слоя	Метод	Характеристики модели		
			R	RMSE _t	RMSE _v
Производные пирена, фенантрена, флуоренона	d ₁ , d ₂ , d ₃ , d ₅ , d ₆ , d ₇	ИНС	0.90	0.76	0.96
		МЛР	0.75	1.45	1.94
Замещенные бифенилы	d ₁ , d ₄	ИНС	0.97	0.59	0.13
		МЛР	0.80	1.21	1.34
Все соединения	d ₁ , d ₄ , d ₅	ИНС	0.87	1.30	1.57
		МЛР	0.75	1.45	1.94

где R - коэффициент корреляции между предсказанной и экспериментальной величинами числа ревертантов для соединений обучающей выборки; RMSE_t - среднеквадратичная ошибка воспроизведения числа ревертантов для соединений обучающей выборки (ln единицы); RMSE_v - среднеквадратичная ошибка предсказания числа ревертантов для соединений контрольной выборки (ln единицы).

Результаты нейросетевого моделирования для двух подвыборок родственных соединений также приведены в Табл. 25. Первая подвыборка содержала 30 соединений в обучающей выборке и 3 соединения в контрольной; вторая - 19 в обучающей и 2 в контрольной. Как видно из Табл. 25, наилучшая модель была получена для замещенных бифенилов, представляющих собой единый

массив структурно-родственных соединений, действующих по одному механизму.

Нам удалось значительно улучшить результаты нейросетевого прогноза, полученные для обобщенной выборки с использованием метода структурного подобию, реализованного в программном комплексе «NASAWIN». Для каждого соединения из контрольной выборки было найдено ближайшее структурно-родственное соединение из обучающей выборки для проведения процедуры коррекции нейросетевого прогноза, результаты которого приводятся в Табл. 26. Среднеквадратичная ошибка нейросетевого прогноза с последующей коррекцией по методу структурного подобию составила 0,30 логарифмических единиц.

Таким образом, примененный нами подход, основанный на введении в модель дескрипторов, отобранных экспертным путем, может иметь свою область применения в качестве проверки выдвинутой гипотезы о механизме действия группы структурно-родственных соединений. Кроме того, полученные нами зависимости могут быть использованы для предварительного прогноза мутагенной активности новых соединений, которые по своей химической структуре близки к соединениям из анализируемой выборки.

Табл. 26. Результаты применения метода структурного подобию для коррекции нейросетевых прогнозов

Соединение	Ближайший структурный сосед			Экспериментальное значение	Результат нейросетевого прогноза	Результат коррекции по методу структурного подобию
	№	Экспериментальное значение	Расчетное значение			
9	10	5,11	4,96	4,53	3,90	4,05
11	16	2,83	1,97	3,74	2,48	3,34
37	51	0,00	3,02	0,00	2,95	-0,07
49	50	0,00	0,29	0,00	0,33	0,04
54	53	0,00	0,32	0,00	0,13	-0,19

7.1.4. Прогнозирование констант заместителей с использованием искусственных нейронных сетей и квантово-химических дескрипторов

В данном исследовании мы изучали возможность прогнозирования значений констант заместителей (двух констант Гаммета σ^m и σ^p ; двух констант Свейна и Лаптона - полевой F и резонансной R ; стерической константы Тафта E_s) при помощи искусственных нейронных сетей с использованием квантово-химических дескрипторов, вычисляемых для модельных соединений, получаемых присоединением заместителей к определенным общим фрагментам (водороду и к метильной, фенильной, пара-нитрофенильной, пара-оксифенильной и орто-диалкилфенильным группам, см. Рис. 51). Для всех получаемых таким образом соединений проводился квантово-химический расчет при помощи полуэмпирического метода PM3 с полной оптимизацией геометрии. В качестве дескрипторов использовались рассчитанные для модельных соединений значения теплоты образования, энергий граничных (высшей занятой и низшей свободной) молекулярных орбиталей, а также зарядов на определенных атомах.

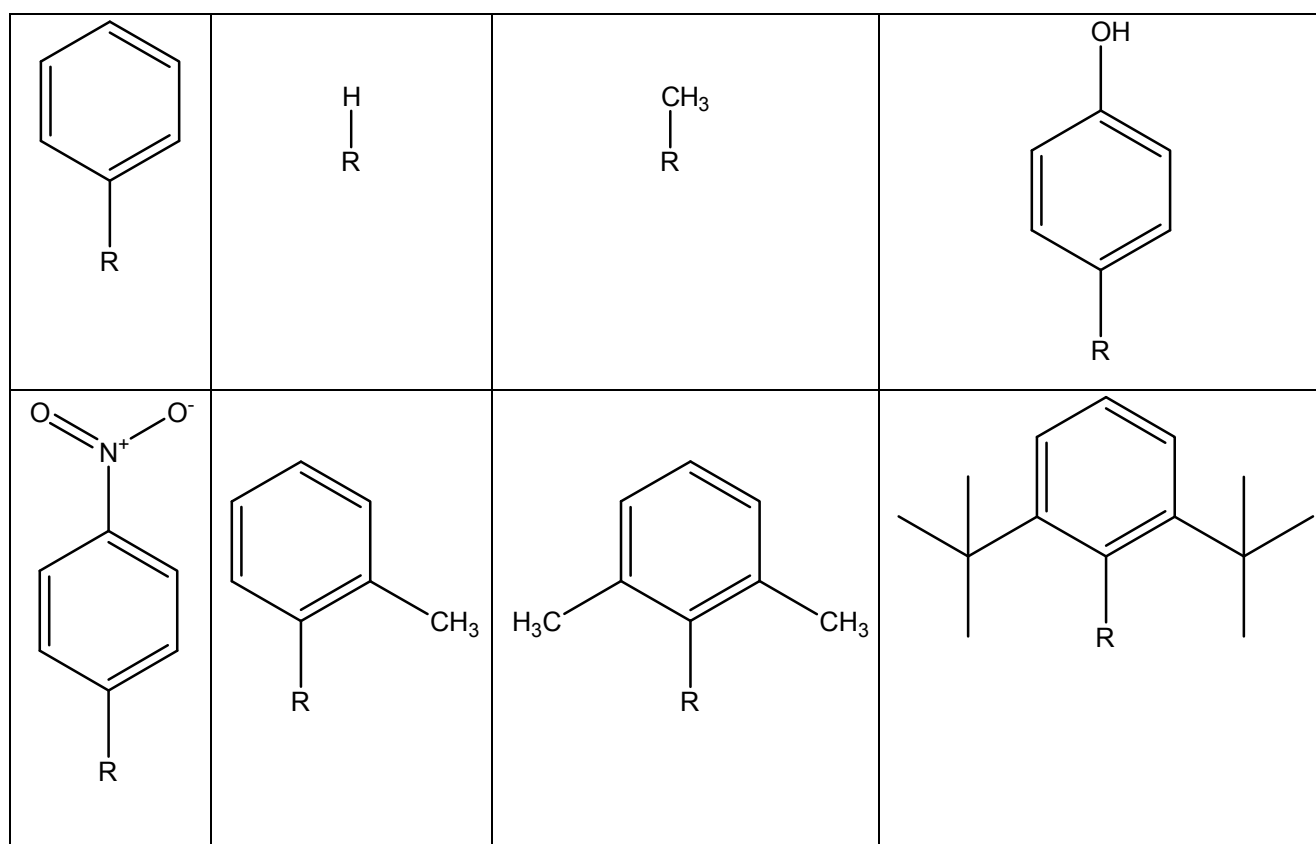


Рис. 51. Модельные соединения, использованные для расчета квантово-химических дескрипторов

Использованная в работе выборка включала данные для 160 наиболее распространенных заместителей. Для контроля прогнозирующей способности построенных моделей она была разбита на 2 выборки: обучающую и контрольную. Статистические параметры построенных нейросетевых моделей представлены в Табл. 27.

Табл. 27. Статистические характеристики нейросетевых моделей для прогнозирования констант заместителей

Прогнозируемая константа заместителя	Коэффициент корреляции	Средне-квадратичная ошибка на обучающей выборке	Средне-квадратичная ошибка на контрольной выборке	Размер обучающей выборки	Размер контрольной выборки
σ^m	0.9686	0.0594	0.1279	144	16
σ^p	0.9589	0.1014	0.1592	144	16
F	0.9403	0.0717	0.1391	143	16
R	0.85	0.1254	0.1452	144	16
E_s	0.9794	0.6555	0.3935	42	4

Полученные низкие среднеквадратичные ошибки прогнозирования на контрольных выборках (0.13 для σ^m , 0.16 для σ^p , 0.14 для F, 0.15 для R, 0.39 для E_s) свидетельствуют о работоспособности данного подхода к прогнозированию констант заместителей.

7.2. Корреляции структура-условия-свойство

7.2.1. Концепция построения нейросетевых зависимостей структура – условия – свойство

В настоящее время, когда накоплен значительный объем экспериментальных данных практически во всех областях химии, особенно остро встает вопрос о возможности обобщения и математической обработки больших мас-

сивов разрозненных данных с целью прогнозирования тех или иных свойств новых веществ, представляющих практический интерес. Одним из наиболее перспективных методов такой обработки является нейросетевое моделирование. К главным достоинствам искусственных нейронных сетей можно отнести возможность построения на их основе нелинейных многопараметровых моделей даже в тех случаях, когда заранее неизвестен точный вид аналитической зависимости структура-свойство.

Классический подход к построению моделей «структура-свойство» основан на аппроксимации зависимости исследуемого свойства от дескрипторов, описывающих структуры химических соединений, при фиксированных «стандартных» условиях, накладываемых на его измерение. Такими условиями могут являться, например, температура, давление, ионная сила раствора и т.д. Это, однако, оставляет открытым вопрос о предсказании значений этого же свойства при других условиях, а также значительно снижает объем доступных для обработки экспериментальных данных. Хотя для этой цели могут быть использованы формулы из арсенала физической химии, однако они не всегда обеспечивают максимально возможную точность прогноза, поскольку часто бывают основаны на использовании «усредненных» эмпирических параметров.

Поскольку, как правило, зависимость свойств химических соединений от условий, в которых они измерены, также носит нелинейный характер, мы предположили, что с помощью методологии искусственных нейронных сетей можно расширить классический подход путем добавления характеристик внешних условий к входным параметрам нейросети [450, 451]. В качестве характеристик среды могут использоваться такие параметры, как температура, давление, концентрация, наличие того или иного растворителя, дескрипторы, характеризующие свойства растворителя, и т.д.

Общая схема предлагаемого подхода к построению зависимостей структура – условия – свойство изображена на Рис. 52.



Рис. 52. Общая схема нахождения зависимостей структура – условия – свойство

Принципиальная возможность получения нейросетевых зависимостей «структура – условия – свойство» проиллюстрирована нами построением моделей для прогнозирования физико-химических свойств углеводородов произвольной структуры, содержащих от 1 до 40 атомов углерода (строились зависимости температур кипения от структуры при различных значениях давления, динамической вязкости и плотности при различных температурах; см. подраздел 7.2.2), а также констант скорости кислотного гидролиза сложных эфиров карбоновых кислот при различной температуре и различных составах растворителей (подраздел 7.2.3).

7.2.2. Построение и анализ нейросетевых зависимостей структура-условие-свойство для физико-химических свойств углеводородов

Для демонстрации возможностей предложенного подхода мы остановили свой выбор на классе углеводородов, поскольку имеются большие массивы экспериментальных данных по свойствам этих соединений, измеренных в различных условиях. Помимо этого, углеводороды, являющиеся важнейшими компонентами нефти, природного газа и продуктов их переработки, широко

используются как топливо, в качестве сырья для получения многих химических продуктов и т.д.

Моделирование зависимости физико-химических свойств углеводородов от их структуры уже проводилось рядом исследователей с использованием линейного регрессионного анализа и топологических индексов [376, 452-460], линейного регрессионного анализа и квантово-химических параметров [461], множественной линейной регрессии и топологических индексов [462-464], а также нейросетевых методов [198, 406, 465, 466] (подробнее о нейросетевых методах см. раздел 1.2). Моделирование проводилось, как правило, для узких серий структурных аналогов. В большинстве случаев использования топологических индексов для описания структур углеводородов моделирование служило лишь иллюстрацией возможности применения предложенных авторами новых индексов.

Экспериментальные данные [467], на базе которых нами были созданы приведенные ниже нейросетевые модели, были получены для углеводородов с длиной цепи от 1 до 40 углеродных атомов: насыщенных, с кратными связями и ароматических; разветвленных и неразветвленных; ациклических, циклических и полициклических; и т.д.

Для исследования каждого свойства была создана своя структурная база данных с помощью компьютерной программы MOLED (см. раздел 8.1). Дескрипторы были рассчитаны с помощью дескрипторного блока FRAGMENT (см. раздел 8.3), входящего в программный комплекс NASAWIN (см. раздел 8.2).

7.2.2.1. Моделирование зависимости структура - давление - температура кипения

В лабораторной практике часто бывает необходимо определить температуру кипения вещества при определенном давлении, как при пониженном, так и при повышенном. Для этой цели обычно используют классическую номограмму «давление – температура», номограммы для отдельных типов соединений, приближенные таблицы или эмпирические уравнения, применимые лишь к от-

дельным конкретным соединениям [468, С.41-46]. Существует несколько компьютерных программ, которые выполняют расчет температур кипения, например, по уравнениям Клаузиуса – Клапейрона [469] и по правилу Тротона [470].

К основным недостаткам вышеперечисленных методов вычисления можно отнести их приближенный характер и ограниченную область применения. Помимо этого, для получения данных по номограмме необходимо располагать дополнительной парой значений температура – давление.

При построении нейросетевой модели были использованы значения температур кипения углеводородов разнообразной структуры при давлениях от 0,001 мм рт.ст. до 10 атмосфер [467, С. 200 - 258]. Для расчета исходных дескрипторов были найдены все фрагменты структур с максимальной длиной 4 атома. Дескриптором служило число повторений данного фрагмента в структурной формуле соединения. Из общего набора дескрипторов было отобрано 354 наиболее значимых дескриптора с помощью пошаговой множественной линейной регрессии. Исходная выборка соединений, содержащая 14346 записи «структура – давление – температура кипения», была разбита случайным образом на обучающую (12911 записи) и контрольную (1434 записи) подвыборки.

В работе была использована трехслойная нейросеть с 10 скрытыми нейронами. Нейросеть обучалась по методу обобщенного дельта-правила со скоростью обучения 0,25 и моментом 0,9. После 1000 итераций, когда статистические показатели модели стабилизировались, обучение было остановлено. Параметры нейросетевой модели в момент прерывания обучения нейросети были следующими: $R = 0,8581$, $RMSE_t = 57,88$, $RMSE_v = 58,15$, где R – коэффициент корреляции между спрогнозированными и экспериментальными значениями, $RMSE_t$ и $RMSE_v$ –среднеквадратичные ошибки для обучающей и контрольной выборок (°C).

Для улучшения статистических показателей мы решили использовать процедуру предварительной модификации дескрипторов. Все рассчитанные фрагментные дескрипторы были подвергнуты модификациям «квадрат величины» и «величина, деленная на количество неводородных атомов в молекуле». Значения давлений, при которых были измерены температуры кипения, также

были модифицированы с использованием функций квадратного корня, квадрата, логарифма, обратной величины и величины, деленной на количество неводородных атомов в молекуле. Объединенный набор дескрипторов, состоящий из полученных в результате модификаций значений наряду с исходными значениями дескрипторов и давления, был подвергнут процедуре отбора наиболее существенных параметров с помощью процедуры БПМЛР. В результате отбора осталось 346 дескрипторов, причем в этот набор вошли все 5 модификаций параметра «давление».

Все характеристики нейросети (за исключением количества входных нейронов) были такими же, что и для модели, использующей немодифицированные дескрипторы в качестве исходных параметров нейросети. Таким образом, была получена нейросетевая модель зависимости температуры кипения углеводородов от их структур и давления со следующими параметрами (см. также Рис. 53): число итераций 2374, $R = 0.9996$, $S_t = 2.7997$, $S_v = 2.8003$.

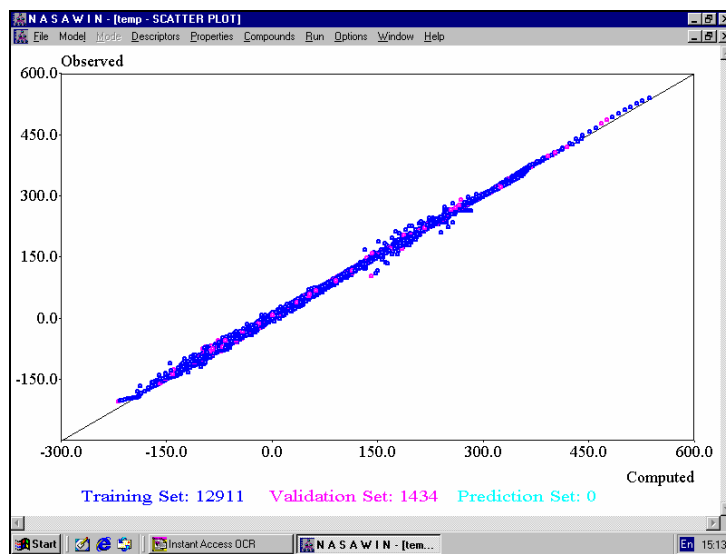


Рис. 53. Результаты нейросетевого моделирования зависимости структура – давление – температура кипения.

Для оценки качества прогноза нейросетевой модели произвольно отобрали несколько соединений из контрольной выборки, для которых определили температуры кипения в зависимости от давления с помощью номограммы давление – температура [468, стр. 46] (см. Табл. 28). Соединения и давления были

отобраны так, чтобы они попадали в область действия номограммы: температуры кипения при атмосферном давлении должны были принадлежать интервалу от 100 до 700°С; давления - интервалу от 0,01 до 700 мм рт. ст.

Таким образом, нейросетевая модель позволяет определять и прогнозировать температуру кипения углеводородов произвольной структуры с более высокой степенью точности, чем широко используемая номограмма, причем для более широкого интервала значений давления.

Табл. 28. Сравнительные результаты определения температур кипения (°С)

Название соединения	Давление, мм рт. ст.	Экспериментальные значения	Нейросетевая модель	Номограмма
<i>n</i> -октатриаконтан	0,01	219,700	218,224	206 (524,900 ¹⁾)
2-метилгептан	14	18,050	17,894	15 (117,647 ¹⁾)
<i>n</i> -гептакозан	25	290,400	290,629	287 (427,300 ¹⁾)
<i>цис</i> -декагидро-нафталин	30	93,000	91,231	97 (195,700 ¹⁾)
псевдокумол	100	103,355	101,851	100 (169,351 ¹⁾)
3,4-диметил-гексан	300	87,230	86,035	85 (117,725 ¹⁾)
1,2-диэтил-бензол	500	167,151	166,022	165 (183,423 ¹⁾)

¹⁾ – экспериментальные температуры кипения при 760 мм рт. ст., использованные при определении температур кипения по номограмме.

Хочется отметить также, что значения температур кипения в области глубокого вакуума определяются по номограмме с большой погрешностью, в то время как небольшая ошибка нейросетевого прогноза постоянна во всем диапазоне значений давления.

Когда данная работа была полностью завершена, в печати появилась статья [471], в которой авторы приводят результаты нейросетевого моделирования

зависимости давления насыщенных паров от температуры и структуры для 274 углеводородов (структуры описывались топологическими индексами и значением молекулярного веса). Однако число нейронов в скрытом слое нейросети, использованной для проведенного моделирования (29 нейронов), свидетельствует об избыточном количестве настраиваемых параметров модели и ставит под сомнение прогнозирующую способность этой модели.

7.2.2.2. Моделирование зависимости «структура - температура – плотность»

При построении нейросетевой модели были использованы значения плотности углеводородов разнообразной структуры в температурном интервале от -180 до 300°C [467, С. 87 - 110]. Для всех структур из исходной выборки были выделены фрагменты с максимальной длиной, равной 4 атомам. Все рассчитанные фрагментные дескрипторы затем были подвергнуты модификациям «квадратный корень из величины», «квадрат величины», «логарифм величины», «обратная величина» и «величина, деленная на количество неводородных атомов в молекуле». В качестве входных параметров были также использованы значения температур, возведенные в квадрат, и значения температур, деленные на количество неводородных атомов в молекуле. Объединенный набор дескрипторов, состоящий из полученных в результате модификаций значений наряду с исходными значениями дескрипторов и температур, был подвергнут процедуре отбора наиболее существенных параметров с помощью метода БПМЛР (см. подраздел 4.1.5). В результате отбора осталось 478 дескрипторов.

Исходная выборка соединений, содержащая 3056 записей структура – температура - плотность, была разбита случайным образом на обучающую (2751 запись) и контрольную (305 записей) подвыборки. В работе была использована трехслойная нейросеть с 10 скрытыми нейронами. Она обучалась по методу устойчивого распространения ошибки с коэффициентами уменьшения и увеличения шага, равными соответственно 0,5 и 1,2. Таким образом, была получена нейросетевая модель зависимости плотности углеводородов от их структуры и температуры со следующими параметрами (см. также Рис.

54): число итераций 9950, $R = 0.9977$, $RMSE_t = 0,0063$, $RMSE_v = 0,0063$, где R – коэффициент корреляции между предсказанными и экспериментальными значениями, $RMSE_t$ и $RMSE_v$ – абсолютные среднеквадратичные ошибки для обучающей и контрольной выборок (г/мл).

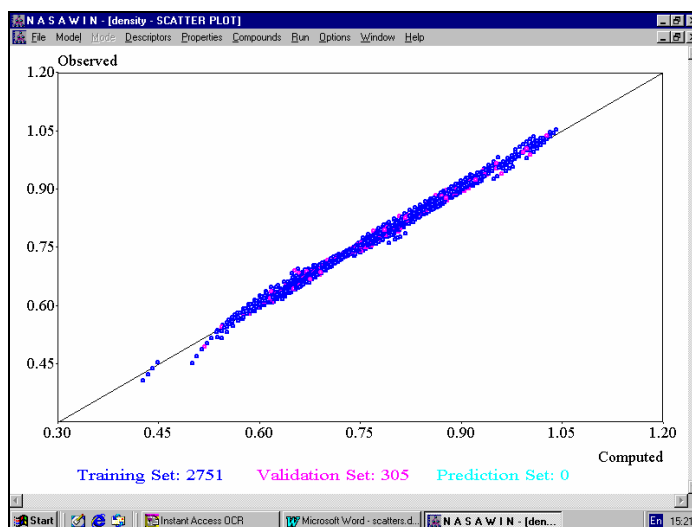


Рис. 54. Результаты нейросетевого моделирования зависимости структура – температура - плотность

Таким образом, нейросетевая модель позволяет определять и прогнозировать плотность углеводов произвольной структуры при произвольных значениях температуры с высокой степенью точности.

7.2.2.3. Моделирование зависимости «структура - температура – динамическая вязкость»

При построении нейросетевой модели были использованы значения динамической вязкости (в сантипуазах) углеводов разнообразной структуры в температурном интервале от -180 до 300°C [467, С. 136 - 159]. Для всех структур из исходной выборки были выделены фрагменты с максимальной длиной, равной 4 атомам. Все рассчитанные фрагментные дескрипторы и значения температур затем были подвергнуты модификациям «квадрат величины», «логарифм величины», «обратная величина» и «величина, деленная на количество неводородных атомов в молекуле». Значения динамической вязкости углеводов были прологарифмированы.

Объединенный набор дескрипторов, состоящий из полученных в результате модификаций значений наряду с исходными значениями дескрипторов и температур, был подвергнут процедуре отбора наиболее существенных параметров с помощью метода БПМЛР (см. подраздел 4.1.5). В результате отбора осталось 307 дескрипторов. Исходная выборка соединений, содержащая 3426 записей, была разбита случайным образом на обучающую (3084 записи) и контрольную (342 записи) подвыборки.

В работе была использована трехслойная нейросеть с 10 скрытыми нейронами. Она обучалась по методу устойчивого распространения ошибки с коэффициентами уменьшения и увеличения шага, равными соответственно 0,5 и 1,2. Таким образом, была получена нейросетевая модель зависимости динамической вязкости углеводородов от их структур и температуры со следующими параметрами (см. также Рис 55): число итераций 10387, $R = 0,9949$, $S_t = 0,1411$, $S_v = 0,1642$, где R – коэффициент корреляции между спрогнозированными и экспериментальными значениями, $RMSE_t$ и $RMSE_v$ – абсолютные среднеквадратичные ошибки для обучающей и контрольной выборок (натуральный логарифм значения динамической вязкости).

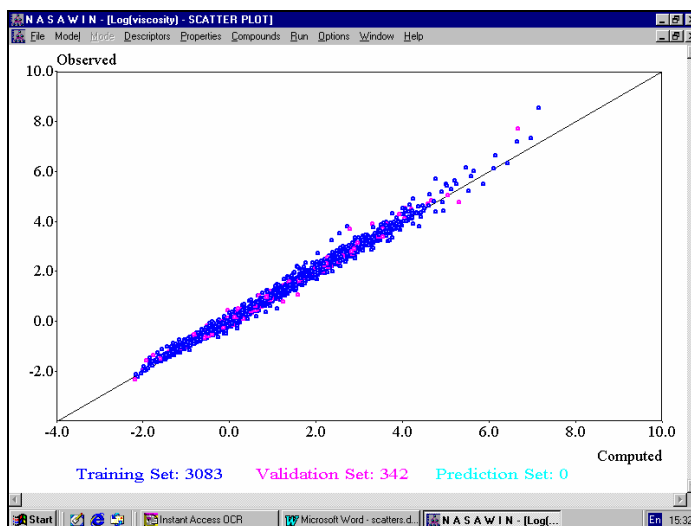


Рис. 55. Результаты нейросетевого моделирования зависимости структура – температура - логарифм динамической вязкости.

Таким образом, нейросетевая модель позволяет определять и прогнозировать динамическую вязкость углеводородов произвольной структуры при произвольных значениях температуры с высокой степенью точности.

7.2.3. Построение и анализ нейросетевых зависимостей «структура – условия реакции – константы скорости» для реакции кислотного гидролиза сложных эфиров карбоновых кислот

Кинетика и механизмы кислотного гидролиза сложных эфиров карбоновых кислот уже давно привлекают внимание исследователей. На базе именно этой реакционной серии Гамметтом было предложено уравнение для количественного описания электронных эффектов заместителей для пара- и мета-замещенных фенилов. Позднее Тафт использовал эту серию для описания стерических эффектов заместителей. К настоящему времени накоплен большой объем данных по константам скоростей реакций и их зависимости от температуры и параметров среды для широкого спектра структур сложных эфиров [397]. В литературе [397] были предложены следующие возможные механизмы реакции гидролиза - ацильный ($A_{Ac}2$) и алкильный ($A_{Alk}1$) (Рис. 56).

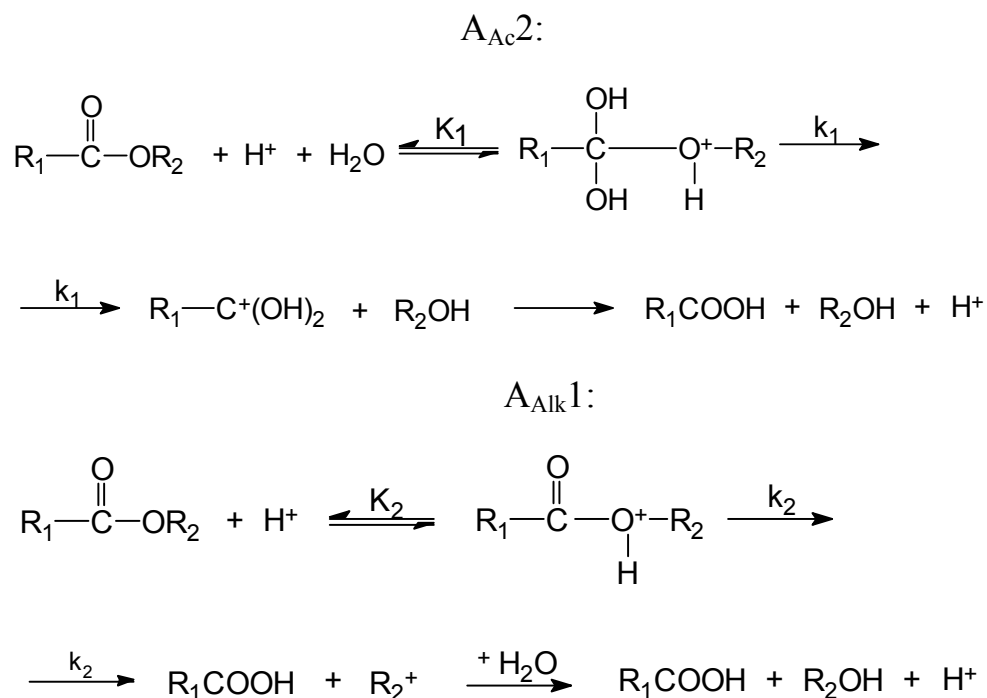


Рис. 56. Возможные механизмы реакции кислотного гидролиза сложных эфиров карбоновых кислот

Оба механизма включают 2 стадии: 1-ая стадия - равновесная, 2-ая стадия - необратимая. На первой обратимой стадии реакции происходит протонирование эфирного атома кислорода, а на 2-ой стадии – гетеролитический разрыв связей. Предложенные механизмы различаются типом разрыва связей: для ацильного механизма характерен разрыв связи между карбонильным углеродом и эфирным кислородом, в то время как при протекании реакции по алкильному механизму разрыв происходит по связи эфирный кислород – атом углерода эфирного остатка. В обоих механизмах присутствует стадия гидратации, но для ацильного механизма гидратация происходит на стадии образования реакционного комплекса, а для алкильного механизма – на стадии, следующей за распадом реакционного комплекса.

Общая скорость процесса зависит от исходной концентрации сложного эфира и значения pH. В качестве катализатора выступает соляная кислота. Константы скоростей для предложенных механизмов вычисляются следующим образом:

$$A_{Ac2}: \quad k_{AB} = K_1 k_1 a_{H_2O}$$

$$A_{Alk1}: \quad k_{AB} = K_2 k_2$$

Оба рассмотренных механизма вносят свой вклад в протекание реакции, и описываются суммарной схемой: $k_{AB} = K_1 k_1 a_{H_2O} + K_2 k_2$, где: $A = R^1COOR^2$; $B = H^+$; k_{AB} – наблюдаемая константа скорости реакции; K_1 и K_2 – константы равновесия; k_1 и k_2 – константы скорости для элементарных стадий реакции.

Несмотря на большой объем экспериментальных данных, ранее были предприняты лишь разрозненные попытки линейного моделирования констант скоростей гидролиза эфиров для узких серий соединений при постоянных условиях проведения реакции. Для такого моделирования с успехом может быть применено уравнение Гаммета. Так, с помощью этого уравнения были описаны константы скоростей кислотного гидролиза для серии бензгидрил-*p*-нитробензоатов в среде вода-ацетон [472]. Каждое уравнение строилось для серии соединений (число соединений варьировалось от 4 до 7) при заданной температуре и концентрации ацетона. Полученные коэффициенты корреляции ле-

жали в интервале от 0.898 до 0.999. Уравнение Гамметта также было успешно применено для описания констант гидролиза серии фенилбензоатов с различными заместителями в орто- и пара- положениях обоих бензольных колец [473] и для серии зависимостей констант скоростей щелочного гидролиза для 5 водных растворов фенилтрифторацетатов с различными заместителями в бензольном кольце ($R^2 = 0.797 - 0.991$); каждому уравнению соответствовало постоянное значение рН из интервала от 5.00 до 9.91.

Интересной представляется работа, в которой приводится корреляция констант гидролиза с химическими сдвигами ^{13}C для серии из 8 замещенных фенилдихлорацетатов [474]. Гидролиз проводился в 20%-ном водном растворе ацетонитрила при температуре 298.2К. Для одного соединения, которое не участвовало в построении модели, было рассчитано значение логарифма константы скорости гидролиза, хорошо согласующееся с экспериментальными данными.

Нас заинтересовала возможность обобщения накопленной информации о кинетике гидролиза сложных эфиров карбоновых кислот. Мы использовали данные [397] по константам скорости реакции при различных температурах и в различных бинарных смесях вода-растворитель, а также в чистой водной среде для нейросетевого моделирования. В данном случае нейросетевое моделирование представляется нам наиболее подходящим инструментом, так как позволяет получать нелинейные многопараметровые модели «структура – свойство», для которых не всегда заранее известен вид аналитической зависимости.

Так как нашей задачей было построение модели для набора очень разнородных соединений, то для описания химических структур исследуемых соединений мы решили воспользоваться не экспериментальными константами заместителей, полученными для конечного числа соединений, а расчетными величинами. Для этой цели нами был рассчитан ряд локальных и глобальных квантово-химических дескрипторов с использованием дескрипторного блока «QUANT».

Для описания условий реакции мы использовали в качестве дескрипторов значения температуры реакции и концентрации органического компонента рас-

творителя, а также параметры, предложенные В.А.Пальмом [475, С. 106] для описания эффектов реакционной среды и основанные на допущении, что всю специфическую сольватацию можно свести к образованию водородных или аналогичных им донорно-акцепторных связей между молекулами растворителя и растворенного вещества и описать двумя параметрами - общей кислотностью (электрофильностью) (E) и общей основностью (нуклеофильностью) (B) растворителя. Неспецифическая сольватация, в свою очередь, может быть описана двумя независимыми свойствами среды - полярностью (Y) и поляризуемостью (P). Перечисленные выше характеристики имеют следующее физическое выражение:

$$Y = (\varepsilon - 1) / (2\varepsilon + 1), \text{ где } \varepsilon\text{-диэлектрическая проницаемость;}$$

$$P = (n^2 - 1) / (2n^2 + 1), \text{ где } n\text{ – показатель преломления.}$$

Шкала значений общей кислотности (E) получена исходя из величин сольватохромных сдвигов, выраженных в энергетической шкале π - π^* -перехода N-[(3,5-дифенил-4-окси)фенил]пиридиний-бетаинов. Шкала значений общей основности (B) рассчитана как разница ИК-частот колебаний ОН-группы связанного с основанием и свободного фенола в среде CCl_4 .

Экспериментальные данные для констант скорости реакции гидролиза были измерены в следующих средах [397, С.7-85]: вода; вода-метанол; вода-этанол; вода-этиленгликоль; вода-ацетон; вода - 1,4-диоксан; вода – диметилсульфоксид; вода –глицерин. Так как мы обрабатывали данные только для водных растворов, то в качестве параметров среды брались лишь значения для неводного компонента смеси.

В настоящей работе нами был использован массив данных, содержащий 2092 записи. Каждой записи соответствовало значение логарифма наблюдаемой константы скорости k_{AB} реакции гидролиза, 4 дескриптора для характеристик растворителя, по одному дескриптору для температуры и молярной концентрации бинарного растворителя, а также набор из 114 дескрипторов, включающий 86 глобальных и 24 локальных квантово-химических дескриптора, описывающих структуру эфира. Для расчета локальных дескрипторов был выбран мак-

симальный общий для всех соединений фрагмент (Рис. 57). Атомы этого фрагмента были помечены, и на каждом атоме вычислялся набор из 6 дескрипторов.

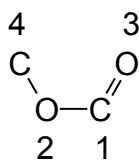


Рис. 57. Максимальный общий фрагмент для структур сложных эфиров.

Весь массив записей был разбит случайным образом на обучающую (содержащую 1883 соединения) и контрольную (содержащую 209 соединений) выборки. Нейросетевое моделирование было проведено с использованием компьютерной программы NASAWIN (см. раздел 8.2). Была использована трехслойная нейросеть с 10 скрытыми нейронами. Нейросеть обучалась по обобщенному дельта-алгоритму со скоростью обучения 0.25 и моментом, равным 0.9. Прогнозирующая способность нейросети оценивалась по величине среднеквадратичной ошибки для записей из контрольной выборки. Обучение прекращалось в момент наступления переучивания.

В результате обучения нейросети была получена модель со следующими параметрами: число итераций 64824, $R = 0.9669$; $RMSE_t = 0.2710$; $RMSE_v = 0.3417$; где R – коэффициент корреляции, $RMSE_t$ и $RMSE_v$ – соответственно ошибки для обучающей и контрольной выборок (логарифм наблюдаемой константы скорости).

Далее для каждого входного нейрона для обученной нейросети были рассчитаны характеристики, введенные нами для интерпретации нейросетевых моделей (см. раздел 4.2). Полученные величины для наиболее важных дескрипторов приведены в Табл. 29.

Табл. 29. Характеристики значимости основных дескрипторов

Дескриптор	M_x	D_x
Температура	868.687	371.282
LocalCharge4	747.253	141.944
LocalLUMODensity3	663.265	232.108
LocalSuperEleDeloc3	-542.794	223.966
Поляризуемость (P)	341.332	94.895
Кислотность (E)	124.892	102.700
Основность (B)	-111.211	29.624
Полярность (Y)	65.212	65.050
Концентрация органического компонента в бинарном растворителе в смеси с водой	-64.999	35.126

Анализ таблицы значимостей дескрипторов показывает, что наибольшее влияние на константу скорости оказывает величина температуры реакции. Повышение температуры приводит к ускорению реакции, а увеличение концентрации неводного компонента растворителя снижает скорость гидролиза. Аналогичный эффект влияния концентрации бинарного растворителя уже отмечался ранее [476]. Этот факт можно объяснить замедлением первой стадии реакции – стадии протонирования – при уменьшении в растворе концентрации протонирующих ионов.

Из четырех параметров, описывающих влияние растворителя, наибольший вклад вносит значение поляризуемости растворителя (P); далее идут значения общей кислотности (E), общей основности (B) и полярности (Y), причем вклад поляризуемости, кислотности и полярности положителен, в то время как вклад основности отрицателен. Таким образом, можно сделать заключение о том, что растворитель, характеризующийся большими значениями поляризуемости и кислотности, облегчает прохождение 2-ой необратимой стадии реакции, заключающейся в разрыве C-O связи и способствует стабилизации получающегося в результате этого разрыва карбонил-иона.

Анализ квантово-химических параметров показал значительное положительное влияние величины локального заряда на α -атоме углерода заместителя R_1 (LocalCharge4), что свидетельствует о преобладании механизма A_{Alk1} и хорошо согласуется с опубликованными данными о механизмах гидролиза эфиров. Также существенно влияние величины электронной плотности нижней незанятой орбитали (LocalLUMODensity3), рассчитанной для карбонильного атома кислорода (вклад положителен), и величины индекса электрофильной суперделокализуемости (LocalSuperEleDeloc3) для этого же атома (вклад отрицателен), что может служить доказательством присутствия ацильного механизма гидролиза.

Таким образом, благодаря использованию аппарата искусственных нейронных сетей оказывается возможным предсказывать константы скоростей кислотного гидролиза сложных эфиров достаточно произвольного строения при произвольной температуре и составе растворителя, а также проанализировать полученную зависимость. Результаты проведенных исследований демонстрируют возможность применения предложенного нами подхода к количественному моделированию реакционной способности органических соединений.

7.3. Индуктивный перенос знаний при интеграции моделей «структура-свойство»

В настоящее время развитие методологии построения моделей «структура-свойство/структура-активность» (QSPR/QSAR) по пути совершенствования дескрипторного описания химических соединений и применения все более совершенных методов анализа данных вошло в стадию насыщения и достигло того уровня, когда существующими методами из базы данных удается извлечь практически всю информацию, полезную для прогнозирования. Как отмечается в работе [477], в большинстве случаев прогнозирующая способность моделей, построенных с использованием «достаточно хороших» наборов дескрипторов и «достаточно хороших» методов анализа данных, уже очень слабо зависит и от

набора дескрипторов и от применяемого метода, а практически целиком определяется базой данных, использованной для построения модели. Таким образом, дальнейшее совершенствование дескрипторного описания химически соединений и внедрение все более новых методов машинного обучения способно будет привести лишь к очень незначительным успехам, а для настоящего прорыва в этом направлении требуется выработка принципиально новых идей, которые позволили бы преодолевать ограничения, связанные с недостаточным объемом содержащейся в химических базах данных полезной информации.

Между тем известно, что имеется принципиальная разница между методами машинного обучения и теми способами, которыми пользуется при обучении человек [477]. Если при машинном анализе данных для надежного построения сколько-нибудь сложной статистической модели требуется очень большой объем данных, то для человека для обучения значительно более сложным концепциям требуется удивительно мало примеров. Одна из причин этого заключается в том, что в настоящее время при машинном анализе данных каждая новая статистическая модель строится практически «с нуля», и получаемые таким образом модели оказываются изолированными друг от друга. Человек же, решая какую-нибудь задачу, всегда опирается на опыт, полученный при решении других задач. При освоении даже принципиально нового материала человек всегда пользуется аналогиями и метафорами, взятыми из ранее усвоенных знаний. Наконец, компоненты полученного знания тесно переплетены между собой в человеческом мозгу, что многократно ускоряет и облегчает процесс получения нового знания. Осознание этого привело в последние годы к формированию нового направления в теории машинного обучения, условно называемого “индуктивным переносом знаний”, которое занимается изучением того, как связывание между собой различных задач анализа данных приводит к улучшению качества получаемых моделей [477].

Таким образом, один из путей преодоления ограничений, связанных с недостаточным объемом содержащихся в отдельных химических базах данных информации, видится в том, чтобы рассматривать разнообразные свойства химических соединений в их тесной взаимной связи и с учетом этого строить мо-

дели «структура-свойство» не изолированными, а связанными друг с другом. Можно ожидать, что в этом случае будет происходить интеграция данных, при которой объем полезной информации для каждого из свойств будет существенно увеличен за счет эффективного использования информации, касающейся других свойств, тесно с ним связанным. Также можно предположить, что чем меньше экспериментальных данных имеется по данному свойству и чем больше экспериментальных данных имеется по связанным с ним другим свойствам, тем более эффективно будет происходить перенос необходимой информации при построении модели для прогнозирования этого свойства. Такой перенос информации возможен между моделями, расположенными внутри сети взаимосвязанных моделей как последовательно (см. раздел 7.4.1), так и параллельно друг относительно друга (см. раздел 7.4.2).

Можно предвидеть, что в перспективе развития методологии QSPR/QSAR место разрозненных и независимых друг от друга одноуровневых моделей «структура-свойство»/«структура-активность» займет организованная в виде «химического мозга» сеть тесно связанных между собой моделей, позволяющая интегрировать внутри себя значительный объем как экспериментальных данных, так и теоретических знаний, что позволит значительно улучшить качество прогнозирования разнообразных свойств химических соединений.

7.3.1. Многоуровневый принцип построения моделей «структура-свойство»

Суть предлагаемого нами многоуровневого подхода к прогнозированию свойств органических соединений в рамках методологии QSAR/QSPR заключается в следующем. Прогнозирование свойств органических соединений проводится в рамках фрагментного подхода [110, 116]. Это дает возможность воспользоваться всеми такими преимуществами фрагментного подхода как быстрота и однозначность вычислений, а также естественный характер интерпретации моделей на языке элементов структурных формул органических соединений. Кроме того, благодаря своему базисному характеру, фрагментные дескрипторы должны обеспечить возможность аппроксимировать любые сколь

угодно сложные зависимости «структура-свойство». В то же время, вместо изолированных одноуровневых моделей, берущих на входе значения фрагментных дескрипторов и выдающей на выходе значения прогнозируемых свойств, предлагается использовать организованную в виде нескольких слоев сеть моделей, в которой выходы моделей предыдущих слоев являются входами для моделей последующих. Заметим, что подобная организация моделей напоминает поэтапный процесс обработки информации, происходящий в многослойных структурах коры головного мозга. От каждой из промежуточных моделей требуется, чтобы на выходе они давали либо экспериментально измеряемые величины, либо расчетные величины, имеющие очевидную интерпретацию. Это дает возможность для каждой промежуточной модели использовать свою базу данных «структура-свойство», которая и должна применяться для ее построения. В этом случае многоуровневая организация моделей дает возможность эффективно проводить индуктивный перенос знаний от моделей предыдущего слоя к моделям последующего, что должно приводить к улучшению качества последних за счет использования дополнительной информации, взятой в неявном виде из других баз данных. Можно предположить, что для эффективности этого процесса необходимо, чтобы модели предыдущего уровня обучались на базах существенно большего размера, чем последующего. На Рис. 58 показана схема традиционного одноуровневого подхода, основанного на т.н. «однозадачном обучении», при котором модели (в данном случае нейросетевые) для прогнозирования разных свойств не связаны друг с другом. В противоположность этому, на Рис. 59 (стр. 266) показана схема многоуровневого подхода, в рамках которого за счет последовательного соединения моделей происходит перенос информации из моделей нижнего уровня в модели верхнего, что приводит к повышению предсказательной способности последних.

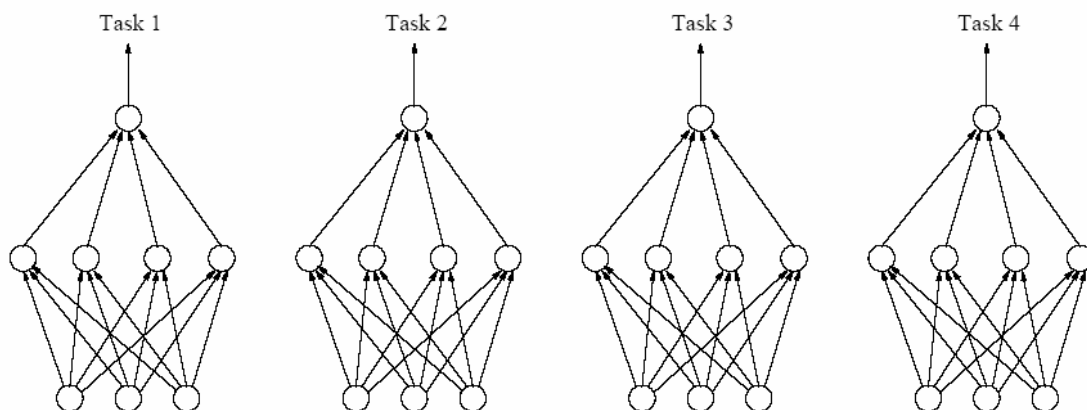


Рис. 58. Традиционный одноуровневый подход, в котором отдельные нейросетевые модели не связаны друг с другом

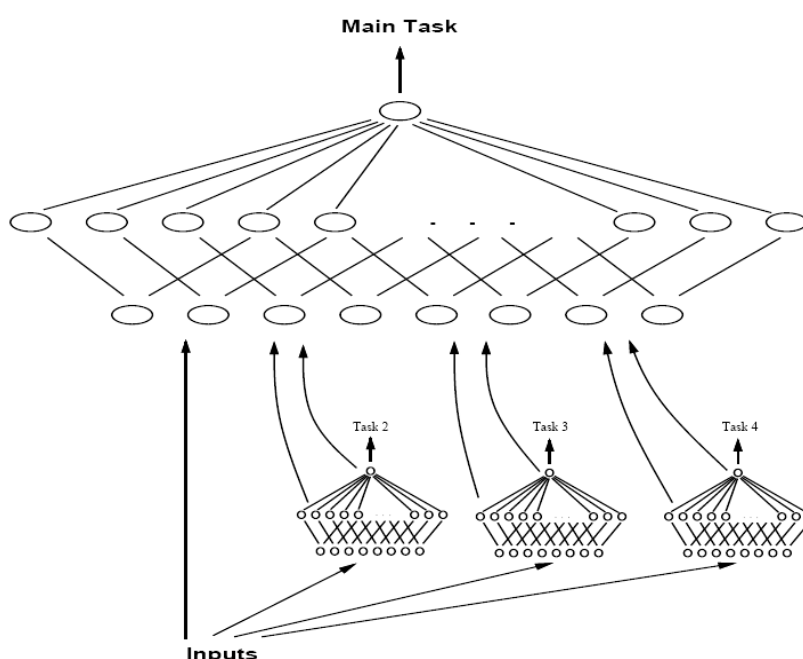


Рис. 59. Схема многоуровневого подхода, в рамках которого за счет последовательного соединения моделей происходит перенос информации из моделей нижнего уровня в модели верхнего

Естественными кандидатами на роль выходных свойств для промежуточных моделей являются физико-химические свойства, связанные с фундаментальными типами взаимодействий (гидрофобность, поляризуемость, характеристики силы водородных связей и т.д.), разнообразные константы заместителей, а также квантово-химические характеристики (ВЗСО, НСМО, заряды на атомах). Заметим, что для большинства из этих величин имеются дескрипторы, которые уже давно успешно используются при построении количественных зави-

симостей «структура-свойство». Принципиальным отличием и преимуществом многоуровневого подхода перед непосредственным использованием для построения моделей физико-химических и квантово-химических дескрипторов является то, что при этом не теряется интерпретируемость моделей через фрагментные дескрипторы на языке структурных формул. Кроме того, сохраняется свойственная фрагментным дескрипторам универсальность и эффективность расчета, что дает возможность использовать многоуровневые сети моделей при высокопроизводительном виртуальном скрининге.

Следует отметить, что кроме улучшения качества прогноза, многоуровневый подход способен преодолеть то, что иногда называется недостатками фрагментного подхода, а именно отсутствие физико-химической интерпретации и проблема “отсутствующих фрагментов” [116]. Прежде всего, благодаря тому, что промежуточные модели дают на выходе экспериментально измеримые или легко интерпретируемые физические величины, сама конечная модель получает очевидную физико-химическую интерпретацию в терминах этих величин. Для такой интерпретации при использовании нейросетевых моделей может быть использован подход, предложенный нами ранее [478]. Что же касается «отсутствующих фрагментов», которые отсутствуют в обучающей выборке но присутствуют в тестовой, то острота этой проблемы смягчается благодаря тому, что эти фрагменты имеют шансы присутствовать в химических структурах, входящих в выборки существенно большего размера, используемые для обучения моделей предыдущих слоев.

Рассмотрим два примера, показывающие преимущества использования многоуровневого подхода. В первом случае на основе опубликованных данных [479] была сформирована выборка 1, содержащая количественные данные по значению логарифма коэффициента сорбции в почве ($\log K_{oc}$) для 568 органических соединений. Во втором случае для создания выборки 2 были взяты из статей [479, 480] данные по значению логарифма растворимости ($\log S$) фуллерена C_{60} в 113 органических растворителях, включая 45 алканов, 36 производных бензола, 7 производных нафталина, 14 кислород, 21 хлор и 15 бромсодержащих соединений. При построении количественных моделей «структура-свойство» в

рамках одноуровневого подхода для описания химических соединений были использованы наборы фрагментных дескрипторов [481] размером до шести неводородных атомов. Предварительный отбор дескрипторов проводился по методу быстрой пошаговой множественной линейной регрессии (БПМЛР) [482]. Отобранные наборы дескрипторов использовались для построения нейросетевых моделей «структура-свойство» при помощи многослойных персептронов [39]. При построении моделей в рамках двухуровневого подхода были точно таким же образом с применением фрагментных дескрипторов и комбинации БПМЛР и многослойных персептронов модели первого уровня, позволяющие прогнозировать значения липофильности $\log P$ и четырех констант Абрахама A , B , E и S , характеризующих, соответственно, кислотность и основность по отношению к образованию водородной связи, избыточную молярную рефракцию и дипольность/поляризуемость. Для построения модели для липофильности была использована выборка 3, включающая 7805 соединений [483], а для констант Абрахама – выборка 4, состоящая из 457 соединений и приведенная в работе [484]. В Табл. 30 представлены статистические характеристики моделей первого уровня. На втором этапе результаты прогноза, полученные с помощью моделей первого уровня для соответствующих выборок органических соединений по логарифму коэффициента сорбции в почве и логарифма растворимости фуллерена C_{60} , были использованы в качестве дескрипторов при построении нейросетевых моделей второго уровня для расчета этих свойств. В всех случаях для оценки прогнозирующей способности моделей была применена процедура двойного 5x4-кратного скользящего контроля [482]. Построение QSPR-моделей осуществляли с помощью программного комплекса NASAWIN [194]. Значения параметра Q^2_{DCV} и среднеквадратичной ошибки прогноза $RMSE_{DCV}$ для моделей, полученных с использованием одноуровневого и многоуровневого подходов для расчета логарифма коэффициента сорбции органических соединений в почве и логарифма растворимости фуллерена C_{60} , приведены в Табл. 31 на стр. 270. Как видно из представленного материала, прогнозирующая способность QSPR моделей, полученных в рамках многоуровневого подхода, значительно превышает прогнозирующую способность одноуровневых моде-

лей, хотя все модели построены на основе одинаковых наборов фрагментных дескрипторов при помощи одного и того же метода машинного обучения. Диаграммы экспериментальных и рассчитанных значений $\log K_{oc}$ и $\log S$, полученных на основе нейросетевых моделей, построенных с использованием многоуровневого подхода, представлены на Рис. 60 на стр. 270.

Табл. 30. Статистические характеристики моделей ‘структура/свойство’ первого уровня для расчета липофильности и констант Абрахама для органических соединений, соответственно включенных в выборки 3 и 4

Свойство	Число соединений в выборке	Коэффициент корреляции	<i>RMSE</i> на обучающей выборке	<i>RMSE</i> на контрольной выборке (1/10 выборки)
Log P	7805	0.980	0.345	0.395
Абрахам А	457	0.983	0.051	0.058
Абрахам В	457	0.971	0.066	0.081
Абрахам Е	457	0.997	0.040	0.074
Абрахам S	457	0.987	0.072	0.137

Преимущество использования многоуровневого подхода продемонстрировано нами также на примере прогнозирования констант устойчивости комплексов циклодекстрина с органическими молекулами [400]. Таким образом, объединение в сеть всего лишь нескольких моделей может привести к заметному улучшению прогнозирующей способности моделей более высокого уровня за счет использования информации, содержащейся в дополнительных базах данных, использованных при построении моделей более низкого уровня. Есть основания считать, что многоуровневый подход может дать значительный эффект не только при прогнозировании физико-химических свойств, как было показано на двух примерах в рамках данного подраздела, но и биологической активности.

Табл. 31. Сравнительные статистические характеристики моделей ‘структура-свойство’, для расчета логарифма коэффициент сорбции органических соединений в почве (выборка 1) и растворимости фуллеренов C₆₀ в органических соединениях (выборка 2), полученных в рамках одноуровневого и многоуровневого подходов QSPR/QSAR

Свойство	Одноуровневый подход		Многоуровневый подход	
	Q^2_{DCV}	$RMSE_{DCV}$	Q^2_{DCV}	$RMSE_{DCV}$
Логарифм коэффициента сорбции в почве	0.598	0.759	0.800	0.534
Логарифм растворимости фуллерена C ₆₀	0.448	0.912	0.637	0.739

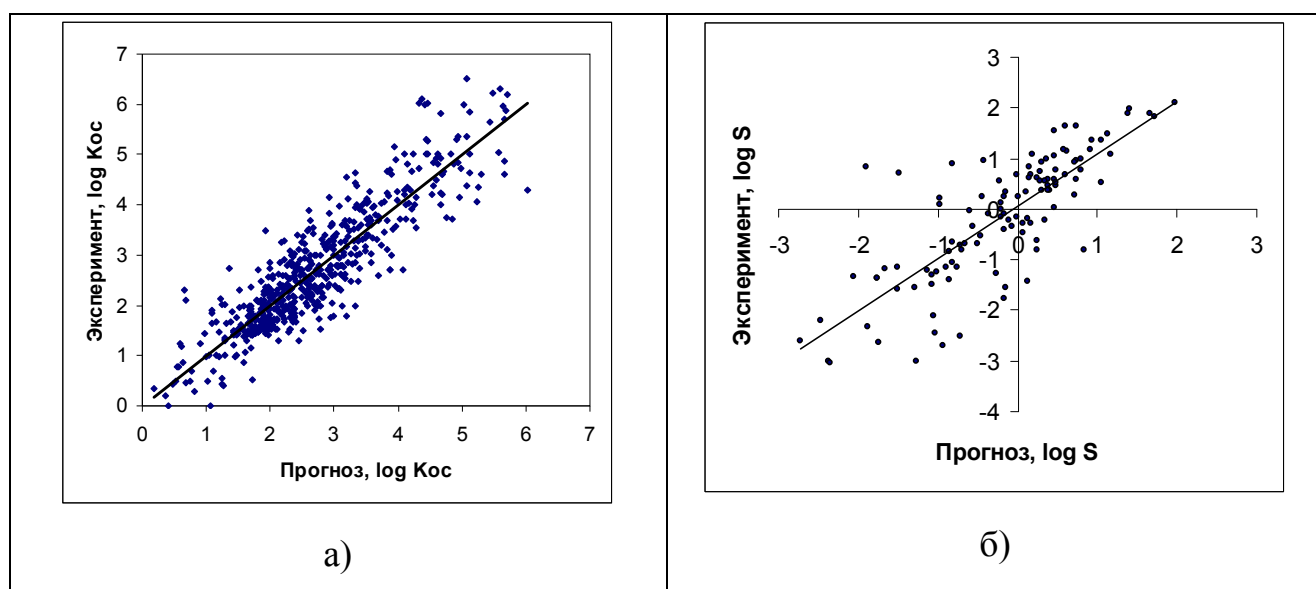


Рис. 60. Диаграммы разброса экспериментальных и рассчитанных значений: (а) логарифма коэффициента сорбции в почве (log Koc) и (б) логарифма растворимости фуллерена C₆₀ (log S), полученных на основе нейросетевых QSPR моделей, построенных с использованием многоуровневого подхода для выборок 1 и 2, включающих, соответственно, 568 и 113 органических соединений.

7.3.2. Параллельный принцип построения моделей «структура-свойство». Многозадачное обучение.

Многозадачным называется такой вид обучения, когда проводится одновременное построение моделей, связь между которыми осуществляется за счет использования общих промежуточных данных. Многозадачное обучение может быть, например, осуществлено при помощи нейросети обратного распространения (см. подраздел 1.2.4), имеющей несколько выходных нейронов по числу

одновременно решаемых задач, связь между которыми осуществляется за счет совместного использования промежуточных данных, формируемых на общем для этих задач скрытом слое нейронов (см. Рис. 61). Это резко отличается от традиционного однозадачного обучения, когда задачи по построению моделей решаются полностью независимо друг от друга (см. Рис. 58 на стр. 266).

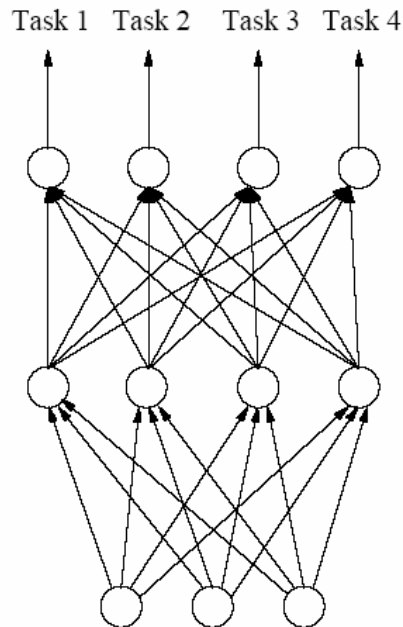


Рис. 61. Многозадачное обучение, при котором проводится одновременное построение взаимосвязанных моделей. Обмен информацией между моделями происходит за счет формирования единого внутреннего представления данных в общем слое скрытых нейронов

Впервые термин «многозадачное обучение» был введен в математическую литературу Р. Каруаной (R. Caruana), который в середине 90-ых годов провел первые систематические исследования в этом направлении [485]. В частности, им было показано, что использование многозадачного обучения приводит к улучшению прогнозирующей способности статистических моделей в том случае, если они являются взаимосвязанными [485]. Следует, однако, подчеркнуть, что само понятие взаимосвязанности моделей в данном случае не имеет ничего общего с фактом наличия корреляции между выходными свойствами: условием взаимосвязанности моделей является хотя бы частичное наличие общих или скоррелированных промежуточных данных, тогда как корреляция между выходными значениями может отсутствовать. В частности, линейно-

регрессионные модели, построенные с использованием одних и тех же входных данных для разных выходов, не считаются взаимосвязанными даже при наличии сильной корреляции между выходными данными, поскольку при их построении не формируется общее для них представление данных. Вследствие этого для множественной линейной регрессии многозадачное обучение эквивалентно однозадачному. В то же время, нейросети обратного распространения, благодаря наличию промежуточного слоя скрытых нейронов, оказываются способными реализовывать многозадачное обучение, осуществляя тем самым более глубокую обработку и интеграцию данных.

Впервые принципиальная возможность построения взаимосвязанных моделей «структура-свойство» была, однако, продемонстрировано нами еще в 1993 г. на примере искусственной нейронной сети с семью выходами, которая способна была одновременно предсказывать семь физических свойств алканов (см. раздел 6.1). Поскольку наше исследование было проведено еще до появления вышеупомянутых первых математических работ по многозадачному обучению, тогда нами не было предпринято систематическое изучение того, какой эффект дает одновременное прогнозирование нескольких свойств нейросетью с несколькими выходами по сравнению с их прогнозированием изолированными нейросетями с одним выходом. Подобное систематическое изучение было, однако, предпринято в нашей недавней работе по прогнозированию 11 констант распределения «ткань-воздух» [477], которая была осуществлена совместно с А.Варнеком, С.Годеном и Ж.Марку из лаборатории хемоинформатики Университета им. Л.Пастера (г. Страсбург, Франция) и И.Тетко и Анил Кумар Пандеем Центра им. Гельгольца (Мюнхен, Германия). В этом исследовании для построения моделей был использован ансамбль нейросетей обратного распространения, реализованный в рамках программы ASNN [342] (а также метод PLS) и фрагментные дескрипторы. В Табл. 32 на стр. 273 для каждого сочетания типа ткани и организма приведен размер выборки, а также значения Q^2 и MAE (средняя абсолютная ошибка), полученные в результате однозадачного (11 нейросетей с одним выходом) и многозадачного (одна нейросеть с 11 выходами) обучения.

Как видно приведенных в таблице данных, во всех случаях, когда имеется лишь небольшой объем экспериментальных данных, применение многозадачного обучения приводит к существенному улучшению прогнозирующей способности при недостатке экспериментальных данных. Эта тенденция особенно хорошо видна на Рис. 62, на котором показан тренд зависимости увеличения показателя Q^2 при переходе к многозадачному обучению от размера выборки. На приведенной диаграмме четко видно, что при размере выборки меньше 90 соединений применение многозадачного обучения приводит к заметному росту прогнозирующей способности, которое происходит за счет неявного переноса информации, использованной для построения моделей для связанных с ними свойств, для которых выборки содержат почти 100 и больше соединений. Для этих же последних свойств применение многозадачного обучения не приводит ни к какому статистически значимому эффекту.

Табл. 32. Статистические характеристики нейросетевых моделей, полученных при однозадачном и многозадачном обучении для констант распределения «ткань-воздух»

Ткань / организм	Число соединений	Однозадачное обучение		Многозадачное обучение	
		Q^2	<i>MAE</i>	Q^2	<i>MAE</i>
Жир человека	42	0.20	0.46	<u>0.57</u>	<u>0.32</u>
Мозг человека	35	0.48	0.48	<u>0.59</u>	<u>0.35</u>
Печень человека	30	0.20	0.38	<u>0.55</u>	<u>0.27</u>
Почки человека	34	0.23	0.60	<u>0.55</u>	<u>0.35</u>
Мышцы человека	38	0.37	0.55	<u>0.51</u>	<u>0.43</u>
Кровь человека	138	0.66	0.48	<u>0.68</u>	<u>0.42</u>
Жир крысы	99	0.70	0.73	<u>0.73</u>	<u>0.70</u>
Мозг крысы	59	0.25	0.25	<u>0.43</u>	<u>0.43</u>
Печень крысы	100	<u>0.72</u>	<u>0.72</u>	0.67	0.67
Почки крысы	27	0.12	0.12	<u>0.27</u>	<u>0.27</u>
Мышцы крысы	97	<u>0.72</u>	<u>0.72</u>	0.67	0.67

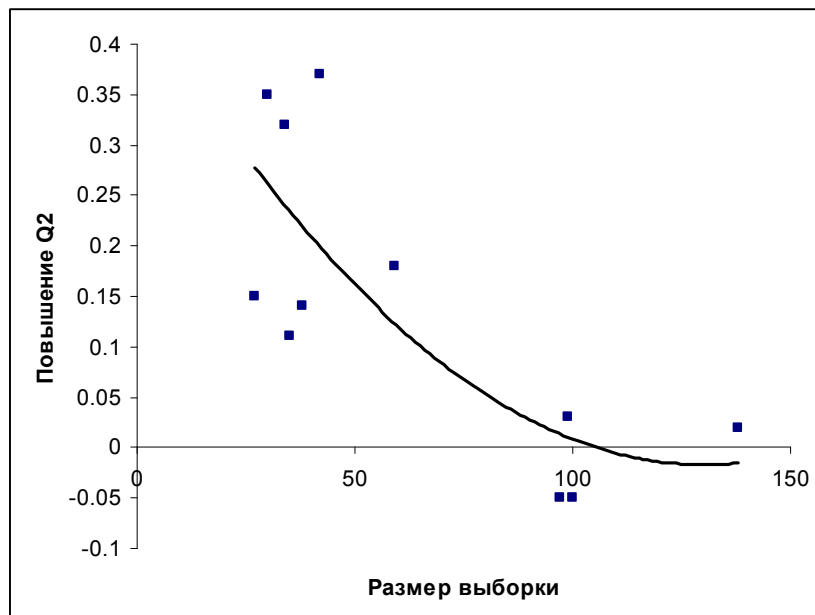


Рис. 62. Повышение Q^2 при переходе к многозадачному обучению в зависимости от размера выборки

Таким образом, при дефиците экспериментальных данных многозадачное обучение приводит к существенному росту прогнозирующей способности моделей «структура-свойство» по сравнению с традиционной методологией построения изолированных моделей.

7.4. Нейронное устройство для проведения прямых корреляций «структура-свойство»

7.4.1. Введение

В настоящее время поиск количественных соотношений между структурами и свойствами органических соединений в значительной мере основан на использовании инвариантов молекулярных графов, базисом которых, как было нами показано выше (см. раздел 3.2), являются ФД. Проблемой, однако, является наличие слишком большого числа ФД, что не дает возможность рассматривать их все в процессе моделирования, В определенной мере процедура БПМЛР (см. подраздел 4.1.5) дает решение этой проблемы за счет предварительного отбора дескрипторов, однако ни одна процедура отбора дескрипторов не может в

принципе гарантировать оптимального решения, поскольку при этом обедняется описание химической структуры. Одним из наиболее перспективных направлений в решении этой проблемы мы видим в том, чтобы вместо отбора дескрипторов из заранее взятого их набора использовать процедуру извлечения непосредственно из структур химических соединений наиболее ценных для моделирования исследуемого свойства дескрипторов*.

Это привело нас к разработке альтернативного подхода к проблеме «структура-свойство», основанного на процедуре поиска зависимости исследуемого свойства непосредственно от элементов матрицы смежности молекулярного графа, однозначно идентифицирующей структуру органического соединения, либо, в более общем случае, от элементов матрицы, описывающей свойства атомов и их пар (например, характеристики связей). В качестве статистического метода анализа зависимости свойств органических соединений от их структуры нами выбран аппарат искусственных нейронных сетей, поскольку с его помощью можно выявлять зависимости между переменными вне рамок каких-либо заранее выбранных моделей. Универсальность аппроксимирующей способности в этом случае обеспечивается промежуточным формированием ФД либо псевдофрагментных дескрипторов в процессе анализа структуры. Принципиальным же отличием от сочетания ИНС с ФД является то, что вместо использования предварительно отобранных дескрипторов, набор которых скорее всего является неоптимальным, происходит направленное «извлечение» наиболее ценных для построения моделей «структура-свойство» дескрипторов непосредственно из первичного описания молекул в виде графа.

Упомянем несколько подходов, связанных с анализом матрицы смежности молекулярного графа при помощи ИНС. Эльрод, Маггиора и Тренари [486, 487] использовали BE-матрицу Уги-Дугуджи [306] для формального представления химической структуры при нейросетевом прогнозировании реакционной

* В настоящее время вокруг решения подобных задач сформировалось специальное направление в теории машинного обучения, называемое «интеллектуальным анализом структурных данных» (structural data mining), и, как частный случай, «интеллектуальным анализом графов» (graph mining). Рассматриваемая в данном разделе работа была нами опубликована раньше появления первых публикаций в этом направлении в математической литературе.

способности органических соединений. Расширенная форма этой матрицы была также использована Вестом при прогнозировании химических сдвигов на ядрах ^{31}P [488]. Для прогнозирования химических сдвигов на ядрах ^{13}C Квасничка использовал специальную нейронную сеть со структурой, отражающую топологию молекул [489]. Нейронная сеть похожей структуры была также использована Вестом при прогнозировании химических сдвигов на ядрах ^{31}P [490]. Во всех этих исследованиях, однако, проводился анализ только локальных свойств (т.е. свойств, отнесенных к определенному атому в молекуле), и эти подходы вряд ли могут быть обобщены на прогнозирование глобальных свойств. Единственным исключением является разработанная Киреевым сеть ChemNet [491], в которой, как и в вышеупомянутых сетях, каждый нейрон соответствует определенному атому в молекуле, а связность нейронов отражает топологию молекул. Хотя эта сеть позволяет прогнозировать, строго говоря, лишь локальные свойства, однако автор принял, что атомный инвариант отражает и свойства молекулы в целом, и поэтому использовал локальные инварианты, вычисляемые нейросетью для определенных атомов в молекуле, для корреляции с глобальными молекулярными свойствами. Тем не менее, полученные в работе корреляции [491] оказались значительно худшего качества по сравнению даже с линейно-регрессионными моделями при использовании топологических индексов в качестве дескрипторов.

7.4.2. Описание нейронного устройства

В настоящей работе мы предлагаем схему нейронного устройства, специально предназначенного для поиска зависимости свойств органических соединений непосредственно от их структур без предварительного вычисления заранее заданного набора инвариантов молекулярных графов. Вместо молекулярных дескрипторов, инвариантных к перенумерации атомов в химической структуре, мы используем числа, описывающие характеристики атомов и атомных пар (в том числе связей) внутри молекулы. В предлагаемом подходе необходимая инвариантность прогнозируемых значений свойства относительно перену-

мерации атомов достигается не за счет предварительного сворачивания матрицы смежности молекулярного графа в набор инвариантов, а благодаря особенностям конструкции нейронного устройства.

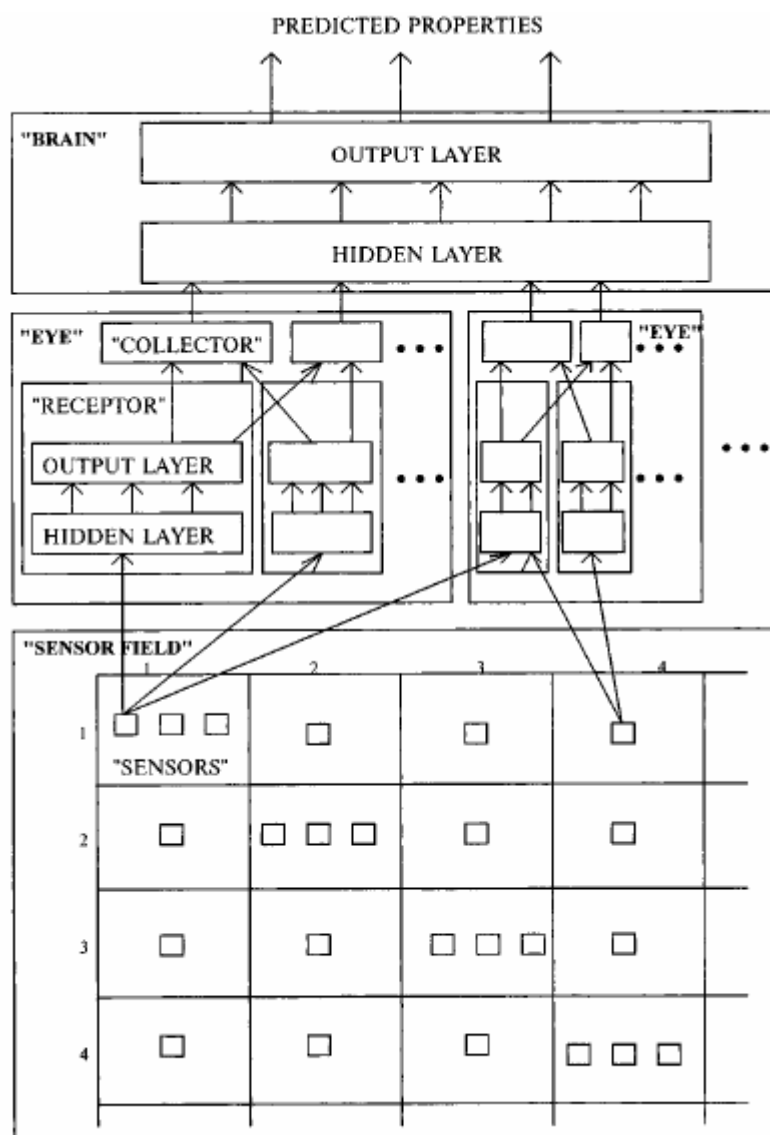


Рис. 63. Архитектура нейронного устройства для проведения прямых корреляций структура-свойство

Как и всякая нейронная сеть, предлагаемое устройство состоит из нейронов (функциональных устройств, осуществляющих преобразования сигналов) и сети связей между ними (синапсов), через которые они посылают друг другу сигналы. Так же, как и в других типах искусственных нейронных сетей, каждый синапс характеризуется числом (весом синапса), на которое умножается сигнал по прохождению через него. Каждый нейрон совершает 2 операции: 1) сумми-

рует все пришедшие к нему сигналы; 2) формирует выходной сигнал путем функционального преобразования полученной суммы:

$$o_i = f(-\theta_i + \sum_j o_j \omega_{ij}), \quad (90)$$

где o_i – выходной сигнал нейрона i , o_j – выходной сигнал нейрона j , ω_{ij} – вес синапса, через который проходит сигнал, θ_i – порог активации нейрона, f – активационная функция, имеющая обычно сигмоидный вид:

$$f(x) = 1/(1 + \exp(-x)). \quad (91)$$

Обучение нейронного устройства заключается в нахождении таких весов синапсов и порогов активации нейронов, чтобы по предъявлению на ее вход сигналов, описывающих химическую структуру, на ее выходе формировались сигналы, соответствующие прогнозируемым значениям свойств этого соединения.

Предлагаемая нами нейросетевая конструкция состоит из трех основных функциональных блоков: 1 – «мозга», 2 – набора «глаз» и 3 – единого «сенсорного поля» (см. Рис. 63). Информация о химической структуре воспринимается «сенсорным полем». Далее она поступает в «глаза» и подвергается обработке, в процессе которой формируется набор сигналов, числовые значения которых не зависят от нумерации атомов в молекуле. Эти сигналы далее поступают в «мозг» для окончательной обработки, на выходе из которого формируются сигналы, соответствующие прогнозируемым значениям свойств для данного химического соединения. Таким образом, как и в традиционном подходе, промежуточно формируется набор инвариантов молекулярных графов, однако эти инварианты не являются заранее заданными и фиксированными, а «подстраиваются» под прогнозируемые свойства в процессе обучения.

Рассмотрим отдельные части этой системы. «Сенсорное поле» представляет собой квадратную матрицу, номера строк и столбцов которой соответствуют номерам атомов в химическом соединении. Размер «сенсорного поля» определяется максимальным числом неводородных атомов в рассматриваемых химических структурах. Сенсоры, расположенные на диагонали матрицы (атомные сенсоры) на пересечении i -ой строки с i -ым столбцом формируют на-

бор сигналов, соответствующих характеристикам i -ого атома в химической структуре. Такими характеристиками могут быть номер строки и столбца соответствующего элемента в Периодической системе Менделеева, заряд атома, число присоединенных к нему атомов водорода, значение электроотрицательности и т.д. Сенсоры, расположенные вне диагонали на пересечении i -ой строки с j -ым столбцом ($i \neq j$) сенсорной матрицы, формируют сигналы, характеризующие отношение между атомами i и j в химической структуре: порядок связи (если эти атомы связаны), вхождение связи в цикл, расстояние (топологическое либо геометрическое) между атомами и т.д.

Каждый «глаз» состоит из: 1) одного или нескольких «коллекторов» и 2) набора идентичных «рецепторов». Сформированные в «сенсорном поле» сигналы поступают на вход «рецепторов». Принципиально важным моментом здесь является то, что все «рецепторы» внутри «глаза» обладают одинаковыми значениями весов связей и порогов активации, т.е. они представляют собой копии одного «рецептора». Каждый из «рецепторов» обрабатывает сигналы «сенсоров», поступающие с небольшого «рецептивного поля», под которым подразумевается часть «сенсорного поля», охватывающая лишь несколько атомов и связей. Внутри «глаза» каждый «рецептор» может быть однозначно идентифицирован упорядоченным вектором $(v_1, v_2, \dots, v_i, \dots, v_n)$, где n – число атомов в рецептивном поле, v_i – порядковый номер соответствующего атома в молекуле. Такой вектор назовем *идентификатором рецептора*. В общем случае число «рецепторов» внутри «глаза» должно быть равно числу способов построения таких векторов $N!/(N - n)!$, где N – число неводородных атомов в химическом соединении, n – число атомов внутри «рецептивного поля» (такие «рецепторы» назовем *n-атомными*). Например, в общем случае трехатомная молекула может быть «воспринята» при помощи трех одноатомных «рецепторов» с идентификаторами (1), (2), (3), шести двухатомных «рецепторов» с идентификаторами (1,2), (2,1), (1,3), (3,1), (2,3), (3,2) и шести трехатомных «рецепторов» с идентификаторами (1,2,3), (1,3,2), (2,1,3), (3,1,2) и (3,2,1). Целиком нейронное устройство со всеми «рецепторами», необходимыми для «восприятия» данной молекулы, образует его *конфигурацию* для нее. Конфигурация с одним «рецепто-

ром» внутри каждого «глаза», содержащая только взаимно независимые подгонные параметры (веса связей и пороги активации нейронов), называется *минимальной*. Минимальная конфигурация может не соответствовать какой-либо конкретной молекуле – ее можно рассматривать как шаблон, с помощью которого можно образовать конфигурацию для любой конкретной молекулы путем размножения рецепторов внутри глаз. Следует отметить, что понятие минимальной конфигурации играет ключевую роль при эмуляции работы данного нейронного устройства на компьютере, поскольку только минимальная конфигурация сети с относительно малым и фиксированным числом нейронов и синапсов может быть размещена в компьютерной памяти и эффективно обработана. При обучении нейронного устройства как только какой-либо настроечный параметр (вес связи либо порог активации) внутри рецептора принимает новое значение, соответствующие параметры во всех других рецепторах внутри этого же «глаза» принимают то же самое значение. Благодаря этому, обучение всего нейронного устройства может быть представлено как минимизация функции ошибки в пространстве подстроечных параметров, относящихся к минимальной конфигурации. Таким образом, минимальной конфигурации достаточно для хранения всех подстроечных параметров нейронного устройства и для его воспроизведения в любой из необходимых конфигураций.

На практике для больших n это число может быть существенно сокращено путем введения фильтров, представляющих собой дополнительное условие на использование «рецепторов» (например, требование наличия внутри рецепторного поля определенной подструктуры). Например, наложение требования наличия внутри «рецептивного поля» определенных подструктур хотя может и увеличить число «глаз» (в соответствии с количеством таких подструктур), но значительно уменьшает число «рецепторов» внутри каждого из них.

Обработанные сигналы со всех «рецепторов» внутри «глаза» накапливаются в «коллекторах», которые определяются как нейроны, суммирующие и, возможно, трансформирующие сигналы, получаемые со всех «рецепторов» внутри «глаза». Таким образом, в поле зрения «глаза» попадает либо все «сенсорное поле», либо, в случае простейшего «глаза», воспринимающего сигналы

только с атомных сенсоров, - диагональ «сенсорного поля». При произвольной перенумерации атомов, при которой атом i получает номер $P(i)$, «рецептор» $(v_1, v_2, \dots, v_i, \dots, v_n)$ получает новый идентификатор $(P(v_1), P(v_2), \dots, P(v_i), \dots, P(v_n))$. Если в глазе присутствуют «рецепторы» со всеми возможными идентификаторами, которые могут быть получены таким образом, то результатом подобной перенумерации станет лишь перестановка «рецепторов» внутри «глаза». Поскольку при суммировании в «коллекторах» сигналов, сформированных «рецепторами», при перестановке слагаемых сумма не меняется, то полная идентичность строения и характеристик всех «рецепторов» внутри «глаза» обеспечивает инвариантность формируемых «коллекторами» сигналов относительно перенумерации атомов в химической структуре.

Каждый «рецептор» внутри нейронного устройства представляет собой многослойную нейронную сеть с обратным распространением ошибки при обучении (т.н. многослойный персептрон), состоящую из одного скрытого и одного выходного слоя. Число скрытых нейронов (т.е. принадлежащих этому скрытому слою) неограниченно, тогда как число выходных нейронов равно числу «коллекторов» внутри «глаза». Каждый скрытый нейрон принимает сигналы от сенсоров, расположенных в соответствующем «рецептивном поле», обрабатывает их и передает результат на каждый из выходных нейронов. Каждый выходной нейрон, в свою очередь, тоже обрабатывает свои входные сигналы и передает результат на соответствующий «коллектор».

«Мозг» также является многослойной нейронной сетью с обратным распространением сигнала при обучении (многослойным персептроном), содержащий один выходной и, возможно, один скрытый слой нейронов. Сигналы, формируемые «коллекторами», поступают на скрытый слой «мозга», откуда на слой выходных нейронов, каждый из которых формирует сигнал, соответствующий прогнозируемому свойству химического соединения. Распространение сигналов внутри сети и динамика ее обучения описываются точно такими же математическими выражениями, как и в случае обычной многослойной нейронной сети с обратным распространением ошибки при обучении [42].

Таким образом, можно выделить 4 этапа обработки структурной информации в рассматриваемом нейронном устройстве: 1) формирование в «сенсорном поле» сигналов, соответствующих характеристикам атомов и связей; 2) обработка в каждом из «рецепторов» сигналов, собранных со своего «рецептивного поля»; 3) формирование в «коллекторах» сигналов, инвариантных к перенумерации атомов в молекуле; 4) окончательная обработка инвариантных сигналов в «мозгу» (см. Рис. 63). Заметим, что для корректной работы сети последний этап не является обязательным.

Следует отметить, что воплощенная в нейронном устройстве идея «рецептивных полей», собранная с которых первичная информация подвергается дальнейшей обработке в последующих слоях нейронов, в результате чего формируются сигналы, инвариантные к возможным трансформациям входных сигналов, составляют основу парадигмы *неокогнитрона* [492-495], разработанного в соответствии с нейрофизиологическими данными о том, как визуальная информация обрабатывается в зрительной коре головного мозга [496-498].

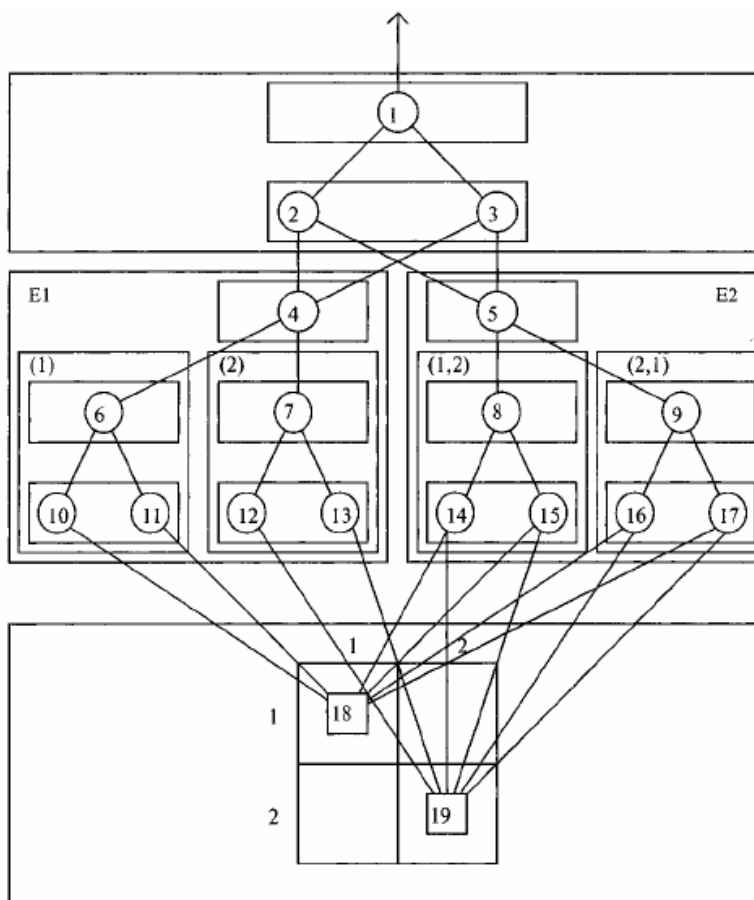


Рис. 64. Конфигурация нейронного устройства для молекулы этана

7.4.3. Примеры разных конфигураций нейронного устройства

Рассмотрим в качестве примера нейронное устройство, состоящее из «мозга» и двух «глаз» (которые мы обозначим как $E1$ и $E2$). Возьмем простейшее «сенсорное поле», содержащее только атомные сенсоры, каждый из которых формирует сигнал, равный числу атомов водорода, присоединенных к соответствующему атому (обозначим этот тип сенсора NH). Пусть каждый из «рецепторов» внутри «глаза» $E1$ получает сигнал только с одного атомного сенсора. Таким образом, «глаз» $E1$ «смотрит» на неводородные атомы в молекуле. Пусть также каждый из «рецепторов» внутри «глаза» $E2$ принимает сигналы от двух атомных «рецепторов», соответствующих атомам, образующих химическую связь между собой. Таким образом, глаз $E2$ «смотрит» на связи между неводородными атомами внутри молекулы. Пусть также каждый «рецептор» внутри обоих глаз содержит два скрытых и один выходной нейрон. В соответствии с числом выходных нейронов, каждый глаз также содержит по одному «коллектору», чей выходной сигнал передается на «мозг».

На Рис. 64 представлена конфигурация описанного выше нейронного устройства для молекулы этана. В этом случае, инвариантность предсказываемых нейронным устройством свойств химических соединений относительно перенумерации атомов обеспечивается следующими ограничениями, налагаемыми на значения весов связей ω' и порогов активации θ' : $\omega'_{4,6} = \omega'_{4,7}$; $\omega'_{6,10} = \omega'_{7,12}$; $\omega'_{6,11} = \omega'_{7,13}$; $\omega'_{5,8} = \omega'_{5,9}$; $\omega'_{8,14} = \omega'_{9,16}$; $\omega'_{8,15} = \omega'_{9,17}$; $\omega'_{10,18} = \omega'_{12,19}$; $\omega'_{11,18} = \omega'_{13,19}$; $\omega'_{14,18} = \omega'_{16,19}$; $\omega'_{14,19} = \omega'_{16,18}$; $\omega'_{15,18} = \omega'_{17,19}$; $\omega'_{15,19} = \omega'_{17,18}$; $\theta'_6 = \theta'_7$; $\theta'_8 = \theta'_9$; $\theta'_{10} = \theta'_{12}$; $\theta'_{11} = \theta'_{13}$; $\theta'_{14} = \theta'_{16}$; $\theta'_{15} = \theta'_{17}$.

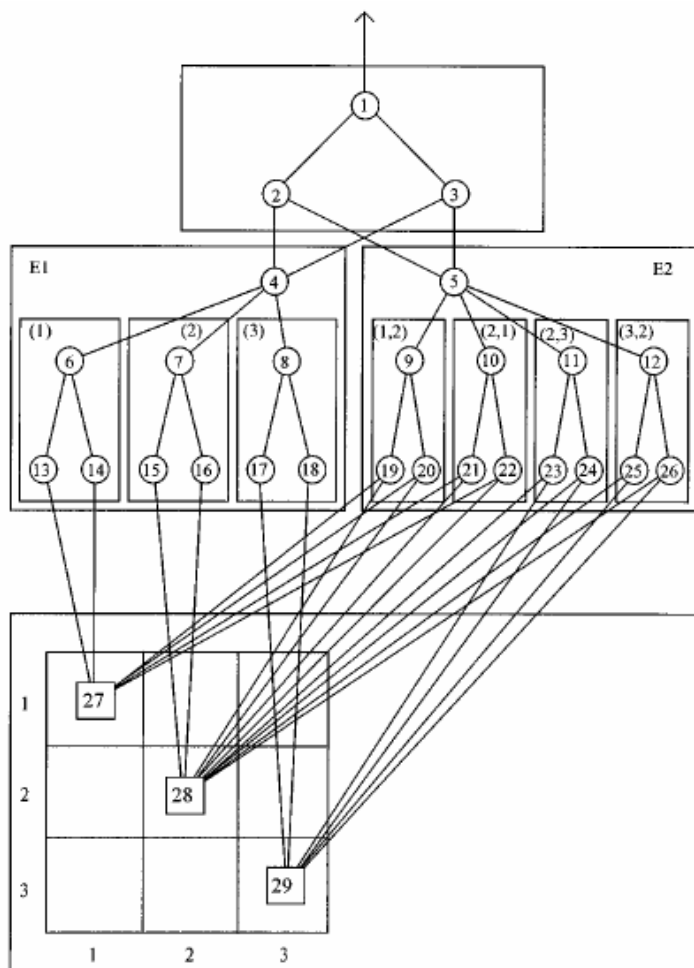


Рис. 65. Конфигурация нейронного устройства для молекулы пропана

В качестве еще одного примера, на Рис. 65 представлена конфигурация этого же нейронного устройства уже в применении к молекуле пропана. В этом случае, инвариантность предсказываемых нейронным устройством свойств химических соединений относительно перенумерации атомов обеспечивается следующими ограничениями, налагаемыми на значения весов связей ω'' и порогов активации θ'' :

$$\begin{aligned} \omega''_{4,6} = \omega''_{4,7} = \omega''_{4,8}; & \quad \omega''_{6,13} = \omega''_{7,15} = \omega''_{8,17}; & \quad \omega''_{6,14} = \omega''_{7,16} = \omega''_{8,18}; \\ \omega''_{13,27} = \omega''_{15,28} = \omega''_{17,29}; & \quad \omega''_{14,27} = \omega''_{16,28} = \omega''_{18,29}; & \quad \omega''_{5,9} = \omega''_{5,10} = \omega''_{5,11} = \omega''_{5,12}; \\ \omega''_{9,19} = \omega''_{10,21} = \omega''_{11,23} = \omega''_{12,25}; & \quad \omega''_{9,20} = \omega''_{10,22} = \omega''_{11,24} = \omega''_{12,26}; & \quad \omega''_{19,27} = \omega''_{21,28} = \omega''_{23,28} = \omega''_{25,29}; \\ \omega''_{19,28} = \omega''_{21,27} = \omega''_{23,29} = \omega''_{25,28}; & \quad \omega''_{20,27} = \omega''_{22,28} = \omega''_{24,28} = \omega''_{26,29}; & \quad \omega''_{20,28} = \omega''_{22,27} = \omega''_{24,29} = \omega''_{26,28}; \\ \theta''_6 = \theta''_7 = \theta''_8; & \quad \theta''_9 = \theta''_{10} = \theta''_{11} = \theta''_{12}; & \quad \theta''_{13} = \theta''_{15} = \theta''_{17}; & \quad \theta''_{14} = \theta''_{16} = \theta''_{18}; & \quad \theta''_{19} = \theta''_{21} = \theta''_{23} = \theta''_{25}; \\ \theta''_{20} = \theta''_{22} = \theta''_{24} = \theta''_{26}. & & & & \end{aligned}$$

Обе эти конфигурации могут быть получены путем размножения «рецепторов» из представленной на Рис. 66 минимальной конфигурации.

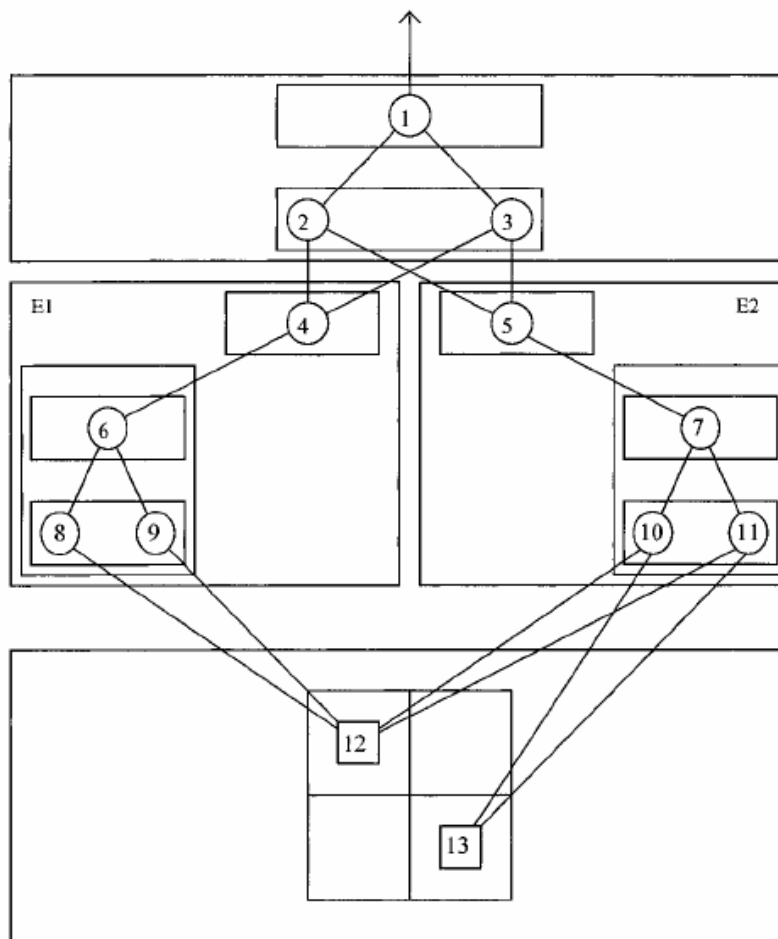


Рис. 66. Минимальная конфигурация нейронного устройства

7.4.4. Применение нейронного устройства в исследованиях «структурное свойство» для органических соединений

Температура кипения алканов. Для первого вычислительного эксперимента с программным эмулятором нейронного устройства было выбрано прогнозирование температуры кипения алканов при нормальных условиях, поскольку по этой теме известно большое число публикаций, что дает возможность осуществить объективное сравнение с результатами, достигнутыми другими авторами (искусственные нейронные сети применялись для прогнозирования температуры кипения алканов в публикациях [198, 406, 421, 465, 491, 499-503]). Выборка, состоящая из 74 алканов C_2-C_9 (данные были взяты из статьи [409]), была случайным образом разбита на две части – обучающую (67 соединений) и контрольную (7 соединений) выборку. Описанное выше нейронное устройство с минимальной конфигурацией, приведенной на Рис. 66, было ис-

пользовано в этом вычислительном эксперименте. Для обучения нейронного устройства было применено «обобщенное дельта-правило» [41] в рассмотренной выше модификации, параметр скорости обучения был взят равным 0.05, а параметр момента – 0.9 (см. [41]). Обучение было прекращено по достижению значения коэффициента корреляции между предсказанными и экспериментальными значениями температуры кипения равного 0.994 (когда значения коэффициента корреляции и среднеквадратичных ошибок на обеих выборках перестали меняться). Эта величина больше любого коэффициента корреляции, который был достигнут на этой же выборке при использовании какого-либо одного топологического индекса в качестве молекулярного дескриптора, но сравнима с коэффициентами корреляции, которые на этой выборке могут быть получены при помощи множественной линейной регрессии сразу с несколькими топологическими индексами. Этот же вывод можно было бы сделать и при рассмотрении среднеквадратичных ошибок на обеих выборках (5.2 градуса на обучающей выборке и 5.1 градус на контрольной выборке). Тем не менее, этот результат хуже того, который был ранее нами достигнут для температуры кипения алканов при использовании стандартного многослойного персептрона и набора топологических индексов либо фрагментных дескрипторов [406]. Кроме того, обучение нейронного устройства происходило крайне медленно (несколько часов на компьютере Pentium-100).

Для поиска путей улучшения работы нейронного устройства мы изучили влияние его архитектуры на производительность. Оказалось, что число «коллекторов» внутри каждого «глаза» существенным образом влияет на время (несколько минут вместо часов на компьютере Pentium-100) и на качество обучения. При использовании той же самой архитектуры нейронного устройства, но с пятью «коллекторами» внутри каждого из двух «глаз», было достигнуто очень высокое значение коэффициента корреляции, равное 0.9994, и очень низкие среднеквадратичные ошибки на обучающей (1.6 градусов) и на контрольной (2.4 градуса) выборках. Подобная прогнозирующая способность находится на уровне лучших результатов, которые были когда-либо достигнуты для температуры кипения алканов. Тем не менее, дальнейшее увеличение числа «коллекто-

ров» ведет к ухудшению прогнозирующей способности нейронного устройства, по-видимому, вследствие неоправданного увеличения числа настраиваемых параметров при небольшом размере обучающей выборки.

При изучении работы нейронного устройства мы также обнаружили, что скорость и стабильность обучения можно существенно улучшить при использовании отдельных значений параметра скорости обучения для «мозга» и «глаз». Как оказалось, для стабильного обучения необходимо, чтобы параметр скорости обучения для «глаз» был на порядок меньше, чем для «мозга». Что же касается абсолютных значений параметра фактора обучения для «мозга», то 0.25 в начале и 0.05 в конце обучения являются, по-видимому, оптимальными.

Вязкость углеводов. В следующем примере выборка, состоящая из 81 представителя разнообразных углеводов C_6-C_{21} [504] (циклических и ациклических, насыщенных и ненасыщенных, ароматических и алифатических) была использована для построения нейросетевой модели, позволяющей прогнозировать их вязкость при 40 °С. Как и в предыдущем случае, выборка была разбита на обучающую (65 соединений) и контрольную (16 соединений) выборки. Нейронное устройство, содержащее «мозг» с двумя скрытыми нейронами и один глаз $E2$ с тремя скрытыми нейронами в каждом «рецепторе» и пятью «коллекторами», было выбрано для этого исследования. После 1100 эпох обучения коэффициент корреляции стал 0.996, среднеквадратичная ошибка на обучающей выборке достигла 0.15 сантипуаз, а на контрольной выборке – 0.18 сантипуаз. Подобные статистические показатели по предсказанию вязкости углеводов до сих пор являются, по-видимому, одними из лучших.

Теплота испарения углеводов. В следующем примере выборка из 267 углеводов C_4-C_{26} [505] (как и в предыдущем примере, циклических и ациклических, насыщенных и ненасыщенных, ароматических и алифатических) была использована для обучения нейронного устройства прогнозированию теплоты испарения. Из этой выборки 54 соединения были случайным образом отобраны в контрольную выборку, тогда как оставшиеся 213 соединений образовали обучающую выборку. Нейронное устройство, содержащее «мозг» с тремя скрытыми нейронами и два «глаза», $E1$ и $E2$, каждый из которых содержит по

три скрытых нейрона и три «коллектора», было применено в этом исследовании. Как и во всех предыдущих примерах, «сенсорное поле» состояло только из сенсоров NH , формирующих сигналы, соответствующие числу атомов водорода, присоединенных к соответствующему атому. Поскольку не наблюдалось «переучивание», обучение было остановлено через 2600 эпох, когда статистические показатели модели перестали меняться, и обученное нейронное устройство обеспечивало коэффициент корреляции 0.996 и среднеквадратичную ошибку 1.44 кДж/моль на обучающей выборке и 1.26 кДж/моль на контрольной выборке. В данном случае прогнозирующая способность нейронного устройства оказалась лучше, чем у ранее опубликованной модели, построенной по этим же данным [505].

Плотность жидких углеводородов. В следующем примере нейронное устройство было обучено прогнозировать плотность жидких углеводородов. Выборка из 141 углеводорода C_5 - C_8 [504] (насыщенные и ненасыщенные, циклические и ациклические, ароматические и алифатические) была случайно разбита на обучающую выборку из 113 соединений и контрольную выборку из 28 соединений. В этом исследовании было применено нейронное устройство, содержащее «мозг» с пятью скрытыми нейронами и два «глаза», $E1$ и $E2$, каждый из которых содержит по пять скрытых нейронов в каждом из рецепторов и пять коллекторов. Тип «сенсоров» был тот же, что и в предыдущих примерах. После 1700 циклов обучения значение коэффициента корреляции достигло 0.971, а среднеквадратичная ошибка стала 0.018 г/см³ на обучающей выборке и 0.019 г/см³ на контрольной выборке. Подобная прогнозирующая способность является очень неплохой.

Теплота сольватации разнообразных органических соединений в циклогексане. В отличие от предыдущих случаев, в выборку для данного примера (взятую с работы [505]) вошли органические соединения, принадлежащие к различным классам. В соответствии с общей методикой проведения исследований, исходная выборка была разбита на обучающую, содержащую 112 соединений, и контрольную, содержащую 28 соединений. Кроме того, в соответствии с результатами предварительных исследований, одно из соединений, перфтор-

бензол, было отброшено как «аутлайер». В данной работе было использовано нейронное устройство, содержащее 3 «глаза»: $E1$, $E2$ и $E3$. «Глаза» $E1$ и $E2$ аналогичны рассмотренным выше, тогда как «глаз» $E3$ содержит рецепторы, принимающие сигналы от трех атомов, соединенных двумя связями. Чтобы различать типы атомов, в дополнение к использованному во всех предыдущих примерах «сенсор»у NH добавлен еще «сенсор» PQN , который определяет главное квантовое число соответствующего атома. Как для «мозга» нейронного устройства, так и для всех его «рецепторов», было задано по три скрытых нейрона. В каждое из трех глаз было помещено по три «коллектора». После 10000 циклов обучения значение коэффициента корреляции составило 0.990, среднеквадратичная ошибка составила 1.77 кДж/моль на обучающей выборке и 2.46 кДж/моль на контрольной выборке. В работе [505] на этой же выборке было ранее показано, что обычно используемые топологические индексы неспособны обеспечить хорошую корреляцию с теплотой сольватации в циклогексане, и поэтому был разработан специальный сольватационный индекс [505], хорошо коррелирующий с теплотой сольватации в циклогексане (в рамках линейной регрессии коэффициент корреляции на всей выборке составляет 0.985, а стандартная ошибка 2.1 кДж/моль). Таким образом, в приведенном примере нейронное устройство оказалось способным составить конкуренцию специально разработанному под свойство топологическому индексу.

Поляризуемость разнообразных химических соединений. В следующем примере для обучения нейронного устройства была использована выборка [361], содержащая как разнообразные органические соединения (размером до 26 неводородных атомов), относящиеся к разным классам, так и простейшие неорганические соединения, например, N_2O , SO_2 , H_2S , O_2 , N_2 , NH_3 , Cl_2 и т.д. Исходная выборка была случайно разбита на обучающую (235 соединений) и контрольную (58 соединений). После серии компьютерных экспериментов мы остановились на нейронном устройстве, содержащем «мозг» с тремя скрытыми нейронами и всего один «глаз» $E1$, содержащий, в свою очередь, «рецепторы» с тремя скрытыми нейронами и пять «коллекторов». В «сенсорное поле» были помещены три вида «сенсоров»: NH , AR и NE . «Сенсор» NH формирует сигнала-

лы, соответствующие числу атомов водорода, присоединенных к данному атому, «сенсор» *AR* формирует сигналы, соответствующие атомному радиусу, а «сенсор» *NE* формирует сигналы в соответствии с числом электронов в атоме. После 2000 эпох обучения нейронного устройства значение коэффициента корреляции составило 0.995, среднеквадратичная ошибка на обучающей выборке 0.86 см³ и на контрольной выборке 0.71 см³. Этот результат значительно лучше всего того, что удастся получить на этой выборке при использовании стандартных топологических индексов в качестве дескрипторов. Хотя в данном случае хороших результатов можно добиться также и при помощи аддитивных схем [506], однако применимость последних ограничена лишь молекулами, все типы атомов или группы которых достаточно представлены в обучающей выборке, тогда как для нейронного устройства это условие необязательно. Именно этим последним обстоятельством выражается потенциальное преимущество применения разработанного нами нейронного устройства для прогнозирования свойств химических соединений по сравнению с использованием аддитивных схем или фрагментных дескрипторов.

Анестетическое давление газов. Целью данного примера является иллюстрация того, что разработанное нейронное устройство может быть использовано для прогнозирования не только физико-химических свойств органических соединений, но и их биологической активности. Мы воспользовались взятой из обзорной статьи [355] базой данных, содержащей углеводороды, галогенированные углеводороды, а также некоторые неорганические газы, такие как молекулярный азот, SF₆, N₂O, а также благородные (инертные) газы. Как и во всех предыдущих примерах, база данных была разбита на обучающую выборку (24 соединения) и контрольную выборку (шесть соединений). Для проведения исследования была построена нейронная сеть, содержащая «мозг» с тремя скрытыми нейронами и один «глаз» *E1*, содержащий «рецепторы», имеющие по три скрытых нейрона и «видящие» только по одному атому, и пять коллекторов. В данном примере мы использовали три типа атомных «сенсоров»: *NH*, *PQN* и *VE*. Первые два типа «сенсоров» (*NH* и *PQN*) описаны выше, а «сенсор» *VE* формирует сигнал в соответствии с числом валентных электронов

на атому. После 4000 эпох обучения нейронного устройства коэффициент корреляции составил 0.990, среднеквадратичная ошибка на обучающей выборке составила 0.18 логарифмических единиц ($\log(1/p)$), а на контрольной выборке – 0.26 логарифмических единиц. Эти статистические параметры значительно превосходят все то, что удается построить на этой выборке с использованием как топологических индексов, так и фрагментных дескрипторов.

В Табл. 33 в сжатом виде представлены результаты рассмотренных выше вычислительных экспериментов по проведению прямых корреляций «структура-свойство» при помощи разработанного нами нейронного устройства.

Табл. 33. Результаты применения нейронного устройства при построении корреляций «структура-свойство»

Свойство	Класс соединений	Коэффициент корреляции	Среднеквадратичная ошибка на обучающей выборке	Среднеквадратичная ошибка на контрольной выборке	Глаза	Сенсоры
Температура кипения при нормальном давлении	алканы	0.9994	1.6 град.в	2.4 град.	<i>E1, E2</i>	<i>NH</i>
вязкость при 40 °С	углеводороды	0.996	0.15 сантипуаз	0.18 сантипуаз	<i>E2</i>	<i>NH</i>
теплота испарения	углеводороды	0.996	1.44 кДж/моль	1.26 кДж/моль	<i>E1, E2</i>	<i>NH</i>
плотность	углеводороды	0.971	0.018 г/см ³	0.019 г/см ³	<i>E1, E2</i>	<i>NH</i>
теплота сольватации в циклогексане	разнообразные соединения	0.990	1.77 кДж/моль	2.46 кДж/моль	<i>E1, E2, E3</i>	<i>NH, PQN</i>
поляризуемость	разнообразные соединения	0.995	0.86 см ³	0.71 см ³	<i>E1</i>	<i>NH, AR, NE</i>
анестетическое давление газов	разнообразные газы	0.990	0.18 лог.ед. ($\log(1/p)$)	0.26 лог.ед. ($\log(1/p)$)	<i>E1</i>	<i>NH, PQN, VE</i>

7.4.5. Выводы

Выше была продемонстрирована способность данного нейронного устройства осуществлять поиск прямых корреляций между структурами органических соединений и их свойствами без необходимости в предварительном выборе и вычислении значений каких-либо топологических индексов, чисел встречаемости определенных фрагментов либо каких-нибудь других типов глобальных молекулярных дескрипторов (инвариантов молекулярных графов). Вместо этого, мы используем локальные дескрипторы, относящиеся к атомам и связям в молекулах. Во всех вышеприведенных примерах использовались лишь простейшие атомные дескрипторы (формируемые атомными сенсорами), значение которых непосредственно связано с элементами матрицы смежности соответствующего молекулярного графа, а потому такую корреляцию вполне справедливо можно считать «прямой» корреляцией между структурой и свойством. Таким образом, эта методология представляет собой альтернативу применению глобальных молекулярных дескрипторов при поиске корреляций «структура-свойство».

С другой стороны, работа данного нейронного устройства вполне сочетается с применением дескрипторов. Во-первых, наряду с рассмотренными выше простейшими атомными сенсорами, возможно введение сенсоров, воспринимающих значения более сложных локальных дескрипторов, требующих специальных вычислений, например, зарядов на атомах либо межатомных расстояний. Во-вторых, в рамках этого подхода вполне возможно использование и глобальных дескрипторов (что для ряда свойств может оказаться даже необходимым), что может быть достигнуто путем непосредственного ввода в «мозг» нейронного устройства сигналов, соответствующих глобальным молекулярным дескрипторам.

Возможен и совсем другой взгляд на данное нейронное устройство. Поскольку выходные сигналы как всего нейронного устройства, так и каждого из его коллекторов, не зависят от нумерации атомов и, следовательно, могут рассматриваться как молекулярные дескрипторы (инварианты молекулярных гра-

фов), то и все нейронное устройство можно рассматривать как инструмент для изобретения молекулярных дескрипторов, максимально приспособленных для построения корреляции с данным свойством. И действительно, в процессе обучения нейронное устройство пытается таким образом скомбинировать значения локальных атомных и межатомных дескрипторов, чтобы значения результирующего дескриптора были максимальным образом приближены к значениям данного свойства.

ГЛАВА 8. РАЗРАБОТКА ПРОГРАММНЫХ СРЕДСТВ

8.1. История разработки программных средств

История разработки программных средств, использовавшихся на разных этапах выполнения данной диссертационной работы, начинается с создания на ПЭКВМ (Персональной Электронной Клавишной Вычислительной Машине) «Искра-226» в 1985-1986 гг. автором диссертационной работы под руководством С.С.Трача и Н.С.Зефирова универсальной программы молекулярной графики для целей органической химии «Модель» [507, 508] как части первой версии компьютерной программы SYMBEQ [509], предназначенной для поиска новых типов реагирования органических соединений. В рамках SYMBEQ «Модель» использовалась для интерактивного ввода графов топологий перераспределения связей и для графического вывода сгенерированных уравнений химических реакций.

В 1986-1987 гг. автором диссертационной работы вместе с М.И.Станкевич и под руководством Н.С.Зефирова была создана первая программа, позволяющая осуществлять поиск структурных фрагментов в молекулярных графах [510, 511]. Эта программа первоначально использовалась нами для расчета фрагментных дескрипторов, пока не был создан для этой цели значительно более совершенный дескрипторный блок FRAGMENT.

В 1988-1989 гг. автором диссертационной работы вместе с М.И.Станкевич и Р.О.Девдариани и под руководством Н.С.Зефирова был создан на ПЭКВМ «Искра-226» программный комплекс STAR (Structure-Activity Relationships) для нахождения корреляций «структура-свойство» на основе топологических индексов и простой линейной регрессии [512]. Комплекс включал: 1) управляющую программу; 2) программу интерактивного ввода химических структур «Модель», отделенную от SYMBEQ и наделенную возможностью создавать базы данных «структура-свойство»; 3) несколько дескрипторных блоков для расчета топологических индексов; 4) статистический блок для

проведения линейного регрессионного анализа. Интересным компонентом комплекса STAR явился дескрипторный блок для вычисления взвешенного индекса Рандича и позволяющий находить для этого оптимальный набор весов путем оптимизации функционала ошибки в пространстве весов при помощи симплекс-метода. Таким путем удалось, например, построить модель для прогнозирования температуры плавления ароматических соединений [513].

Следующим важным этапом в разработке программных средств явилось создание в 1990-1992 гг. программного комплекса для поиска количественных корреляций «структура-свойство» «EMMA», предназначенного для работы в среде MS-DOS на IBM PC-совместимых персональных компьютерах первых поколений. В рамках комплекса EMMA автором диссертационной работы были созданы:

1) программа интерактивного ввода химических структур и ведения баз данных «структура-свойство» MOLED (в сущности, программа «Модель» из комплекса STAR была переписана под среду MS-DOS и дополнена новыми возможностями);

2) дескрипторный блок FRAGMENT для расчета фрагментных дескрипторов (см. разделы 5.1 и 8.3);

3) дескрипторный блок НМО (описание не включено в данную диссертационную работу), предназначенный для проведения квантово-химических расчетов молекул непредельных соединений с использованием стандартного метода Хюккеля и вычисления по результатам расчетов набора квантово-химических дескрипторов;

4) дескрипторный блок FRAGPROP (см. разделы 5.4. и 8.4) для расчета псевдофрагментных дескрипторов;

5) дескрипторные блоки, предназначенные для расчета разнообразных типов топологических индексов, в частности, CONNECT, KAPPA, BALABAN, BASAK, ELEM, VX, LOUSE и др. (описание этих блоков не включено в диссертационную работу).

В разработку комплекса «ЭММА» наиболее существенный вклад также внесли Д.В.Сухачев (управляющая программа, блок построения статистической

модели при помощи пошагового варианта множественной линейной регрессии и блок прогноза, которые совместно образуют программу «ЭММА» - головную программу комплекса), Д.Е.Петелин (дескрипторные блоки для расчета топологических индексов и физико-химических дескрипторов, в частности ETS, NB, NFORM, INDPAR, STERIC, VW и др.), О.Ломова (генератор химических структур GOLD [514, 515], который сейчас вполне обосновано можно назвать генератором виртуальных комбинаторных библиотек для виртуального скрининга) и А.Ю.Зотов (блок управления расчетом дескрипторов и некоторые дескрипторные блоки). Работы по созданию комплекса «ЭММА» проводились под руководством В.А.Палюлина и Н.С.Зефирова.

В 1993-1995 гг. автором диссертационной работы (под руководством В.А.Палюлина и Н.С.Зефирова) была разработана для среды MS-DOS программа-эмулятор искусственных нейронных сетей, специально приспособленная для построения количественных моделей «структура-свойство», NASA (Neural Approach to Structure-Activity) [516]. При помощи этой программы были получены результаты, изложенные в подразделах 4.4.1 и 6.1 данной диссертационной работы.

В 1996 г. автором диссертационной работы вместе с Н.М.Гальберштам (под руководством В.А.Палюлина и Н.С.Зефирова) была создана для среды Windows 3.1 первая версия программного комплекса NASAWIN (Neural Approach to Structure-Activity for Windows) [194, 517], и с тех пор он находится в постоянном развитии. Первоначально NASAWIN включала только эмулятор многослойной нейронной сети обратного распространения, перенесенный из программы NASA, и набор дескрипторных блоков, перенесенный из программного комплекса «ЭММА», при этом практически все перенесенные компоненты были перепрограммированы. Возможности дескрипторного блока FRAGMENT были существенно расширены по сравнению с версией, работавшей в комплексе «ЭММА» (работа по расширению возможностей этого блока велась вместе с Н.В.Артеменко). В ходе своего дальнейшего развития в NASAWIN было включено множество дополнительных методов статистического анализа, дескрипторных блоков, методов визуализации и архитектур нейронных сетей, в резуль-

тате чего NASAWIN превратился в мощный универсальный программный комплекс для построения моделей «структура-свойство» и прогнозирования свойств органических соединений.

8.2. Программный комплекс «NASAWIN»

Отсутствие удобного для химика-органика инструмента, позволяющего получать, анализировать и использовать для прогноза нейросетевые модели зависимости структура-свойство, побудило нас к разработке компьютерной программы, базирующейся на методологии искусственных нейронных сетей и ориентированной на работу с химической информацией.

Программный комплекс «NASAWIN» позволяет:

- 1) загружать и просматривать базы данных, содержащие структуры химических соединений и их свойства;
- 2) вычислять наборы дескрипторов, описывающих химические структуры, и отбирать наиболее значимые;
- 3) выявлять и интерпретировать количественные зависимости между значениями дескрипторов и свойств химических соединений при помощи многослойной нейронной сети прямого распространения;
- 4) статистически оценивать полученные модели;
- 5) использовать полученные нейросетевые модели для прогнозирования свойств произвольных химических соединений.

Программные коды написаны на языке C++ с использованием компилятора MVC++ 6.0. Программа содержит около 80000 строк. Наряду с общепринятыми алгоритмами работы с нейронными сетями, «NASAWIN» обладает множеством характерных черт, которые делают этот комплекс уникальным инструментом для исследования зависимости «структура-свойство» в химии. Рассмотрим основные возможности, которые предоставляет программа «NASAWIN» для получения нейросетевых моделей структура-свойство.

8.2.1. Представление химической информации

NASAWIN может работать с химическими базами данных, записанными как в стандартном SDF-формате, поддерживаемом основными существующими коммерческими программами, так и в SET-STR-формате, который поддерживается рядом программ и программных комплексов, разработанных на химическом факультете МГУ (в частности молекулярный редактор «MOLED», программный комплекс «EMMA», генератор химических структур «GOLD», многочисленные дескрипторные блоки и т.д.). При необходимости комплекс «NASAWIN» без явного вмешательства пользователей сам производит конвертацию между необходимыми форматами, благодаря чему обеспечивается его интегрированная работа с многочисленным ориентированным на химию программным обеспечением. Кроме того, «NASAWIN» содержит и самостоятельные средства просмотра используемых баз данных.

Также важно отметить, что «NASAWIN» позволяет работать и с «разреженными» базами данных. Такие базы очень часто встречаются в химии, т.к. часто не для всех соединений, представленных в базе данных, измерены все значения свойств или получены все значения дескрипторов.

8.2.2. Интеграция с программными компонентами, осуществляющими расчет дескрипторов химических структур

Управляющая программа «NASAWIN» обеспечивает согласованную работу с гибким набором многочисленных автономных программных компонент, проводящих расчет разнообразных дескрипторов химических структур: подструктурных, топологических, позиционных, физико-химических и квантово-химических. Кроме того, «NASAWIN» предоставляет встроенную библиотеку, облегчающую разработку новых дескрипторных блоков.

8.2.3. Химически-ориентированная визуализация

При обработке химических баз данных очень важно знать, какая химическая структура скрывается за каждой записью в базе данных и за каждой точкой на графиках зависимостей, из-за чего использование для этой цели статистических либо нейросетевых пакетов общего назначения часто оказывается крайне неудобным и неэффективным. «NASAWIN» позволяет абсолютно на всех этапах взаимодействия пользователя с программой видеть структурные формулы химических соединений прямо в диалоговых окнах или в окнах визуализации хода и результатов обучения, что резко повышает удобство и эффективность работы с программой.

8.2.4. Модификация дескрипторов и свойств

В настоящее время «NASAWIN» поддерживает 8 типов модификаций дескрипторов. Кроме общеупотребительных типов модификаций дескрипторов (взятие квадрата, квадратного корня, логарифма, обратного числа и порогового индикатора) предусмотрены и специфические для химии типы, вычисляемые с учетом количества неводородных атомов в молекуле («деление на число атомов», «умножение на число атомов» и «обратная величина, умноженная на число атомов»).

Также возможны следующие 3 вида модификаций для исследуемых свойств: взятие обратного числа, взятие логарифма, а также использование специфического типа модификации «логарифм от обратной величины», что часто бывает необходимо при обработке данных по биологической активности химических соединений.

8.2.5. Предварительный отбор дескрипторов

При использовании подструктурных дескрипторов при поиске соотношений «структура-свойство» практически всегда оказывается, что их значения линейно взаимосвязаны. Для этого случая в «NASAWIN» специально предусмотр-

рена возможность формирования такого поднабора дескрипторов, внутри которого отсутствует линейная попарная зависимость между ними, что часто позволяет резко сократить число используемых дескрипторов. Кроме того, во многих задачах прогнозирования физико-химических свойств химических соединений степень нелинейности их зависимости от значений дескрипторов оказывается не очень высокой, хотя и существенной для максимально точного прогнозирования, что дает возможность использовать быстрые линейно-регрессионные методы отбора дескрипторов. Хотя в общем случае сформированный таким образом набор отобранных дескрипторов может оказаться неоптимальным, в реальных задачах по изучению зависимости «структура-свойство» (когда число подструктурных дескрипторов может составить тысячи и даже десятки тысяч, что делает проблематичным использование чистых нейросетевых методов отбора дескрипторов) такой подход часто оказывается единственно возможным. Для обеспечения этого в «NASAWIN» предусмотрена специальная интерактивная процедура пошаговой линейной регрессии (БПМЛР, см. подраздел 4.1.5), которая позволяет пользователю быстро сформировать небольшой набор ценных дескрипторов, который в дальнейшем может быть использован для обучения нейронной сети.

8.2.6. Построение классификационных моделей структура-активность

Очень часто, особенно при работе с биологическими данными, значения свойств представлены на качественном уровне (1 - есть активность, 0 - нет активности). Программа «NASAWIN» способна самостоятельно различать типы представления исходных данных и в зависимости от этого строить классификационные либо регрессионные. Кроме того, предусмотрена возможность ручного разбиения массива исследуемых соединений по каждому конкретному свойству на активные и неактивные, с последующим построением классификационных моделей. Пользователь может изменять пороговую величину для такого разбиения. Подчеркнем, что в данную диссертационную работу включено использование только регрессионных методов.

8.2.7. Нейросетевые парадигмы

Программный комплекс «NASAWIN» основан главным образом на использовании нейросетей обратного распространения (см. подраздел 1.2.4). Основные алгоритмы обучения, реализованные в NASAWIN, это «обобщенное дельта-правило» (см. пункт 1.2.4.4) и метод эластичного распространения (см. пункт 1.2.4.5). Поскольку последний метод обучения проявил себя при эксплуатации программы значительно лучше первого, то именно он и используется по умолчанию. Для уменьшения «переучивания» при обучении может быть включен один из четырех типов регуляризаторов. Кроме того, в «NASAWIN» реализованы также самоорганизующиеся карты Кохонена (см. пункт 1.2.5.1), которые могут быть использованы кластеризации базы данных, а также специальная динамически наращиваемая сеть для решения классификационных задач распознавания образов. Использование последних двух нейросетевых парадигм выходит за рамки данной диссертационной работы.

8.2.8. Интерпретация нейросетевых моделей

В ходе построения нейросетевых моделей рассчитываются все описанные выше статистические параметры (см. раздел 4.2), предназначенные для анализа вкладов входных параметров нейросети в получаемые модели. Эти данные представляются в числовом виде в диалоговых окнах, а также графически: на каждой итерации обучения нейросеть перерисовывается в соответствии с данными о значимости дескрипторов и величинах весовых коэффициентов связей.

8.2.9. Отбор дескрипторов в ходе обучения нейросети

Рассчитанные характеристики значимости дескрипторов могут использоваться для отбора наиболее важных дескрипторов в ходе обучения нейросети. Для более четкого выявления значимых дескрипторов предусмотрена дополнительная возможность отсева малозначимых весовых коэффициентов. Для того,

чтобы выявить малозначимые весовые коэффициенты, используется процедура «забывания», т.е. на каждой итерации каждый весовой коэффициент уменьшается на некую величину, пропорциональную его значению. Для вычисления этой пропорциональной величины в программе «NASAWIN» используются линейные, квадратичные и логарифмические функции, а также функция Гаусса. Таким образом, несущественные весовые коэффициенты сводятся к нулю, что позволяет сократить размерность нейросети путем удаления нейронов с нулевыми синапсами.

8.2.10. Определение момента начала «переучивания» нейросети

С целью определения момента перехода обучения нейросети из «обобщающей» в «запоминающую» фазу, то есть того момента, когда среднеквадратичная ошибка для контрольных соединений начинает возрастать и обучение нейросети должно быть прервано, в программе «NASAWIN» предусмотрена следующая процедура. Вся выборка соединений разбивается автоматически или вручную на 3 подвыборки – обучающую, контрольную и выборку прогноза. На соединениях из обучающей выборки строится нейросетевая модель. Точка перехода обучения нейросети из одной фазы в другую определяется автоматически и соответствует моменту начала увеличения среднеквадратичной ошибки для соединений из контрольной выборки. Прогнозирующая способность нейросети оценивается по величине средней ошибки, вычисленной для соединений из выборки прогноза, в момент начала «переучивания». Предложенный трехвыборочный метод позволяет, таким образом, оценить прогнозирующую способность нейросети с использованием соединений, не участвующих в процессе обучения нейросети, что является существенным преимуществом по сравнению со стандартным методом перекрестной оценки.

8.2.11. Кластеризация баз данных

Реальные химические базы данных часто бывают неоднородными и содержат несколько групп соединений, различных по типу строения или по механизму действия, и в этом случае построение единой нейросетевой модели не всегда оправдано.

Для подразделения базы данных на кластеры в «NASAWIN» используется анализ активностей скрытых нейронов. Для этого строятся графики зависимостей выходных сигналов для всех возможных пар скрытых нейронов, причем оба нейрона должны принадлежать одному и тому же скрытому слою нейросети. Было отмечено, что на таких графиках соединения, характеризующиеся близким строением, располагаются близко друг к другу и образуют таким образом отдельные группы. Пользователю предоставляется возможность вручную выделить интересующие его кластеры, а затем построить отдельные нейросетевые модели для каждого найденного кластера.

8.2.12. Динамическая визуализация хода обучения нейросети

Для удобства работы с программой «NASAWIN» предусмотрена возможность наблюдения за ходом обучения в режиме реального времени. Пользователь может выбирать интересующие его свойства и режимы визуализации, а также влиять на ход обучения нейросети путем динамического изменения параметров. В программе реализованы следующие виды графической интерпретации обучения нейросетей:

- графики изменения рассчитанных величин свойств по отношению к их экспериментальным значениям (Scatter Plot);
- графики изменения абсолютных среднеквадратичных ошибок свойств в ходе обучения (History Plot);
- динамическое отображение нейросети, показывающее подстройку весовых коэффициентов и распределение значимостей дескрипторов (Network Plot);
- отображение основной статистической информации о модели и ходе обучения нейросети (Model Info);

- карты кластеризации базы данных (Factors Plot).

8.2.13. Определение области применимости модели

Для решения этой проблемы «NASAWIN» всегда сохраняет в файлах с построенными нейросетевыми моделями информацию о распределении дескрипторов по обучающей выборке (максимальные и минимальные значения), а также допустимый коэффициент «растяжки» этих границ. Эта информация затем используется на этапе прогноза для принятия решения о том, какие из соединений принадлежат областям применимости соответствующих моделей.

8.2.14. Химически-ориентированный блок прогноза

Полученные нейросетевые модели могут затем использоваться для оценки свойств новых соединений. Для прогноза отбираются только соединения, структурно родственные соединениям из выборки, для которой была построена нейросетевая модель.

Основная особенность прогнозирования в задачах выявления зависимостей «структура-свойство» в химии заключается в тесной взаимосвязи с решением «обратной задачи», заключающейся в направленном дизайне химических соединений с заранее заданными свойствами. Для обеспечения этого в «NASAWIN» предусмотрены специальные средства представления результатов прогноза и интерактивные средства взаимодействия с ними, которые специально направлены на решение «обратной задачи».

8.3. Дескрипторный блок «FRAGMENT»

Дескрипторный блок FRAGMENT предназначен для расчета фрагментных дескрипторов. Первая версия этого блока [356] была разработана как компонент программного комплекса ЭММА и предназначалась для работы в среде MS-DOS. В дальнейшем нами была создана значительно усовершенствованная версия, ориентированная на работу в среде Windows, которая используется в

нейросетевом программном комплексе NASAWIN (см. раздел 8.2) для расчета фрагментных дескрипторов. Кроме того, специальная версия этого дескрипторного блока FRAGMDLL, реализованная в виде библиотеки динамического связывания (dll), входит в состав автономной программы-прогнозатора (см. раздел 8.5) и дескрипторных блоков, реализующих нейросетевые модели «структурасвойство» в рамках многоуровневого подхода (см. раздел 7.3.1). Принципы построения и генерации фрагментных дескрипторов, реализованные в данном дескрипторном блоке, описаны в разделе 5.1. Дескрипторный блок FRAGMENT написан на языке Delphi и содержит около 18,5 тысяч строк исходного текста.

Программный комплекс NASAWIN использует дескрипторный блок FRAGMENT в двух случаях: 1) при построении новой модели и 2) при прогнозировании свойств на основе уже подготовленной модели.

В первом случае управление генерацией фрагментов осуществляется при помощи диалоговых окон, которые позволяют пользователю специфицировать: 1) максимальный размер (число атомов) генерируемых фрагментов; 2) типы фрагментов (цепочечные, циклические, разветвленные, би- и трициклические); 3) уровни обобщения для каждого вида фрагментов; 4) необходимость отбрасывания «редких» фрагментов (а также задать минимальное число структур, в которых должен встречаться каждый из генерируемых фрагментов); 5) необходимость оставлять из группы статистически эквивалентных фрагментных дескрипторов (т.е. принимающих одинаковые либо пропорциональные друг другу значения для всех соединений выборки) только один; 6) необходимость генерации фрагментов с «выделенными» атомами (см. раздел 5.3); 6) необходимость использования файла масок для подробной спецификации типов генерируемых фрагментов; 7) необходимость использования файла, содержащего структуры нестандартных фрагментов произвольной сложности, для которых FRAGMENT должен осуществлять расчет фрагментных дескрипторов.

При работе дескрипторного блока FRAGMENT в режиме прогноза управление генерацией фрагментов осуществляется при помощи специального текстового файла-маски, содержащего список кодов необходимых фрагментов.

8.4. Дескрипторный блок «FRAGPROP»

В Табл. 34 представлены вычисляемые дескрипторным блоком FRAGPROP дескрипторы для фрагментов с размером от 1 до 5 атомов:

Табл. 34. Дескрипторы FRAGPROP

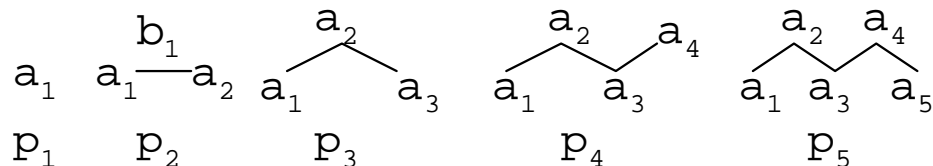
1	$p1_{Na=N_a}$	Число атомов в молекуле
2	$p1_{Ne=N_e} = \sum_{i=1}^{N_a} n_{ei}$	Общее число электронов в молекуле
3	$p1_{ANe} = N_e / N_a$	Среднее число электронов в атоме.
4	$p1_{SR2} = \sum_{i=1}^{N_a} R_i^2$	Сумма квадратов атомных радиусов в молекулы
5	$p1_{AR2} = \frac{1}{N_a} \sum_{i=1}^{N_a} R_i^2$	Среднее значение квадрата атомного радиуса
6	$p1_{SR3} = \sum_{i=1}^{N_a} R_i^3$	Сумма кубов атомных радиусов в молекуле
7	$p1_{AR3} = \frac{1}{N_a} \sum_{i=1}^{N_a} R_i^3$	Среднее значение куба атомного радиуса
8	$p1_{SR3E} = \sum_{i=1}^{N_a} \frac{R_i^3}{n_{ei}}$	Сумма отношений кубов атомных радиусов к числу электронов в этих атомах
9	$p1_{AR3E} = \frac{1}{N_a} \sum_{i=1}^{N_a} \frac{R_i^3}{n_{ei}}$	Среднее значение отношения куба атомного радиуса к числу электронов на атоме
10	$p1_{SE} = \sum_{i=1}^{N_a} \chi_i$	Сумма электроотрицательностей всех атомов в молекуле.
11	$p1_{AE} = \frac{1}{N_a} \sum_{i=1}^{N_a} \chi_i$	Среднее значение электроотрицательности.
12	$p1_{LE} = \min(\chi_i)$	Минимальная электроотрицательность атома в молекуле.
13	$p1_{HE} = \max(\chi_i)$	Максимальная электроотрицательность атома в молекуле.
14	$p1_{SI} = \sum_{i=1}^{N_a} I_i$	Сумма атомных потенциалов ионизации в молекуле.
15	$p1_{AI} = \frac{1}{N_a} \sum_{i=1}^{N_a} I_i$	Средний потенциал ионизации атомов в молекуле.
16	$p1_{LI} = \min_i(I_i)$	Минимальный потенциал ионизации атома в молекуле.
17	$p1_{HI} = \max_i(I_i)$	Максимальный потенциал ионизации атома в молекуле.

18	$p1_Nlp$	Количество неподеленных электронных пар в молекуле.
19	$p1_Npi = N_{\pi}$	Количество π -электронов в молекуле
20	$p1_SC = \sum_{i=1}^{N_a} q_i$	Суммарный заряд молекулы
21	$p1_SC2 = \sum_{i=1}^{N_a} q_i^2$	Сумма квадратов формальных зарядов на атомах
22	$p1_Nb = N_b$	Количество химических связей в молекуле.
23	$p1_Nbc = N_{bc}$	Количество входящих в циклы химических связей в молекуле
24	$p2_SDEH = \sum_{p2 a_2 \neq H} \chi(a_1) - \chi(H) $	Сумма модулей разностей электроотрицательностей для всех связей X-H в молекуле.
25	$p2_SDEHnc = \sum_{p2 a_2 \neq H, a_1 \neq C} \chi(a_1) - \chi(H) $	Сумма модулей разностей электроотрицательностей для всех связей X-H в молекуле, где X-гетероатом
26	$p2_SDE = \sum_{p2} \chi(a_1) - \chi(a_2) \cdot n_b$	Сумма по всем связям в молекуле произведений модулей разностей электроотрицательностей атомов на порядок связи между ними
27	$p2_SDEnh = \sum_{p2 a_1 \neq H, \wedge a_2 \neq H} \chi(a_1) - \chi(a_2) \cdot n_b$	Сумма по всем связям в молекуле произведений модулей разностей электроотрицательностей неводородных атомов на порядок связи между ними
28	$p2_ADE = \frac{1}{N_b} \sum_{p2} \chi(a_1) - \chi(a_2) \cdot n_b$	Среднее значение произведений модулей разностей электроотрицательностей атомов на порядок связи между ними
29	$p2_ADEnh = \frac{1}{N_b} \sum_{p2 a_1 \neq H, \wedge a_2 \neq H} \chi(a_1) - \chi(a_2) \cdot n_b$	Среднее значение произведений модулей разностей электроотрицательностей неводородных атомов на порядок связи между ними
30	$p2_HDE = \max_{p2} (\chi(a_1) - \chi(a_2) \cdot n_b)$	Максимальное значение произведения модуля разности электроотрицательностей для всех связей в молекуле на порядок соответствующей связи
31	$p2_SPE = \sum_{p2} \chi(a_1) \cdot \chi(a_2)$	Сумма произведений электроотрицательности атомов для всех связей в молекуле.

32	$p2_APE = \frac{1}{N_b} \sum_{p^2} \chi(a_1) \cdot \chi(a_2)$	Среднее значение произведения электроотрицательностей атомов для всех связей в молекуле.
33	$p2_HPE = \max_{p^2} (\chi(a_1) \cdot \chi(a_2))$	Максимальное значение произведения электроотрицательностей атомов для всех связей в молекуле.
34	$p2_SPR = \sum_{p^2} R(a_1) \cdot R(a_2)$	Сумма произведений атомных радиусов для всех связей в молекуле.
35	$p2_APR = \frac{1}{N_b} \sum_{p^2} R(a_1) \cdot R(a_2)$	Среднее значение произведений атомных радиусов для всех связей в молекуле.
36	$p2_HPR = \max_{p^2} (R(a_1) \cdot R(a_2))$	Максимальное значение произведения атомных радиусов для всех связей в молекуле.
37	$p3_SPR = \sum_{p^3} R(a_1) \cdot R(a_3)$	Сумма произведений радиусов атомов, разделенных двумя связями.
38	$p3_APR = \frac{1}{N_{p^3}} \sum_{p^3} R(a_1) \cdot R(a_3)$	Среднее значение произведений радиусов атомов, разделенных двумя связями.
39	$p3_SPDE = \sum_{p^3} (\chi(a_1) - \chi(a_2)) \cdot (\chi(a_3) - \chi(a_2))$	Сумма произведений разностей электроотрицательности атомов в положениях 1-2 и 3-2 для всех трехатомных связных фрагментов.
40	$p3_APDE = \frac{1}{N_{p^3}} \sum_{p^3} (\chi(a_1) - \chi(a_2)) \cdot (\chi(a_3) - \chi(a_2))$	Среднее значение произведений разностей электроотрицательности атомов в положениях 1-2 и 3-2 для всех трехатомных связных фрагментов.
41	$p3_SPDEnh = \sum_{p^3 a_1 \neq H \wedge a_2 \neq H} (\chi(a_1) - \chi(a_2)) \cdot (\chi(a_3) - \chi(a_2))$	Сумма произведений разностей электроотрицательности неводородных атомов в положениях 1-2 и 3-2 для всех трехатомных связных фрагментов
42	$p3_APDEnh = \frac{1}{N_{p^3}} \sum_{p^3 a_1 \neq H \wedge a_2 \neq H} (\chi(a_1) - \chi(a_2)) \cdot (\chi(a_3) - \chi(a_2))$	Среднее значение произведений разностей электроотрицательности неводородных атомов в положениях 1-2 и 3-2 для всех трехатомных связных фрагментов

43	$p4_SPR = \sum_{p4} R(a_1) \cdot R(a_4)$	Сумма произведений атомных радиусов в положениях 1-4.
44	$p4_APR = \frac{1}{N_{p4}} \sum_{p4} R(a_1) \cdot R(a_2)$	Среднее значение произведений атомных радиусов в положениях 1-4 по всем 4-атомным цепочкам.
45	$p4_SPDE = \sum_{p4} (\chi(a_1) - \chi(a_2)) \cdot (\chi(a_4) - \chi(a_3))$	Сумма произведений разностей электроотрицательности атомов в положениях 1-2 и 4-3 для всех 4-атомных цепочек.
46	$p4_APDE = \frac{1}{N_{p4}} \sum_{p4} (\chi(a_1) - \chi(a_2)) \cdot (\chi(a_4) - \chi(a_3))$	Среднее значение произведений разностей электроотрицательности атомов в положениях 1-2 и 4-3 для всех 4-атомных цепочек.
47	$p5_SPR = \sum_{p5} R(a_1) \cdot R(a_5)$	Сумма произведений атомных радиусов в положениях 1-5.
48	$p5_APR = \frac{1}{N_{p5}} \sum_{p5} R(a_1) \cdot R(a_5)$	Среднее значение произведений атомных радиусов в положениях 1-5 по всем 5-атомным цепочкам.
49	$p5_SPDE = \sum_{p5} (\chi(a_1) - \chi(a_2)) \cdot (\chi(a_5) - \chi(a_4))$	Сумма произведений разностей электроотрицательности атомов в положениях 1-2 и 5-4 для всех 5-атомных цепочек.
50	$p5_APDE = \frac{1}{N_{p5}} \sum_{p5} (\chi(a_1) - \chi(a_2)) \cdot (\chi(a_5) - \chi(a_4))$	Среднее значение произведений разностей электроотрицательности атомов в положениях 1-2 и 5-4 для всех 5-атомных цепочек.

где χ - электроотрицательность, r_a - атомный ковалентный радиус, n_e - количество электронов, I - потенциал ионизации, q - формальный заряд на атоме, N_π - количество π -электронов в молекуле, n_{e_s} - количество электронов в атоме N_{lp} - количество неподеленных пар электронов в молекуле, N_{pn} - число цепочек длиной n в молекуле, a_n (атомы), b_n (связи) и p_n (цепочки атомов) определяются следующим образом:



8.5. Автономные прогнозаторы свойств органических соединений

Кроме «химически-ориентированного блока прогнозы», встроенного в основную программу NASAWIN и фактически являющегося одним из ее режимов работы, программный комплекс NASAWIN включает три типа автономных прогнозаторов свойств органических соединений: 1) интерактивный; 2) запускаемый с командной строки; 3) встроенный в дескрипторный блок. Работа всех трех вышеперечисленных типов программ основана на том, что в NASAWIN предусмотрена возможности записи построенной модели (нейросетевой либо линейно-регрессионной) в виде файла, содержащего исходный код процедуры на языке C++, осуществляющей необходимые для прогнозирования вычисления. Этот файл предназначен для того, чтобы его оттранслировали при помощи компилятора MS Visual Studio C++ и связали со специальными библиотеками в составе NASAWIN, осуществляющими целый ряд необходимых для прогнозирования операций: расчет, преобразования и шкалирование дескрипторов, проверка области применимости модели и др. В результате получается упакованный в dll-файл динамической библиотеки «вычислительный сервер», осуществляющий все необходимые для осуществления прогноза вычисления. Три вышеупомянутые типа автономных прогнозаторов работают с одними и теми же «вычислительными серверами», с которыми связываются «вычислительные клиенты» соответствующих типов: 1) интерактивный, 2) запускаемый с командной строки и 3) встроенный в дескрипторный блок.

Интерактивный «вычислительный клиент» представляет собой программу, работающую под управлением операционной системы Windows, которая позволяет пользователю в интерактивном режиме загружать MOL- и SDF-файлы, содержащие структуры соединений для прогноза, просматривать эти структуры, выбирать нужные свойства для прогноза (из числа заранее подготовленных «вычислительных серверов»), осуществлять прогноз и выводить на экран либо в файл результаты прогноза. Программа также проверяет область применимости моделей, однако эта возможность может быть отключена. Прогноз осуществляется как для регрессионных, так и для классификационных мо-

делей. Программа также позволяет прогнозировать одно и то же свойство (активность) по нескольким моделям, и выдавать результат в виде усредненного значения по нескольким регрессионным моделям, либо в виде консенсуса предсказаний, сделанных по нескольким классификационным моделям. Выдаваемая при этом информация позволяет также оценивать надежность прогнозирования.

Работающий из-под командной строки «вычислительный клиент» предоставляет возможность осуществлять прогноз для выборки соединений, заданной в виде файла, с записью результатов прогноза тоже в файл. Этот тип прогнозаторов предназначен для работы из-под Web-сервера, обеспечивая взаимодействие с пользователем через Интернет.

Наконец, третий тип «вычислительного клиента» позволяет использовать программу прогнозирования в качестве дескрипторного блока, что предоставляет возможность осуществления многоуровневого подхода к прогнозированию свойств органических соединений (см. раздел 7.4.1). Следует отметить, что встроенные в дескрипторные блоки программы-прогнозаторы могут быть использованы рекурсивно, т.е. дескрипторные блоки, вызываемые из программ-прогнозаторов, также могут представлять собой программы-прогнозаторы.

ВЫВОДЫ

1. Теоретически обоснован и разработан универсальный подход к прогнозированию свойств органических соединений на основе комбинированного использования искусственных нейронных сетей и фрагментных дескрипторов.
2. В рамках развития нейросетевых подходов разработаны: а) трехвыборочный подход и на его основе - процедуры трехвыборочного и двойного скользящего контроля, позволяющие эффективно предотвращать «переучивание» нейросетей и объективно оценивать прогнозирующую способность нейросетевых моделей; б) статистический метод быстрой пошаговой множественной линейной регрессии, позволяющий эффективно осуществлять отбор дескрипторов для построения нейросетевых моделей; в) метод интерпретации нейросетевых регрессионных моделей, позволяющий описывать характер найденных зависимостей; г) концепция «обучаемой симметрии», позволяющая улучшать прогнозирующую способность моделей «структура-свойство» за счет корректного учета в них свойств симметрии.
3. В рамках развития фрагментных подходов разработаны: а) иерархическая система классификации типов атомов, входящих в состав фрагментов, а также структура и алгоритм генерации фрагментных дескрипторов, ориентированных на прогнозирование свойств органических соединений; б) концепция фрагментов с «выделенными» атомами, позволяющая прогнозировать: локальные свойства органических соединений; константы заместителей и скоростей реакций; свойства полимерных и супрамолекулярных соединений; биологическую активность внутри рядов органических соединений с учетом стереохимической информации; в) концепция псевдофрагментных дескрипторов как средство повышения прогнозирующей способности моделей «структура-свойство» за счет решения проблемы «редких» фрагментов.
4. В рамках развития интегрированных подходов разработаны: а) методы интеграции нейросетевого и молекулярного моделирования, ведущие к значительному улучшению прогнозирующей способности построенных моделей; б) концепция построения нейросетевых моделей «структура-условия-свойство»,

позволяющая прогнозировать разнообразные свойства и реакционную способность органических соединений при различных внешних условиях; в) методы объединения нейросетевых моделей на основе концепций многоуровневого и многозадачного обучения, позволяющие повышать прогнозирующую способность моделей за счет интеграции разнородных экспериментальных данных; г) концепция проведения прямых корреляций «структура-свойство» и на ее основе специальные архитектуры нейронных сетей, позволяющие осуществлять прогнозирование свойств органических соединений непосредственно из описания молекулярного графа без предварительного вычисления молекулярных дескрипторов.

5. Разработан программный комплекс, позволяющий в полном объеме осуществить весь цикл работ по построению моделей «структура-свойство» и «структура-условия-свойство», и с их помощью осуществлять прогнозирование самых разнообразных свойств органических соединений.
6. Построены модели для прогнозирования 62 разнообразных свойств органических соединений: а) температуры кипения и плавления, молярного объема, молярной рефракции, теплоты испарения, критической температуры, критического давления и поверхностного натяжения алканов; б) октанового числа, вязкости, теплоты испарения и плотности углеводородов; в) динамической вязкости и плотности углеводородов при разной температуре; г) температуры кипения, вязкости, плотности, давления насыщенных паров, поляризуемости, магнитной восприимчивости, энтальпии сублимации, энтальпии парообразования, температуры вспышки, теплоты сольватации в циклогексане, анестетического давления газов, липофильности, значений 4 констант Абрахама, коэффициента сорбции в почве и растворимости фуллерена C_{60} для разнообразных соединений, принадлежащих к разным классам; д) констант ионизации фенолов, карбоновых кислот и азотсодержащих соединений; е) положения длинноволновой полосы поглощения спиртового раствора симметричных цианиновых красителей; ж) энтальпии образования алифатических полинитросоединений; з) сродства азо- и антрахиноновых красителей к целлюлозному волокну; и) химических сдвигов в ^{31}P ЯМР спектрах производных монофосфинов; й) температуры плавления

ления ионных жидкостей, представляющих собой бромиды производных пиридинов, имидазолов, бензимидазолов и четвертичных солей аммония; к) показателя преломления, плотности и температуры стеклования аморфных полимеров; л) константы скорости гидролиза сложных эфиров карбоновых кислот при разной температуре и разном составе растворителя; м) констант заместителей σ^m , σ^p , F , R , E_s ; н) 11 констант распределения «ткань-воздух» для произвольных органических соединений; о) мутагенной активности нитропроизводных гетероциклических аналогов полициклических углеводородов и бифенила; п) блокирующей способности дигидропиридинов по отношению к ионным каналам L-типа; р) галлюциногенной активности фенилалкиламинов; с) способности аналогов НЕРТ ингибировать обратную транскриптазу вируса ВИЧ-1; т) эмбриотоксичности синтетических аналогов биогенных аминов.

ЛИТЕРАТУРА

1. Гиллер С.А.; Глаз А.Б.; Растригин Л.А.; Розенблит А.Б. Распознавание физиологической активности химических соединений на перцептроне со случайной адаптацией структуры. // ДАН СССР. - 1971. - Т. 199, № 4. - С. 851-853.
2. Hiller S.A.; Golender V.E.; Rosenblit A.B.; Rastrigin L.A.; Glaz A.B. Cybernetic methods of drug design. I. Statement of the problem--the perceptron approach. // Comput. Biomed. Res. - 1973. - V. 6, № 5. - P. 411-421.
3. Zupan J.; Gasteiger J. Neural networks: a new method for solving chemical problems or just a passing phase? // Anal. Chim. Acta. - 1991. - V. 248, № 1. - С. 1-30.
4. McCulloch W.S.; Pitts W. A logical calculus of the ideas immanent in nervous activity. // Bull. Math. Biophys. - 1943. - V. 5. - P. 115-133.
5. Розенблатт Ф. Принципы нейродинамики. - Мир: М. - 1964. - 480 с.
6. Нильсен Н. Обучающиеся машины. - Мир: М. - 1967. - 506 с.
7. Минский М.; Пейперт С. Перцептроны. - Мир: М. - 1971. - 261 с.
8. Мкртчян С.О. Нейроны и нейронные сети (Введение в теорию формальных нейронов и нейронных сетей). - Энергия: М. - 1971. - 232 с.
9. Галушкин А.И. Синтез многослойных систем распознавания образов. - Энергия: М. - 1974. - 376 с.
10. Rumelhart D.E.; McClelland J.L. Parallel Distributed Processing. - MIT Press: Cambridge, MA. - 1986. - Vol. 1,2.
11. Горбань А.Н. Обучение нейронных сетей. - ПараГраф: М. - 1990. - 160 с.
12. Freeman J.A.; Skapura D.M. Neural networks: algorithms, applications, and programming techniques. - Addison-Wesley Publishing Company: Menlo Park, CA - 1991. - 414 p.
13. Уоссерман Ф. Нейрокомпьютерная техника. - Мир: М. - 1992. - 240 с.
14. Ritter H.; Martinetz T.; Schulten K. Neural Computation and Self-Organizing Maps – An Introduction. - Addison-Wesley: New York. - 1992. - 293 p.
15. Vealanturf L.P.J. Analysis and Applications of Artificial Neural Networks. - Prentice Hall: NY - 1995. - 242 p.

16. Горбань А.Н.; Россиев Д.А. Нейронные сети на персональном компьютере. - Наука: Новосибирск. - 1996. - 276 с.
17. Bigus J.P. Data mining with neural networks: solving business problems – from application development to decision support. - McGraw-Hill: NY. - 1996. - 221 p.
18. Ежов А.А.; Шумский С.А. Нейрокомпьютинг и его приложения в экономике и бизнесе. - МИФИ: М. - 1998. - 224 с.
19. Галушкин А.И. Теория нейронных сетей. Кн. 1. - ИПРЖР: М. - 2000. - 416 с.
20. Kohonen T. Self-Organizing Maps. - Springer: - 2001. - 260 p.
21. Головкин В.А. Нейронные сети: обучение, организация и применение. - ИПРЖР: М. - 2001. - 256 с.
22. Круглов В.В.; Борисов В.В. Искусственные нейронные сети. Теория и практика. - Горячая линия – Телеком: М. - 2001. - 382 с.
23. Каллан Р. Основные концепции нейронных сетей. - Издательский дом «Вильямс»: М. - 2001. - 291 с.
24. Rabunal J.R.; Dorrado J. Artificial Neural Networks in Real-Life Applications. - IGP: Hershey, London, Melbourne, Singapore. - 2006. - 395 p.
25. Агеев А.Д.; Балухто А.Н.; Бычков А.В.; др. Нейроматематика. Кн. 6: Учебное пособие для вузов. - ИПРЖР: М. - 2002. - 448 с.
26. Мкртчян С.О. Проектирование логических устройств ЭВМ на нейронных элементах. - Энергия: М. - 1977. - 482 с.
27. Кирсанов Э.Ю. Цифровые нейрокомпьютеры: Архитектура и схемотехника. - Изд-во Казан. гос. техн. ун-та: Казань. - 1995. - 131 с.
28. Галушкин А.И. Нейрокомпьютеры. Кн. 3: Учебное пособие для вузов. - ИПРЖР: М. - 2000. - 528 с.
29. Комарцова Л.Г.; Максимов А.В. Нейрокомпьютеры: Учебное пособие для вузов. - Изд-во МГТУ им. Н.Э. Баумана: М. - 2002. - 320 с.
30. Gasteiger J.; Zupan J. Neural Networks in Chemistry. // Angew. Chem. Int. Ed. Engl. - 1993. - V. 105, № 4. - P. 503-527.

31. *Aoyama T.; Ichikawa H.* Neural Networks Applied to Pharmaceutical Problems. IV. Basic Operating Characteristics of Neural Networks When Applied to Structure-Activity Studies. // *Chem. Pharm. Bull.* - 1991. - V. 39, № 2. - P. 358-366.
32. *Burns J.A.; Whitesides G.M.* Feed-forward neural networks in chemistry: mathematical systems for classification and pattern recognition. // *Chem. Rev.* - 1993. - V. 93, № 8. - P. 2583-2601.
33. *Devillers J.* Neural Networks in QSAR and Drug Design. - Academic Press: London. - 1996. - 284 p.
34. *Zupan J.; Gasteiger J.* Neural Networks in Chemistry. - Wiley-VCH: Weinheim. - 1999. - 380 p.
35. *Баскин И.И.; Палюлин В.А.; Зефирова Н.С.* Применение искусственных нейронных сетей в химических и биохимических исследованиях. // *Вестн. Моск. ун-та. Сер. 2. Хи-мия.* - 1999. - Т. 40, № 5. - С. 323-326.
36. *Kovesdi I.; Dominguez-Rodriguez M.F.; Orfi L.; Naray-Szabo G.; Varro A.; Papp J.G.; Matyus P.* Application of neural networks in structure-activity relationships. // *Med Res Rev.* - 1999. - V. 19, № 3. - P. 249-269.
37. *Гальберштам Н.М.; Баскин И.И.; Палюлин В.А.; Зефирова Н.С.* Нейронные сети как метод поиска зависимостей структура – свойство органических соединений. // *Успехи химии.* - 2003. - Т. 72, № 7. - С. 706-727.
38. *Баскин И.И.; Палюлин В.А.; Зефирова Н.С.* Применение искусственных нейронных сетей для прогнозирования свойств химических соединений. // *Нейрокомпьютеры: разработка, применение.* - 2005. - Т. № 1-2. - С. 98-101.
39. *Баскин И.И.; Палюлин В.А.; Зефирова Н.С.* Многослойные персептроны в исследовании зависимостей «структура-свойство» для органических соединений. // *Рос. хим. ж. (Ж. Рос. хим. об-ва им. Д.И. Менделеева).* - 2006. - Т. 50, № - С. 86-96.
40. *Baskin I.I.; Palyulin V.A.; Zefirov N.S.* Neural networks in building QSAR models. // *Methods in molecular biology (Clifton, N.J.).* - 2008. - V. 458. - P. 137-158.
41. *Rumelhart D.E.; Hinton G.E.; Williams R.J.* Learning Internal Representations by Back-Propagating Errors. // *Nature.* - 1986. - V. 323, № 6088 - P. 533-536.

42. *Rumelhart D.E.; Hinton G.E.; Williams R.J.* Learning internal representation by error propagation. // *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 1: Foundations.*, Rumelhart D.E.; McClelland J.L., Eds. MIT Press: Cambridge, MA. - 1986. - P. 318-362.
43. *Widrow B.; Hoff M.E.* Adaptive switching circuits. // 1960 IREWESCON Convention Record, IRE: New York. - 1960. - P. 96-104.
44. *Lehtokangas M.; Saarinen J.* Weight initialization with reference patterns. // *Neurocomputing.* - 1998. - V. 20, № 1-3. - P. 265-278.
45. *Yam J.Y.F.; Chow T.W.S.* A weight initialization method for improving training speed in feedforward neural network. // *Neurocomputing.* - 2000. - V. 30, № 1-4. - P. 219-232.
46. *Patnaik L.M.; Rajan K.* Target detection through image processing and resilient propagation algorithms. // *Neurocomputing.* - 2000. - V. 35, № 1-4. - P. 123-135.
47. *Riedmiller M.; Braun H.* A direct adaptive method for faster backpropagation learning: The RPROP algorithm. // *Proceedings of the IEEE International Conference on Neural Networks.* - 1993. - P. 586-591.
48. *Hagan M.T.; Demuth H.B.; Beale M.H.* *Neural Network Design.* - PWS Publishing: Cambridge, MA. - 1996. - 252 p.
49. *Медведев В.С.; Потемкин В.Г.* Нейронные сети. МАТЛАБ 6. - ДИАЛОГ-МИФИ: М. - 2002. - 496 с.
50. *Charalambous C.* Conjugate gradient algorithm for efficient training of artificial neural networks. // *IEEE Proceedings.* - 1992. - V. 139, № 3. - P. 301-310.
51. *Fletcher R.; Reeves C.M.* Function minimization by conjugate gradients. // *Computer Journal.* - 1964. - V. 7. - P. 149-154.
52. *Dennis J.; Schnabel R.B.* *Numerical Methods for Unconstrained Optimization and Nonlinear Equations.* - Prentice-Hall: Englewood Cliffs, NJ. - 1983. - 378 p.
53. *Hagan M.T.; Menhaj M.* Training feedforward networks with the Marquardt algorithm. // *IEEE Transactions on Neural Networks.* - 1994. - V. 5, № 6. - P. 989-993.

54. *Karelsen M.; Dobchev D.A.; Kulshyn O.V.; Katritzky A.R.* Neural networks convergence using physicochemical data. // *J. Chem. Inf. Model.* - 2006. - V. 46, № 5. - P. 1891-1897.
55. *Kohonen T.* The self-organizing map. // *Neurocomputing.* - 1998. - V. 21, № 1-3. - P. 1-6.
56. *Linde Y.; Buzo A.; Gray R.M.* An algorithm for vector quantization. // *IEEE Trans. Communication.* - 1980. - P. 28, № 1. - P. 84-95.
57. *Gray R.M.* Vector quantization. // *IEEE ASSP Mag.* - 1984. - V. 1, № 2. - P. 4-29.
58. *Gersho A.* On the structure of vector quantizers. // *IEEE Trans. Inform. Theory.* - 1979. - V. 25, № 4. - P. 373-380.
59. *Martinez T.M.; Berkovich S.G.; Schulten K.J.* "Neural-Gas" network for vector quantization and its applications to time-series prediction. // *IEEE Trans. Neural Networks.* - 1993. - V. 4, № 4. - P. 558-569.
60. *Questier F.; Guo Q.; Walczak B.; Massart D.L.; Boucon C.; de Jong S.* The Neural Gas network for classifying analytical data. // *Chemom. Intel. Lab. Sys.* - 2002. - V. 61, № 1-2. - P. 105-121.
61. *Daszykowski M.; Walczak B.; Massart D.L.* On the Optimal Partitioning of Data with K-Means, Growing K-Means, Neural Gas, and Growing Neural Gas. // *J. Chem. Inf. Comput. Sci.* - 2002. - V. 42, № 6. - P. 1378-1389.
62. *Fritzke B.* A growing neural gas network learns topologies. // *Advances in neural information processing systems*, Tesauro G.; Touretzky D.S.; Leen T.K., Eds. MIT Press: Cambridge, MA. - 1995. - V. 7. - P. 625-632.
63. *Kohonen T.* The Self-Organizing Map. // *Proc. IEEE.* - 1990. - V. 78, № 9. - P. 1464-1480.
64. *Baurin N.; Mozziconacci J.-C.; Arnoult E.; Chavatte P.; Marot C.; Morin-Allory L.* Two Dimensional QSAR Consensus Prediction for High-Throughput Virtual Screening. An Application to COX-2 Inhibition Modeling and Screening of the NCI Database. // *J. Chem. Inf. Comput. Sci.* - 2004. - V. 44, № 1. - P. 276-285.
65. *Hecht-Nielsen R.* Counterpropagation networks. // *Applied Optics.* - 1987. - V. 26, № 23. - P. 4979-4984.

66. *Grossberg S.* Some networks that can learn, remember and reproduce any number of complicated space-time patterns. // *Journal of Mathematics and Mechanics.* - 1969. - V. 19, № 1. - P. 53-91.
67. *Moody J.; Darken C.* Learning in networks of locally-tuned processing units. // *Neural Comput.* - 1989. - V. 1, № 2. - P. 281-294.
68. *Bishop C.* *Neural Networks for Pattern Recognition.* - Oxford University Press: Walton Street, Oxford OX2 6DP. - 1995. - 251 p.
69. *Hartman E.; Keeler J.D.; Kowalski J.M.* Layered neural networks with Gaussian hidden units as universal approximations. // *Neural Comput.* - 1990. - V. 2, № 2. - P. 210-215.
70. *MacQueen J.B.* Some Methods for classification and Analysis of Multivariate Observations. // *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press: Berkeley. - 1967. - V. 1. - P. 281-297.
71. *Likas A.; Vlassis N.; Verbeek J.J.* The Global K-Means Clustering Algorithm. // *Pattern Recognit.* - 2003. - V. 36, № 2. - P. 451-461.
72. *Golub G.H.; Kahan W.* Calculating the singular values and pseudoinverse of a matrix. // *J. SIAM Numer. Anal. Ser. B* - 1965. - V. 2, № 3. - P. 205-224.
73. *Specht D.* Probabilistic Neural Networks. // *Neural Networks.* - 1990. - V. 3, № 1. - P. 109-118.
74. *Specht D.* A General Regression Neural Network. // *IEEE Trans. Neural Networks.* - 1991. - V. 2, № 6. - P. 568-576.
75. *Nadaraya E.A.* On Non-Parametric Estimates of Density Functions and Regression Curves. // *Theory. Probability Its Appl.* - 1965. - V. 10, № 1. - P. 186-190.
76. *Watson G.S.* Smooth regression analysis. // *Sankhya, Ser. A.* - 1964. - V. 26, № 4. - P. 359-372.
77. *Parzen E.* On estimation of a probability density function and mode. // *Annals of Mathematical Statistics.* - 1962. - V. 33, № 3. - P. 1065-1076.
78. *Carpenter G.; Grossberg S.* Neural dynamics of category learning and recognition: Attention, memory consolidation and amnesia. // *Brain Structure, Learning and Memory (AAAS Symposium Series)*, Davis J.; Newburgh R.; Wegman E., Eds. Westview Press. - 1987. - P. 233-290.

79. *Carpenter G.A.; Grossberg S.* A massively parallel architecture for a self-organizing neural pattern recognition machine. // *Comput. Vision Graph. Image Process.* - 1987. - V. 37, № 1. - P. 54-115.
80. *Carpenter G.; Grossberg S.* ART-2: Self-organization of stable category recognition codes for analog input patterns. // *Applied Optics.* - 1987. - V. 26, № 23. - P. 4919-4930.
81. *Grossberg S.* Competitive learning: From interactive activation to adaptive resonance. // *Cognitive Science.* - 1987. - V. 11, № 1. - P. 23-63.
82. *Carpenter G.A.; Grossberg S.; Marcuzon N.; Reynolds J.H.; Rosen D.B.* Fuzzy ARTMAP: A Neural Network Architecture for Incremental Supervised Learning of Analogue Multidimensional Maps. // *IEEE Trans. Neural Networks.* - 1992. - V. 3, № 5. - P. 698-713.
83. *Carpenter G.A.; Grossberg S.* Fuzzy ARTMAP: A synthesis of neural networks and fuzzy logic for supervised categorization and nonstationary prediction. // *Fuzzy Sets, Neural Networks, and Soft Computing*, Yager R.R.; Zadeh L.A., Eds. Van Nostrand Reinhold: New York. - 1994. - P. 126-165.
84. *Эйген М.; Шустер П.* Гиперцикл. Принципы самоорганизации макромолекул. - Мир: М. - 1982. - 272 с.
85. *Кольцова Э.М.; Гордеев Л.С.* Методы синергетики в химии и химической технологии. - "Химия": М. - 1999. - 256 с.
86. *Гарел Д.; Гарел О.* Колебательные химические реакции. - Мир: М. - 1986. - 148 с.
87. *Белоусов Б.П.* Периодически действующая реакция и ее механизм. // Сб. рефератов по радиационной медицине за 1958 г., Медгиз: М. - 1959. - С. 145-148.
88. *Жаботинский А.М.* Колебательные процессы в биологических и химических системах. - Наука: М. - 1974. - 178 с.
89. *Рамбиди Н.Г.* Биомолекулярные нейрокомпьютеры. // *Нейрокомпьютеры: разработка, применение.* - 1998. - № 1-2. - С. 27-33.

90. *Petrosian A.; Prokhorov D.; Homan R.; Dasheiff R.; Wunsch D.I.* Recurrent Neural Network based Prediction of Epileptic Seizures in Intra- and Extracranial EEG // *Нейрокомпьютеры: разработка, применение.* - 1998. - № 1-2. - С. 47-59.
91. *Cohen M.A.; Grossberg S.G.* Absolute stability of global pattern formation and parallel memory storage by competitive neural networks. // *IEEE Transactions on Systems, Man and Cybernetics.* - 1983. - V. 13. - P. 815-826.
92. *Hopfield J.J.* Neural networks and physical systems with emergent collective computational abilities. // *Proc Natl Acad Sci U S A.* - 1982. - V. 79, № 8. - P. 2554-2558.
93. *Hopfield J.J.* Neurons with graded response have collective computational properties like those of two-state neurons. // *Proc Natl Acad Sci U S A.* - 1984. - V. 81, № 10. - P. 3088-3092.
94. *Hopfield J.J.; Tank D.W.* "Neural" computation of decisions in optimization problems. // *Biol Cybern.* - 1985. - V. 52, № 3. - P. 141-152.
95. *Hopfield J.J.; Tank D.W.* Computing with neural circuits: a model. // *Science.* - 1986. - V. 233, № 4764. - P. 625-633.
96. *Hebb D.O.* *The Organization of Behavior.* - Wiley: New York. - 1949. - 335 p.
97. *Farhat N.H.; Psaltis D.; Prata A.; Paek E.* Optical implementation of the Hopfield model. // *Applied Optics.* - 1985. - V. 24, № 10. - P. 1469-1475.
98. *Hopfield J.J.; Feinstein D.I.; Palmer R.G.* 'Unlearning' has a stabilizing effect in collective memories. // *Nature.* - 1983. - T. 304, № 5922. - C. 158-159.
99. *Abu-Mostafa Y.S.; St. Jacques J.* Information capacity of the Hopfield model. // *IEEE Transactions on Information Theory.* - 1985. - V. 31, № 4. - P. 461-464.
100. *Crick F.; Mitchison G.* The function of dream sleep. // *Nature.* - 1983. - V. 304, № 5922. - P. 111-114.
101. *Ezhov A.A.; Vvedensky V.L.* Object generation with neural networks (when spurious memories are useful). // *Neural Networks.* - 1996. - V. 9, № 9. - P. 1491-1495.
102. *Ezhov A.A.* Empty classes, predictive and clustering thinking networks. // *Neural Network World.* - 1994. - V. 4. - P. 671-688.

103. *Ежов А.А.; Токаев А.Г.; Чечеткин В.Р.* NEGATRON: Нейросетевой пакет программ для анализа скрытых повторов в геномных последовательностях ДНК. // Научная сессия МИФИ – 99. Всероссийская научно-техническая конференция «Нейроинформатика-99». Сборник научных трудов. В 3 частях. Ч. 3., МИФИ: М. - 1999. - С. 182-188.
104. *Hinton G.E.; Sejnowski T.J.* Learning and relearning in Boltzmann machines. // Parallel distributed processing, MIT Press: Cambridge, MA. - 1986. - V. 1. - P. 282-317.
105. *Todeschini R.; Consonni V.* Handbook of Molecular Descriptors. - Wiley-VCH Publishers: Weinheim. - 2000. - 668 p.
106. *Aoyama T.; Suzuki Y.; Ichikawa H.* Neural networks applied to structure-activity relationships. // J. Med. Chem. - 1990. - V. 33, № 3. - P. 905-908.
107. *Aoyama T.; Suzuki Y.; Ichikawa H.* Neural networks applied to pharmaceutical problems. III. Neural networks applied to quantitative structure-activity relationship (QSAR) analysis. // J. Med. Chem. - 1990. - V. 33, № 9. - P. 2583-2590.
108. *Andrea T.A.; Kalayeh H.* Applications of neural networks in quantitative structure-activity relationships of dihydrofolate reductase inhibitors. // J. Med. Chem. - 1991. - V. 34, № 9. - P. 2824-2836.
109. *Baskin I.I.; Palyulin V.A.; Zefirov N.S.* Chapter 8. Neural Networks in Building QSAR Models. // Artificial Neural Networks: Methods and Protocols, Livingstone D.S., Ed. Humana Press, a part of Springer Science + Business Media. - 2008. - P. 139-160.
110. *Zefirov N.S.; Palyulin V.A.* Fragmental Approach in QSPR. // J. Chem. Inf. Comput. Sci. - 2002. - V. 42, № 5. - P. 1112-1122.
111. *Japertas P.; Didziapetris R.; Petrauskas A.* Fragmental methods in the design of new compounds. Applications of The Advanced Algorithm Builder. // Quant. Struct.-Act. Relat. - 2002. - V. 21, № 1. - P. 23-37.
112. *Артеменко Н.В.; Баскин И.И.; Палюлин В.А.; Зефирова Н.С.* Искусственные нейронные сети и фрагментный подход в прогнозировании физико-химических свойств органических соединений. // Изв. РАН, Сер. хим. - 2003. - № 1. - С. 19-28.

113. *Merlot C.; Domine D.; Church D.J.* Fragment analysis in small molecule discovery. // *Curr. Opin. Drug Discov. Devel.* - 2002. - V. 5, № 3. - P. 391-399.
114. *Varnek A.; Fourches D.; Hoonakker F.; Solov'ev V.P.* Substructural fragments: an universal language to encode reactions, molecular and supramolecular structures. // *J. Comput. Aided Mol. Des.* - 2005. - V. 19, № 9-10. - P. 693-703.
115. *Baskin I.; Varnek A.* Building a chemical space based on fragment descriptors. // *Comb. Chem. High Throughput Screening.* - 2008. - V. 11, № 8. - P. 661-668.
116. *Baskin I.; Varnek A.* Fragment Descriptors in SAR/QSAR/QSPR Studies, Molecular Similarity Analysis and in Virtual Screening. // *Chemoinformatics Approaches to Virtual Screening* Varnek A.; Tropsha A., Eds. RSC Publisher: Cambridge. - 2008. - P. 1-43.
117. *Vogel A.I.* Atomic parachors of carbon and hydrogen. // *Chemistry & Industry (London, United Kingdom).* - 1934. - P. 85.
118. *Zahn C.T.* The Significance of Chemical Bond Energies. // *J. Chem. Phys.* - 1934. - V. 2. - P. 671-680.
119. *Souders M.; Matthews C.S.; Hurd C.O.* Relationship of Thermodynamic Properties to Molecular Structure. Heat Capacities and Heat Contents of Hydrocarbon Vapors. // *Ind. Eng. Chem.* - 1949. - V. 41, № 5. - P. 1037-1048.
120. *Souders M.; Matthews C.S.; Hurd C.O.* Entropy and Heat of Formation of Hydrocarbon Vapors. // *Ind. Eng. Chem.* - 1949. - V. 41, № 5. - P. 1048-1056.
121. *Franklin J.L.* Prediction of Heat and Free Energies of Organic Compounds. // *Ind. Eng. Chem.* - 1949. - V. 41, № 5. - P. 1070-1076.
122. *Franklin J.L.* Calculation of the Heats of Formation of Gaseous Free Radicals and Ions. // *J. Chem. Phys.* - 1953. - V. 21, № 11. - P. 2029-2033.
123. *Татевский В.М.* Химическое строение углеводородов и их теплоты образования. // *ДАН СССР.* - 1950. - Т. 25, № 6. - С. 819-822.
124. *Bernstein H.J.* The Physical Properties of Molecules in Relation to Their Structure. I. Relations between Additive Molecular Properties in Several Homologous Series. // *J. Chem. Phys.* - 1952. - V. 20, № 2. - P. 263-269.
125. *Laidler K.J.* System of Molecular Thermochemistry for Organic Gases and Liquids. // *Canadian J. Chem.* - 1956. - V. 34. - P. 626-648.

126. *Benson S.W.; Buss J.H.* Additivity Rules for the Estimation of Molecular Properties. Thermodynamic Properties. // *J. Chem. Phys.* - 1958. - V. 29, № 3. - P. 546-572.
127. *Allen T.L.* Bond Energies and the Interactions between Next-Nearest Neighbors. I. Saturated Hydrocarbons, Diamond, Sulfanes, S₈, and Organic Sulfur Compounds. // *J. Chem. Phys.* - 1959. - V. 31, № 4. - P. 1039-1049.
128. *Смоленский Е.А.* Применение теории графов для вычисления структурно-аддитивных свойств углеводородов. // *Журн. физ. химии.* - 1964. - Т. 38, № 5. - С. 1288-1291.
129. *Free S.M., Jr.; Wilson J.W.* A Mathematical Contribution to Structure-Activity Studies. // *J. Med. Chem.* - 1964. - V. 7, № 4 - P. 395-399.
130. *Golender V.E.; Rozenblit A.B.* Logico-structural approach to computer-assisted drug design. // *Med. Chem. (Academic Press).* - 1980. - V. 11, №. 9. - P. 299-337.
131. *Avidon V.V.; Pomerantsev I.A.; Golender V.E.; Rozenblit A.B.* Structure-Activity Relationship Oriented Languages for Chemical Structure Representation. // *J. Chem. Inf. Comput. Sci.* - 1982. - V. 22, № 4. - P. 207-214.
132. *Cramer R.D., 3rd; Redl G.; Berkoff C.E.* Substructural analysis. A novel approach to the problem of drug design. // *J. Med. Chem.* - 1974. - V. 17, № 5. - P. 533-535.
133. *Brugger W.E.; Stuper A.J.; Jurs P.C.* Generation of Descriptors from Molecular Structures. // *J. Chem. Inf. Model.* - 1976. - V. 16, № 2. - P. 105-110.
134. *Stuper A.J.; Jurs P.C.* ADAPT: A Computer System for Automated Data Analysis Using Pattern Recognition Techniques. // *J. Chem. Inf. Model.* - 1976. - V. 16, № 2. - P. 99-105.
135. *Hodes L.; Hazard G.F.; Geran R.I.; Richman S.* A statistical-heuristic methods for automated selection of drugs for screening. // *J. Med. Chem.* - 1977. - V. 20, № 4. - P. 469-475.
136. *Adamson G.W.* Automatic methods of handling chemical structure and property information. // *Proc. Analyt. Div. Chem. Soc.* - 1977. - V. 14, № 2. - P. 26-28.
137. *Adamson G.W.; Bush J.A.* Method for relating the structure and properties of chemical compounds. // *Nature.* - 1974. - V. 248, № 5447. - P. 406-407.

138. *Adamson G.W.; Bawden D.* Method of structure-activity correlation using Wiswesser line notation. // *J. Chem. Inf. Comput. Sci.* - 1975. - V. 15, № 4. - P. 215-220.
139. *Adamson G.W.; Bush J.A.* Evaluation of an empirical structure-activity relation for property prediction in a structurally diverse group of local anesthetics. // *J. Chem. Soc., Perkin Trans. 1.* - 1976. - № 2. - P. 168-172.
140. *Adamson G.W.; Bawden D.* A substructural analysis method for structure-activity correlation of heterocyclic compounds using Wiswesser line notation. // *J. Chem. Inf. Comput. Sci.* - 1977. - V. 17, № 3. - P. 164-171.
141. *Adamson G.W.; Bawden D.* An empirical method of structure-activity correlation for polysubstituted cyclic compounds using Wiswesser Line Notation. // *J. Chem. Inf. Comput. Sci.* - 1976. - V. 16, № 3. - P. 161-165.
142. *Milne M.; Lefkowitz D.; Hill H.; Powers R.* Search of CA Registry (1.25 Million Compounds) with the Topological Screens System. // *J. Chem. Doc.* - 1972. - V. 12, № 3. - P. 183-189.
143. *Adamson G.W.; Cowell J.; Lynch M.F.; McLure A.H.W.; Town W.G.; Yapp A.M.* Strategic Considerations in the Design of a Screening System for Substructure Searches of Chemical Structure Files. // *J. Chem. Doc.* - 1973. - V. 13, № 3. - P. 153-157.
144. *Feldman A.; Hodes L.* An Efficient Design for Chemical Structure Searching. I. The Screens. // *J. Chem. Inf. Comput. Sci.* - 1975. - V. 15, № 3. - P. 147-152.
145. *Willett P.* A Screen Set Generation Algorithm. // *J. Chem. Inf. Comput. Sci.* - 1979. - V. 19, № 3. - P. 159-162.
146. *Willett P.* The Effect of Screen Set Size on Retrieval from Chemical Substructure Search Systems. // *J. Chem. Inf. Comput. Sci.* - 1979. - V. 19, № 4. - P. 253-255.
147. *Willett P.; Winterman V.; Bawden D.* Implementation of nearest-neighbor searching in an online chemical structure search system. // *J. Chem. Inf. Comput. Sci.* - 1986. - V. 26, № 1. - P. 36-41.

148. *Fisanick W.; Lipkus A.H.; Rusinko A.* Similarity searching on CAS Registry substances. 2. 2D structural similarity. // *J. Chem. Inf. Comput. Sci.* - 1994. - V. 34, № 1. - P. 130-140.
149. *Hodes L.* Clustering a large number of compounds. 1. Establishing the method on an initial sample. // *J. Chem. Inf. Comput. Sci.* - 1989. - V. 29, № 2. - P. 66-71.
150. *McGregor M.J.; Pallai P.V.* Clustering of Large Databases of Compounds: Using the MDL "Keys" as Structural Descriptors. // *J. Chem. Inf. Comput. Sci.* - 1997. - V. 37, № 3. - P. 443-448.
151. *Turner D.B.; Tyrrell S.M.; Willett P.* Rapid Quantification of Molecular Diversity for Selective Database Acquisition. // *J. Chem. Inf. Comput. Sci.* - 1997. - V. 37, № 1. - P. 18-22.
152. *Durant J.L.; Leland B.A.; Henry D.R.; Nourse J.G.* Reoptimization of MDL Keys for Use in Drug Discovery. // *J. Chem. Inf. Comput. Sci.* - 2002. - V. 42, № 6. - P. 1273-1280.
153. *Tong W.; Lowis D.R.; Perkins R.; Chen Y.; Welsh W.J.; Goddette D.W.; Heritage T.W.; Sheehan D.M.* Evaluation of Quantitative Structure-Activity Relationship Methods for Large-Scale Prediction of Chemicals Binding to the Estrogen Receptor. // *J. Chem. Inf. Model.* - 1998. - V. 38, № 4. - P. 669-677.
154. *Cramer R.D.* BC(DEF) parameters. 1. The intrinsic dimensionality of intermolecular interactions in the liquid state. // *J. Am. Chem. Soc.* - 1980. - V. 102, № 6. - P. 1837-1849.
155. *Cramer R.D.* BC(DEF) parameters. 2. An empirical structure-based scheme for the prediction of some physical properties. // *J. Am. Chem. Soc.* - 1980. - V. 102, № 6. - P. 1849-1859.
156. *Klopman G.* Artificial intelligence approach to structure-activity studies. Computer automated structure evaluation of biological activity of organic molecules. // *J. Am. Chem. Soc.* - 1984. - V. 106, № 24. - P. 7315-7321.
157. *Klopman G.; Rosenkranz H.S.* Structural requirements for the mutagenicity of environmental nitroarenes. // *Mutat. Res.* - 1984. - V. 126, № 3. - P. 227-238.
158. *Klopman G.; Kalos A.N.* Causality in structure-activity studies. // *J. Comput. Chem.* - 1985. - V. 6, № 5. - P. 492-506.

159. *Rosenkranz H.S.; Mitchell C.S.; Klopman G.* Artificial intelligence and Bayesian decision theory in the prediction of chemical carcinogens. // *Mutat. Res.* - 1985. - V. 150, № 1-2. - P. 1-11.
160. *Klopman G.; Frierson M.R.; Rosenkranz H.S.* Computer analysis of toxicological data bases: mutagenicity of aromatic amines in Salmonella tester strains. // *Environmental Mutagenesis.* - 1985. - V. 7, № 5. - P. 625-644.
161. *Rosenkranz H.S.; Klopman G.* Mutagens, carcinogens, and computers. // *Progress in Clinical and Biological Research.* - 1986. - V. 209. Pt. A. - P. 71-104.
162. *Klopman G.; Namboodiri K.; Kalos A.N.* Computer automated evaluation and prediction of the Iball Index of carcinogenicity of polycyclic aromatic hydrocarbons. // *Progress in Clinical and Biological Research.* - 1985. - V. 172, Pt. A. - P. 287-298.
163. *Klopman G.* Predicting toxicity through a computer automated structure evaluation program. // *Environmental Health Perspectives.* - 1985. - V. 61. - P. 269-274.
164. *Klopman G.; Macina O.T.* Use of the computer automated structure evaluation program in determining quantitative structure-activity relationships within hallucinogenic phenylalkylamines. // *J. Theor. Biol.* - 1985. - V. 113, № 4. - P. 637-648.
165. *Klopman G.; Contreras R.* Use of artificial intelligence in structure-activity correlations of anticonvulsant drugs. // *Mol. Pharmacol.* - 1985. - V. 27, № 1. - P. 86-93.
166. *Klopman G.; Venegas R.E.* CASE study of in vitro inhibition of sparteine monooxygenase. // *Acta Pharmaceutica Jugoslavica.* - 1986. - V. 36, № 2. - P. 189-209.
167. *Klopman G.; Kalos A.N.* Quantitative structure-activity relationships of beta-adrenergic agents. Application of the computer automated structure evaluation (CASE) technique of molecular fragment recognition. // *J. Theor. Biol.* - 1986. - V. 118, № 2. - P. 199-214.
168. *Klopman G.; Macina O.T.; Simon E.J.; Hiller J.M.* Computer automated structure evaluation of opiate alkaloids. // *J. Mol. Struct. Theochem.* - 1986. - V. 27, № 3-4. - P. 299-308.

169. *Klopman G.; Macina O.T.; Levinson M.E.; Rosenkranz H.S.* Computer automated structure evaluation of quinolone antibacterial agents. // *Antimicrobial Agents and Chemotherapy*. - 1987. - V. 31, № 11. - P. 1831-1840.
170. *Klopman G.; Macina O.T.* Computer-automated structure evaluation of antileukemic 9-anilinoacridines. // *Mol. Pharmacol.* - 1987. - V. 31, № 4. - P. 457-476.
171. *Artemenko N.V.; Baskin I.I.; Palyulin V.A.; Zefirov N.S.* Artificial neural network and fragmental approach in prediction of physicochemical properties of organic compounds. // *Russ. Chem. Bull.* - 2003. - V. 52, № 1. - P. 20-29.
172. *Smolenskii E.A.* On Some Aspects of the Structure–Property Problem // *Dokl. Chem.* - 1999. - V. 365, № 4-6. - P. 93-98.
173. *Smolenskii E.A.; Slovokhotova O.L.; Chuvaeva I.V.; Zefirov N.S.* Information Significance of Topological Indices. // *Dokl. Chem.* - 2004. - V. 397, № 2. - P. 173.
174. *Nutt C.W.* The correlation and prediction of the optical and thermodynamic properties of saturated liquid hydrocarbons by the group contribution method. // *Transactions of the Faraday Society*. - 1957. - V. 53. - C. 1538-1544.
175. *Ghose A.K.; Crippen G.M.* Atomic Physicochemical Parameters for Three-Dimensional Structure-Directed Quantitative Structure-Activity Relationships I. Partition Coefficients as a Measure of Hydrophobicity. // *J. Comput. Chem.* - 1986. - V. 7, № 4. - P. 565-577.
176. *Ghose A.K.; Crippen G.M.* Atomic Physicochemical Parameters for Three-Dimensional-Structure-Directed Quantitative Structure-Activity Relationships. 2. Modeling Dispersive and Hydrophobic Interactions. // *J. Chem. Inf. Comput. Sci.* - 1987. - V. 27, № 1. - P. 21-35.
177. *Ghose A.K.; Pritchett A.; Crippen G.M.* Atomic physicochemical parameters for three dimensional structure directed quantitative structure-activity relationships III: Modeling hydrophobic interactions. // *J. Comput. Chem.* - 1988. - V. 9, № 1. - P. 80-90.
178. *Viswanadhan V.N.; Ghose A.K.; Revankar G.R.; Robins R.K.* Atomic physicochemical parameters for three dimensional structure directed quantitative structure-activity relationships. 4. Additional parameters for hydrophobic and dispersive interactions and their application for an automated superposition of certain naturally oc-

- curing nucleoside antibiotics. // *J. Chem. Inf. Comput. Sci.* - 1989. - V. 29, № 3. - P. 163-172.
179. *Ghose A.K.; Viswanadhan V.N.; Wendoloski J.J.* Prediction of hydrophobic (lipophilic) properties of small organic molecules using fragmental methods: An analysis of ALOGP and CLOGP methods. // *J. Phys. Chem. A.* - 1998. - V. 102, № 21. - P. 3762-3772.
180. *Wildman S.A.; Crippen G.M.* Prediction of Physicochemical Parameters by Atomic Contributions. // *J. Chem. Inf. Comput. Sci.* - 1999. - V. 39, № 5. - P. 868-873.
181. *Suzuki T.; Kudo Y.* Automatic log P estimation based on combined additive modeling methods. // *J. Comput. Aided. Mol. Des.* - 1990. - V. 4, № 2. - P. 155-198.
182. *Convard T.; Dubost J.-P.; Le Solleu H.; Kummer E.* SMILOGP: A Program for a fast evaluation of theoretical log-p from the smiles code of a molecule. // *Quant. Struct.-Act. Relat.* - 1994. - V. 13. - P. 34-37.
183. *Wang R.; Fu Y.; Lai L.* A New Atom-Additive Method for Calculating Partition Coefficients. // *J. Chem. Inf. Comput. Sci.* - 1997. - V. 37, № 3. - P. 615-621.
184. *Wang R.; Gao Y.; Lai L.* Calculating partition coefficient by atom-additive method. // *Persp. Drug Discov. Design.* - 2000. - V. 19. - P. 47-66.
185. *Hou T.J.; Xia K.; Zhang W.; Xu X.J.* ADME Evaluation in Drug Discovery. 4. Prediction of Aqueous Solubility Based on Atom Contribution Approach. // *J. Chem. Inf. Comput. Sci.* - 2004. - V. 44, № 1. - P. 266-275.
186. *Winkler D.A.; Burden F.R.; Watkins A.J.R.* Atomistic topological indices applied to benzodiazepines using various regression methods. // *Quantitative Structure-Activity Relationships.* - 1998. - V. 17, № 1. - P. 14-19.
187. *Bernstein H.J.* Bond energies in hydrocarbons. // *Trans. Faraday Soc.* - 1962. - V. 58 - P. 2285-2306.
188. *Kalb A.J.; Chung A.L.H.; Allen T.L.* Bond Energies and the Interactions between Next-Nearest Neighbors. III. Gaseous and Liquid Alkanes, Cyclohexane, Alkylcyclohexanes, and Decalins. // *J. Am. Chem. Soc.* - 1966. - V. 88, № 13. - P. 2938-2942.

189. *Nilakantan R.; Bauman N.; Dixon J.S.; Venkataraghavan R.* Topological Torsion: A New Molecular Descriptor for SAR Applications. Comparison with Other Descriptors. // *J. Chem. Inf. Comput. Sci.* - 1987. - V. 27, № 2. - P. 82-85.
190. *Kearsley S.K.; Sallamack S.; Fluder E.M.; Andose J.D.; Mosley R.T.; Sheridan R.P.* Chemical Similarity Using Physiochemical Property Descriptors. // *J. Chem. Inf. Comput. Sci.* - 1996. - V. 36, № 1. - P. 118-127.
191. *Klopman G.* MULTICASE. 1. A Hierarchical computer automated structure evaluation program. // *Quant. Struct.-Act. Relat.* - 1992. - V. 11, № 2. - P. 176-184.
192. *Klopman G.* The MultiCASE Program II. Baseline Activity Identification Algorithm (BAIA). // *J. Chem. Inf. Comput. Sci.* - 1998. - V. 38, № 1. - P. 78-81.
193. *Артеменко Н.В.; Баскин И.И.; Палюлин В.А.; Зефирова Н.С.* Прогнозирование физических свойств органических соединений при помощи искусственных нейронных сетей в рамках подструктурного подхода. // *Докл. РАН.* - 2001. - Т. 381, № 2. - С. 203-206.
194. *Baskin I.I.; Halberstam N.M.; Artemenko N.V.; Palyulin V.A.; Zefirov N.S.* NASAWIN – a universal software for QSPR/QSAR studies. // *EuroQSAR 2002 Designing Drugs and Crop Protectants: processes, problems and solutions.*, Ford M., Ed. Blackwell Publishing. - 2003. - С. 260-263.
195. *Кумсков М.И.* Перспективы использования программной системы BIBIGON для предсказания физико-химических свойств фторсодержащих органических соединений. // *Журн. орг. химии.* - 1995. - Т. 31, № 10. - С. 1495-1498.
196. *Solov'ev V.P.; Varnek A.; Wipff G.* Modeling of Ion Complexation and Extraction Using Substructural Molecular Fragments. // *J. Chem. Inf. Comput. Sci.* - 2000. - V. 40, № 3. - P. 847-858.
197. *Varnek A.; Wipff G.; Solovev V.P.* Towards an information system on solvent extraction. // *Solvent Extraction and Ion Exchange.* - 2001. - V. 19, № 5. - P. 791-837.
198. *Gakh A.A.; Gakh E.G.; Sumpter B.G.; Noid D.W.* Neural Network-Graph Theory Approach to the Prediction of the Physical Properties of Organic Compounds. // *J. Chem. Inf. Comput. Sci.* - 1994. - V. 34, № 4. - P. 832-839.

199. *Rucker G.; Rucker C.* Counts of all walks as atomic and molecular descriptors. // *J. Chem. Inf. Comput. Sci.* - 1993. - V. 33, № 5. - P. 683-695.
200. *Adamson G.W.; Cowell J.; Lynch M.F.; Town W.G.; Yapp A.M.* Analysis of structural characteristics of chemical compounds in a large computer-based file. Part IV. Cyclic fragments. // *J. Chem. Soc., Perkin Trans. 1.* - 1973. - V. № 8. - P. 863-865.
201. *Adamson G.W.; Creasey S.E.; Eakins J.P.; Lynch M.F.* Analysis of structural characteristics of chemical compounds in a large computer-based file. Part V. More detailed cyclic fragments. // *J. Chem. Soc., Perkin Trans. 1.* - 1973. - V. № 19. - P. 2071-2076.
202. *Wiswesser W.J.* How the WLN began in 1949 and how it might be in 1999. // *J. Chem. Inf. Comput. Sci.* - 1982. - V. 22, № 2. - P. 88-93.
203. *Weininger D.* SMILES, A Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. // *J. Chem. Inf. Comput. Sci.* - 1988. - V. 28, № 1. - P. 31-36.
204. *Weininger D.; Weininger A.; Weininger J.L.* SMILES: 2. Algorithm for generation of unique SMILES notation. // *J. Chem. Inf. Comput. Sci.* - 1989. - V. 29, № 2. - P. 97-101.
205. *Adamson G.W.; Bawden D.* Substructural Analysis Techniques for Empirical Structure-Property Correlation. Application to Stereochemically Related Molecular Properties. // *J. Chem. Inf. Comput. Sci.* - 1980. - V. 20, № 2. - P. 97-100.
206. *Adamson G.W.; Bawden D.* Automated Additive Modeling Techniques Applied to Thermochemical Property Estimation. // *J. Chem. Inf. Comput. Sci.* - 1980. - V. 20, № 4. - P. 242-246.
207. *Adamson G.W.; Bawden D.* Comparison of Hierarchical Cluster Analysis Techniques for Automatic Classification of Chemical Structures. // *J. Chem. Inf. Comput. Sci.* - 1981. - V. 21, № 4. - P. 204-209.
208. *Vidal D.; Thormann M.; Pons M.* LINGO, an efficient holographic text based method to calculate biophysical properties and intermolecular similarities. // *J. Chem. Inf. Model.* - 2005. - V. 45, № 2. - P. 386-393.

209. *Татевский В.М.* Классическая теория строения молекул и квантовая механика. - Химия: М. - 1973. - 520 с.
210. *Степанов Н.Ф.; Ерлыкина М.Е.; Филиппов Г.Г.* Методы линейной алгебры в физической химии. - Изд-во Моск. ун-та: М. - 1976. - 359 с.
211. *Benson S.W.; Cruickshank F.R.; Golden D.M.; Haugen G.R.; O'Neal H.E.; Rodgers A.S.; Shaw R.; Walsh R.* Additivity rules for the estimation of thermochemical properties. // *Chem. Rev.* - 1969. - V. 69, № 3. - P. 279-324.
212. *Adamson G.W.; Lynch M.F.; Town W.G.* Analysis of Structural Characteristics of Chemical Compounds in a Large Computer-based File. Part II. Atom-Centered Fragments. // *J. Chem. Soc. C.* - 1971. - P. 3702-3706.
213. *Adamson G.W.; Lambourne D.R.; Lynch M.F.* Analysis of structural characteristics of chemical compounds in a large computer-based file. Part III. Statistical association of fragment incidence. // *J. Chem. Soc., Perkin Trans. 1.* - 1972. - P. 2428 - 2433.
214. *Hodes L.* Selection of molecular fragment features for structure-activity studies in antitumor screening. // *J. Chem. Inf. Comput. Sci.* - 1981. - V. 21, № 3. - P. 132-136.
215. *Poroikov V.V.; Filimonov D.A.; Borodina Y.V.; Lagunin A.A.; Kos A.* Robustness of biological activity spectra predicting by computer program PASS for noncongeneric sets of chemical compounds. // *J. Chem. Inf. Comput. Sci.* - 2000. - V. 40, № 6. - P. 1349-1355.
216. *Filimonov D.; Poroikov V.; Borodina Y.; Glorizova T.* Chemical Similarity Assessment through Multilevel Neighborhoods of Atoms: Definition and Comparison with the Other Descriptors. // *J. Chem. Inf. Comput. Sci.* - 1999. - V. 39, № 4. - P. 666-670.
217. *Xing L.; Glen R.C.* Novel methods for the prediction of logP, pKa, and logD. // *J. Chem. Inf. Comput. Sci.* - 2002. - V. 42, № 4. - P. 796-805.
218. *Bender A.; Mussa H.Y.; Glen R.C.; Reiling S.* Molecular Similarity Searching Using Atom Environments, Information-Based Feature Selection, and a Naive Bayesian Classifier. // *J. Chem. Inf. Comput. Sci.* - 2004. - V. 44, № 1. - P. 170-178.

219. *Bender A.; Mussa H.Y.; Glen R.C.; Reiling S.* Similarity Searching of Chemical Databases Using Atom Environment Descriptors (MOLPRINT 2D): Evaluation of Performance. // *J. Chem. Inf. Comput. Sci.* - 2004. - V. 44, № 5. - P. 1708-1718.
220. *Glen R.C.; Bender A.; Arnby C.H.; Carlsson L.; Boyer S.; Smith J.* Circular fingerprints: Flexible molecular descriptors with applications from physical chemistry to ADME. // *IDrugs.* - 2006. - V. 9, № 3. - P. 199-204.
221. *Rodgers S.; Glen R.C.; Bender A.* Characterizing bitterness: Identification of key structural features and development of a classification model. // *J. Chem. Inf. Model.* - 2006. - V. 46, № 2. - P. 569-576.
222. *Cannon E.O.; Amini A.; Bender A.; Sternberg M.J.E.; Muggleton S.H.; Glen R.C.; Mitchell J.B.O.* Support vector inductive logic programming outperforms the Naive Bayes Classifier and inductive logic programming for the classification of bioactive chemical compounds. // *J. Comput. Aided Mol. Des.* - 2007. - V. 21, № 5. - P. 269-280.
223. *Faulon J.-L.; Visco D.P., Jr.; Pophale R.S.* The Signature Molecular Descriptor. 1. Using Extended Valence Sequences in QSAR and QSPR Studies. // *J. Chem. Inf. Comput. Sci.* - 2003. - V. 43, № 3. - P. 707-720.
224. *Faulon J.-L.; Churchwell C.J.; Visco D.P., Jr.* The Signature Molecular Descriptor. 2. Enumerating Molecules from Their Extended Valence Sequences. // *J. Chem. Inf. Comput. Sci.* - 2003. - V. 43, № 3. - P. 721-734.
225. *Churchwell C.J.; Rintoul M.D.; Martin S.; Visco D.P., Jr.; Kotu A.; Larson R.S.; Sillerud L.O.; Brown D.C.; Faulon J.L.* The signature molecular descriptor. 3. Inverse-quantitative structure-activity relationship of ICAM-1 inhibitory peptides. // *J. Mol. Graph. Model.* - 2004. - V. 22, № 4. - P. 263-273.
226. *Bremser W.* Hose -- a novel substructure code. // *Analytica Chimica Acta.* - 1978. - V. 103, № 4. - P. 355-365.
227. *Dubois J.-E.; Panaye A.; Attias R.* DARC System: Notions of Defined and Generic Substructures. Filiation and Coding of FREL Substructure (SS) Classes. // *J. Chem. Inf. Comput. Sci.* - 1987. - V. 27, № 2. - P. 74-82.
228. *Dubois J.E.; Doucet J.P.; Panaye A.; Fan B.T.* DARC Site Topological Correlations : Ordered Structural Descriptors and Property Evaluation. // *Topological Indi-*

- ces and Related Descriptors in QSAR and QSPR, Devillers J.; Balaban A.T., Eds. Gordon and Breach Sciences Publishers: Amsterdam. - 1999. - P. 613-673.
229. *Xiao Y.; Qiao Y.; Zhang J.; Lin S.; Zhang W.* A Method for Substructure Search by Atom-Centered Multilayer Code. // *J. Chem. Inf. Comput. Sci.* - 1997. - V. 37, № 4. - P. 701-704.
230. *Bender A.; Young D.W.; Jenkins J.L.; Serrano M.; Mikhailov D.; Clemons P.A.; Davies J.W.* Chemogenomic Data Analysis: Prediction of Small-Molecule Targets and the Advent of Biological Fingerprints. // *Comb. Chem. High Throughput Screen.* - 2007. - V. 10, № 8. - P. 719-731.
231. *Nidhi M.G.; Davies J.W.; Jenkins J.L.* Prediction of Biological Targets for Compounds Using Multiple-Category Bayesian Models Trained on Chemogenomics Databases. // *J. Chem. Inf. Model.* - 2006. - V. 46, № 3. - P. 1124-1133.
232. *Adamson G.W.; Bush J.A.; McLure A.H.W.; Lynch M.F.* An Evaluation of a Substructure Search Screen System Based on Bond-Centered Fragments. // *J. Chem. Doc.* - 1974. - V. 14, № 1. - P. 44-48.
233. MDL Information Systems, Inc. // MDL Information Systems, Inc. www.mdli.com.
234. *Ahrens E.K.F.* Customization for Chemical Database Applications. // *Chemical Structures*, Warr W.A., Ed. - 1988. - P. 97-111.
235. *Raymond J.W.; Willett P.* Maximum common subgraph isomorphism algorithms for the matching of chemical structures. // *J Comput Aided Mol Des.* - 2002. - V. 16, № 7. - P. 521-533.
236. *Розенблит А.Б.; Голендер В.Е.* Логико-комбинаторные методы в конструировании лекарств. - Зинатне: Рига. - 1983. - 352 с.
237. *Hagadone T.R.* Molecular substructure similarity searching: efficient retrieval in two-dimensional structure databases. // *J. Chem. Inf. Model.* - 1992. - V. 32, № 5. - P. 515-521.
238. *Ruiz I.L.; Garcia C.G.; Gomez-Nieto M.A.* Clustering Chemical Databases Using Adaptable Projection Cells and MCS Similarity Values. // *J. Chem. Inf. Model.* - 2005. - V. 45, № 5. - P. 1178-1194.

239. *Stahl M.; Mauser H.* Database Clustering with a Combination of Fingerprint and Maximum Common Substructure Methods. // *J. Chem. Inf. Model.* - 2005. - V. 45, № 3. - P. 542-548.
240. *Bacha P.A.; Gruver H.S.; Den Hartog B.K.; Tamura S.Y.; Nutt R.F.* Rule Extraction from a Mutagenicity Data Set Using Adaptively Grown Phylogenetic-like Trees. // *J. Chem. Inf. Model.* - 2002. - V. 42, № 5. - P. 1104-1111.
241. *Sheridan R.P.* Finding Multiactivity Substructures by Mining Databases of Drug-Like Compounds. // *J. Chem. Inf. Comput. Sci.* - 2003. - V. 43, № 3. - P. 1037-1050.
242. *Авидон В.В.; Лексина Л.А.* Дескрипторный язык для анализа сходства химических структур органических соединений. // *НТИ.* - Сер. 2. - 1974. - № 3. - С. 22-25.
243. *Carhart R.E.; Smith D.H.; Venkataraghavan R.* Atom Pairs as Molecular Features in Structure-Activity Studies: Definition and Applications. // *J. Chem. Inf. Comput. Sci.* - 1985. - V. 25, № 2. - P. 64-73.
244. *Horvath D.* High Throughput Conformational Sampling & Fuzzy Similarity Metrics: A Novel Approach to Similarity Searching and Focused Combinatorial Library Design and its Role in the Drug Discovery Laboratory. // *Combinatorial Library Design and Evaluation: Principles, Software Tools and Applications*, Ghose A.; Viswanadhan V., Eds. Marcel Dekker: New York. - 2001. - P. 429-472.
245. *Horvath D.; Jeandenans C.* Neighborhood Behavior of in Silico Structural Spaces with Respect to in Vitro Activity Spaces-A Novel Understanding of the Molecular Similarity Principle in the Context of Multiple Receptor Binding Profiles. // *J. Chem. Inf. Comput. Sci.* - 2003. - V. 43, № 2. - P. 680-690.
246. *Bonachera F.; Parent B.; Barbosa F.; Froloff N.; Horvath D.* Fuzzy Tricentric Pharmacophore Fingerprints. 1. Topological Fuzzy Pharmacophore Triplets and Adapted Molecular Similarity Scoring Schemes. // *J. Chem. Inf. Model.* - 2006. - V. 46, № 6. - P. 2457-2477.
247. *Schuffenhauer A.; Floersheim P.; Acklin P.; Jacoby E.* Similarity Metrics for Ligands Reflecting the Similarity of the Target Proteins. // *J. Chem. Inf. Comput. Sci.* - 2003. - V. 43, № 2. - P. 391-405.

248. MOE, Molecular Operating Environment, Chemical Computing Group Inc., Montreal, Canada. // MOE, Molecular Operating Environment, Chemical Computing Group Inc., Montreal, Canada. - www.chemcomp.com,
249. *Franke L.; Byvatov E.; Werz O.; Steinhilber D.; Schneider P.; Schneider G.* Extraction and visualization of potential pharmacophore points using support vector machines: application to ligand-based virtual screening for COX-2 inhibitors. // *J. Med. Chem.* - 2005. - V. 48, № 22. - P. 6997-7004.
250. *Byvatov E.; Sasse B.C.; Stark H.; Schneider G.* From virtual to real screening for D3 dopamine receptor ligands. // *ChemBioChem.* - 2005. - V. 6, № 6. - P. 997-999.
251. *Hansch C.; Fujita T.* p- ρ -Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. // *J. Am. Chem. Soc.* - 1964. - V. 86, № 8. - P. 1616-1626.
252. *Hansch C.; Muir R.M.; Fujita T.; Maloney P.P.; Geiger F.; Streich M.* The Correlation of Biological Activity of Plant Growth Regulators and Chloromycetin Derivatives with Hammett Constants and Partition Coefficients. // *J. Am. Chem. Soc.* - 1963. - V. 85, № 18. - P. 2817-2824.
253. *Fleischer R.; Froberg P.; Büge A.; Nuhn P.; Wiese M.* QSAR Analysis of Substituted 2-Phenylhydrazonoacetamides Acting as Inhibitors of 15-Lipoxygenase. // *Quant. Struct.-Act. Relat.* - 2000. - V. 19, № 2. - P. 162-172.
254. *Hatrik S.; Zahradnik P.* Neural Network Approach to the Prediction of the Toxicity of Benzothiazolium Salts from Molecular Structure. // *J. Chem. Inf. Comput. Sci.* - 1996. - V. 36, № 5. - P. 992-995.
255. *Bemis G.W.; Murcko M.A.* The properties of known drugs. 1. Molecular frameworks. // *J. Med. Chem.* - 1996. - V. 39, № 15. - P. 2887-2893.
256. *Bemis G.W.; Murcko M.A.* Properties of known drugs. 2. Side chains. // *J. Med. Chem.* - 1999. - V. 42, № 25. - P. 5095-5099.
257. *Randic M.* Representation of molecular graphs by basic graphs. // *J. Chem. Inf. Comput. Sci.* - 1992. - V. 32, № 1. - P. 57-69.
258. *Мнухин В.Б.* Базис алгебры инвариантов графов. // Математический анализ и его приложения. - Ростов-на-Дону. - 1983. - С. 55-60.

259. *Baskin I.I.; Skvortsova M.I.; Stankevich I.V.; Zefirov N.S.* On the Basis of Invariants of Labeled Molecular Graphs. // *J. Chem. Inf. Comput. Sci.* - 1995. - V. 35, № 3. - P. 527-531.
260. *Skvortsova M.I.; Baskin I.I.; Skvortsov L.A.; Palyulin V.A.; Zefirov N.S.; Stankevich I.V.* Chemical graphs and their basis invariants. // *J. Mol. Struct. Theor. Chem.* - 1999. - V. 466. - P. 211-217.
261. *Скворцова М.И.; Федяев К.С.; Баскин И.И.; Палюлин В.А.; Зефирова Н.С.* Новый способ кодирования химических структур на основе базисных фрагментов. // *Докл. РАН.* - 2002. - Т. 382, № 5. - С. 645-648.
262. *Скворцова М.И.; Федяев К.С.; Палюлин В.А.; Зефирова Н.С.* Моделирование связи между структурой и свойствами углеводородов на основе базисных топологических дескрипторов. // *Изв. РАН, Сер. хим.* - 2004. - № 8. - С. 1527-1535.
263. *Estrada E.* Spectral Moments of the Edge Adjacency Matrix in Molecular Graphs. 1. Definition and Applications to the Prediction of Physical Properties of Alkanes. // *J. Chem. Inf. Comput. Sci.* - 1996. - V. 36, № 4. - P. 844-849.
264. *Estrada E.* Spectral Moments of the Edge-Adjacency Matrix of Molecular Graphs. 2. Molecules Containing Heteroatoms and QSAR Applications. // *J. Chem. Inf. Comput. Sci.* - 1997. - V. 37, № 2. - P. 320-328.
265. *Estrada E.* Spectral Moments of the Edge Adjacency Matrix in Molecular Graphs. 3. Molecules Containing Cycles. // *J. Chem. Inf. Comput. Sci.* - 1998. - V. 38, № 1. - P. 23-27.
266. *Estrada E.; Pena A.; Garcia-Domenech R.* Designing sedative/hypnotic compounds from a novel substructural graph-theoretical approach. // *J Comput Aided Mol Des.* - 1998. - V. 12, № 6. - P. 583-595.
267. *Estrada E.; Gutierrez Y.* Modeling chromatographic parameters by a novel graph theoretical sub-structural approach. // *Journal of Chromatography A.* - 1999. - V. 858, № 2. - P. 187-199.
268. *Estrada E.; Gutierrez Y.; Gonzalez H.* Modeling Diamagnetic and Magneto-optic Properties of Organic Compounds with the TOSS-MODE Approach. // *J. Chem. Inf. Comput. Sci.* - 2000. - V. 40, № 6. - P. 1386-1399.

269. Estrada E.; Gonzalez H. What Are the Limits of Applicability for Graph Theoretic Descriptors in QSPR/QSAR? Modeling Dipole Moments of Aromatic Compounds with TOPS-MODE Descriptors. // J. Chem. Inf. Comput. Sci. - 2003. - V. 43, № 1. - P. 75-84.
270. Gonzalez M.P.; Helguera A.M.; Diaz H.G. A TOPS-MODE approach to predict permeability coefficients. // Polymer. - 2004. - V. 45, № 6. - P. 2073-2079.
271. Estrada E.; Molina E.; Perdomo-Lopez I. Can 3D Structural Parameters Be Predicted from 2D (Topological) Molecular Descriptors? // J. Chem. Inf. Comput. Sci. - 2001. - V. 41, № 4. - P. 1015-1021.
272. Estrada E.; Uriarte E.; Montero A.; Teijeira M.; Santana L.; De Clercq E. A novel approach for the virtual screening and rational design of anticancer compounds. // J Med Chem. - 2000. - V. 43, № 10. - P. 1975-1985.
273. Estrada E.; Vilar S.; Uriarte E.; Gutierrez Y. In Silico Studies toward the Discovery of New Anti-HIV Nucleoside Compounds with the Use of TOPS-MODE and 2D/3D Connectivity Indices. 1. Pyrimidyl Derivatives. // J. Chem. Inf. Comput. Sci. - 2002. - V. 42, № 5. - P. 1194-1203.
274. Estrada E.; Patlewicz G.; Gutierrez Y. From Knowledge Generation to Knowledge Archive. A General Strategy Using TOPS-MODE with DEREK To Formulate New Alerts for Skin Sensitization. // J. Chem. Inf. Comput. Sci. - 2004. - V. 44, № 2. - P. 688-698.
275. Gonzalez M.P.; Diaz H.G.; Ruiz R.M.; Cabrera M.A.; de Armas R.R. TOPS-MODE based QSARs derived from heterogeneous series of compounds. Applications to the design of new herbicides. // J. Chem. Inf. Comput. Sci. - 2003. - V. 43, № 4. - P. 1192-1199.
276. Gonzalez M.P.; Moldes M.D.T. QSAR study of N-6-(substituted-phenylcarbamoyl) adenosine-5'-uronamides as agonist for A(1) adenosine receptors. // Bull. Math. Biol. - 2004. - V. 66, № 4. - P. 907-920.
277. Gonzalez M.P.; Dias L.C.; Helguera A.M.; Rodriguez Y.M.; de Oliveira L.G.; Gomez L.T.; Diaz H.G. TOPS-MODE based QSARs derived from heterogeneous series of compounds. Applications to the design of new anti-inflammatory compounds. // Bioorganic & Medicinal Chemistry. - 2004. - V. 12, № 16. - P. 4467-4475.

278. *Molina E.; Gonzales Diaz H.; Gonzalez M.P.; Rodriguez E.; Uriarte E.* Designing Antibacterial Compounds through a Topological Substructural Approach. // *J. Chem. Inf. Comput. Sci.* - 2004. - V. 44, № 2. - P. 515-521.
279. *Gonzalez M.P.; Diaz H.G.; Cabrera M.A.; Ruiz R.M.* A novel approach to predict a toxicological property of aromatic compounds in the *Tetrahymena pyriformis*. // *Bioorg. Med, Chem.* - 2004. - V. 12, № 4. - P. 735-744.
280. *Helguera A.M.; Gonzalez M.P.; Briones J.R.* TOPS-MODE approach to predict mutagenicity in dental monomers. // *Polymer.* - 2004. - V. 45, № 6. - P. 2045-2050.
281. *Gonzalez M.P.; Dias L.C.; Helguera A.M.* A topological sub-structural approach to the mutagenic activity in dental monomers. 2. Cycloaliphatic epoxides. // *Polymer.* - 2004. - V. 45, № 15. - P. 5353-5359.
282. *Gonzalez M.P.; Moldes M.d.C.T.; Fall Y.; Dias L.C.; Helguera A.M.* A topological sub-structural approach to the mutagenic activity in dental monomers. 3. Heterogeneous set of compounds. // *Polymer.* - 2005. - V. 46, № 8. - P. 2783-2790.
283. *Kramer S.; De Raedt L.; Helma C.* In *Molecular feature mining in HIV data*, Seventh ACM SIGKDD international conference on Knowledge discovery and data mining, San Francisco, California, August 26 - 29, 2001, 2001; ACM Press, New York, NY: San Francisco, California. - 2001. - P. 136-143.
284. *De Raedt L.; Kramer S.* In *The Levelwise Version Space Algorithm and its Application to Molecular Fragment Finding*, The Seventeenth International Joint Conference on Artificial Intelligence, 2001; Morgan Kaufmann: San Francisco. - 2001. - P. 853-862.
285. *Kramer S.; De Raedt L.* In *Feature construction with version spaces for biochemical applications*, The eighteenth International Conference on Machine Learning, 2001; Morgan Kaufmann: San Francisco, CA. - 2001. - P. 258-265.
286. *Inokuchi A.* Mining Generalized Substructures from a Set of Labeled Graphs. // *Proceedings of the Fourth IEEE International Conference on Data Mining (ICDM'04)* - IEEE Computer Society. - 2004. - P. 415-418
287. *Yan X.; Han J.* gspan: Graph-based substructure pattern mining. // *Proceedings of the 2002 IEEE International Conference on Data Mining.* - 2002. - P. 721-724.

288. *Saigo H.; Kadowaki T.; Tsuda K.* In A Linear Programming Approach for Molecular QSAR analysis, International Workshop on Mining and Learning with Graphs 2006. - 2006. - P. 85-96.
289. *Asai T.; Abe K.; Kawasoe S.; Arimura H.; Satamoto H.; Arikawa S.* Efficient Substructure Discovery from Large Semi-structured Data. // SIAM SDM'02. - 2002.
290. *Chi Y.; Muntz R.R.; Nijssen S.; Kok J.N.* Frequent subtree mining -- an overview. // Fundamenta Informaticae - 2005. - V. 66, № 1-2. - P. 161-198.
291. *Inokuchi A.; Washio T.; Motoda H.* In An Apriori-Based Algorithm for Mining Frequent Substructures from Graph Data, 4th European Conf. on Principles and Practice of Knowledge Discovery in Databases (PKDD), Lyon, France, September 2000, 2000; Lyon, France, September 2000. - 2000. - P. 13-23.
292. *Kuramochi M.; Karypis G.* In Frequent Subgraph Discovery, 1st IEEE Conference on Data Mining, 2001. - 2001. - P. 313-320.
293. *Borgelt C.; Meinel T.; Berthold M.* MoSS: A Program for Molecular Substructure Mining. // Proceedings of the 1st international Workshop on Open Source Data Mining: Frequent Pattern Mining Implementations ACM Press, New York, NY: Chicago, Illinois, August 21 - 21, 2005. - 2005. - P. 6-15.
294. *Zaki M.J.* Efficiently mining frequent trees in a forest. // Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM Press: Edmonton, Alberta, Canada. - 2002. - P. 71-80
295. *Chi Y.; Yang Y.; Muntz R.R.* HybridTreeMiner: An efficient algorithm for mining frequent rooted trees and free trees using canonical forms. // The 16th International Conference on Scientific and Statistical Database Management (SSDBM'04), June 2004. - 2004.
296. *Chi Y.; Yang Y.; Xia Y.; Muntz R.R.* CMTreeMiner: Mining both closed and maximal frequent subtrees. // The Eighth Pacific Asia Conference on Knowledge Discovery and Data Mining (PAKDD'04), May 2004. - 2004.
297. *Dehaspe L.; Toivonen H.; King R.D.* Finding frequent substructures in chemical compounds. // 4th International Conference on Knowledge Discovery and Data Mining, Agrawal R.; Stolorz P.; Piatetsky-Shapiro G., Eds. AAAI Press. - 1998. - P. 30-36.

298. *Deshpande M.; Kuramochi M.; Karypis G.* Frequent sub-structure based approaches for classifying chemical compounds. // Proceedings of the Third IEEE international Conference on Data Mining (November 19 - 22, 2003). ICDM., IEEE Computer Society, Washington, DC. - 2003. - P. 35-49.
299. *Demiriz A.; Bennett K.P.; Shawe-Taylor J.* Linear Programming Boosting via Column Generation. // Mach. Learn. - 2002. - V. 46, № 1-3. - P. 225-254.
300. *Graham D.J.; Malarkey C.; Schulmerich M.V.* Information Content in Organic Molecules: Quantification and Statistical Structure via Brownian Processing. // J. Chem. Inf. Comput. Sci. - 2004. - V. 44, № 5. - P. 1601-1611.
301. *Batista J.; Godden J.W.; Bajorath J.* Assessment of molecular similarity from the analysis of randomly generated structural fragment populations. // J. Chem. Inf. Model. - 2006. - V. 46, № 5. - P. 1937-1944.
302. *Batista J.; Bajorath J.* Chemical database mining through entropy-based molecular similarity assessment of randomly generated structural fragment populations. // J. Chem. Inf. Model. - 2007. - V. 47, № 1. - P. 59-68.
303. *Sanderson D.M.; Earnshaw C.G.* Computer prediction of possible toxic action from chemical structure; the DEREK system. // Hum. Exp. Toxicol. - 1991. - V. 10, № 4. - P. 261-273.
304. *Takeuchi K.; Kuroda C.; Ishida M.* Prolog-based functional group perception and calculation of 1-octanol/water partition coefficients using Rekker's fragment method. // J. Chem. Inf. Model. - 1990. - V. 30, № 1. - P. 22-26.
305. *Chen L.* Reaction Classification and Knowledge Acquisition. // Handbook of Chemoinformatics, Gasteiger J., Ed. Wiley-VCH: Weinheim. - 2003. - V. 1. - P. 348-388.
306. *Dugundji J.; Ugi I.* An Algebraic Model of Constitutional Chemistry as a Basis for Chemical Computer Programs. // Topics Curr. Chem. - 1973. - V. 39 - P. 19-64.
307. *Zefirov N.S.; Trach S.S.* Systematization of tautomeric processes and formal-logical approach to the search for new topological and reaction types of tautomerism. // Chemica Scripta. - 1980. - V. 15, № 1. - P. 4-12.
308. *Zefirov N.S.* An approach to systematization and design of organic reactions. // Accounts of Chemical Research. - 1987. - V. 20, № 7. - P. 237-243.

309. *Vladutz G.* Modern Approaches to Chemical Reaction Searching. // Approaches to Chemical Reaction Searching, Willett P., Ed. Gower: London. - 1986. - P. 202-220.
310. *Fujita S.* Description of Organic Reactions Based on Imaginary Transition Structures. 1. Introduction of New Concepts. // J. Chem. Inf. Comput. Sci. - 1986. - V. 26, № 4. - P. 205-212.
311. *Fujita S.* 'Structure-Reaction Type' Paradigm in the Conventional Methods of Describing Organic Reactions and the Concept of Imaginary Transition Structures Overcoming This Paradigm. // J. Chem. Inf. Comput. Sci. - 1987. - V. 27, № 3. - P. 120-126.
312. *Borodina Y.; Rudik A.; Filimonov D.; Kharchevnikova N.; Dmitriev A.; Bli-nova V.; Poroikov V.* A New Statistical Approach to Predicting Aromatic Hydroxyla-tion Sites. Comparison with Model-Based Approaches. // J. Chem. Inf. Comput. Sci. - 2004. - V. 44, № 6. - P. 1998-2009.
313. *Kier L.B.; Hall L.H.* Molecular Connectivity in Chemistry and Drug Research. - Academic Press: New York (NY). - 1976. - 257 p.
314. *Knuth D.* Section 6.4: Hashing. // The Art of Computer Programming, Volume 3: Sorting and Searching, Second Edition ed.; Addison-Wesley: Reading, MA. - 1988. - V. 3. - P. 513-558.
315. *Cormen T.H.; Leiserson C.E.; Rivest R.L.; Stein C.* Chapter 11: Hash Tables. // Introduction to Algorithms, Second ed.; MIT Press and McGraw-Hill. - 2001. - P. 224-228.
316. *Ash S.; Cline M.A.; Homer R.W.; Hurst T.; Smith G.B.* SYBYL Line Notation (SLN): A Versatile Language for Chemical Structure Representation. // J. Chem. Inf. Comput. Sci. - 1997. - V. 37, № 1. - P. 71-79.
317. *Knuth D.E.* Sorting and searching. // The art of computer programming. - Ad-dison-Wesley: Reading, MA. - 1973. - V. 3. - P. 490-493.
318. *Tarasov V.A.; Mustafaev O.N.; Abilev S.K.; Mel'nik V.A.* Use of compound structural descriptors for increasing the efficiency of QSAR study. // Russian Journal of Genetics. - 2005. - V. 41, № 7. - P. 814-821.

319. *Кадыров Ч.Ш.; Тюрина Л.А.; Симонов В.Д.; Семенов В.А.* Машинный поиск химических препаратов с заданными свойствами. - Фан: Ташкент. - 1989. - 164 с.
320. *Gillet V.J.; Willett P.; Bradshaw J.* Similarity Searching Using Reduced Graphs. // *J. Chem. Inf. Comput. Sci.* - 2003. - V. 43, № 2. - P. 338-345.
321. *Barker E.J.; Gardiner E.J.; Gillet V.J.; Kitts P.; Morris J.* Further Development of Reduced Graphs for Identifying Bioactive Compounds. // *J. Chem. Inf. Comput. Sci.* - 2003. - V. 43, № 2. - P. 346-356.
322. *Tetko I.V.; Bruneau P.; Mewes H.-W.; Rohrer D.C.; Poda G.I.* Can We Estimate the Accuracy of ADMET Predictions? // *Drug Discovery Today.* - 2006. - V. 11, № 15/16. - P. 700-707.
323. *Leo A.J.; Hoekman D.* Calculating log P(oct) with no missing fragments; The problem of estimating new interaction parameters. // *Persp. Drug Discov. Des.* - 2000. - V. 18. - P. 19-38.
324. *Honorio K.M.; Garratt R.C.; Andricopulo A.D.* Hologram quantitative structure-activity relationships for a series of farnesoid X receptor activators. // *Bioorg. Med. Chem. Lett.* - 2005. - V. 15, № 12. - P. 3119-3125.
325. *Judson P.N.* Rule Induction for Systems Predicting Biological Activity. // *J. Chem. Inf. Comput. Sci.* - 1994. - V. 34, № 1. - P. 148-153.
326. *Станкевич М.И.; Станкевич И.В.; Зефирова Н.С.* Топологические индексы в органической химии. // *Успехи химии.* - 1988. - Т. 57, № 3. - С. 337-366.
327. *Rouvray D.H.* Should We Have Designs on Topological Indexes? // *Chemical Applications of Topology and Graph Theory*, King R.B., Ed. Elsevier: Amsterdam. - 1983. - P. 159-177.
328. *Balaban A.* Chemical Graphs. XXXIV. Five New Topological Indices for the Branching of Tree-like Graphs. // *Theor. Chim. Acta.* - 1979. - V. 53, № 4. - P. 355-375.
329. *Seybold P.G.; May M.; Bagal U.A.* Molecular Structure-Property Relationships. // *J. Chem. Educ.* - 1987. - V. 64. - P. 575-581.
330. *Randic M.* Generalized Molecular Descriptors. // *J. Math. Chem.* - 1991. - V. 7. - P. 155-168.

331. *Rouvray D.H.* Predicting Chemistry from Topology. // *Sci. Am.* - 1986. - Т. 254, № 3 - С. 36-43.
332. *Мнухин В.Б.* Базис алгебры инвариантов графов. // *Математический анализ и его приложения.* - Ростов-на-Дону. - 1983. - С. 55-60.
333. *Колмогоров А.Н.* О представлении непрерывных функций нескольких переменных в виде суперпозиций непрерывных функций одного переменного и сложения. // *Докл. АН СССР.* - 1957. - Т. 114, № 5. - С. 953-956.
334. *Hecht-Nielsen R.* Kolmogorov's Mapping Neural Network Existence Theorem. // *IEEE First Annual Int. Conf. on Neural Networks, San Diego, IEEE Press: New York.* - 1987. V. 3. - P. 11-13.
335. *Sprecher D.A.* A numerical implementation of Kolmogorov's superpositions. // *Neural Networks.* - 1996. - V. 9, № 5. - P. 765-772.
336. *Sprecher D.A.* A numerical implementation of Kolmogorov's superpositions II. // *Neural Networks.* - 1997. - V. 10, № 3. - P. 447-457.
337. *Kůrková V.* Kolmogorov's theorem and multilayer neural networks. // *Neural Networks.* - 1992. - V. 5, № 3. - P. 501-506.
338. *Weigend A.S.; Huberman B.A.; Rumelhart D.E.* Predicting the future: a connectionist approach. // *Int. J. Neural Systems.* - 1990. - V. 1, № 3. - P. 193-209.
339. *Tetko I.V.; Livingstone D.J.; Luik A.I.* Neural network studies. 1. Comparison of overfitting and overtraining. // *J. Chem. Inf. Comput. Sci.* - 1995. - V. 35, № 5. - P. 826-833.
340. *Bishop C.M.* *Pattern Recognition and Machine Learning.* - Springer: New York. - 2006. - 738 p.
341. *Baskin I.I.; Skvortsova M.I.; Palyulin V.A.; Zefirov N.S.* Quantitative Chemical Structure-Property/Activity Studies Using Artificial Neural Networks. // *Foundations of Computing and Decision Sciences.* - 1997. - V. 22, № 2. - P. 107-116.
342. *Tetko I.V.* Neural Network Studies. 4. Introduction to Associative Neural Networks. // *J. Chem. Inf. Comput. Sci.* - 2002. - V. 42, № 3. - P. 717-728.
343. *Baskin I.I.; Zhokhova N.I.; Palyulin V.A.; Ivanova A.A.; Zefirov A.N.; Zefirov N.S.* Labeled Fragmental Descriptors and Their Use in QSAR/QSPR Studies. // *Book of Abstracts of the XVI European Symposium on Quantitative Structure-Activity Re-*

lationships and Molecular Modelling, 10-17 September 2006, Mediterranean Sea, Italy. - 2006. - P. 206.

344. *Manallack D.T.; Ellis D.D.; Livingstone D.J.* Analysis of linear and nonlinear QSAR data using neural networks. // *J. Med. Chem.* - 1994. - V. 37, № 22. - P. 3758-3767.

345. *Гилев С.Е.; Коченов Д.А.; Миркес Е.М.; Россиев Д.А.* Контрастирование, оценка значимости параметров, оптимизация их значений и интерпретация в нейронных сетях. // Доклады III Всероссийского семинара «Нейроинформатика и ее приложения», Красноярск. - 1995. - С. 66-78.

346. *Горбань А.Н.; Миркес Е.М.* Логически прозрачные нейронные сети для производства знаний из данных. Вычислительный центр СО РАН в г. Красноярске. Рукопись деп. в ВИНТИ 17.07.97, № 2434-B97 ed.; - Красноярск. - 1997. - 12 с.

347. *Царегородцев В.Г.* Производство полуэмпирических знаний из таблиц данных с помощью обучаемых искусственных нейронных сетей. // Методы нейроинформатики, КГТУ: Красноярск. - 1998. - С. 176-198.

348. *Царегородцев В.Г.* Технология производства явных знаний из таблиц данных при помощи нейронных сетей. // Нейроинформатика и ее приложения : Тезисы докладов VI Всероссийского семинара, 1998, КГТУ: Красноярск. - 1998. - С. 186-188.

349. *Горбань А.Н.; Царегородцев В.Г.* Производство явных знаний из таблиц данных с помощью обучаемых разреживаемых нейронных сетей. // Научная сессия МИФИ – 99. Всероссийская научно-техническая конференция «Нейроинформатика-99». Сборник научных трудов. В 3 частях. Ч. 1, МИФИ: М. - 1999. - С. 32-39.

350. *Davis G.W.* Sensitivity analysis in neural net solutions. // *IEEE Transactions on Systems, Man, and Cybernetics.* - 1989. - V. 19. - P. 1078-1082.

351. *Goldblum A.; Yoshimoto M.; Hansch C.* Quantitative structure-activity relationship of phenyl N-methylcarbamate. Inhibition of acetylcholinesterase. // *J. Agric. Food. Chem.* - 1981. - V. 29. - P. 277-288.

352. Корн Г.; Корн Т. Справочник по математике для научных работников и инженеров. Определения, теоремы, формулы. - Наука: М. - 1968. - 720 с.
353. Viswanadhan V.N.; Mueller G.A.; Basak C.; Weinstein J.N. A new QSAR algorithm combining principal component analysis with a neural network: application to calcium channel antagonists *Periodical* [Online]. - 1996. <http://www.netsci.org/Science/Compchem/feature07.html>.
354. Coburn R.A.; Wierzba M.; Suto M.J.; Solo A.J.; Triggle A.M.; Triggle D.J. 1,4-Dihydropyridine antagonist activities at the calcium channel: a quantitative structure-activity relationship approach. // *J. Med. Chem.* - 1988. - V. 31, № 11. - P. 2103-2107.
355. Gupta S.P. QSAR studies on drugs acting at the central nervous system. // *Chem. Rev.* - 1989. - V. 89, № 8. - P. 1765-1800.
356. Баскин И.И.; Палюлин В.А.; Зефирова Н.С. Программа генерации наборов подграфов для молекулярных графов. // Тезисы докладов межвузовской конференции "Молекулярные графы в химических исследованиях", Калинин, 1990. - 1990. - С. 5.
357. Баскин И.И.; Палюлин В.А.; Зефирова Н.С. Метод автоматического отбора структурных фрагментов для поиска регрессионной зависимости "структура-свойство" на основе их иерархической классификации. // Тезисы докладов I-ой Всесоюзной конференции по теоретической органической химии, Волгоград, 1991. - 1991. - С. 557.
358. Бацанов С.С. Структурная рефрактометрия. - Высшая школа: Москва. - 1976. - 304 с.
359. Stout J.M.; Dykstra C.E. Static Dipole Polarizabilities of Organic Molecules. Ab Initio Calculations and a Predictive Model. // *J. Am. Chem. Soc.* - 1995. - V. 117, № 18. - P. 5127-5132.
360. Applequist J.; Carl J.R.; Fung K.-K. Atom dipole interaction model for molecular polarizability. Application to polyatomic molecules and determination of atom polarizabilities. // *J. Am. Chem. Soc.* - 2002. - V. 94, № 9. - P. 2952-2960.
361. Miller K.J. Additivity methods in molecular polarizability. // *J. Am. Chem. Soc.* - 1990. - V. 112, № 23. - P. 8533-8542.

362. *Miller K.J.* Calculation of the molecular polarizability tensor. // *J. Am. Chem. Soc.* - 1990. - V. 112, № 23. - P. 8543-8551.
363. *Bosque R.; Sales J.* Polarizabilities of Solvents from the Chemical Composition. // *J. Chem. Inf. Comput. Sci.* - 2002. - V. 42, № 5. - P. 1154-1163.
364. *Лебедев Ю.А.; Мирошнченко Е.А.; Кнобель Ю.К.* Термохимия нитросоединений. - Наука: М. - 1970. - 168 с.
365. Физическая энциклопедия. - БСЭ: М. - 1988-1998. - Т. 3.
366. *Pascal P.* // *Ann.Chim.Phys.* - 1910. - V. 19. - P. 5-70.
367. *Schmalz T.G.; Klein D.J.; Sandleback B.L.* Chemical graph-theoretical cluster expansion and diamagnetic susceptibility. // *J. Chem. Inf. Comput. Sci.* - 2002. - V. 32, № 1. - P. 54-57.
368. *O'Sullivan P.S.; Hamerka H.F.* Semiempirical theory of diamagnetic susceptibilities with particular emphasis on oxygen-containing organic molecules. // *J. Am. Chem. Soc.* - 2002. - V. 92, № 1. - P. 25-32.
369. *de Luca G.; Russo N.; Sicilia E.; Toscano M.* Molecular quadrupole moments, second moments, and diamagnetic susceptibilities evaluated using the generalized gradient approximation in the framework of Gaussian density functional method. // *J. Chem. Phys.* - 1996. - V. 105, № 8. - P. 3206-3210.
370. *Li L.-F.; Zhang Y.; You X.-Z.* Molecular Topological Index and Its Application. 4. Relationships with the Diamagnetic Susceptibilities of Alkyl-IVA Group Organometallic Halides. // *J. Chem. Inf. Comput. Sci.* - 2002. - V. 35, № 4. - P. 697-700.
371. *Katritzky A.R.; Barczynski P.; Musumarra G.; Pisano D.; Szafran M.* Aromaticity as a quantitative concept. 1. A statistical demonstration of the orthogonality of classical and magnetic aromaticity in five- and six-membered heterocycles. // *J. Am. Chem. Soc.* - 2002. - V. 111, № 1. - P. 7-15.
372. *Пожарский А.Ф.* Теоретические основы химии гетероциклов. - Химия: М. - 1985. - 559 с.
373. *Weast R.C., CRC Handbook of Chemistry and Physics.* 64 ed.; CRS Press: Boca Raton, Florida, 1983.

374. *Abraham M.H.; McGowan J.C.* The Use of Characteristic Volumes to Measure Cavity Terms in Reversed Phase Liquid Chromatography. // *Chromatographia*. - 1987. - V. 23, № 4. - P. 243-246.
375. *Сагдеев Е.В.; Барабанов В.П.* Зависимость энтальпии парообразования органических соединений от температуры кипения. // *Журн. физ. химии*. - 2004. - Т. 78, № 12. - С. 2119-2125.
376. *Toropov A.; Toropova A.; Ismailov T.; Bonchev D.* 3D weighting of molecular descriptors for QSPR/QSAR by the method of ideal symmetry (MIS). 1. Application to boiling points of alkanes. // *J. Mol. Struct. THEOCHEM*. - 1998. - V. 424, № 3. - P. 237-247.
377. *Ivanciuc O.; Ivanciuc T.; Klein D.J.; Seitz W.A.; Balaban A.T.* Wiener index extension by counting even/odd graph distances. // *J. Chem. Inf. Comput. Sci*. - 2001. - V. 41, № 3. - P. 536-549.
378. *Chalk A.J.; Beck B.; Clark T.* A Temperature-Dependent Quantum Mechanical/Neural Net Model for Vapor Pressure. // *J. Chem. Inf. Comput. Sci*. - 2001. - V. 41, № 4. - P. 1053-1059.
379. *Wei W.; Han J.; Wen X.* Group Vector Space Method for Estimating Enthalpy of Vaporization of Organic Compounds at the Normal Boiling Point. // *J. Chem. Inf. Comput. Sci*. - 2004. - V. 44, № 4. - P. 1436-1439.
380. *Лебедев Ю.А.; Мирошнченко Е.А.* Термохимия парообразования органических веществ. - Наука: М. - 1981. - 215 с.
381. *Charlton M.H.; Docherty R.; Hutchings M.G.* Quantitative structure-sublimation enthalpy relationship studied by neural networks, theoretical crystal packing calculations and multilinear regression analysis. // *J. Chem. Soc., Perkin Trans. 2*. - 1995. - № 11. - P. 2023 - 2030.
382. *Gavezzotti A.* Molecular packing and other structural properties of crystalline oxohydrocarbons. // *J. Phys. Chem*. - 2002. - V. 95, № 22. - P. 8948-8955.
383. *Gavezzotti A.* Statistical analysis of some structural properties of solid hydrocarbons. // *J. Am. Chem. Soc*. - 2002. - V. 111, № 5. - P. 1835-1843.
384. *Puri S.; Chickos J.S.; Welsh W.J.* Three-Dimensional Quantitative Structure-Property Relationship (3D-QSPR) Models for Prediction of Thermodynamic Proper-

- ties of Polychlorinated Biphenyls (PCBs): Enthalpy of Sublimation. // J. Chem. Inf. Comput. Sci. - 2002. - V. 42, № 1. - P. 109-116.
385. *Евланов С.Ф.* Температура вспышки в открытом тигле и нижний температурный предел воспламенения жидкостей. // Журн. прикл. химии. - 1991. - Т. 64, № 4. - С. 832-836.
386. *Tetteh J.; Suzuki T.; Metcalfe E.; Howells S.* Quantitative Structure-Property Relationships for the Estimation of Boiling Point and Flash Point Using a Radial Basis Function Neural Network. // J. Chem. Inf. Comput. Sci. - 1999. - V. 39, № 3. - P. 491-507.
387. *Katritzky A.R.; Petrukhin R.; Jain R.; Karelson M.* QSPR Analysis of Flash Points. // J. Chem. Inf. Comput. Sci. - 2001. - V. 41, № 6. - P. 1521-1530.
388. *Timofei S.; Schmidt W.; Kurunczi L.; Simon Z.* A review of QSAR for dye affinity for cellulose fibres. // Dyes and Pigments. - 2000. - V. 47, № 1-2. - P. 5-16.
389. *Funar-Timofei S.; Schueuermann G.* Comparative molecular field analysis (CoMFA) of anionic azo dye-fiber affinities I: Gas-phase molecular orbital descriptors. // J. Chem. Inf. Comput. Sci. - 2002. - V. 42, № 4. - P. 788-795.
390. *Schueuermann G.; Funar-Timofei S.* Multilinear Regression and Comparative Molecular Field Analysis (CoMFA) of Azo Dye-Fiber Affinities. 2. Inclusion of Solution-Phase Molecular Orbital Descriptors. // J. Chem. Inf. Comput. Sci. - 2003. - V. 43, № 5. - P. 1502-1512.
391. *Timofei S.; Fabian W.M.F.* Comparative Molecular Field Analysis of Heterocyclic Monoazo Dye-Fiber Affinities. // J. Chem. Inf. Comput. Sci. - 1998. - V. 38, № 6. - P. 1218-1222.
392. *Fabian W.M.F.; Timofei S.* Comparative molecular field analysis (CoMFA) of dye-fibre affinities. Part 2. Symmetrical bisazo dyes. // J. Mol. Struct.: THEOCHEM. - 1996. - V. 362, № 2. - P. 155-162.
393. *Fabian W.M.F.; Timofei S.; Kurunczi L.* Comparative molecular field analysis (CoMFA), semiempirical (AM1) molecular orbital and multiconformational minimal steric difference (MTD) calculations of anthraquinone dye-fibre affinities. // J. Mol. Struct.: THEOCHEM. - 1995. - V. 340, № 1-3. - P. 73-81.

394. *Polanski J.; Gieleciak R.; Wyszomirski M.* Comparative molecular surface analysis (CoMSA) for modeling dye-fiber affinities of the azo and anthraquinone dyes. // *J. Chem. Inf. Comput. Sci.* - 2003. - V. 43, № 6. - P. 1754-1762.
395. *Bosque R.; Sales J.* A QSPR Study of the ^{31}P NMR Chemical Shifts of Phosphines. // *J. Chem. Inf. Comput. Sci.* - 2001. - V. 41, № 1. - P. 225-232.
396. *Hannongbua S.; Nivesanond K.; Lawtrakul L.; Pungpo P.; Wolschann P.* 3D-Quantitative Structure-Activity Relationships of HEPT Derivatives as HIV-1 Reverse Transcriptase Inhibitors, Based on Ab Initio Calculations. // *J. Chem. Inf. Comput. Sci.* - 2001. - V. 41, № 3. - P. 848-855.
397. *Пальм В.А.* Таблицы констант скорости и равновесия гетеролитических органических реакций. - ВИНТИ: М. - 1975. - Т. 1(2). - 299 с.
398. *Halberstam N.M.; Baskin I.I.; Palyulin V.A.; Zefirov N.S.* Quantitative structure-conditions-property relationship studies. Neural network modelling of the acid hydrolysis of esters. // *Mendeleev Communications.* - 2002. - № 5. - P. 185-186.
399. *Ингольд К.* Теоретические основы органической химии. - Мир: М. - 1973. - 1055 с.
400. *Жохова Н.И.; Бобков Е.В.; Баскин И.И.; Палюлин В.А.; Зефиоров А.Н.; Зефиоров Н.С.* Расчет стабильности комплексов органических соединений с β -циклодекстрином с помощью метода QSPR. // *Вестн. Моск. ун-та. Сер. 2. Химия.* - 2007. - Т. 48, № 5. - С. 329-332.
401. *Varnek A.; Kireeva N.; Tetko I.V.; Baskin I.I.; Solov'ev V.P.* Exhaustive QSPR studies of a large diverse set of ionic liquids: How accurately can we predict melting points? // *J. Chem. Inf. Model.* - 2007. - V. 47, № 3. - P. 1111-1122.
402. *van Krevelen D.V.* Properties of Polymers. Second ed. - Elsevier: Amsterdam. - 1976. - 264 p.
403. *Аскадский А.А.; Матвеев Ю.И.* Химическое строение и физические свойства полимеров. - Химия: М. - 1983. - 248 с.
404. *Vicserano J.* Prediction of polymer properties. Second ed.- Marcel Dekker, Inc.: New York. - 1996. - 528 p.
405. *Баскин И.И.; Бузников Г.А.; Кабанкин А.С.; Ландау М.А.; Лексина Л.А.; Ордуханян А.А.; Палюлин В.А.; Зефиоров Н.С.* Компьютерное изучение зависи-

- мости между эмбриотоксичностью и структурами синтетических аналогов биогенных аминов. // Изв. РАН. Сер. биол. - 1997. - Т. № 4. - С. 407-413.
406. *Баскин И.И.; Палюлин В.А.; Зефирова Н.С.* Вычислительные нейронные сети как альтернатива линейному регрессионному анализу при изучении количественных соотношений "структура-свойство" на примере физико-химических свойств углеводородов. // Докл. РАН. - 1993. - Т. 332, № 6. - С. 713-716.
407. *Selected Values of Physical and Thermodynamic Properties of Hydrocarbons and Related Compounds.* - Carnegie Press: Pittsburgh. - 1953.
408. *Balaban A.T.; Kier L.B.; Joshi N.* Structure-property analysis of octane numbers for hydrocarbons (alkanes, cycloalkanes, alkenes). // MATCH. - 1992. - V. 28. - С. 13-27.
409. *Needham D.E.; Wei I.C.; Seybold P.G.* Molecular modeling of the physical properties of alkanes. // J. Am. Chem. Soc. - 1988. - V. 110, № 13. - P. 4186-4194.
410. *Ivanciuc O.; Ivanciuc T.; Filip P.A.; Cabrol-Bass D.* Estimation of the liquid viscosity of organic compounds with a quantitative structure-property model. // J. Chem. Inf. Comput. Sci. - 1999. - V. 39, № 3. - P. 515-524.
411. *Ванник В.Е.; Червоненкис А.Я.* Теория распознавания образов. - Наука: М. - 1979. - 237 с.
412. *Rissanen J.* A universal prior for the integers and estimation by minimum description length. // Annals of Statistics. - 1983. - V. 11, № 2. - P. 416-431.
413. *Rissanen J.* Universal coding, information, prediction, and estimation. // IEEE Trans. Inf. Theory. - 1984. - V. 30 - P. 629-636.
414. *Katritzky A.R.; Chen K.; Wang Y.L.; Karelson M.; Lucic B.; Trinajstić N.; Suzuki T.; Schuurmann G.* Prediction of Liquid Viscosity for Organic Compounds by a Quantitative Structure-Property Relationship. // J. Phys. Org. Chem. - 2000. - V. 13, № 1. - P. 80-86.
415. *Flukalog Database*, Fluka Chemie AG: 1995.
416. *Katritzky A.R.; Maran U.; Lobanov V.S.; Karelson M.* Structurally Diverse Quantitative Structure-Property Relationship Correlations of Technologically Relevant Physical Properties. // J. Chem. Inf. Comput. Sci. - 2000. - V. 40, № 1. - P. 1-18.

417. *Goll E.S.; Jurs P.C.* Prediction of Vapor Pressures of Hydrocarbons and Halo-hydrocarbons from Molecular Structure with a Computational Neural Network Model. // *J. Chem. Inf. Comput. Sci.* - 1999. - V. 39, № 6. - P. 1081-1089.
418. *Hall L.H.; Story C.T.* Boiling Point and Critical Temperature of a Heterogeneous Data Set: QSAR with Atom Type Electrotopological State Indices Using Artificial Neural Networks. // *J. Chem. Inf. Comput. Sci.* - 1996. - V. 36, № 5. - P. 1004-1014.
419. *Egolf L.M.; Jurs P.C.* Prediction of boiling points of organic heterocyclic compounds using regression and neural network techniques. // *J. Chem. Inf. Comput. Sci.* - 1993. - V. 33, № 4. - P. 616-625.
420. *Egolf L.M.; Wessel M.D.; Jurs P.C.* Prediction of boiling points and critical temperatures of industrially important organic compounds from molecular structure. // *J. Chem. Inf. Comput. Sci.* - 1994. - V. 34, № 4. - P. 947-956.
421. *Hall L.H.; Story C.T.* Boiling point of a set of alkanes, alcohols and chloroalkanes: QSAR with atom type electrotopological state indices using artificial neural networks. // *SAR and QSAR in Environmental Research.* - 1997. - V. 6, № 3-4. - P. 139-161.
422. *Лифшиц Э.Б.; Райхина Р.Д.; Куркина Л.Г.; Ушомирский М.Н.* Применение методов корреляционного анализа для изучения зависимости свойств полиметиновых красителей от их строения. // *Журн. науч. и прикл. фото- и кинематографии.* - 1996. - V. 41, № 1. - P. 43-62.
423. *Киприанов А.И.* Избранные труды. - Наук. Думка: Киев. - 1979. - 649 с.
424. *Левкоев И.И.* Избранные труды. - Наука: М. - 1982. с.
425. *Фэрстер Т.* Окраска и строение органических соединений с точки зрения современной физической теории. // *Успехи химии.* - 1940. - Т. 9, № 1. - С. 71-104.
426. *Dewar M.I.S.* Colour and Constitution. Part I. Basic Dyes. // *J. Chem. Soc.* - 1950. - № 3. - P. 2329-2334.
427. *Knott E.B.* The Colour of Organic Compounds. Part I. A General Colour Rule. // *J. Chem. Soc.* - 1951. - № 2. - P. 1024-1028.

428. Дядюша Г.Г.; Качковский А.Д. Длины волн первых электронных переходов симметричных цианиновых красителей. // Укр. хим. журн. - 1975. - Т. 41, № 11. - С. 1176-1181.
429. Ait A.O.; Baskin I.I.; Varachevsky V.A.; Alfimov M.V. In Client/Server Molecular Modeling System for Automatic Construction of 3D Data Base on Photochromic Compounds, International Symposium CACR-96, Moscow, December 17-18, 1996. - Moscow, 1996. - P. 37.
430. Tehan B.G.; Lloyd E.J.; Wong M.G.; Pitt W.R.; Montana J.G.; Manallack D.T.; Gancia E. Estimation of pK_a Using Semiempirical Molecular Orbital Methods. Part 1: Application to Phenols and Carboxylic Acids. // Quant. Struct.-Act. Relat. - 2002. - V. 21, № 5. - P. 457-472.
431. Tehan B.G.; Lloyd E.J.; Wong M.G.; Pitt W.R.; Gancia E.; Manallack D.T. Estimation of pK_a Using Semiempirical Molecular Orbital Methods. Part 2: Application to Amines, Anilines and Various Nitrogen Containing Heterocyclic Compounds. // Quant. Struct. - Act. Relat. - 2002. - V. 21, № 5. - P. 473-485.
432. Gross K.C.; Seybold P.G.; Peralta-Inga Z.; Murray J.S.; Politzer P. Comparison of Quantum Chemical Parameters and Hammett Constants in Correlating pK_a Values of Substituted Anilines. // J. Org. Chem. - 2001. - V. 66, № 21. - P. 6919-6925.
433. Liptak M.D.; Gross K.C.; Seybold P.G.; Feldgus S.; Shields G.C. Absolute pK_a Determinations for Substituted Phenols. // J. Am. Chem. Soc. - 2002. - V. 124, № 22. - P. 6421-6427.
434. Liptak M.D.; Shields G.C. Accurate pK_a Calculations for Carboxylic Acids Using Complete Basis Set and Gaussian-n Models Combined with CPCM Continuum Solvation Methods. // J. Am. Chem. Soc. - 2001. - V. 123, № 30. - P. 7314-7319.
435. Gargallo R.; Sotriffer C.A.; Liedl K.R.; Rode B.M. Application of multivariate data analysis methods to comparative molecular field analysis (CoMFA) data: proton affinities and pK_a prediction for nucleic acids components. // J Comput Aided Mol Des. - 1999. - V. 13, № 6. - P. 611-623.
436. Баскин И.И.; Палюлин В.А.; Зефиоров Н.С. MODEL - программа интерактивного ввода молекулярных графов. // Тезисы докладов межвузовской конфе-

- ренции “Молекулярные графы в химических исследованиях”, Калинин, 1990. - 1990. - С. 6.
437. *Rosenkranz H.S.; Klopman G.* CASE, the computer-automated structure evaluation system, as an alternative to extensive animal testing. // *Toxicol Ind Health.* - 1988. - V. 4, № 4. - P. 533-540.
438. *You Z.; Brezzell M.D.; Das S.K.; Espadas-Torre M.C.; Hooberman B.H.; Sinsheimer J.E.* Ortho-Substituent Effects on the in Vitro and in Vivo Genotoxicity of Benzidine Derivatives. // *Mutation Res.* - 1994. - V. 319 - P. 19-30.
439. *You Z.; Brezzell M.D.; Das S.K.; Hooberman B.H.; Sinsheimer J.E.* Substituent Effects on the in Vitro and in Vivo Genotoxicity of 4-Aminobiphenyl and 4-Aminostilbene Derivatives. // *Mutation Res.* - 1994. - V. 320 - P. 45-58.
440. *Абилев С.К.; Любимова И.К.; Мигачев Г.И.* Влияние структурных особенностей нитропроизводных флуоренона и бифенила на фреймшифт-мутагенез в тестерных штаммах *Salmonella typhimurium*. // *Генетика.* - 1993. - Т. 29, № 10. - С. 1640-1645.
441. *Любимова И.К.; Абилев С.К.; Мигачев Г.И.* Взаимосвязь между мутагенной активностью и химической структурой в ряду производных бифенила. // *Генетика.* - 1995. - Т. 31, № 2. - С. 268-272.
442. *Любимова И.К.; Абилев С.К.; Мигачев Г.И.* Влияние некоторых структурных особенностей в молекулах производных пирена и его гетероциклических аналогов на мутагенную активность. // *Генетика.* - 1995. - Т. 31, № 1. - С. 128-132.
443. *Баскин И.И.; Любимова И.К.; Абилев С.К.; Зефиоров Н.С.* Исследование количественной связи между мутагенной активностью химических соединений и их структурой. Замещенные бифенилы. // *Докл. РАН.* - 1993. - Т. 332, № 5. - С. 587-589.
444. *Баскин И.И.; Палюлин В.А.; Любимова И.К.; Абилев С.К.; Зефиоров Н.С.* Количественная связь между мутагенной активностью гетероциклических аналогов пирена и фенантрена и их структурой. // *Докл. РАН.* - 1994. - Т. 339, № 1. - С. 106-108.

445. Любимова И.К.; Абилов С.К.; Гальберштам Н.М.; Баскин И.И.; Палюлин В.А.; Зефирова Н.С. Компьютерное предсказание мутагенной активности замещенных полициклических соединений. // Изв. РАН, Сер. биол. - 2001. - Т. № 2. - С. 180-186.
446. Любимова И.К. Зависимость мутагенной активности полициклических ароматических соединений от их структуры. Автореферат диссертации на соискание ученой степени кандидата биологических наук. - М. - 1994.
447. Дьячков П.Н. Квантовохимические расчеты в изучении механизма действия и токсичности чужеродных веществ. // Итоги науки и техн. ВИНТИ. Сер. Токсикология. - 1990. - Т. 16, № - С. 1-280.
448. Vance W.; Levin D. Structural Features of Nitroaromatics That Determine Mutagenic Activity in Salmonella Typhimurium. // Environ. Mutagen. - 1984. - V. 6. - P. 797-811.
449. Hirayama T.; Kusakabe H.; Watanabe T.; Ozasa S.; Fujioka Y.; Fukui S. Relationship Between Mutagenic Potency in Salmonella Strains and the Chemical Structure of Nitrobiphenyls. // Mutat. Res. - 1986. - V. 163, № 2. - P. 101-107.
450. Гальберштам Н.М.; Баскин И.И.; Палюлин В.А.; Зефирова Н.С. Прогнозирование констант скоростей реакций кислотного гидролиза сложных эфиров с использованием искусственных нейронных сетей. // Труды VII Всероссийской конференции «Нейрокомпьютеры и их применение». НКП-2001 с международным участием. - 2001. - С. 423-424.
451. Гальберштам Н.М.; Баскин И.И.; Палюлин В.А.; Зефирова Н.С. Применение методологии искусственных нейронных сетей для прогнозирования констант скорости реакции кислотного гидролиза сложных эфиров. // Труды 2-ой Всероссийской конференции «Молекулярное моделирование». - 2001. - С. 62.
452. Lukovits I. The detour index. // Croat. Chem. Acta. - 1996. - V. 69, № 3. - P. 873-882.
453. Toropov A.A.; Toropova A.P.; Ismailov T.T.; Voropaeva N.L.; Ruban I.N. Extended molecular connectivity: prediction of boiling points of alkanes. // J. Struct. Chem. - 1998. - V. 38, № 6. - P. 965-969.

454. *Kobakhidze N.; Gverdtsiteli M.* Algebraic study of cycloalkanes. // Bull. Georgian Acad. Sci. - 1996. - V. 153, № 1. - P. 55-56.
455. *Plavsic D.; Trinajstic N.; Amic D.; Soskic M.* Comparison between the structure-boiling point relationships with different descriptors for condensed benzenoids. // New J. Chem. - 1998. - V. 22, № 10. - P. 1075-1078.
456. *Ren B.* A New Topological Index for QSPR of Alkanes. // J. Chem. Inf. Comput. Sci. - 1999. - V. 39, № 1. - P. 139-143.
457. *Castro E.A.; Tueros M.; Toropov A.A.* Maximum topological distances based indices as molecular descriptors for QSPR. Application to aromatic hydrocarbons. // Comput. Chem. - 2000. - V. 24, № 5. - P. 571-576.
458. *Ivanciuc O.; Ivanciuc T.; Balaban A.T.* The complementary distance matrix, a new molecular graph metric. // ACH – Models Chem. - 2000. - V. 137, № 1. - P. 57-82.
459. *Randic M.* High quality structure-property regressions. Boiling points of smaller alkanes. // New J. Chem. - 2000. - V. 24, № 3. - P. 165-171.
460. *Gutman I.; Tomovic Z.* On the application of line graphs in quantitative structure-property studies. // J. Serb. Chem. Soc. - 2000. - V. 65, № 8. - P. 577-580.
461. *Thanikaivelan P.; Subramanian V.; Raghava R.J.; Unni N.B.* Application of quantum chemical descriptor in quantitative structure activity and structure property relationship. // Chem. Phys. Lett. - 2000. - V. 323, № 1-2. - P. 59-70.
462. *Randic M.* Quantitative structure-property relationship. Boiling points of planar benzenoids. // New J. Chem. - 1996. - V. 20, № 10. - P. 1001-1009.
463. *Lucic B.; Trinajstic N.* New developments in QSPR/QSAR modeling based on topological indices. // SAR QSAR Environ. Res. - 1997. - V. 7. - P. 45-62.
464. *Liu S.; Cao C.; Li Z.* Approach to Estimation and Prediction for Normal Boiling Point (NBP) of Alkanes Based on a Novel Molecular Distance-Edge (MDE) Vector, I. // J. Chem. Inf. Comput. Sci. - 1998. - V. 38, № 3. - P. 387-394.
465. *Espinosa G.; Yaffe D.; Cohen Y.; Arenas A.; Giralt F.* Neural Network Based Quantitative Structural Property Relations (QSPRs) for Predicting Boiling Points of Aliphatic Hydrocarbons. // J. Chem. Inf. Comput. Sci. - 2000. - V. 40, № 3. - P. 859-879.

466. *Goll E.S.; Jurs P.C.* Prediction of the Normal Boiling Points of Organic Compounds from Molecular Structures with a Computational Neural Network Model. // *J. Chem. Inf. Comput. Sci.* - 1999. - V. 39, № 6. - P. 974-983.
467. *Татевский В.М.* Физико-химические свойства индивидуальных углеводов. - Гостоптехиздат: М. - 1960. - 412 с.
468. *Гордон А.; Форд Р.* Спутник химика. - Мир: М. - 1976. - 541 с.
469. *Kreiger A.G.; K. W.C.* Computer Iteration of Handbook Data. // *J. Chem. Educ.* - 1971. - V. 48. - P. 457.
470. Goodman J.M.; Kirby P.D.; Haustedt L.O. Some Calculations for Organic Chemists: Boiling Point Variation, Boltzman Factors and Eyring Equation *Periodical* [Online]. <http://preprint.chemweb.com/orgchem/0009006>.
471. *Yaffe D.; Cohen Y.* Neural Network Based Temperature-Dependent Quantitative Structure Property Relations (QSPRs) for Predicting Vapor Pressure of Hydrocarbons. // *J. Chem. Inf. Comput. Sci.* - 2001. - V. 41, № 2. - P. 463-477.
472. *Silver M.S.* The Effect of the Nature of the Leaving Group upon Relative Solvolytic Reactivity. // *J. Am. Chem. Soc.* - 1961. - V. 83, № 2. - P. 404-408.
473. *Seoud O.A.; Martins M.F.* Kinetics and Mechanism of the Hydrolysis of Substituted Phenyl Benzoates Catalyzed by the o-Iodosobenzoate Anion. // *J. Phys. Org. Chem.* - 1995. - V. 8, № 10. - P. 637-646.
474. *Neuvonen H.; Neuvonen K.* Correlation Analysis of Carbonyl Carbon ¹³C NMR Chemical Shifts, IR Absorption Frequencies and Rate Coefficients of Nucleophilic Acyl Substitutions. A Novel Explanation for the Substituent Dependence of Reactivity. // *J. Chem. Soc., Perkin Trans. 2.* - 1999. - № 7. - P. 1497-1502.
475. *Пальм В.А.* Основы количественной теории органических реакций. - Химия: Л. - 1977. - 359 с.
476. *Stimson V.R.* The Kinetics of Alkyl-Oxygen Fission in Ester Hydrolysis. Part II. tert.-Butyl 2:4:6-Trimethylbenzoate in Aqueous Acetone. // *J. Chem. Soc.* - 1955. - P. 2010-2013.
477. *Varnek A.; Gaudin C.; Marcou G.; Baskin I.; Pandey A.K.; Tetko I.V.* Inductive Transfer of Knowledge: Application of Multi-Task Learning and Feature Net

- Approaches to Model Tissue-Air Partition Coefficients. // *J. Chem. Inf. Model.* - 2009. - Т. 49, № 1. - С. 133-144.
478. *Baskin I.I.; Ait A.O.; Halberstam N.M.; Palyulin V.A.; Zefirov N.S.* An approach to the interpretation of backpropagation neural network models in QSAR studies. // *SAR and QSAR in Env. Res.* - 2002. - V. 13, № 1. - P. 35-41.
479. *Huuskonen J.* Prediction of Soil Sorption Coefficient of a Diverse Set of Organic Chemicals From Molecular Structure. // *J. Chem. Inf. Comput. Sci.* - 2003. - V. 43, № 5. - P. 1457-1462.
480. *Marcus Y.; Smith A.L.; Korobov M.V.; Mirakyan A.L.; Avramenko N.V.; Stukalin E.B.* Solubility of C60 Fullerene. // *The Journal of Physical Chemistry B.* - 2001. - V. 105, № 13. - P. 2499-2506.
481. *Артеменко Н.В.; Баскин И.И.; Палюлин В.А.; Зефирова Н.С.* Прогнозирование физических свойств органических соединений при помощи искусственных нейронных сетей в рамках подструктурного подхода. // *Докл. РАН.* - 2001. - Т. 381, № 2. - С. 203-206.
482. *Жохова Н.И.; Баскин И.И.; Палюлин В.А.; Зефирова А.Н.; Зефирова Н.С.* Фрагментные дескрипторы с «выделенными» атомами и их применение в исследованиях количественных соотношений «структура-активность» / «структура-свойство». // *Докл. РАН.* - 2007. - Т. 417, № 5. - С. 639-641.
483. *Артеменко Н.В.; Палюлин В.А.; Зефирова Н.С.* Нейросетевая модель липофильности органических соединений на основе фрагментных дескрипторов. // *Докл. РАН.* - 2002. - Т. 383, № 6. - С. 771-773.
484. *Jover J.; Bosque R.; Sales J.* Determination of Abraham Solute Parameters from Molecular Structure. // *J. Chem. Inf. Comput. Sci.* - 2004. - V. 44, № 3. - P. 1098-1106.
485. *Caruana R.* Multitask Learning. // *Machine Learning.* - 1997. - V. 28, № 1. - P. 41-75.
486. *Elrod D.W.; Maggiora G.M.; Trenary R.G.* Applications of Neural Networks in Chemistry. 1. Prediction of Electrophilic Aromatic Substitution Reactions. // *J. Chem. Inf. Comput. Sci.* - 1990. - V. 30, № 4. - P. 477-484.

487. *Elrod D.W.; Maggiora G.M.; Trenary R.G.* Application of Neural Networks in Chemistry. 2. A General Connectivity Representation for the Prediction of Regiochemistry. // *Tetrahedron Comput. Methodol.* - 1990. - V. 3 - P. 163-174.
488. *West G.* Empirical ^{31}P Spectrum Prediction by Neural Networks. // *NATO-ASI Molecular Spectroscopy: Recent Experimental and Computational Advances*, Ponta Delgada. - 1992.
489. *Kvasnička V.* An Application of Neural Networks in Chemistry. Prediction of ^{13}C NMR Chemical Shifts. // *J. Math. Chem.* - 1991. - V. 6. - P. 63-76.
490. *West G.M.J.* Predicting Phosphorus NMR Shifts Using Neural Networks. // *J. Chem. Inf. Comput. Sci.* - 1993. - V. 33, № 4. - P. 577-589.
491. *Kireev D.B.* ChemNet: A Novel Neural Network Based Method for Graph/Property Mapping. // *J. Chem. Inf. Comput. Sci.* - 1995. - V. 35, № 2. - P. 175-180.
492. *Fukushima K.* Neocognitron: A Self-Organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position. // *Biol. Cybernetics.* - 1980. - V. 36 - P. 193-202.
493. *Fukushima K.; Miyake S.* Neocognitron: A New Algorithm for Pattern Recognition Tolerant of Deformations and Shifts in Position. // *Pattern Recognition.* - 1982. - V. 15. - P. 455-469.
494. *Fukushima K.* A Hierarchical Neural Network Model for Associative Memory. // *Biol. Cybernetics.* - 1984. - V. 50. - P. 105-113.
495. *Fukushima K.* A Neural Network Model for Selective Attention in Visual Pattern Recognition. // *Biol. Cybernetics.* - 1986. - V. 55. - P. 5-15.
496. *Hubel D.H.; Wiesel T.N.* Receptive Fields, Binocular Interaction and Functional Architecture in the Cat's Visual Cortex. // *J. Physiol.* - 1962. - V. 160. - P. 106-154.
497. *Hubel D.H.; Wiesel T.N.* Receptive Fields and Functional Architecture in Two Nonstriate Visual Areas (18 and 19) of the Cats. // *J. Neurophysiol.* - 1965. - V. 28. - P. 229-289.
498. *Hubel D.H.; Wiesel T.N.* Functional Architecture of Macaque Monkey Visual Cortex. // *Proc. Roy. Soc. London Ser. B.* - 1977. - V. 98 - P. 1-59.

499. *Lohninger H.* Evaluation of Neural Networks Based on Radial Basis Functions and Their Application to the Prediction of Boiling Points from Structural Parameters. // *J. Chem. Inf. Comput. Sci.* - 1993. - V. 33, № 5. - P. 736-744.
500. *Cherqaoui D.; Villemin D.* Use of a neural network to determine the boiling point of alkanes. // *J. Chem. Soc., Faraday Trans.* - 1994. - V. 90, № 1. - P. 97-102.
501. *Cherqaoui D.; Villemin D.; Kvasnicka V.* Application of neural network approach for prediction of some thermochemical properties of alkanes. // *Chemometrics and Intelligent Laboratory Systems.* - 1994. - V. 24, № 2. - P. 117-128.
502. *Yan L.; Chen N.* Quantitative structure-activity relationship study of octane number and boiling point of alkanes using artificial neural networks method. // *Jisunji Yu Yingyong Huaxue.* - 1994. - V. 11. - P. 286-287.
503. *Bianucci A.M.; Micheli A.; Sperduti A.; Starita A.* A novel approach to QSPR/QSAR based on neural networks for structures. // *Studies in Fuzziness and Soft Computing.* - 2003. - V. 120. - P. 265-296.
504. *Rossini F.D.; Pitzer K.S.; Arnett R.L.; Braun R.M.; Pimentel G.C.* Selected Values of Physical and Thermodynamic Properties of Hydrocarbons and Related Compounds. - Carnegie Press: Pittsburgh, PA. - 1953. p.
505. *Антипин И.С.; Арсланов Н.А.; Палюлин В.А.; Коновалов А.И.; Зефирова Н.С.* Сольватационный топологический индекс. Топологическая модель описания дисперсионных взаимодействий. // *ДАН СССР.* - 1991. - Т. 316, № 4 - С. 925-928.
506. *Miller K.J.* Additivity methods in molecular polarizability. // *J. Am. Chem. Soc.* - 1990. - V. 112, № 23. - P. 8533-8542.
507. *Баскин И.И.; Зефирова Н.С.; Трач С.С.* “Модель” - универсальная программа графики для целей органической химии. // Тезисы докладов 7 Всесоюзной конференции “Использование вычислительных машин в химических исследованиях и спектроскопии молекул”, Рига, 1986. - 1986. - С. 27-28.
508. *Зефирова Н.С.; Баскин И.И.; Трач С.С.* Универсальная программа машинной графики для целей органической химии. // *Журн. Всес. хим. о-ва им. Д.И. Менделеева.* - 1987. - Т. 32, № 1. - С. 112-113.

509. *Баскин И.И.; Трач С.С.; Зефирова Н.С.* Программа поиска новых типов реагирования органических соединений на ПЭКВМ “Искра-226”. // Тезисы докладов 7 Всесоюзной конференции “Использование вычислительных машин в химических исследованиях и спектроскопии молекул”, Рига, 1986. - 1986. - С. 27-28.
510. *Станкевич М.И.; Баскин И.И.; Зефирова Н.С.* Комбинаторные модели и алгоритмы в химии. Поиск структурных фрагментов. - Деп. ВИНТИ, №4288-В: - 1986. - 28 с.
511. *Станкевич М.И.; Баскин И.И.; Зефирова Н.С.* Автоматизированный поиск структурных фрагментов. Алгоритм и программа. // Журн. структ. химии. - 1987. - Т. 28, № 6. - С. 136-137.
512. *Баскин И.И.; Станкевич М.И.; Девдариани Р.О.; Зефирова Н.С.* Комплекс программ для нахождения корреляций “структура-свойство” на основе топологических индексов. // Журн. структ. химии. - 1989. - Т. № 6. - С. 145-147.
513. *Баскин И.И.; Девдариани Р.О.; Палюлин В.А.; Скворцова М.И.; Зефирова Н.С.* Прогнозирование температур плавления ароматических соединений некоторых классов на основе использования взвешенных топологических индексов. // Тезисы докладов VIII Всесоюзной конференции “Использование вычислительных машин в спектроскопии молекул и химических исследованиях”, Новосибирск, 1989. - 1989. - С. 251.
514. *Lomova O.A.; Sukhachev D.V.; Kumskov M.I.; Palyulin V.A.; Zefirov N.S.* The Generation of Molecular Graphs for QSAR Studies by the Acyclic Fragment Combining. // MATCH. - 1992. - V. 27. - P. 153-174.
515. *Tratch S.S.; Lomova O.A.; Sukhachev D.V.; Palyulin V.A.; Zefirov N.S.* Generation of molecular graphs for QSAR studies: an approach based on acyclic fragment combinations. // J. Chem. Inf. Comput. Sci. - 1992. - V. 32, № 2. - P. 130-139.
516. *Baskin I.I.; Palyulin V.A.; Zefirov N.S.* NASA. A computer program for performing QSAR/QSPR studies using artificial neural networks. // QSAR and Molecular Modelling: Concepts, Computational Tools and Biological Applications, Prous Science Publishers: Barcelona. - 1995. - P. 30-31.

517. *Halberstam N.M.; Baskin I.I.; Palyulin V.A.; Zefirov N.S.* In *NASAWIN – A Program Simulator of Neural Networks for Structure-Activity Relationship Studies*, International symposium CACR-96. - 1996. - P. 37-38.

БЛАГОДАРНОСТИ

Автор выражает глубокую признательность своему глубокоуважаемому учителю академику РАН Зефирову Н.С., а также всем сотрудникам, принимавшим участие в проведении исследований: в.н.с. Палюлину В.А., проф. Скворцовой М.И., с.н.с. Жоховой Н.И., д.б.н. Абилеву С.К., к.б.н. Любимовой И.К., к.ф-м.н. Айту А.О, н.с. Зефирову А.Н., к.ф-м.н. Кештовой С.В., prof. Varnek A. (University of Strasbourg, France), Tetko I.V. (Institute of Bioinformatics and Systems Biology, Neuherberg, Germany), аспирантам Гальберштам Н.М., Артеменко Н.В., Ивановой А.А.

СПИСОК ОБОЗНАЧЕНИЙ И СОКРАЩЕНИЙ

- MAE – Mean Absolute Error (средняя абсолютная ошибка)
- MAE_t – средняя абсолютная ошибка на обучающей выборке
- MAE_v – средняя абсолютная ошибка на внутренней контрольной выборке
- MAE_p – средняя абсолютная ошибка на внешней контрольной выборке
- MAE_{DCV} – средняя абсолютная ошибка, оцененная в условиях двойного скользящего контроля
- Q² – коэффициент детерминации, вычисленный в условиях скользящего контроля
- Q_{DCV}² – коэффициент детерминации, вычисленный в условиях двойного скользящего контроля
- QSAR – Quantitative Structure-Activity Relationships (количественные корреляции структура-активность)
- QSPR – Quantitative Structure-Property Relationships (количественные корреляции структура-свойство)
- R² – коэффициент детерминации
- RMSE – Root Mean Squared Error (среднеквадратичная ошибка)
- RMSE_t – среднеквадратичная ошибка на обучающей выборке
- RMSE_{v, s_v} – RMSE на внутренней контрольной выборке
- RMSE_{p, s_p} – RMSE на внешней контрольной выборке
- RMSE_{DCV} – RMSE, вычисленная в условиях двойного скользящего контроля
- БПМЛР – быстрая пошаговая множественная линейная регрессия
- ГСДЦ – граф связности дескрипторных центров
- ИНС – искусственные нейронные сети
- МОП – максимальный общий подграф
- ПФД – псевдофрагментные дескрипторы
- ТИ – топологические индексы
- ФД – фрагментные дескрипторы
- ФКСП – фрагментарный код суперпозиции подструктур
- ЦАФ – центрированные на атомах фрагменты