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

АВТОРЕФЕРАТ

диссертации на соискание ученой степени

доктора физико-математических наук

Москва - 2009

Работа выполнена в лаборатории органического синтеза кафедры органической химии Химического факультета Московского государственного университета имени М.В.Ломоносова

Официальные оппоненты: доктор физико-математических наук, профессор
Жидомиров Георгий Михайлович

доктор физико-математических наук
Кумсков Михаил Иванович

доктор химических наук, профессор
Пивина Татьяна Степановна

Ведущая организация: Институт физиологически-активных веществ
Российской академии наук
(г. Черноголовка)

Защита состоится 18 марта 2010 г. в 15 часов на заседании диссертационного совета Д 501.001.50 по химическим и физико-математическим наукам при Московском государственном университете имени М.В.Ломоносова по адресу: 119991, г. Москва, Ленинские горы, МГУ имени М.В.Ломоносова, д. 1, стр. 3, Химический факультет, ауд. 446.

С диссертацией можно ознакомиться в библиотеке Химического факультета Московского государственного университета им. М.В.Ломоносова

Автореферат разослан «11» февраля 2010 г.

Ученый секретарь
диссертационного совета, к.х.н. Матушкина Н.Н.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность работы. Современный этап развития нашей цивилизации характеризуется, прежде всего, беспрецедентным ростом мощности и распространенности компьютерной техники, и, вслед за этим, проникновением информатики во все сферы человеческой деятельности. Роботы, всевозможные устройства и компьютерные программы, оснащенные искусственным интеллектом, который уже в ближайшее время превзойдет по своим возможностям человеческий, начинают играть доминирующую роль не только в быту и промышленном производстве, но и в научных исследованиях.

Процессы информатизации быстро проникают и в химию. Этому особенно способствует то, что на протяжении многих лет химия развивалась как преимущественно эмпирическая наука, и потому в ней накоплено огромное количество экспериментальных данных, проведение глубокого анализа которых уже невозможно без применения средств современной информатики. Как результат, на стыке химии и информатики возникает и быстро оформляется в самостоятельную научную дисциплину хемоинформатика, методы которой начинают активно внедряться во все области химии, и, прежде всего, в органическую химию. Ранее этому процессу препятствовало отсутствие универсальной и строго обоснованной методологии и реализующего ее программного обеспечения, которые позволили бы химику на основе обработки экспериментальных данных осуществлять прогнозирование самых разнообразных свойств химических соединений и материалов.

На первом этапе выполнения настоящей диссертационной работы нами было теоретически обосновано, что такой универсальной методологией является сочетание искусственных нейронных сетей (ИНС) и фрагментных дескрипторов (ФД). Однако методология применения ИНС для прогнозирования свойств химических соединений была в это время практически неразвита, а в литературе имелись лишь единичные публикации в этом направлении. Известные ранее типы ФД, как правило, были нацелены на решение узкого круга задач и никак не могли быть положены в основу универсальной методологии поиска зависимостей между структурой органических соединений и их физико-химическими свойствами (QSPR), а также биологической активностью (QSAR). 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 свойства химических соединений.

Таким образом, весьма актуальным является усовершенствование и интеграция нейросетевых и фрагментных подходов для моделирования и прогнозирования свойств органических соединений.

Цель работы. Целью настоящей диссертационной работы является создание универсальной методологии на базе ИНС и ФД, а также реализующего ее программного комплекса, позволяющего находить и анализировать количественные

зависимости между структурами органических соединений и их свойствами (с учетом и без учета влияния внешних условий), и на основе этого прогнозировать свойства еще неизученных соединений.

Научная новизна работы.

1. Впервые применен аппарат искусственных нейронных сетей для количественного прогнозирования физико-химических свойств органических соединений и их реакционной способности.
2. Впервые разработан и применен универсальный подход к прогнозированию свойств органических соединений на основе комбинированного использования искусственных нейронных сетей и фрагментных дескрипторов.
3. Впервые предложена методика построения нелинейных зависимостей «структура-условия-свойства».
4. Впервые предложен метод интерпретации нейросетевых количественных зависимостей свойств органических соединений от их структуры.
5. Впервые разработаны и применены методы интеграции нейросетевых моделей «структура-свойство» на основе многоуровневого и многозадачного принципов их построения.
6. Впервые предложена концепция проведения прямых корреляций «структура-свойство» и на ее основе разработаны специальные архитектуры нейронных сетей, позволяющие осуществлять прогнозирование свойств органических соединений непосредственно из описания молекулярного графа без промежуточного вычисления вектора молекулярных дескрипторов. Тем самым впервые было осуществлено построение статистических регрессионных моделей с использованием невекторных (структурных, графовых) данных.
7. Впервые построены QSPR-модели «структура-свойство», позволяющие прогнозировать спектральные свойства красителей, а также кинетические константы гомогенных органических реакций.

Результатом работы явилось создание нового научного направления – **нейросетевого моделирования свойств органических соединений на основе фрагментного подхода.**

Практическая значимость работы. Предложенные методики позволяют расширить область традиционного моделирования «структура-свойство», улучшить прогнозирующую способность получаемых моделей, интерпретировать нейросетевые модели. Разработанный программный комплекс является универсальным инструментом для изучения зависимостей «структура-свойство», «структура-условия-свойство» и может широко использоваться для моделирования и прогноза широкого спектра свойств химических соединений. Построенные нейросетевые модели позволяют прогнозировать ряд физико-химических свойств, реакционную способность и биологическую активность органических соединений.

Личный вклад автора. Все результаты диссертации получены лично автором или в соавторстве при его непосредственном участии. Автору принадлежит выбор стратегии работы, постановка задач, математическое обоснование выбранного подхода, планирование расчетов и анализа, необходимых для решения поставленных задач, а также разработка необходимых для этого компьютерных программ.

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. Основные вклады соавторов указаны в соответствующих разделах диссертации и автореферата.

. Апробация работы. Основные результаты работы были представлены на 28 всесоюзных, российских и международных научных конференциях, в том числе, на межвузовской конференции “Молекулярные графы в химических исследованиях” в Калинине в 1990 г., на I-ой Всесоюзной конференции по теоретической органической химии в Волгограде в 1991 г., на 10-ом европейском симпозиуме “QSAR and Molecular Modelling” в Барселоне (Испания) в 1994 г., на II Российском национальном конгрессе "Человек и лекарство" в Москве в 1995 г., на втором международном симпозиуме по приобретению, представлению и обработке знаний «KARP-95» в Оборне (США, штат Алабама) в 1995 г., на 7-ом международном симпозиуме по наукам об окружающей среде «QSAR-96» в Эльсиноре (Дания) в 1996 г., на Международном симпозиуме по применению компьютеров в химических исследованиях «CACR-96» в Москве в 1996 г., на IV Российском национальном конгрессе «Человек и лекарство» в Москве в 1997 г., на 5-ом Европейском конгрессе по интеллектуальным и мягким вычислениям «EUFIT’97» в Аахене (Германия) в 1997 г., на XVI Менделеевском съезде по общей и прикладной химии в Санкт-Петербурге в 1998 г., на I Всероссийской конференции “Молекулярное моделирование” в Москве в 1998 г., на первом индо-американском симпозиуме по математической химии в приложении к молекулярному дизайну и оценке токсичности химикатов в Сантиникетане (Индия, западная Бенгалия) в 1998 г., на 12-ом европейском симпозиуме по количественным соотношениям структура-активность «Molecular Modelling and Prediction of Bioactivity» в Копенгагене (Дания) в 1998 г., на V Всероссийской конференции «Нейрокомпьютеры и их применение» в Москве в 1999 г., на международной школе-семинаре по компьютерной автоматизации и информатизации в науке и технике «ACS’2000» в Москве в 2000 г., на 9-ом международном симпозиуме по количественным соотношениям «структура-активность» в науках об окружающей среде «Crossroads to the XXI Century» в Бургасе (Болгария) в 2000 г., на VII Всероссийской конференции «Нейрокомпьютеры и их применение» в Москве в 2001 г., на II Всероссийской конференции «Молекулярное моделирование» в Москве в 2001 г., на 3-ей Всероссийской школе-конференции по квантовой и вычислительной химии им. В.А.Фока в

Москве в 2001 г., на международной конференции по фотохимии в Москве в 2001 г., на 14-ом Европейском симпозиуме по количественным соотношениям «структура-активность» «EuroQSAR-2002» в Борнемуте (Великобритания) в 2002 г., на 1-ой Российской школе-конференции «Молекулярное моделирование в химии, биологии и медицине» в Саратове в 2002 г., на II Российской школе-конференции «Молекулярное моделирование в химии, биологии и медицине» в Саратове в 2004 г., на XVI Европейском симпозиуме по количественным соотношениям «структура-активность» и молекулярному моделированию на Средиземном море в Италии в 2006 г., на 2-ой германской конференции по химической информатике в Госляре (Германия) в 2006 г., на 5-ой Всероссийской конференции «Молекулярное моделирование» в Москве в 2007 г., на XVIII Менделеевском съезде по общей и прикладной химии в Москве в 2007 г., в Страсбургской летней школе по хемоинформатике «CheminfoS3» в Оберне (Франция) в 2008 г., на 4-ой германской конференции по химической информатике в Госляре (Германия) в 2008 г.

Публикации. Содержание диссертации изложено в 54 публикациях, включая 2 главы в монографиях, 41 оригинальную статью в российских и международных журналах, в том числе 40 в журналах, рекомендованных ВАК, и 11 статей в сборниках.

Структура и объем работы. Диссертация изложена на 365 страницах машинописного текста, состоит из введения, 2 глав обзора литературы, 6 глав обсуждения результатов, выводов и списка цитированной литературы (517 ссылок), содержит 34 таблиц и 66 рисунков.

СОДЕРЖАНИЕ РАБОТЫ

Главным содержанием настоящей работы является создание универсальной методологии, позволяющей с единых позиций осуществлять количественный прогноз самых разнообразных свойств органических соединений на основе обработки экспериментальных данных. Математически обоснован и на множестве примеров продемонстрирован центральный тезис диссертационной работы: такой универсальной методологией является сочетание многослойных искусственных нейронных сетей (ИНС) персептронного типа и фрагментных дескрипторов (ФД).

Первая и вторая главы диссертационной работы являются литературным обзором, главы с третьей по восьмую – обсуждением результатов.

Глава 1. Искусственные нейронные сети

В данной главе рассматривается математический аппарат ИНС – современного метода машинного обучения, в основе работы которого лежит имитация функционирования клеток головного мозга человека. Основное преимущество ИНС перед классическими методами статистического анализа состоит в возможности аппроксимации по экспериментальным данным любых сколь угодно сложных нелинейных зависимостей произвольного и заранее неизвестного вида.

После краткого введения в **разделе 1.2** рассмотрены основные принципы нейросетевого моделирования. ИНС 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. В процессе работы нейросети осуществляется преобразование сигналов (кодирующих обрабатываемые данные) внутри нейронов и их передача между соседними нейронами.

Архитектура ИНС определяется топологией соединений нейронов между собой. Нейроны внутри сети, как правило, организованы в группы, называемые слоями. Нейроны, принимающие внешние данные для последующей обработки, называются входными; нейроны, выводящие уже обработанные данные, называются выходными. Остальные нейроны, участвующие в промежуточной обработке данных, называются скрытыми.

Подобно сетям биологических нейронов, ИНС способны обучаться на примерах путем подстройки весов связей между нейронами. В главе подробно рассматриваются методы обучения многослойных нейронных сетей – самой популярной архитектуры ИНС, имитирующей послойную организацию коры головного мозга человека. Все эти методы основаны на использовании алгоритма «обратного распространения (backpropagation) ошибки» для вычисления производных, вследствие чего такие ИНС часто называют нейросетями обратного распространения. Альтернативное название – многослойные персептроны. Важнейшее свойство ИНС этого типа заключается в способности обучаться апроксимации любых сколь угодно сложных нелинейных зависимостей между входными и выходными данными. Именно поэтому они и были выбраны в качестве основного инструмента обработки данных в рамках диссертационной работы.

В **разделе 1.3** рассматриваются основные принципы применения многослойных ИНС для прогнозирования свойств химических соединений. Прежде всего, для построения нейросетевой модели подготавливается база данных, содержащая структуры химических соединений и известные значения тех свойств, которые в дальнейшем предполагается при помощи обученной ИНС прогнозировать. Как правило, эта база разбивается на две части. По первой из них, называемой обучающей выборкой, путем многократного предъявления ее ИНС, производится обучение последней. По второй, называемой контрольной выборкой, производится контроль прогнозирующей способности ИНС. На следующем этапе для всех химических соединений из выборок производится расчет дескрипторов, т.е. чисел, описывающих структуру химических соединений. Далее 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. Число выходных нейронов равно числу одновременно прогнозируемых свойств, причем в качестве прогнозируемого значения каждого из свойств берется выходное значение соответствующего выходного нейрона. Скрытые же нейроны служат для промежуточных вычислений, и их

число подбирается, исходя из критерия максимизации прогнозирующей способности ИНС.

Обучающая выборка в процессе обучения ИНС ей многократно предъявляется. При каждом таком предъявлении значения дескрипторов каждого из соединений устанавливаются на входных нейронах. Далее ИНС запускается на счет, и с выходных нейронов снимаются прогнозируемые значения свойств, которые сравниваются с экспериментальными. На основании найденной разницы между экспериментальными и прогнозируемыми значениями, 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. Обученная таким образом ИНС может быть использована для прогнозирования свойств новых химических соединений. Для этого значения вычисленных для них дескрипторов устанавливаются на входные нейроны, ИНС запускается на счет, и с выходных нейронов снимаются спрогнозированные значения свойств этих соединений.

В разделе 1.4 перечислены основные ограничения ИНС и проблемы, связанные с их применением. Разработка эффективных методов решения этих проблем составила важную часть диссертационной работы (см. Главу 4).

Глава 2. Фрагментные дескрипторы в поиске зависимостей «структура-свойство»

Данная глава посвящена рассмотрению фрагментных дескрипторов (ФД), т.е. чисел, показывающих наличие данного фрагмента внутри химической структуры. К преимуществам ФД обычно относят следующие: 1) простота и эффективность вычисления; 2) простота интерпретации со структурно-химической точки зрения; 3) базисный характер, выражающийся в возможности аппроксимировать с их помощью любую зависимость «структура-свойство» (это было показано в рамках данной диссертационной работы, см. главу 3).

Глава начинается с изложения в **разделе 2.1** истории ФД, берущей начало с появления первых аддитивных схем в 30-40-ых годах прошлого века.

В **разделе 2.2** приведена подробная классификация ФД по следующим категориям: 1) типам молекулярных графов, соответствующих структурным фрагментам; 2) типам молекулярных структур; 3) типам значений дескрипторов; 4) типам дескрипторных наборов; 5) связности фрагментов; 6) уровням детализации молекулярных графов.

В **разделе 2.3** перечислены основные ограничения ФД и проблемы, связанные с их использованием. Разработка способов решения этих проблем составила важную часть диссертационной работы (см. главу 5).

Глава 3. Математическое обоснование выбранного подхода

В данной главе содержится математическое обоснование использования сочетания многослойных ИНС с ФД в качестве универсального подхода к прогнозированию свойств органических соединений на основе анализа эмпирических данных.

Раздел 3.1 посвящен рассмотрению значимости для химии поиска базиса инвариантов помеченных графов. В нем отмечается, что один из наиболее популярных подходов к решению проблемы поиска соотношений «структура-свойство» основан на представлении химической структуры в виде помеченного молекулярного графа. В этом случае молекулярные дескрипторы (т.е. числа, описывающие химические структуры) и функции, аппроксимирующие разнообразные свойства химических соединений, являются инвариантами графов, т.е. числовыми характеристиками, не зависящими от нумерации вершин графа. Следовательно, при известном базисе инвариантов помеченных графов задачу поиска соотношений «структура-свойство» можно решить путем разложения зависимости моделируемого свойства от структуры химического соединения по такому базису (таковой ранее известен не был).

Раздел 3.2 содержит две основные теоремы о базисе инвариантов помеченных графов, впервые сформулированные в ходе совместной работы с М.И.Скворцовой, которая предложила их строгое математическое доказательство.

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

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

где: $H_{V,E}^{(n)}$ - множество всех возможных помеченных графов с максимальным числом вершин n; c_j – некоторые константы, не зависящие от H и зависящие от f; $g_j(H)$ – 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 множества $H_{V,E}^{(n)}$. Суммирование ведется по подграфам H_j, получаемым из H путем удаления ребер всеми неэквивалентными способами.

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

Таким образом, теорема 1 строго определяет, что базисом инвариантов помеченных графов являются числа вложений различных подграфов $g_j(H)$. Единственным отличием $g_j(H)$ от вышеупомянутых ФД является то, что при их вычислении рассматриваются вложения всех подграфов – как связных, так и, главным образом, несвязных, тогда как ФД строятся, как правило, на основе связных подграфов. Несвязных подграфов, однако, чрезвычайно много по сравнению со связными и с ними очень неудобно работать. Теорема 2 как раз и позволяет не рассмат-

ривать несвязные подграфы и устанавливает полиномиальный характер связи между значением произвольного инварианта $f(H)$ и значениями ФД, построенных на основе связных подграфов. Таким образом, теорема 2 устанавливает тип дескрипторов, с помощью которых может быть аппроксимирован любой инвариант помеченного графа и, следовательно, любое скалярное свойство химических соединений. При этом, однако, остается нерешенной проблема о способах нахождения огромного числа коэффициентов, содержащихся в таком полиноме.

В разделе 3.3 рассматривается найденное нами эффективное решение этой проблемы путем применения теоремы Колмогорова о представлении непрерывных функций нескольких переменных в виде суперпозиций непрерывных функций одного переменного и сложения. С использованием нейросетевой интерпретации вышеупомянутой теоремы, данной Р. Хехт-Нильсеном (R.Hecht-Nielsen), а также математических результатов, полученных в работах Куркова (Kůrková), можно сделать вывод о возможности аппроксимации рассматриваемой в теореме 2 полиномиальной зависимости с помощью многослойной ИНС. Это легло в основу центрального положения диссертационной работы: любая сколь угодно сложная зависимость между структурой органического соединения и его свойством может быть аппроксимирована при помощи многослойной ИНС с двумя слоями скрытых нейронов и набора ФД. Отметим, что в большинстве случаев для аппроксимации зависимостей «структура-свойство» достаточно и одного слоя скрытых нейронов.

Глава 4. Разработка нейросетевых подходов

Данная глава содержит описание предложенных нами подходов к решению задач, связанных с применением ИНС для поиска количественных корреляций «структура-свойство».

Раздел 4.1 содержит описание разработанных нами способов решения проблем, связанных с явлением «переучивания» ИНС. **Подраздел 4.1.1** содержит анализ этого явления. Суть его заключается в следующем: процесс обучения нейросети может быть условно разделен на две последовательные фазы – «обобщения» и «запоминания». Для химических соединений, содержащихся в обучающей выборке, среднеквадратичная ошибка прогнозирования свойств постоянно уменьшается по ходу обучения в обеих фазах. В то же время, для соединений, отсутствующих в обучающей выборке, среднеквадратичная ошибка прогнозирования сначала уменьшается 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, содержащихся, например, в контрольных выборках. Эффект «переучивания» схематически показан на Рис. 1.

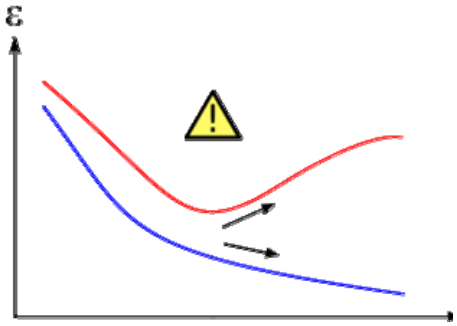


Рис. 1. Эффект "переучивания" нейросети. Нижняя кривая показывает ход изменения среднеквадратичной ошибки прогнозирования для соединений, входящих в обучающую выборку, а верхняя – в контрольную выборку. Восклицательным знаком отмечена точка перехода из фазы «обобщения» в фазу «запоминания»

В **подразделе 4.1.2** рассмотрены известные из литературы способы предотвращения «переучивания» и показано, что наиболее эффективным из них является остановка обучения при достижении наименьшей среднеквадратичной ошибки прогнозирования на контрольной выборке. Тем не менее, при его применении возникает новая проблема, суть которой состоит в следующем. Поскольку контрольная выборка используется для остановки обучения, т.е. для отбора модели, то содержащаяся в ней информация частично попадает в отобранную модель, и поэтому контроль по такой выборке уже не может считаться полностью независимым, а среднеквадратичная ошибка прогнозирования на ней – для объективной оценки прогнозирующей способности этой модели. В **подразделе 4.1.3** изложено предложенное нами эффективное решение этой проблемы.

Для решения вышеизложенной проблемы предлагается использовать трехвыборочный метод, согласно которому производится деление всего набора данных на 3 выборки: обучающую, внутреннюю контрольную и внешнюю контрольную. По обучающей выборке идет построение моделей, внутренняя контрольная выборка используется для отбора оптимальной для прогнозирования модели, а ошибка прогнозирования на внешней контрольной выборке, которая никаким образом не участвует ни в построении, ни в отборе модели, используется для оценки прогнозирующей способности этой модели. Разбивку набора данных на три выборки можно осуществлять либо случайным образом, либо систематически 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. Оборн, штат Алабама, США) и был положительно воспринят сообществом математиков, специализирующихся в области ИНС. Почти одновременно и независимо от нас сходные идеи были также опубликованы И.Тетко с соавторами. С тех пор трехвыборочный метод превратился в обязательный атрибут нейросетевых исследований в данной области. Трехвыборочный метод, в сочетании с идеями ансамблевого подхода к построению моделей «структура-свойство», лег в основу как более ранней методики, изложенной в подразделе 6.3.1 (т.н. трехвыборочного скользящего контроля), так и более поздней разработки – процедуры двойного скользящего контроля, описанной в **подразделе 4.1.4**.

В рамках предложенной нами процедуры двойного скользящего контроля исходная база данных систематически разбивается на 3 части: обучающую, внутреннюю контрольную и внешнюю контрольную выборки в соотношении (N-2):1:1. Внутренняя контрольная выборка используется для отбора моделей с наилучшей прогнозирующей способностью, а внешняя контрольная выборка – для оценки прогнозирующей способности отобранных моделей. Предсказанное значение свойства для каждого химического 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. 2 представлена диаграмма разбиения баз данных для $N = 5$.

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

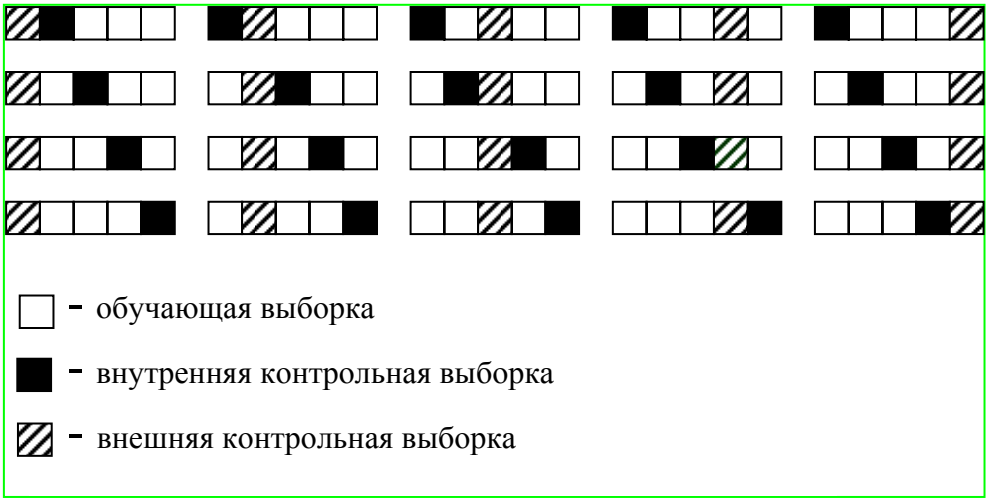


Рис. 2. Схема 5×4-кратного двойного скользящего контроля

Метод двойного скользящего контроля обеспечивает корректную 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.

В **подразделе 4.1.5** описан разработанный нами статистический метод построения линейно-регрессионных моделей, названный методом Быстрой Пошаго-

вой Множественной Линейной Регрессии (БПМЛР), который основан на трехвыборочном подходе, совместим с процедурой двойного скользящего контроля, и позволяет очень эффективно осуществлять предварительный отбор дескрипторов для ИНС. Благодаря его использованию решается проблема невозможности обработки при помощи ИНС выборок, включающих большое число дескрипторов.

В рамках метода БПМЛР внутренняя контрольная выборка используется для определения оптимального числа включаемых в модель дескрипторов. Работа метода основана на использовании текущего вектора ошибок (невязок), который в начале работы 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. 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 модели. Что касается регрессионного коэффициента при самом дескрипторе, 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 среднеквадратичной ошибки прогнозирования на внутренней контрольной выборке, тогда как среднеквадратичная ошибка прогнозирования на внешней контрольной выборке используется для оценки прогнозирующей способности итоговой 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 мощный метод статистического анализа, обладающий очень высокой производительностью и позволяющий даже на персональном компьютере эффективно работать с очень большим числом дескрипторов. Последнее свойство важно при работе с ФД ввиду их очень большого числа.

Раздел 4.2 содержит описание предложенного нами подхода к интерпретации нейросетевых регрессионных моделей. Необходимость его разработки была обусловлена тем, что раньше ИНС рассматривались как «черный ящик», способный осуществлять прогноз, но не предоставляющий никакой возможности описать нейросетевые модели на содержательном уровне. Ранее именно это и считалось основным недостатком ИНС, поскольку для обоснованного использования построенных моделей часто требуется понимание лежащих в их основе физико-химических и биологических явлений. И действительно, наборы весовых коэффи-

циентов не могут быть непосредственно использованы для интерпретации нейросетевых моделей, поскольку их числовые значения в значительной мере меняются при перестроении последних, а также сильно зависят от числа скрытых нейронов, и поэтому нельзя их непосредственно использовать для описания нейросетевых моделей «структура-свойство» на качественном уровне.

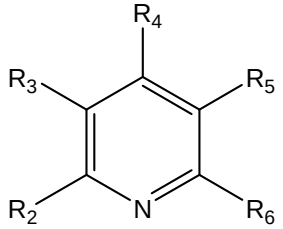
Для решения этой проблемы мы предлагаем использовать специальный набор статистических характеристик, значения которых, в отличие от значений весовых коэффициентов, почти не меняются при перестроении моделей, слабо зависят от числа скрытых нейронов и вполне могут быть использованы для интерпретации нейросетевых моделей. Более того, с их помощью можно анализировать даже такие характеристики соотношений «структура-свойство», которые обычно невозможно извлечь при помощи стандартных статистических подходов и которые могут быть важны для понимания природы соответствующих физико-химических и биологических процессов.

Основная идея предлагаемого подхода состоит в использовании для интерпретации нейросетевых моделей статистических характеристик, основанных на коэффициентах разложения в ряд по Тэйлору-Маклорену функции f, описывающей зависимость выходов ИНС от входов. Итак, предлагаются следующие характеристики: M_x – среднее значение первой частной производной по отношению к значению дескриптора x по выборке; D_x – дисперсия значений первой частной производной по выборке; M_{xx} – среднее значение второй частной производной по выборке; M_{xy} – среднее значение второй смешанной частной производной по отношению к значениям двум дескрипторов (x и y); I_x - сумма квадратов значений первой частной производной. Заметим, что значения M_x являются аналогами регрессионных коэффициентов в линейно-регрессионных моделях; аналогично D_x показывают степень нелинейности нейросетевых моделей, а M_{xx} и M_{xy} служат для анализа нелинейного характера моделей и взаимодействия в них дескрипторов.

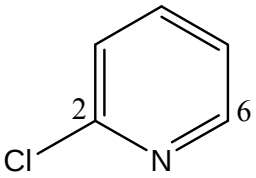
Нами продемонстрировано на нескольких примерах, что при использовании вышеперечисленных статистических характеристик стало возможным извлечь из набора данных не только информацию, которую предоставляют традиционные методы линейного регрессионного анализа (например, о знаке и величине влияния дескрипторов на свойства химических соединений), но и получить дополнительную ценную информацию о нелинейном характере зависимостей «структура-свойство» и взаимодействии дескрипторов.

В **разделе 4.3** рассматривается предложенная нами концепция обучаемой симметрии как пример использования ИНС для решения одной из задач, возникающих при построении корреляций «структура-свойство», которые в принципе не могут быть корректно решены при помощи линейных статистических методов. Как известно, классический подход к выявлению количественной зависимости «структура-свойство» («структура-активность») для узкого ряда соединений, обладающих одинаковым скелетом, предполагает использование в качестве дескрипторов констант заместителей. В этом случае может возникнуть проблема, ко-

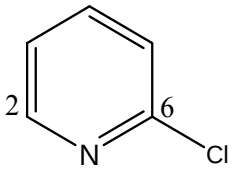
гда несколько положений заместителей топологически эквивалентны. Например, для пиридина (**1**) заместители R_2 и R_6, а также R_3 и R_5 находятся в топологически эквивалентных положениях. В этом случае корректно построенная модель «структура-свойство» должна обеспечить, например, одинаковое значение спрогнозированного свойства для 2-хлорпиридина (**2**) и 6-хлорпиридина (**3**), поскольку это одно и то же соединение.



1



2



3

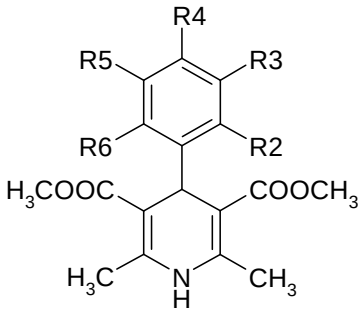
Возникает вопрос: как можно построить такую модель? Нами показано, что такие обычно применяемые для этой цели подходы, как предварительная канонизация структур и использование простейших аддитивных симметрических функций, не дают адекватного решения задачи. Более того, строго математически доказано, что общий вид необходимой для построения такой модели функции, инвариантной относительно перестановки некоторых своих аргументов, должен быть нелинейным относительно этих аргументов. Следовательно, обычно применяемые в «классическом QSAR» средства линейного статистического моделирования не могут в принципе привести к построению оптимальной модели с необходимыми свойствами симметрии. Поэтому в данном случае мы рекомендуем использовать процедуры анализа данных, обеспечивающие возможность построения нелинейных моделей произвольной сложности, например ИНС.

Для решения этой проблемы мы предлагаем концепцию обучаемой симметрии. Согласно этой концепции необходимо: а) расширить обучающую выборку соединений путем добавления копий соединений («клонов») с теми же значениями моделируемого свойства, но различающихся перестановкой топологически эквивалентных позиций присоединения заместителей (например, структура **2** должна быть дополнена структурой **3**); б) использовать ИНС для выявления количественной зависимости «структура-активность». В этом случае ИНС обучаются строить нелинейные зависимости «структура-активность» с необходимыми свойствами симметрии.

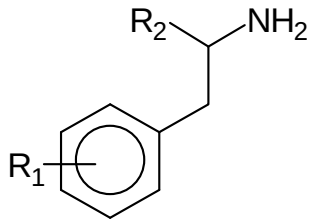
Эффект применения концепции обучаемой симметрии проиллюстрирован в данной диссертационной работе на двух примерах построения количественных моделей «структура – биологическая активность» для блокаторов кальциевых каналов L-типа (**4**) и для обладающих галлюциногенной активностью фенилалкиламинов (**5**). В обоих случаях в качестве дескрипторов использовались константы заместителей (как и в оригинальных работах, откуда выборки были взяты), а в качестве метода анализа данных – ИНС. Модели строились как на исходных базах, так и на базах, расширенных путем добавления «клонов», и при этом использовалась одна и та же разбивка на обучающую и контрольную выборки (второй кон-

трольной выборки не понадобилось из-за отсутствия «переучивания»). В Табл. 1 представлены значения среднеквадратичной ошибки прогнозирования на контрольных выборках для этих двух случаев.

Как видно из Табл. 1, применение концепции обучаемой симметрии в обоих случаях привело к значительному улучшению прогнозирующей способности нейросетевых моделей. Подчеркнем также, что построенные нами нейросетевые количественные модели «структура-активность» существенно лучше по своим статистическим характеристикам опубликованных ранее для этих же наборов данных.



4



5

Табл. 1. Сравнение прогнозирующей способности нейросетевых моделей, построенных без и с добавлением "клонов" в соответствии с концепцией обучаемой симметрии

Моделируемое свойство	Размер выборки	Среднеквадратичная ошибка прогнозирования на контрольной выборке (в логарифмических единицах)	
		без «клонов»	с «клонами»
Блокирующая способность дигидропиридинов **4**	46	1.59	0.71
Галлюциногенная активность фенилалкиламинов **5**	35	0.98	0.47

Глава 5. Разработка фрагментных подходов

Данная глава содержит набор разработанных нами концепций, методов, программ и алгоритмов, нацеленных на превращение фрагментного подхода в мощный инструмент максимально точного моделирования широкого разнообразия свойств органических соединений. В главе не только приводятся способы преодоления существовавших ранее ограничений ФД, но и предлагаются методики, направленные на значительное расширение сферы применения фрагментного подхода.

Раздел 5.1 посвящен описанию принципов построения разработанных нами ФД, а также методов и алгоритмов их генерации при помощи дескрипторного блока Fragment. Отмечается, что основными отличительными особенностями раз-

работанного нами варианта ФД 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: а) большого числа типов генерируемых фрагментов (см. Рис. 4) в сочетании с развитой четырехуровневой классификацией типов атомов (см. подраздел 5.1.2); б) механизма их автоматического обобщения; в) нескольких стратегий комбинирования разных уровней классификации атомов внутри фрагментов. 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 поддержкой динамически меняющегося списка групп статистически эквивалентных дескрипторов. Важными особенностями также является возможность работы с «выделенными» атомами (см. раздел 5.3), полимерными структурами (см. раздел 5.4) и стереохимической информацией. Пример кодировки фрагмента дан на Рис. 5.

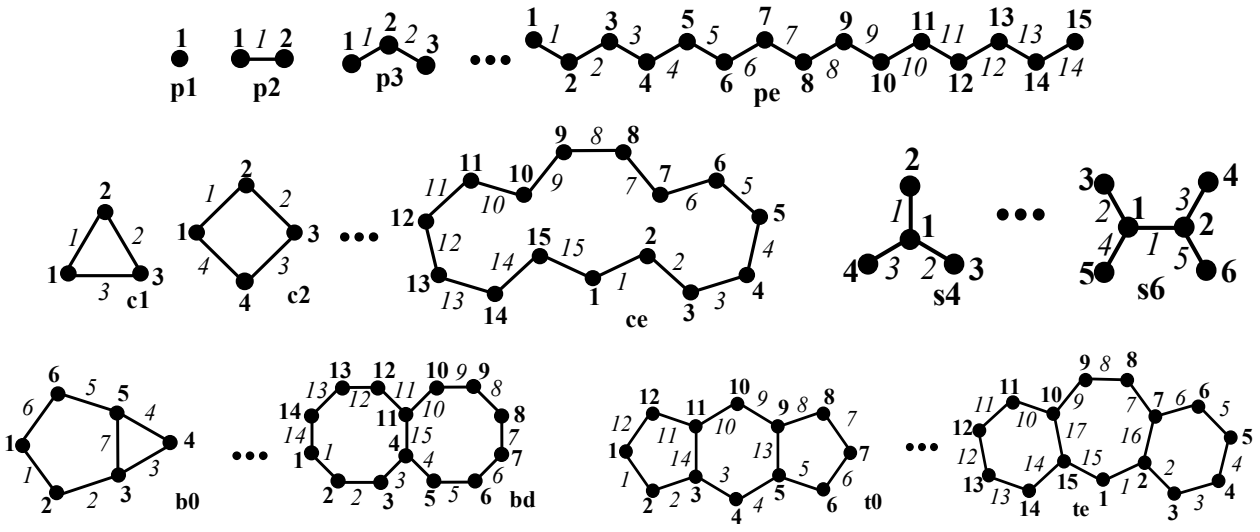


Рис. 3. Типы фрагментных дескрипторов. Коды **p1**…**pe** соответствуют линейным фрагментам, включающим, соответственно, от 1 до 15 атомов; коды **c3**…**cf** соответствуют циклическим фрагментам, включающим от 3 до 15 атомов; коды **s4**…**s6** соответствуют разветвленным фрагментам, включающим от 4 до 6 атомов; коды **b0**…**bd** – 14 типам бициклических фрагментов; коды **t0**…**te** – 15 типам трициклических фрагментов.

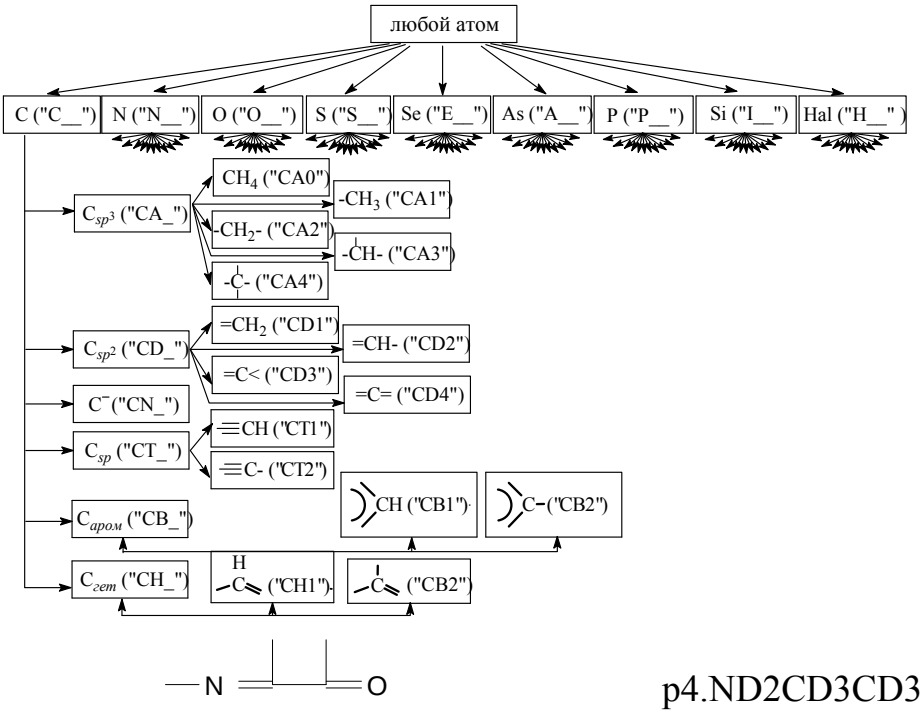


Рис. 4. Иерархическая система классификации атомов во фрагментах. Полностью показана ветка, соответствующая атомам углерода. Переход к более высокому уровню обобщения достигается путем замены в коде атома крайнего правого символа, отличного от символа подчеркивания, на символ подчеркивания.

p4.ND2CD3CD3OD_.212

Рис. 5. Пример кодировки фрагмента. Код фрагмента формируется из разделенных через запятую кода типа фрагмента, сцепленных кодов атомов и сцепленных кодов связей.

В **разделе 5.2** приведены примеры прогнозирования физико-химических свойств органических соединений с использованием ФД и статистического аппарата множественной линейной регрессии. Эффект от перехода к нейросетевому моделированию описан ниже в разделе 6.6. Далее в подразделах 5.2.1 (на примере прогнозирования поляризуемости химических соединений) и 5.2.2 (на примере прогнозирования энтальпии образования алифатических полинитросоединений) показано, что ФД при линейном моделировании являются удобным средством автоматического создания аддитивных схем расчета физико-химических свойств органических соединений. В подразделах от 5.2.3 до 5.2.7 приведены работы (сделанные в соавторстве с Н. И. Жоховой), в которых ФД, в сочетании с множественной линейной регрессией, были успешно использованы для прогнозирования таких видов физико-химических свойств, которые лишь с большим трудом поддаются расчету при помощи методов квантовой химии и молекулярного моделирования. Такими свойствами являются: а) магнитная восприимчивость; б) энтальпия парообразования; в) энтальпия сублимации; г) температура вспышки; д) сродство азо- и антрахиноновых красителей к целлюлозному волокну. В Табл. 2 приведены статистические характеристики построенных моделей с наиболее высокой прогнозирующей способностью. Отметим, что во всех случаях построенные модели превзошли по своим статистическим показателям модели, ранее опубликованные в литературе и построенные на тех же данных.

Табл. 2. Статистические характеристики моделей, основанных на сочетании ФД с аппаратом множественной линейой регрессии

Свойство	Обучающая выборка		Контрольная выборка
	R^2	s	*MAE* или *RMSE**
Поляризуемость, A^3	0.997	0.38	-
Энтальпия образования алифатических полинитросоединений, ккал/моль	0.985	2.65	-
Магнитная восприимчивость $\times 10^{-6}$ единиц	0.985	4.99	7.02
Энтальпия парообразования, ккал/моль	0.993	1.79	1.57
Энтальпия сублимации, ккал/моль	0.845	2.97	2.16
Температура вспышки, ^oC	0.956	11.4	11.8
Сродство азо- и антрахиноновых красителей к целлюлозному волокну, кДж·моль$^{-1}$	0.954	0.76	0.83*

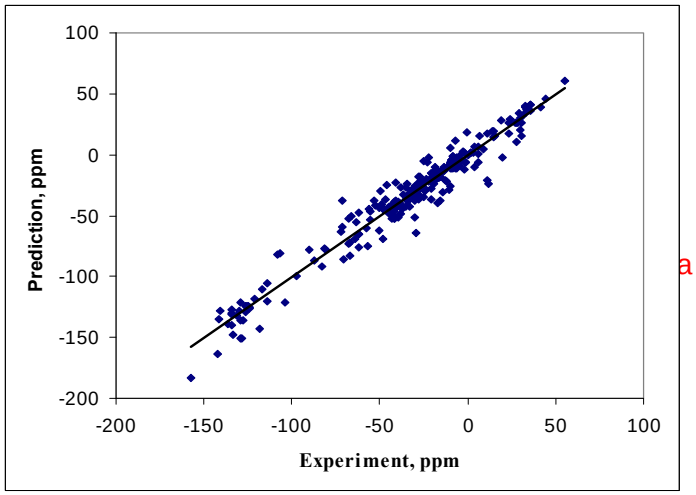
В **разделе 5.3** рассматривается подход, который позволяет значительно расширить круг свойств, для прогнозирования которых можно применять ФД за счет указания специальных «выделенных» атомов, играющих специфическую роль в природе моделируемого свойства. Например, при моделировании константы основности 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 соответствии с тем, по каким причинам этот атом выделен; 2) при генерации ФД каждая такая метка рассматривается как отдельный псевдоатом с именем, соответствующем символу метки; 3) при построении уравнений «структура-свойство» предусмотрена возможность включать в модели только те дескрипторы, которые содержат такой псевдоатом.

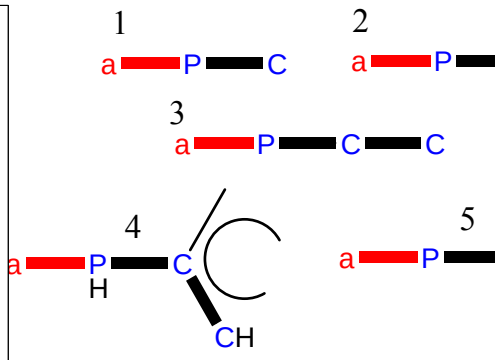
Мы предлагаем использовать ФД с “выделенными” атомами для моделирования широкого круга свойств: 1) при расчете локальных характеристик молекул, таких, например, как химические сдвиги в спектрах ЯМР, либо кислотно-основные свойства определенных атомов в молекулах; 2) при прогнозировании биологической 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 ФД. Таким образом, использование ФД с «выделенными» атомами позволяет значительно расширить сферу применения фрагментного подхода в поиске количественных соотношений «структура-свойство».

Далее на нескольких примерах рассмотрено применение ФД с «выделенными» атомами. Во всех случаях генерация дескрипторов проводилась при помощи блока Fragment. Предварительный отбор дескрипторов осуществлялся с помощью метода БПМЛР, а построение окончательной модели – при помощи трехслойной ИНС. Оценка прогнозирующей способности проводилась с помощью процедуры двойного скользящего контроля.

В подразделе 5.3.1 рассмотрено применение ФД с «выделенными» атомами для моделирования химических сдвигов в ^{31}P ЯМР спектрах производных монофосфинов. Диаграмма разброса, список наиболее важных фрагментов и статистические характеристики построенной модели приведены на Рис. 6. Этот пример иллюстрирует возможность использования дескрипторов данного типа для прогнозирования локальных свойств химических соединений, которые можно приписать определенным атомам или группам атомов внутри молекулы. В этом случае использование цепочечных фрагментов с терминальными «выделенными» атомами позволяет получать легко интерпретируемые модели, наглядно показывающие пути влияния отдельных атомов или групп внутри молекулы на изучаемое свойство. Например, первые три фрагмента на Рис. 6 отражают σ-индукционное влияние алкильных заместителей на атом фосфора, четвертый – эффект сопряжения с ароматическим ядром, пятый – влияние расположенного в орто-положении атома фтора.



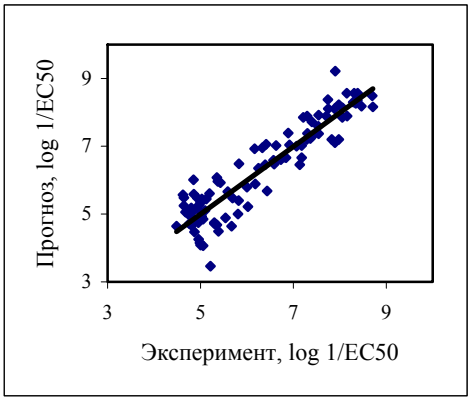


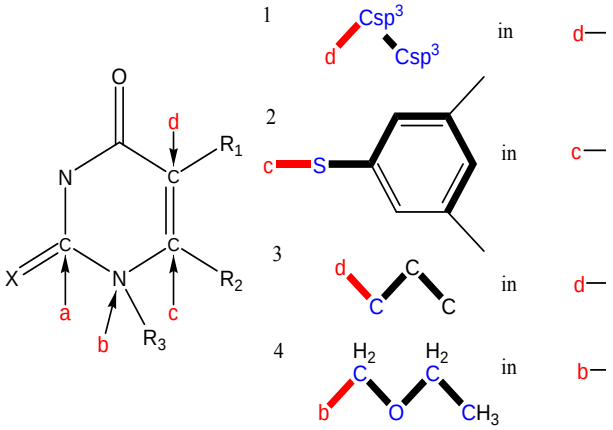
Q^2_{DCV} = 0.8298, $RMSE_{DCV}$ = 5.7 ppm, MAE_{DCV} = 6.1 ppm

Рис. 6. Диаграмма разброса, список наиболее важных фрагментов и статистические характеристики нейросетевой модели для прогнозирования химических сдвигов в ^{31}P ЯМР спектрах производных монофосфинов.

В подразделе 5.3.2 рассмотрено применение ФД с «выделенными» атомами для моделирования способности аналогов 1-[(2-гидроксиэтокси)-метил]–6(фенилтио)тимина (HEPT) ингибировать обратную транскриптазу вируса ВИЧ-1. Соответствующие диаграмма разброса, список наиболее важных фрагментов и статистические характеристики построенной модели приведены на Рис. 7. Данный пример иллюстрирует возможность применения ФД с «выделенными» атомами для количественного прогнозирования биологической активности органических соединений внутри рядов соединений с одинаковым общим фрагментом (скелетом). Следует отметить, что обычно ФД 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 невозможности их интерпретации с целью выявления факторов, влияющих на биологическую активность.

Предлагаемые ФД с «выделенными» атомами полностью решают эту проблему, поскольку позволяют позиционировать все рассматриваемые фрагменты относительно заранее заданных внутри химической структуры «реперных точек». На изображенной (Рис. 7) 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 ФД (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 интерпретируемыми со структурно-химической точки зрения, поскольку значения регрессионных коэффициентов в линейных моделях и введенной нами характеристики M_x для нейросетевых моделей четко показывают, какая группировка атомов в каком положении вносит тот или иной вклад в биологическую активность, и, следовательно, какие изменения нужно внести для ее оптимизации. Более того, рассмотрение характеристик M_{xy} позволяет выявить синергию и диссинергию во влиянии различных группировок атомов на биологическую активность.





$$Q^2_{DCV} = 0.856,\ RMSE_{DCV} = 0.52 \text{ и } MAE_{DCV} = 0.41$$

Рис. 7. Диаграмма разброса, список наиболее важных фрагментов и статистические характеристики нейросетевой модели для прогнозирования способности аналогов НЕРТ ингибировать обратную транскриптазу вируса ВИЧ-1

В подразделе 5.3.3 рассмотрено применение ФД с «выделенными» атомами для прогнозирования констант скорости гидролиза эфиров карбоновых кислот. В данном случае в качестве «выделенных» атомов взяты реакционные центры, включающие атомы углерода, входящие в образующиеся в входе реакции карбоксильную и гидроксильную группы. Кроме ФД с «выделенными» атомами, в соответствии с развиваемой нами методологией построения моделей «структура-условия-свойство» (см. раздел 7.2), мы также использовали дескрипторы, описывающие условия реакции: состав растворителя и температуру. В результате была получена нейросетевая модель со следующими статистическими характеристиками, определенными при помощи процедуры двойного скользящего контроля: $Q^2_{DCV} = 0.9162$, $RMSE_{DCV} = 0.31$ и $MAE_{DCV} = 0.19$. Три наиболее важных фрагмента из вошедших в построенную модель изображены на Рис. 8. Первый из них 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.

Таким образом, данный пример иллюстрирует возможность применения ФД 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.

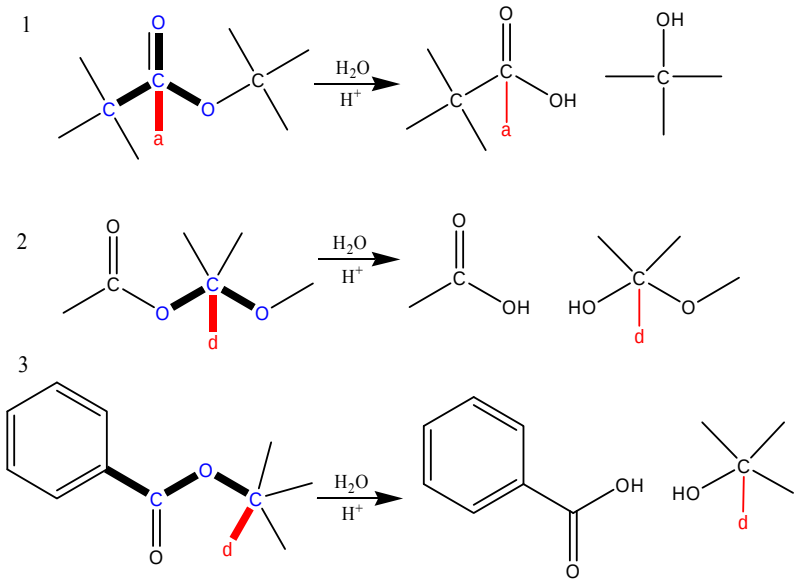


Рис. 8. Наиболее важные фрагменты для прогнозирования констант скоростей гидролиза сложных эфиров

Раздел 5.4 посвящен предложенной нами концепции псевдофрагментных дескрипторов (ПФД) как одного из возможных подходов к решению проблемы «отсутствующих» (или «редких») фрагментов, которые могут отсутствовать (либо быть недостаточно представленными) в обучающей выборке, но присутствовать в соединениях, для которых осуществляется прогноз. Поскольку величины вкладов таких фрагментов не могут быть определены по обучающей выборке, то можно ожидать значительных ошибок прогнозирования для соединений, их содержащих. 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. Для этой цели мы предлагаем использовать особую категорию ФД, значения которых вычисляются путем комбинирования свойств атомов, присутствующих в этих фрагментах. Дескрипторы такого рода мы будем называть псевдофрагментными дескрипторами (ПФД), чтобы их отличать от «настоящих» ФД, 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 ПФД начинает входить в статистические модели вместо многочисленных «настоящих» ФД, в том числе и потенциально редких, выступая тем самым в качестве сжатого обобщения последних. Это в значительной степени и решает проблему редких фрагментов,

если ПФД строятся на основе присутствующих практически во всех молекулах отдельных атомов или небольших цепочек атомов.

В качестве примера простейшего ПФД рассмотрим конструкцию $\frac{1}{N_a}\sum_{i=1}^{N_a} R_i^3$, где: R_i – ковалентный радиус атома, N_a – число атомов в молекуле. 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 удельный объем атома. Можно предположить, что он будет играть существенную роль при прогнозировании волюметрических свойств веществ, например, плотности. При включении такого дескриптора в модель, даже если будет требоваться осуществить прогноз подобного свойства для химического соединения, содержащего редкий элемент (отсутствующий в обучающей выборке), все равно будет дана разумная аппроксимация его вклада в прогнозируемое свойство.

В соответствии с вышеизложенными принципами нами было сконструировано 50 ПФД на основе как отдельных атомов, так и коротких цепочек, включающих до 5 атомов. Для их вычисления нами разработан дескрипторный блок FRAGPROP (в составе созданного нами программного комплекса NASAWIN). Опыт работы с этим блоком показал, что добавление ПФД к «настоящим» ФД практически всегда повышают прогнозирующую способность моделей, предназначенных для прогнозирования физико-химических свойств органических соединений. Приведем в качестве примера прогнозирование трех ключевых физических свойств полимеров на основе структур мономеров при помощи статистических моделей, построенных методом БПМЛР. В Табл. 3 приведено сравнение статистических характеристик для построенных с использованием ФД моделей как с добавлением, так и без добавления ПФД.

Как видно из таблицы, ПФД позволяют в значительной степени улучшать качество моделей, построенных на основе ФД, за счет решения проблемы редких фрагментов. Следует отметить, что хотя ПФД можно применять и без ФД для построения моделей «структура-свойство», наилучшие модели всегда получаются только в сочетании с «настоящими» ФД. Поэтому их применение следует рассматривать как способ улучшения моделей, построенных на базе ФД.

Табл. 3. Статистические характеристики моделей, полученных для прогнозирования физических свойств полимеров с использованием как только ФД, так и с добавлением ПФД

Свойство	Только ФД			ФД с добавлением ПФД		
	Q^2_{DCV}	$RMSE_{DCV}$	MAE_{DCV}	Q^2_{DCV}	$RMSE_{DCV}$	MAE_{DCV}
n	0.782	0.033	0.021	0.872	0.026	0.015
T_g	0.849	45.0	32.0	0.864	42.7	28.0
ρ	0.474	0.159	0.096	0.910	0.066	0.043

где: n – показатель преломления при 298K; T_g – температура стеклования (в градусах Кельвина); ρ – плотность в аморфном состоянии (г/см^3, 298K).

Глава 6. Сочетание ИНС и ФД

Данная глава посвящена изучению эффекта от совместного использования ИНС и ФД. На большом числе примеров проводится сравнение с линейными моделями и делается вывод о преимуществах этого сочетания.

Раздел 6.1 посвящен изложению результатов нашей первой работы по нейросетевому моделированию, опубликованной еще в 1993 г., в которой математические аппараты ИНС и пошаговой множественной линейной регрессии в сочетании с ФД и топологическими индексами (ТИ) были систематически применены для построения моделей, позволяющих прогнозировать разнообразные свойства углеводородов (главным образом, алканов). Для возможности сравнений при построении моделей одна и та же база была одинаковым образом разбита на обучающую и контрольную выборки. Результаты вычислительных экспериментов приведены в Табл. 4. В экспериментах 1-6 прогнозировалось по одному свойству (один выходной нейрон в ИНС), тогда как в моделях 7 и 8 одновременно прогнозировалось шесть различных свойств (шесть выходных нейронов) с помощью единой нейросетевой модели. Все линейно-регрессионные модели строились отдельно для каждого свойства.

Из анализа данных в Табл. 4 можно сделать следующие выводы.

1) Для углеводородов температура кипения, плавления, октановое число, критическая температура и поверхностное натяжение прогнозируются существенно лучше при использовании ИНС по сравнению с линейным регрессионном анализом. Это свидетельствует о нелинейном характере зависимости перечисленных выше свойств от рассматриваемых дескрипторов.

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

3) В большинстве случаев использование ФД приводит к построению моделей с лучшей прогнозирующей способностью по сравнению с топологическими индексами.

4) Сочетание ИНС с ФД чаще всего приводит к построению моделей с наилучшей прогнозирующей способностью.

Именно этот последний вывод и послужил отправным толчком для проведения большой серии разноплановых исследований, которые и легли в основу данной диссертационной работы.

Итак, оценивая рассмотренную в данном разделе работу, можно сказать, что она во многих отношениях явилась пионерной:

1) Она явилась первой работой, в которой аппарат ИНС был применен для прогнозирования физико-химических свойств органических соединений.

2) В ней впервые применено сочетание аппарата ИНС и ФД для прогнозирования свойств органических соединений.

3) В ней впервые было успешно применено многозадачное обучение, позволяющее одновременно осуществлять прогноз нескольких свойств в рамках одной модели.

Табл. 4. Результаты нейросетевого и линейно-регрессионного моделирования физико-химических свойств углеводородов

№	Выборка			Дескрипторы	ИНС			Множественная линейная регрессия		
	Свойство	N_t	N_v		s_t	R	s_v	s_t	R	s_v
1	$bp(a)$	159	18	ТИ	4.08	0.999	2.33	9.44	0.996	10.9
2	$bp(a)$	159	16	ФД	4.74	0.999	2.18	23.0	0.979	22.5
3	$mp(a)$	81	9	ТИ	16.2	0.976	13.8	29.4	0.924	28.5
4	$mp(a)$	81	9	ФД	16.0	0.977	16.8	32.9	0.902	31.8
5	$on(hc)$	138	15	ТИ	10.9	0.841	12.1	13.2	0.761	17.0
6	$on(hc)$	138	15	ФД	5.97	0.954	4.37	10.6	0.858	10.4
7	$V_m(a)$	63	6	ТИ	0.84	0.999	0.89	0.45	1.000	0.64
	$MR(a)$	63	6	ТИ	0.15	1.000	0.18	0.04	1.000	0.09
	$H_e(a)$	63	6	ТИ	0.44	0.994	0.51	0.27	0.999	0.21
	$T_c(a)$	63	6	ТИ	3.80	0.994	3.94	5.25	0.996	2.82
	$P_c(a)$	63	6	ТИ	0.46	0.984	0.39	0.68	0.988	0.39
	$\sigma(a)$	63	6	ТИ	0.18	0.996	0.28	0.28	0.990	0.29
8	$V_m(a)$	63	6	ФД	0.88	0.999	1.10	0.62	1.000	0.42
	$MR(a)$	63	6	ФД	0.20	0.999	0.18	0.04	1.000	0.09
	$H_e(a)$	63	6	ФД	0.44	0.996	0.56	0.18	1.000	0.07
	$T_c(a)$	63	6	ФД	3.37	0.995	3.58	7.52	0.993	4.96
	$P_c(a)$	63	6	ФД	0.44	0.986	0.23	0.79	0.986	0.40
	$\sigma(a)$	63	6	ФД	0.17	0.996	0.17	0.31	0.989	0.23

где для алканов: $bp(a)$ – температура кипения, 1 атм., ^{o}C; $mp(a)$ – температура плавления, ^{o}C; $V_m(a)$ – молярный объем, см^3/моль; $R(a)$ – молярная рефракция, см^3/моль; $H_e(a)$ – теплота испарения, кДж/моль; $T_c(a)$ – критическая температура, ^{o}C; $P_c(a)$ – критическое давление, атм.; $\sigma(a)$ – поверхностное натяжение, дин/см; $on(hc)$ – октановое число углеводородов (алканов, алкенов, циклоалканов); N_t – число соединений в обучающей выборке; N_v – число соединений в контрольной выборке; R – множественный коэффициент корреляции (квадратный корень от коэффициента детерминации); s_t – среднеквадратичная ошибка на обучающей выборке; s_v – среднеквадратичная ошибка на контрольной выборке.

В **разделе 6.2** сравнивается прогнозирующая способность нейросетевых и некоторых из рассмотренных выше линейно-регрессионных моделей (см. Табл.2

на стр. 19), построенных, в отличие от моделей из предыдущего раздела, на выборках существенно большего размера. Эти выборки содержат разнородные органические соединения, принадлежащие разным классам. Результаты сравнения прогнозирующей способности на одних и тех же контрольных выборках представлены в Табл. 5.

Табл. 5. Точность прогноза для линейно-регрессионных и нейросетевых моделей

Свойство	MAE_p или RMSEp* для линейно-регрессионной модели	MAEp или $RMSE_p$* для нейросетевой модели
Магнитная восприимчивость. $\times 10^{-6}$ единиц	7.02	6.25
Энтальпия парообразования, ккал/моль	1.57	1.77
Энтальпия сублимации, ккал/моль	2.16	1.66
Температура вспышки, ^{o}C	15.8*	14.6*

Как видно из Табл. 5, для трех из четырех свойств (т.е. для магнитной восприимчивости, энтальпии сублимации и температуры вспышки) применение ИНС 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. Это вполне согласуется с рассмотренными выше результатами, полученными для углеводородов. Таким образом, в большинстве случаев применение ИНС вместо аппарата множественной линейной регрессии приводит к улучшению прогнозирующей способности количественных моделей «структура-свойство».

Раздел 6.3 посвящен применению сочетания ИНС с ФД для моделирования ряда ключевых и технологически-важных физических свойств органических соединений, как то: температуры кипения, вязкости, плотности и давления насыщенных паров. Для этих свойств модели строились только по разнородным выборкам, содержащим представителей разных классов органических соединений. Исследование проводилось в рамках процедуры трехвыборочного скользящего контроля, которая явилась дальнейшим развитием трехвыборочного подхода и предшественницей процедуры двойного скользящего контроля. Основная идея метода – использование процедуры скользящего контроля и ансамбля нейросетевых моделей вместо единичной модели. Это позволяет сделать прогноз и оценку его качества более обоснованным и не зависящим от конкретной разбивки базы на три выборки - обучающую, внутреннюю и внешнюю контрольные. Статистические показатели построенных моделей представлены в Табл. 6.

Как видно из Табл. 6, нейросетевые модели обладают лучшими статистическими показателями по сравнению с линейно-регрессионными моделями, причем для температуры кипения, плотности и вязкости это различие существенно. Здесь также следует отметить, что полученные нейросетевые модели по этим показателям превосходят все опубликованные ранее в литературе. В данном разделе

также исследуется эффект использования ансамблей нейросетевых моделей, результатом прогноза которых является значение, получаемое путем усреднения прогнозов, выдаваемых индивидуальными моделями. В Табл. 6 также проведено сравнение двух наборов статистических показателей, первый из которых является результатом усреднения соответствующих показателей индивидуальных нейросетевых моделей, а второй описывает прогнозирующую способность их ансамбля. Приведенные данные позволяют сделать вывод о существенных преимуществах использования ансамблей нейросетевых моделей по сравнению с индивидуальными моделями. Можно предположить, что в данном случае два основных фактора вносят вклад в это явление. Во-первых, усреднение по моделям, получаемым при разных разбивках базы данных, позволяет эффективно использовать для обучения информацию из внутренних контрольных выборок, что эквивалентно увеличению эффективного размера обучающих выборок. Во-вторых, наблюдается известное явление подавления «шума» при усреднении.

Табл. 6. Статистические показатели моделей для прогнозирования физических свойств органических соединений

прогнозируемое свойство / статистические показатели		T кип, ^{o} C	lg (η), Па·с	d, г/см^3	lg (VP), Па
Количество соединений		510	367	803	349
Ансамбль ИНС	R	0.9911	0.9904	0.9943	0.9979
	$RMSE_t$	9.1	0.078	0.018	0.095
	$RMSE_v$	16.1	0.177	0.036	0.140
	$RMSE_p$	16.9	0.208	0.043	0.158
Индивидуальные ИНС	R	0.9869	0.9815	0.9911	0.9969
	$RMSE_t$	11.0	0.105	0.034	0.118
	$RMSE_v$	16.1	0.189	0.052	0.143
	$RMSE_p$	17.2	0.219	0.061	0.161
Линейно-регрессионные модели	R	0.9814	0.9794	0.9897	0.9902
	$RMSE_t$	12.9	0.111	0.036	0.198
	$RMSE_v$	16.7	0.195	0.055	0.248
	$RMSE_p$	18.6	0.212	0.067	0.258

где: T кип – температура кипения; η - вязкость; d – плотность; VP – давление насыщенных паров; R – коэффициент корреляции между спрогнозированными и экспериментальными значениями; $RMSE_t$ – среднеквадратичная ошибка на обучающих выборках; $RMSE_v$ – среднеквадратичная ошибка на внутренних контрольных выборках; $RMSE_p$ - среднеквадратичная ошибка на внешних контрольных выборках.

На Рис. 9 представлены диаграммы разброса, полученные для внешних контрольных выборок.

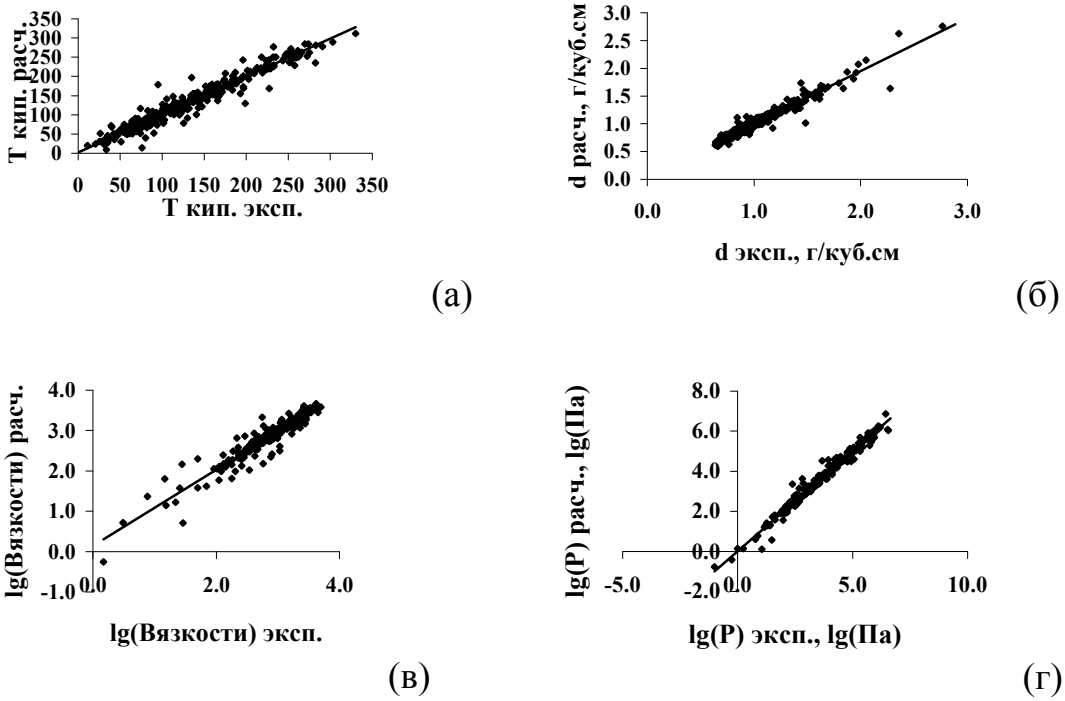


Рис. 9. Диаграммы разброса, полученные для внешних контрольных выборок при прогнозировании: (а) температуры кипения; (б) плотности; (в) вязкости; (г) давления насыщенных паров

Раздел 6.4 посвящен применению ИНС в сочетании с ФД и ПФД для прогнозирования температуры плавления ионных жидкостей, общие структуры которых приведены на Рис. 10. Были построены модели для четырех выборок, включающих: а) 126 бромидов производных пиридинов (PYR, **6** и **7**); б) 384 бромида производных имидазолов и бензимидазолов (IMZ, **8** и **9**); в) 207 бромидов четвертичных аммониев (QUAT, **10**); г) 717 соединений, входящих во все вышеупомянутые наборы (FULL). В Табл. 7 представлены средние абсолютные ошибки прогноза полученных моделей, оцененные при помощи процедуры скользящего контроля с использованием внешних контрольных выборок. В этой же таблице приведены аналогичные показатели, полученные при применении двух линейных методов – БПМЛР и метода частичных наименьших квадратов (PLS). Как видно из таблицы, в большинстве случаев ИНС приводит к построению лучших моделей по сравнению с БПМЛР и PLS.

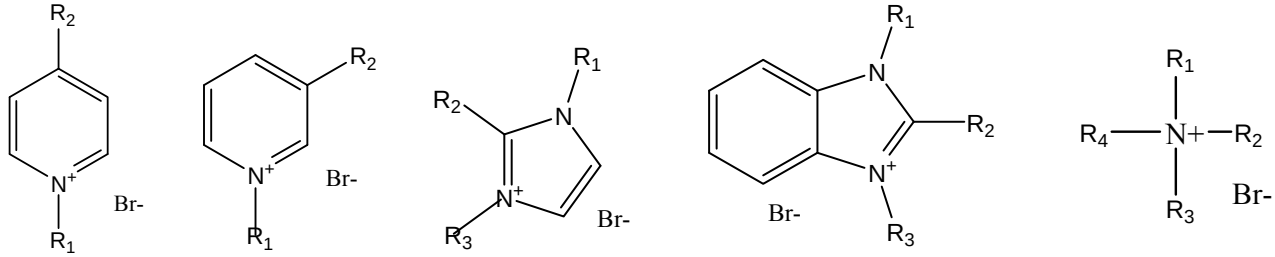


Рис. 10. Структуры ионных жидкостей

Табл. 7. Значения средней абсолютной ошибки прогнозирования температуры плавления ионных жидкостей (в градусах Кельвина)

	PYR	IMZ	QUAT	FULL
ИНС	26.2	32.4	30.3	31.5
БПМЛР	34.8	36.2	36.1	33.7
PLS	32.5	31.9	31.8	31.9

Для того, чтобы провести объективное сравнение развиваемого нами подхода с широким набором существующих в настоящее время методов поиска количественных соотношений «структура-свойство», мы приняли участие в совместном исследовании, проведенном несколькими группами авторов, в ходе которого широкий набор современных методов машинного обучения (ассоциативные нейронные сети ASNN, машины опорных векторов SVM, метод ближайших соседей kNN, метод частичных наименьших квадратов PLS, нейронные сети обратного распространения и множественная линейная регрессия), реализованные в нескольких программных комплексах (VCCLAB, ISIDA и NASAWIN), в сочетании с разнообразными типами дескрипторов (несколько типов ФД, ПФД, дескрипторы на основе электронно-топологических состояний атомов, а также все виды дескрипторов, генерируемых программой DRAGON) были применены для моделирования температуры плавления ионных жидкостей с использованием вышеупомянутых данных. Было проведено сравнение всех построенных моделей и показано, что модели, построенные при помощи программного комплекса NASAWIN на основе ИНС/ФД, заняли первые два места наряду с ASNN/E-counts. Если учесть, что ASNN построена на основе ИНС, а дескрипторы E-counts являются фрагментными, то можно сделать вывод, что именно комбинация ИНС с ФД приводит к построению наилучших моделей для прогнозирования температуры плавления ионных жидкостей.

Глава 7. Разработка интегрированных подходов

В данной главе излагаются предложенные нами подходы, которые включают разного рода интеграцию ИНС: а) с методами молекулярного моделирования; б) с комбинацией дескрипторных описаний химических соединений и внешних условий, а также: в) между собой. Все это ведет к значительному расширению круга свойств химических соединений, поддающихся надежному прогнозированию при помощи разрабатываемых нами методов.

Раздел 7.1 посвящен совместному применению ИНС и методов молекулярного моделирования, включающих молекулярно-механические и квантово-химические расчеты. В нем отмечается, что, несмотря на большие успехи в области молекулярного моделирования, ни одна даже самая совершенная молекулярная модель не способна охватить всего комплекса взаимодействий, в которые вовле-

чена реальная молекулярная система, равно как и учесть эти взаимодействия с достаточно высокой точностью. Это служит серьезным препятствием к практическому применению построенных теоретических моделей. В связи с этим особую актуальность приобретает проблема соотнесения теоретически рассчитываемых характеристик молекулярных систем с проявляемыми 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.

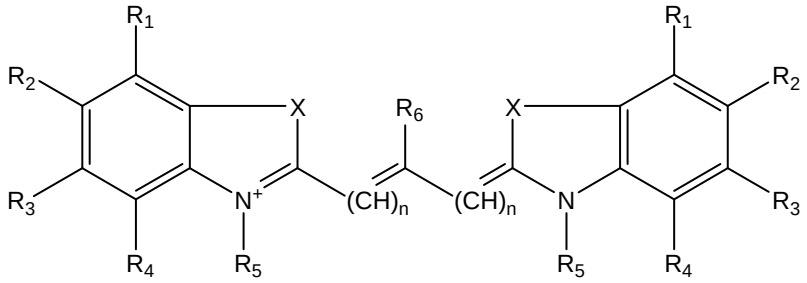
Генеральным направлением в решении указанной проблемы нам видится использование математического аппарата обработки данных, позволяющего выявлять любые сколь угодно сложные зависимости неизвестного вида между теоретически рассчитываемыми молекулярными характеристиками и экспериментальными данными. Именно это является как раз той самой задачей, для решения которой особенно хорошо подходят ИНС (в особенности в сочетании с ФД)! Преимущество применения ИНС заключается в их уникальной способности извлекать из эксперимента и обобщать зависимости, которые крайне трудно вывести из теоретических соображений. Поэтому аппарат ИНС является необходимым дополнением к методам молекулярного моделирования, способным резко повысить их прогнозирующую способность.

Возникает вопрос: если ИНС в сочетании с ФД могут аппроксимировать любое свойство, то зачем понадобилось их комбинировать с методами молекулярного моделирования? Все зависит от объема имеющихся экспериментальных данных (см. Табл. 8). Если данных достаточно много, то этого сочетания действительно достаточно для моделирования любого свойства. Если данных очень мало либо они вообще отсутствуют, то нейросети не могут быть обучены, 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 ФД, 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 нейросетевых моделей.

Табл. 8. Выбор метода моделирования в зависимости от объема данных

Объем экспериментальных данных	**Предпочтительный метод моделирования**
Мало либо отсутствуют	Молекулярное моделирование
Промежуточный объем данных	Сочетание молекулярного и нейросетевого моделирования
Достаточно много	Нейросетевое моделирование

В подразделе 7.1.1 рассматривается применение ИНС для прогнозирования положения длинноволновой полосы поглощения симметричных цианиновых красителей **11**, растворенных в этаноле (работа сделана в соавторстве с А.О. Айтом). В качестве дескрипторов брались энергии граничных молекулярных орбиталей, рассчитанные при помощи квантово-химического метода PM3, а также набор ФД, задающих тип гетероциклов. База данных была случайным образом разбита на обучающую и контрольную выборки. В Табл. 9 представлены статистические характеристики нейросетевых моделей, полученных как при наличии произвольного заместителя R_6, так и при R_6=H. Следует отметить, что достигнутая точность прогнозирования положения полосы поглощения значительно превосходит точность, с которой это свойство может быть предсказано с помощью прецизионных квантово-химических расчетов.



11

Табл. 9. Результаты нейросетевого моделирования положения длинноволновой полосы поглощения симметричных цианиновых красителей **11** в спиртовом растворе

Заместитель R_6	N	R	$RMSE_t$ (в нм)	$RMSE_v$ (в нм)
произвольный	398	0.9928	10.6	7.0
H	174	0.9976	4.4	3.4

где: N – общее число соединений; R – коэффициент корреляции; $RMSE_t$ – среднеквадратичная ошибка на обучающей выборке; $RMSE_v$ – среднеквадратичная ошибка на контрольной выборке

В подразделе 7.1.2 рассматривается применение ИНС для прогнозирования констант ионизации для нескольких классов органических соединений. В работе были использованы данные для 174 фенолов, 238 карбоновых кислот и 268 азотсодержащих соединений. Прежде всего, при помощи полуэмпирического квантово-химического метода PM3 нами были рассчитаны значения набора дескрипторов, описывающих электронные свойства молекул, такие, как: 1) энергии граничных орбиталей; 2) заряд на меченом атоме; 3) максимальный отрицательный заряд на атоме; 4) максимальный заряд на атоме водорода; 5) дипольный момент; 6) электрофильная, нуклеофильная и радикальная суперделокализация; 7) атомная самополяризуемость. Кроме того, нами были еще использованы ФД с «выделенными» атомами. Предварительный отбор дескрипторов проводился с помощью метода БПМЛР. Статистические характеристики полученных моделей приведены в Табл. 10.

Табл. 10. 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.

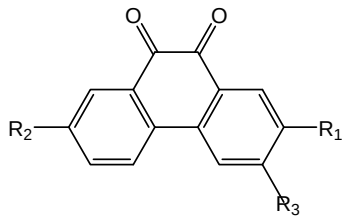
Из анализа Табл. 10 можно сделать следующие выводы. Во-первых, применение ИНС во всех случаях приводит к получению моделей с лучшими статистическими показателями. Во-вторых, сочетание ФД с квантово-химическими дескрипторами приводит к построению моделей с лучшей прогнозирующей способностью по сравнению с использованием одних ФД.

Следующим этапом стало моделирование этого свойства для объединенной базы данных. При этом была получена модель с характеристиками: $R^2 = 0.9938$, $RMSE_t = 0.34$, $RMSE_v = 0.40$. Полученные результаты показали хорошую применимость рассматриваемого нами подхода для прогнозирования данного свойства.

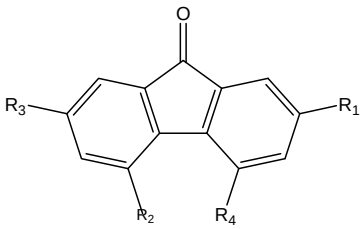
В подразделе 7.1.3 рассматривается моделирование мутагенной активности полициклических нитросоединений **12-20** (это исследование было осуществлено в соавторстве с С.К. Абилевым). Были использованы экспериментальные данные по мутагенной активности в штамме *Salmonella typhimurium* TA 1538 (*hisD3052, rfa, uvr*), регистрирующем мутации сдвига рамки считывания, без метаболической активации фракцией S9 печени млекопитающих.

Особенность этого исследования состоит в том, что в нем исходный набор дескрипторов формировался экспертным путем в соответствии с гипотезами о механизме действия нитроароматических соединений и эмпирическими заключениями о влиянии элементов структуры на мутагенную активность. Как известно, основным путем биотрансформации нитроаренов, приводящим к образованию мутагенных, канцерогенных и токсичных метаболитов, является восстановление нитрогруппы нитроредуктазами клетки. Способность к восстановлению нитроаренов коррелирует с таким параметром, как энергия низшей незанятой молекуляр-

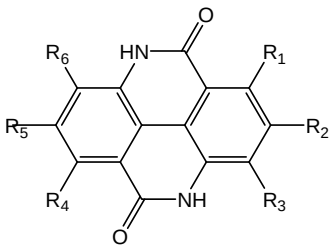
ной орбитали E_{LUMO} (дескриптор d_1). По этой же причине были выбраны и два других квантово-химических дескриптора: максимальный заряд на атоме азота (дескриптор d_2) и максимальный заряд на атоме кислорода (дескриптор d_3). В качестве дескриптора d_4 в модель был включен коэффициент распределения октанол-вода logP (гидрофобность), характеризующий способность молекулы достигать сайтов взаимодействия в живом организме. Поскольку мутагенная активность полициклических нитросоединений в значительной мере определяется положением нитрогруппы относительно общего бифенильного фрагмента, то в качестве ФД были выбраны: наличие нитрогруппы в *пара*-положении - d_5; наличие аминогруппы в *пара*-положении - d_6; наличие *мета*- и *орто*-заместителей - d_7.



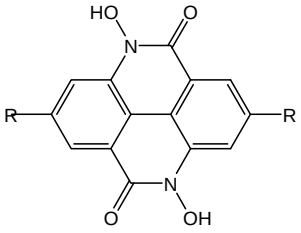
12



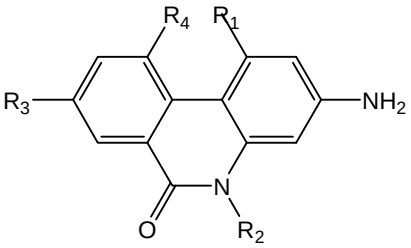
13



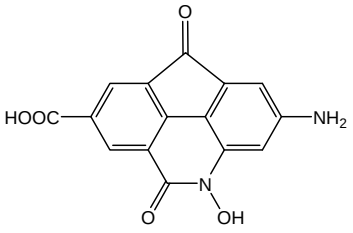
14



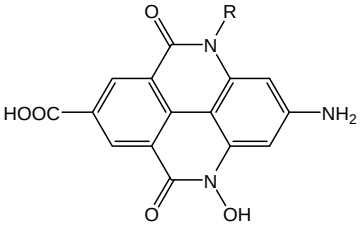
15



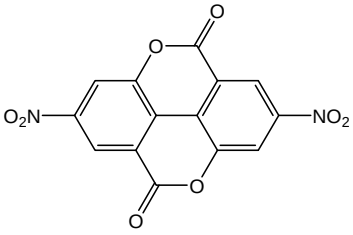
16



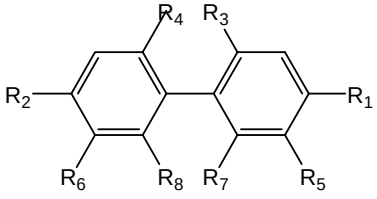
17



18



19



20

Моделирование проводилось как для всей выборки (54 соединения), так и для подвыборок, содержащих нитропроизводные гетероциклических аналогов полициклических углеводородов (пирена, фенантрена, флуорена) **12**-**19** и бифенила **20**. Построение модели проводилось двумя методами: а) пошагового метода множественной линейной регрессии МЛР; б) трехслойной ИНС. Статистические показатели полученных моделей приведены в Табл. 11. Анализ приведенных в ней данных указывает на значительные преимущества нейросетевого по сравнению с линейно-регрессионным моделированием. Следует отметить, что столь большое

различие мы наблюдали всегда при использовании наборов дескрипторов, сформированных экспертным путем с учетом природы моделируемого свойства. В этом случае эксперт может указать лишь на важные дескрипторы, но никак не может специфицировать точный тип функциональной зависимости от них. Именно поэтому ИНС, способные аппроксимировать произвольные зависимости заранее неизвестного вида, значительно лучше подходят для решения этой задачи.

Табл. 11. Статистические показатели нейросетевых и линейно-регрессионных моделей

Выборка соединений	Дескрипторы	Метод	Характеристики модели		
			R^2	$RMSE_t$	$RMSE_v$
Производные пирена, фенантрена, флуоренона	d_1, d_2, d_3, d_5, d_6, d_7	ИНС	0.81	0.76	0.96
		МЛР	0.56	1.45	1.94
Замещенные бифенилы	d_1, d_4	ИНС	0.94	0.59	0.13
		МЛР	0.64	1.21	1.34
Все соединения	d_1, d_4, d_5	ИНС	0.76	1.30	1.57
		МЛР	0.56	1.45	1.94

где: R^2 - коэффициент детерминации; $RMSE_t$, $RMSE_v$ – среднеквадратичная ошибка на обучающей и контрольной выборке (логарифмические единицы).

В подразделе 7.1.4 рассмотрено совместное применение ИНС и методов молекулярного моделирования для прогнозирования пяти 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. Набор использованных дескрипторов включает значения энергий граничных молекулярных орбиталей, зарядов на атомах, а также теплот образования производных бензола, содержащих исследуемые заместители. 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.

Раздел 7.2 посвящен применению ИНС для построения моделей «структура-условия-свойство». Он начинается с обоснования предложенной нами концепции построения нейросетевых моделей «структура-условия-свойство». Отмечается, что 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 прогнозировании этого же свойства при других условиях, а также значительно снижает объем доступных для обработки экспериментальных данных.

Поскольку, как правило, зависимость свойств химических соединений от условий, в которых они измерены, носит нелинейный характер, мы предположили,

что с помощью методологии ИНС можно расширить классический подход путем добавления характеристик внешних условий к входным данным, поступающим на вход нейросети. В качестве характеристик среды могут использоваться такие параметры, как температура, давление, концентрация, наличие того или иного растворителя, дескрипторы, характеризующие свойства растворителя, и т.д. Принцип построения моделей «структура – условия – свойство» при помощи ИНС показан на Рис. 11.

Возможность построения нейросетевых зависимостей «структура – условия – свойство» проиллюстрирована на примере моделей для физико-химических свойств углеводородов произвольной структуры, содержащих от 1 до 40 атомов углерода, а также констант скорости кислотного гидролиза сложных эфиров карбоновых кислот при различной температуре и различных составах растворителей. В случае углеводородов строились зависимости температуры кипения от структуры (при различных значениях давления), а также динамической вязкости и плотности (при различных температурах). В этом случае для описания химической структуры углеводородов были использованы ФД, тогда как для описания условий – значения температуры либо давления. При моделировании реакции гидролиза сложных эфиров их структуры были описаны при помощи квантово-химических дескрипторов. При этом условия проведения реакции были представлены: а) температурой; б) концентрацией органического компонента бинарного растворителя (в смеси с водой); в) значениями четырех параметров, предложенных В.А. Пальмом для описания влияния реакционной среды на скорости органических реакций, как то: общей кислотностью (электрофильностью) (*E*); общей основностью (нуклеофильностью) (*B*); полярностью (*Y*); поляризуемостью (*P*). Любопытно отметить, что осуществленная позже замена квантово-химических дескрипторов на ФД с «выделенными» атомами привела к модели с несколько лучшей прогнозирующей способностью. Статистические показатели построенных моделей представлены в Табл. 12. Они свидетельствуют о работоспособности предложенного подхода к моделированию зависимостей «структура-условия-свойство» при помощи ИНС.

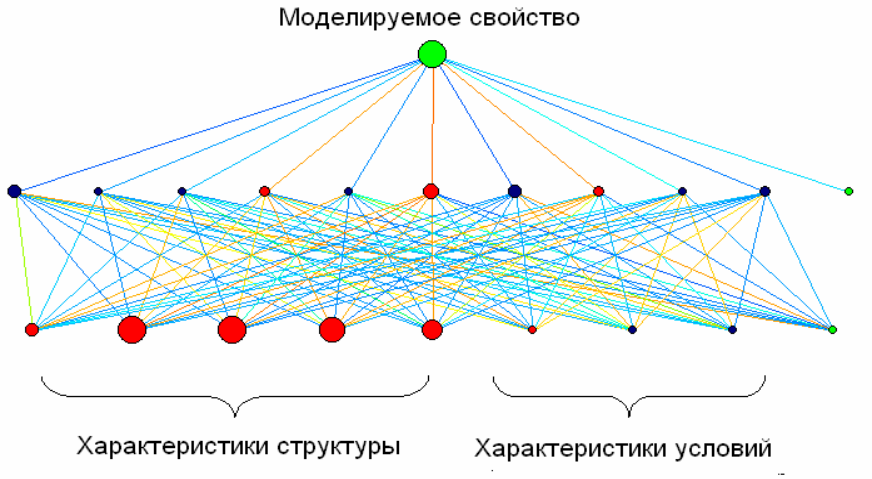


Рис. 11. Принцип построения моделей «структура–условия–свойство» при помощи ИНС

Табл. 12. Статистические показатели моделей «структура-условия-свойство»

Моделируемое свойство	Число пар «структура-условие»	R^2	$RMSE_t$	$RMSE_p$
Температура кипения углеводородов при разном давлении. (^oC)	14346	0.999	2.80	2.80
Динамическая вязкость углеводородов при разной температуре (log сантипуазов)	3426	0.990	0.14	0.16
Плотность углеводородов при разной температуре (г/см^3)	3056	0.995	0.0063	0.0063
Константа скорости гидролиза сложных эфиров карбоновых кислот при разной температуре и разном составе растворителя	2092	0.935	0.27	0.34

Раздел 7.3 посвящен рассмотрению методов, основанных на индуктивном переносе знаний при интеграции нейросетевых моделей «структура-свойство». Он начинается с констатации того, что одним из основных факторов, препятствующих построению моделей «структура-свойство» с высокой прогнозирующей способностью, является недостаток экспериментальных данных. Одним из путей преодоления связанных с этим ограничений нам видится в том, чтобы рассматривать разнообразные свойства химических соединений в их тесной взаимосвязи, и с учетом этого строить модели «структура-свойство» не изолированными, а связанными друг с другом. В этом случае, вследствие т.н. *индуктивного переноса знаний* должна происходить интеграция данных, при которой объем полезной информации для каждого из свойств будет увеличен за счет эффективного использования информации, касающейся других свойств, тесно с ним связанным. Такой перенос информации возможен между моделями, расположенными внутри сети взаимосвязанных моделей как последовательно (см. подраздел 7.3.1), так и параллельно друг относительно друга (см. подраздел 7.3.2). Можно предвидеть, что в перспективе место разрозненных и независимых друг от друга моделей «структура-свойство» займет организованная в виде «химического мозга» сеть тесно связанных между собой моделей, позволяющая интегрировать внутри себя значительный объем как экспериментальных данных, так и знаний, что позволит значительно улучшить качество прогнозирования разнообразных свойств органических соединений.

В подразделе 7.3.1 рассматривается последовательный способ интеграции нейросетевых моделей на основе предложенного нами многоуровневого принципа построения моделей «структура-свойство», суть которого заключается в следующем. Прогнозирование свойств органических соединений проводится в рамках фрагментного подхода, однако вместо изолированных одноуровневых моделей

(см. Рис. 12), берущих на входе значения ФД и выдающих на выходе значения прогнозируемых свойств, предлагается использовать организованную в виде нескольких слоев сеть моделей. Выходы моделей предыдущих слоев являются входами для моделей последующих (см. Рис. 13). В этом случае многоуровневая организация дает возможность проводить индуктивный перенос знаний от моделей предыдущего слоя к моделям последующего, что должно приводить к улучшению качества последних.

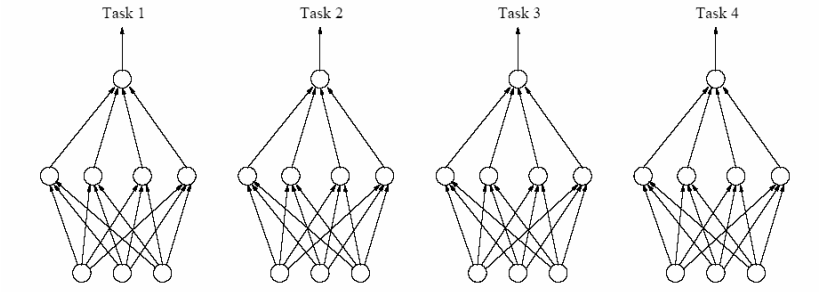


Рис. 12. Традиционный одноуровневый подход (т.н. однозадачное обучение), в котором отдельные нейросетевые модели не связаны друг с другом

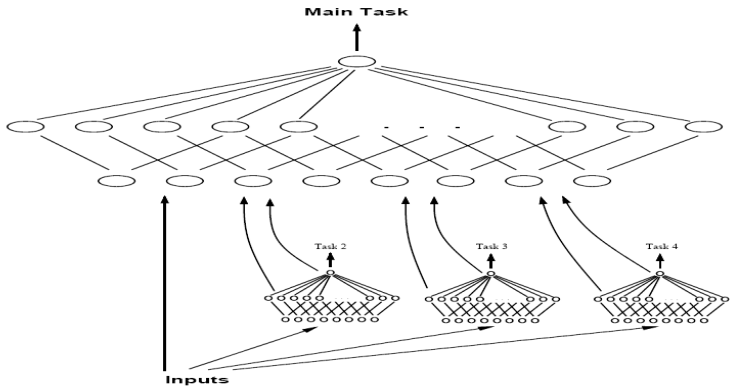


Рис. 13. Схема многоуровневого подхода, в рамках которого за счет последовательного соединения моделей происходит индуктивный перенос знаний из моделей нижнего уровня в модели верхнего

То, что при многоуровневом подходе происходит индуктивный перенос знаний, нами продемонстрировано на двух примерах. Первый из них касается моделирования коэффициента сорбции органических соединений в почве, второй – растворимости фуллерена C_{60} в органических растворителях. Построение моделей проводилось при помощи ИНС и ФД в рамках одноуровневого и многоуровневого подхода. В последнем случае были предварительно построены на том же наборе ФД промежуточные 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 диполярность/поляризуемость. Результаты прогноза первого уровня были после этого использованы в качестве дескрипторов при построении моделей второго уровня. В Табл. 13 представлены статистические характеристики промежуточных моделей первого уровня, а в Табл. 14 – целевых моделей второго уровня. Приведенные в последней таблице данные свидетельствуют о значительном улучшении прогно-

зирующей способности целевых моделей за счет индуктивного переноса знаний, полученных при формировании промежуточных моделей первого уровня.

Табл. 13. Статистические характеристики моделей «структура-свойство» первого уровня для расчета липофильности и констант Абрахама

Свойство	Размер выборки	R	*RMSEt*	*RMSEv*
Log P	7805	0.980	0.345	0.395
Абрахам A	457	0.983	0.051	0.058
Абрахам B	457	0.971	0.066	0.081
Абрахам E	457	0.997	0.040	0.074
Абрахам S	457	0.987	0.072	0.137

Табл. 14. Сравнительные статистические характеристики моделей «структура-свойство», полученных в рамках одноуровневого и многоуровневого подходов

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

Подраздел 7.3.2 посвящен рассмотрению параллельного принципа интеграции нейросетевых моделей «структура-свойство» в рамках т.н. многозадачного обучения, когда проводится одновременное построение моделей, связь между которыми осуществляется за счет использования общих промежуточных данных (см. Рис. 14). При построении моделей «структура-свойство» многозадачное обучение может быть осуществлено, например, при помощи многослойной ИНС, имеющей несколько выходных нейронов по числу одновременно моделируемых свойств, причем индуктивный перенос знаний между моделями осуществляется за счет совместного использования промежуточных данных, формируемых на общем скрытом слое нейронов.

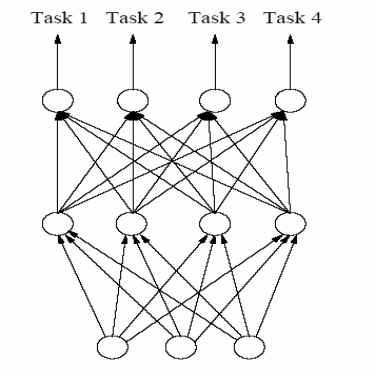


Рис. 14. Многозадачное обучение, при котором проводится одновременное построение взаимосвязанных моделей. Обмен информацией между моделями происходит за счет формирования единого внутреннего представления данных в общем слое скрытых нейронов

Впервые принципиальная возможность построения взаимосвязанных моделей «структура-свойство» была продемонстрирована нами еще в 1993 г. на примере ИНС с шестью выходами, способной одновременно предсказывать шесть физических свойств алканов (см. раздел 6.1). Поскольку исследование было проведено до появления первых математических работ по многозадачному обучению, мы не предпринимали попыток систематического изучения того, какой эффект дает его применение по сравнению с однозадачным обучением (см. Рис. 13), при котором каждое из свойств прогнозируется изолированной нейросетью с одним выходом. Подобное систематическое изучение было предпринято в нашей недавней работе по прогнозированию 11 констант распределения «ткань-воздух», которая была осуществлена совместно с несколькими группами авторов. В этой работе для получения моделей «структура-свойство» использовались ИНС с ФД. Полученные результаты наглядно представлены в виде изображенной на Рис. 15 диаграммы, показывающей зависимость повышения параметра Q^2 от размера выборки при переходе от однозадачного к многозадачному обучению. На диаграмме виден четкий тренд, показывающий, что с уменьшением размера выборки происходит резкое увеличение прогнозирующей способности моделей при переходе к многозадачному обучению за счет индуктивного переноса знаний.

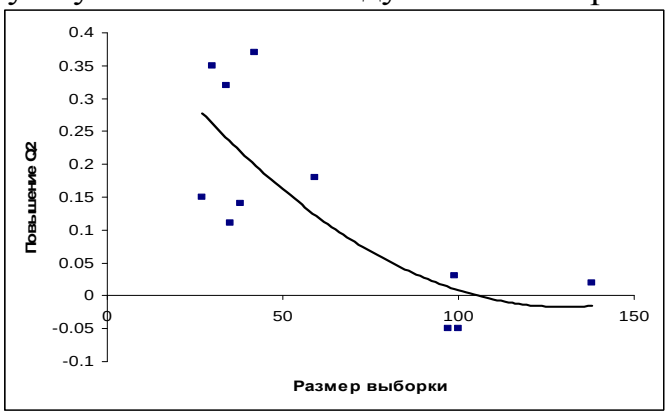


Рис. 15. Зависимость повышения показателя Q^2 от размера выборки при переходе от однозадачного к многозадачному обучению. Каждая точка соответствует одному из 11 моделируемых свойств.

Раздел 7.4 посвящен описанию разработанного нами нейронного устройства для проведения прямых корреляций «структура-свойство». При его применении не требуется предварительного вычисления каких-либо молекулярных дескрипторов. Его универсальная аппроксимирующая способность обеспечивается сочетанием ИНС с ФД либо ПФД, однако 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. Эти дескрипторы формируются промежуточно в процессе работы нейронного устройства и не видны извне. На Рис. 16 представлена принципиальная схема нейронного устройства. Работоспособность его проверена на ряде примеров (см. Табл. 15). Во всех случаях подтверждена высокая прогнозирующая способность построенных моделей.

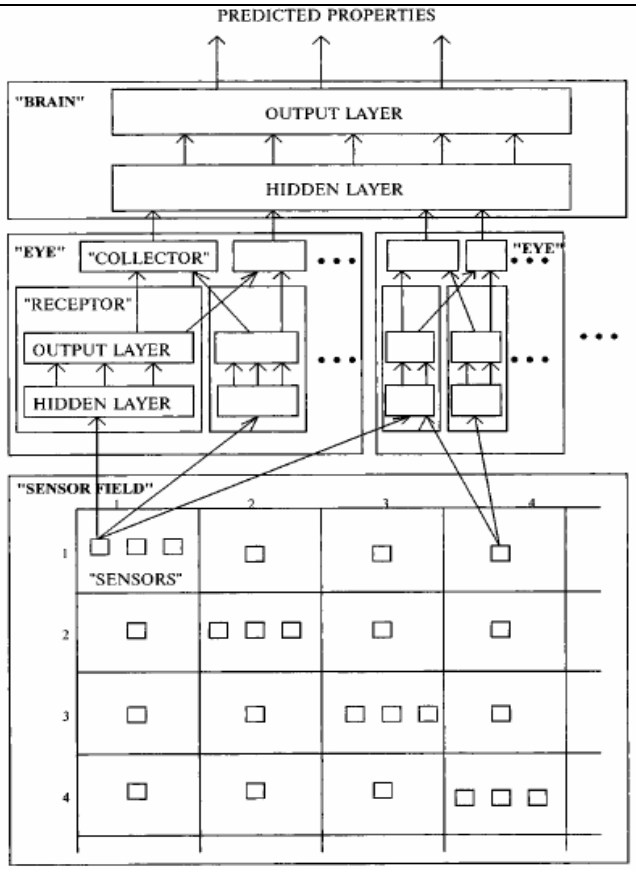


Рис. 16. Принципиальная схема нейронного устройства для осуществления прямых корреляций «структура-свойство». Предлагаемое устройство имитирует процесс обработки человеком зрительной информации. Оно представляет собой сложную интегрированную систему, состоящую из нескольких ИНС, часть из которых («рецепторы») анализируют при помощи «сенсоров» проецируемые на «сетчатку» молекулярные структуры, «коллекторы» по результатам этого анализа формируют промежуточные ФД либо ПФД, которые поступают в «мозг», осуществляющий с их помощью предсказание свойств органических соединений. Набор ИНС, занимающихся формированием промежуточных ФД одного типа, объединены в «глаза».

Табл. 15. Результаты применения нейронного устройства при построении корреляций «структура-свойство»

Свойство	Класс соединений	R^2	$RMSE_t$	$RMSE_t$
Температура кипения при нормальном давлении, град	алканы	0.999	1.6	2.4
Вязкость при 40 ^oC, сантипуазы	углеводороды	0.992	0.15	0.18
Теплота испарения, кДж/моль	углеводороды	0.943	1.44	1.26
Плотность, г/см^3	углеводороды	0.971	0.018	0.019
Теплота сольватации в циклогексане, кДж/моль	разнообразные	0.980	1.77	2.46
Поляризуемость, см^3	разнообразные	0.990	0.86	0.71
Анестетическое давление газов, лог.ед. (log(1/p))	разнообразные	0.980	0.18	0.26

Глава 8. Разработка программных средств

Данная глава посвящена рассмотрению разработанных в рамках диссертационной работы программных средств, центральным из которых является программный комплекс NASAWIN. Указанный комплекс позволяет в полном объеме осуществить весь цикл работ по построению моделей «структура-свойство», и с

их помощью осуществлять прогнозирование самых разнообразных свойств органических соединений. Именно на нем была осуществлена большая часть рассмотренных выше исследований. Основные компоненты комплекса: управляющая программа, набор дескрипторных блоков (программных компонент, позволяющих вычислять разнообразные молекулярные дескрипторы), автономная программа для прогнозирования свойств органических соединений и набор утилит. Общий объем программных средств – более 150,000 строк программного кода.

Раздел 8.1 содержит подробное описание истории создания программных средств, использованных на разных этапах выполнения диссертационной работы, большинство из которых в настоящее время включено в состав комплекса NASAWIN.

Раздел 8.2 содержит описание центрального звена этого комплекса – управляющей программы, в которую интегрировано множество средств статистического анализа химических данных. Центральное место в них принадлежит многослойным ИНС. С помощью этой программы можно:

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;
4) статистически оценивать полученные модели;
5) определять области применимости моделей;
6) использовать полученные нейросетевые модели для прогнозирования свойств химических соединений.

Раздел 8.3 содержит описание дескрипторного блока Fragment, позволяющего рассчитывать ФД в соответствии с методологией, изложенной выше в разделе 5.1.

Раздел 8.4 содержит описание дескрипторного блока FragProp, осуществляющего расчет 50 ПФД (см. раздел 5.4).

Раздел 8.5 содержит описание автономной программы, для прогнозирования свойств органических соединений с помощью нейросетевых моделей, построенных при помощи NASAWIN.

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 многоуровнего 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 предварителного 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 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.

Основное содержание диссертации изложено в опубликованных работах.

Статьи в журналах, рекомендованных ВАК РФ для публикации основных результатов докторской диссертации:

1. Зефиров 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.
2. Станкевич М.И., Баскин И.И., Зефиров Н.С. Автоматизированный поиск структурных фрагментов. Алгоритм и программа. // Журн. структ. химии. – 1987. - Т. 28, № 6. - С. 136-137.
3. Баскин И.И., Станкевич М.И., Девдариани Р.О., Зефиров Н.С. Комплекс программ для нахождения корреляций «структура-свойство» на основе топологических индексов. // Журн. структ. химии. – 1989. - № 6. - С. 145-147.
4. Баскин И.И., Палюлин В.А., Зефиров 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.

5. Баскин И.И., Палюлин В.А., Зефиров Н.С. Методология поиска прямых корреляций между структурами и свойствами органических соединений при помощи вычислительных нейронных сетей. // Докл. РАН. – 1993. - Т. 333, № 2. - С. 176-179.
6. Баскин И.И., Любимова И.К., Абилев С.К., Зефиров Н.С. Исследование количественной связи между мутагенной активностью химических соединений и их структурой. Замещенные бифенилы. // Докл. РАН. – 1993. - Т. 332, № 5. - С. 587-589.
7. Баскин И.И., Палюлин В.А., Любимова И.К., Абилев С.К., Зефиров Н.С. Количественная связь между мутагенной активностью гетероциклических аналогов пирена и фенантрена и их структурой. // Докл РАН. – 1994. - Т. 339, № 1. - С. 106-108.
8. Баскин И.И., Скворцова М.И., Станкевич И.В., Зефиров Н.С. О базисе инвариантов помеченных молекулярных графов. // Докл. РАН. – 1994. - Т. 339, № 3. - С. 346-350.
9. Baskin I.I., Skvortsova M.I., Stankevich I.V., Zefirov N.S. On basis of invariants of labeled molecular graphs. // J. Chem. Inf. Comput. Sci. – 1995. - Vol. 35, № 3. - P. 527-531.
10. Сидорова А.В., Баскин И.И., Палюлин В.А., Петелин Д.Е., Зефиров Н.С. Исследование зависимостей между структурой и октановыми числами углеводородов. // Докл. РАН. – 1996. - Т. 350, № 5. - С. 642-646.
11. Баскин И.И., Айт А.О., Гальберштам Н.М., Палюлин В.А., Алфимов М.В., Зефиров Н.С. Применение методологии искусственных нейронных сетей для прогнозирования свойств сложных молекулярных систем. Предсказание положения длинноволновой полосы поглощения симметричных цианиновых красителей. // Докл. РАН. – 1997. – Т. 357, № 1. – С. 57-59.
12. Баскин И.И., Гальберштам Н.М., Палюлин В.А., Зефиров Н.С. Компьютерная реализация искусственных нейронных сетей для решения задач по выявлению связи “структура-свойство”. // Информационные технологии. – 1997. - № 9. - С. 27-30.
13. Баскин И.И., Палюлин В.А., Зефиров Н.С. Нейроматематика - будущее вычислительной химии. // Нейрокомпьютеры: разработка, применение. – 1997. - № 3-4. - С. 17-23.
14. Баскин И.И., Бузников Г.А., Кабанкин А.С., Ландау М.А., Лексина Л.А., Орудуханян А.А., Палюлин В.А., Зефиров Н.С. Компьютерное изучение зависимости между эмбриотоксичностью и структурами синтетических аналогов биогенных аминов. // Изв. РАН, Сер. биол. – 1997. - № 4. - С. 407-413.
15. 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. (Theochem). – 1999. - V. 466. - P. 211-217.

16. Баскин И.И., Палюлин В.А., Зефиров Н.С. Применение искусственных нейронных сетей в химических и биохимических исследованиях. // Вестн. Моск. ун-та. Сер. 2. Химия. – 1999. - Т. 40, № 5. - С. 323-326.
17. Baskin, I.I.; Halberstam, N.M.; Mukhina, T.V.; Palyulin, V.A.; Zefirov, N.S. The Learned Symmetry Concept in Revealing Quantitative Structure-Activity Relationships with Artificial Neural Networks. // SAR and QSAR in Env. Res. – 2001. - Vol. 12. - P. 401-416.
18. Артеменко Н.В., Баскин И.И., Палюлин В.А., Зефиров 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.
19. Любимова И.К., Абилев С.К., Гальберштам Н.М., Баскин И.И., Палюлин В.А., Зефиров Н.С. Компьютерное предсказание мутагенной активности замещенных полициклических соединений. // Изв. РАН, Сер. биол. – 2001. - № 2. – С. 180-186.
20. Баскин И.И., Палюлин В.А., Зефиров Н.С. Прогнозирование энтальпий образования алифатических полинитросоединений. // Вестн. Моск. ун-та. Сер. 2. Химия. – 2001. - Т. 42, № 6. - С. 387-389.
21. 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. - Vol. 13, № 1. - P. 35-41
22. Halberstam N.M., Baskin I.I., Palyulin V.A., Zefurov N.S. Quantitative Structure – Conditions – Property Relationships Studies. Neural Network Modelling of Acid Hydrolysis of Esters. // Mendeleev Communications. – 2002. - Vol. 1, № 6. - P. 185-186.
23. Гальберштам Н.М., Баскин И.И., Палюлин В.А., Зефиров Н.С. Построение нейросетевых зависимостей структура-условия-свойство. Моделирование физико-химических свойств углеводородов. // Докл. РАН. – 2002. - Т. 384, № 2. - С. 202-205.
24. Гальберштам Н.М., Баскин И.И., Палюлин В.А., Зефиров Н.С. Нейронные сети как метод поиска зависимостей структура – свойство органических соединений. // Успехи химии. – 2003. - Т. 72, №7. – С. 706-727.
25. Артеменко Н.В., Баскин И.И., Палюлин В.А., Зефиров Н.С. Искусственные нейронные сети и фрагментный подход в прогнозировании физико-химических свойств органических соединений. // Изв. РАН, Сер. хим. - 2003. - № 1. - С. 19-28.
26. Жохова Н.И., Баскин И.И., Палюлин В.А., Зефиров А.Н., Зефиров Н.С. Расчет энтальпии сублимации методом QSPR с применением фрагментного подхода. // Журн. прикл. химии. - 2003. - Т. 76, № 12. - С. 1966-1970.
27. Жохова Н.И., Баскин И.И., Палюлин В.А., Зефиров А.Н., Зефиров Н.С. Фрагментные дескрипторы в QSAR: применение для расчета температуры вспышки. // Изв. РАН, Сер. хим. – 2003. - №. 9. - С. 1787-1793.

28. Жохова Н.И., Баскин И.И., Палюлин В.А., Зефиров А.Н., Зефиров Н.С. Фрагментные дескрипторы в QSPR: применение для расчета поляризуемости молекул. // Изв. РАН, Сер. хим. – 2003. - № 5. - С. 1005-1009.
29. Жохова Н.И., Баскин И.И., Палюлин В.А., Зефиров А.Н., Зефиров Н.С. Фрагментные дескрипторы в QSPR: применение для расчета магнитной восприимчивости. // Журн. структ. химии. – 2004. - Т. 45, № 4. - С. 660-669.
30. Баскин И.И., Палюлин В.А., Зефиров Н.С. Применение искусственных нейронных сетей для прогнозирования свойств химических соединений. // Нейрокомпьютеры: разработка, применение. - 2005. - № 1-2. - С. 98-101.
31. Жохова Н.И., Баскин И.И., Палюлин В.А., Зефиров А.Н., Зефиров Н.С. Исследование сродства красителей к целлюлозному волокну в рамках фрагментарного подхода в QSPR. // Журн. прикл. химии. – 2005. - Т. 78, № 6. - С. 1034-1037.
32. Баскин И.И., Палюлин В.А., Зефиров 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. L, № 2. - С. 86-96.
33. Иванова А.А., Баскин И.И., Палюлин В.А., Зефиров Н.С. Оценка значений констант ионизации для различных классов органических соединений с использованием фрагментного подхода к поиску зависимостей «структура-свойство». // Докл. РАН. – 2007. - Т. 413, № 6. - С. 766-770.
34. Жохова Н.И., Баскин И.И., Палюлин В.А., Зефиров А.Н., Зефиров Н.С. Фрагментные дескрипторы с «выделенными» атомами и их применение в исследованиях QSAR/QSPR. // Докл. РАН. – 2007. - Т. 417, № 5. - С. 639-641.
35. Varnek A., Kireeva N., Tetko I.V., Baskin I.I., Solov'ev V.P. Exhaustive QSPR Studies of Large Diverse Set of Ionic Liquids: How Accurately Can We Predict Melting Points? // J. Chem. Inf. Comput. Sci. – 2007. - Vol. 47, № 3. - P. 1111-1122.
36. Жохова Н.И., Палюлин В.А., Баскин И.И., Зефиров А.Н., Зефиров Н.С. Фрагментные дескрипторы в методе QSPR: применение для расчета энтальпии испарения органических соединений. // Журн. физ. химии. – 2007. - Т. 81, № 1. - С. 15-18.
37. Жохова Н.И., Бобков Е.В., Баскин И.И., Палюлин В.А., Зефиров А.Н., Зефиров 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.
38. Baskin I., Varnek A. Building a chemical space based on fragment descriptors. // Comb. Chem. High Throughput Screening. – 2008. - Vol. 11, № 8. - P. 661-668.
39. 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. - Vol. 49, № 1. - P. 133-144.
40. Баскин И.И., Жохова Н.И., Палюлин В.А., Зефиров А.Н., Зефиров Н.С. Многоуровневый подход к прогнозированию свойств органических соединений в

рамках методологии исследования количественных соотношений «структура-свойство/структура-активность». // Докл. РАН. – 2009. - Т. 427, № 3. - С. 335-339.

Главы в монографиях:

41. Baskin I.I., Palyulin V.A., Zefirov N.S. Chapter 8. Neural Networks in Building QSAR Models. // In: Artificial Neural Networks: Methods and Protocols / Livingstone D.S., Ed. – Humana Press, a part of Springer Science + Business Media. - 2008. – P. 139-160.
42. Baskin I.I., Varnek A. Chapter 1. Fragment Descriptors in SAR/QSAR/QSPR Studies, Molecular Similarity Analysis and in Virtual Screening. // In: Chemoinformatics Approaches to Virtual Screening / Varnek A., Tropsha A., Ed. – RCS Publishing. - 2008. – P. 1-43.

Статьи в рецензируемых журналах:

43. Baskin I.I., Skvortsova M.I., Palyulin V.A., Zefirov N.S. Quantitative chemical structure – property/activity relationship studies using artificial neural networks. // Foundations of Computing and Decision Sciences. - 1997. - Vol. 22, № 2. – P. 107-116.

Статьи в сборниках:

44. Palyulin V.A., Baskin I.I., Petelin D.E., Zefirov N.S. Novel descriptors of molecular structure in QSAR and QSPR studies. // QSAR and Molecular Modelling: Concepts, Computational Tools and Biological Applications. - Barcelona: Prous Science Publishers. – 1995. - P. 51-52.
45. 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. - Barcelona: Prous Science Publishers. – 1995. - P. 30-31.
46. Zefirov N.S., Baskin I.I., Halberstam N.M., Palyulin V.A. Artificial Neural Networks Oriented for the Chemical Structure-Property Relationship Modelling. // EUFIT'97 5th European Congress on Intelligent Techniques & Soft Computing. Book of Abstracts. – 1997. – V. 1. – P.552-556.
47. Баскин И.И., Палюлин В.А., Зефиров Н.С. Применение искусственных нейронных сетей в химических и биохимических исследованиях. // V Всероссийская конференция «Нейрокомпьютеры и их применение». Москва, 17-19 февраля 1999 г. Сборник докладов. - С. 28-31.
48. Baskin I.I., Keshtova S.V., Palyulin V.A., Zefirov N.S. Combining Molecular Modelling with the Use of Artificial Neural Networks as an Approach to Predict Substituent Constants and Bioactivity. // Molecular Modeling and Prediction of Bioactivity; K. Gundertofte; F.S. Jorgensen, Eds. - Klumer Academic / Plenum Publishers: New York, Boston, Dordrecht, London, Moscow. – 2000. - P. 468-469.

49. Гальберштам Н.М., Баскин И.И., Палюлин В.А., Зефиров 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.- Москва. – 2001. – С. 423-424.
50. Баскин И.И., Палюлин В.А., Зефиров Н.С. Нейрокомпьютеры и геном человека. // Труды VII Всероссийской конференции «Нейрокомпьютеры и их применение». НКП-2001 с международным участием. Москва, 14-16 февраля 2001 г. – Москва. - С. 13-16.
51. Баскин И.И., Гальберштам Н.М., Палюлин В.А., Зефиров Н.С. NASAWIN – программный комплекс для изучения соотношений структура-свойство в химии. // Труды VII Всероссийской конференции «Нейрокомпьютеры и их применение». НКП-2001 с международным участием. Москва. – 2001. – С. 419-422.
52. Артеменко Н.В., Баскин И.И., Гальберштам Н.М., Палюлин В.А., Зефиров Н.С. Прогнозирование физических свойств органических соединений при помощи нейронных сетей в рамках подструктурного подхода. // Труды VII Всероссийской конференции «Нейрокомпьютеры и их применение» НКП-2001 с международным участием. – Москва. – 2001. - С. 414-418.
53. Айт А.О., Баскин И.И., Гальберштам Н.М. Прогнозирование физико-химических свойств симметричных цианиновых красителей с использованием методологии искусственных нейронных сетей. // Труды VII Всероссийской конференции «Нейрокомпьютеры и их применение». НКП-2001 с международным участием. - Москва. – 2001. – С. 411-413.
54. 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 / Eds., M.Ford et al. - Blackwell Publishing. – 2003. - P. 260-263.