Полные статьи

УДК 541.6

Искусственные нейронные сети и фрагментный подход в прогнозировании физико-химических свойств органических соединений

*Н. В. Артеменко, И. И. Баскин, В. А. Палюлин, Н. С. Зефиров**

Московский государственный университет им. М. В. Ломоносова, Химический факультет, Российская Федерация, 119992 Москва, Ленинские горы.
Факс: (095) 939 0290. E-mail: zefirov@org.chem.msu.su

Для прогнозирования физико-химических свойств органических соединений развит подход на основе фрагментных дескрипторов (числа вхождений подструктурных фрагментов в структуры химических соединений) в сочетании с математическим аппаратом искусственных нейронных сетей. В качестве примера рассмотрено построение нейросетевых моделей для прогнозирования вязкости, плотности и давления насыщенного пара различных классов органических соединений.

Ключевые слова: искусственные нейронные сети, нейросетевое моделирование, фрагментные дескрипторы, физико-химические свойства, вязкость, плотность, давление насыщенного пара.

Несмотря на быстрое развитие квантово-химических методов молекулярного моделирования, в настоящее время при прогнозировании большинства физико-химических свойств основную роль играют эмпирические подходы, основанные на применении различных дескрипторов молекулярной структуры[1]. Аддитивные схемы исторически занимают первое место среди эмпирических подходов для аппроксимации физико-химических свойств. В качестве дескрипторов в этих методах использовали число различных простейших фрагментов в молекуле[2–7], а физико-химические свойства представляли в виде суммы величин, приходящихся на парные взаимодействия атомов[8–14]. В основу работ[5,15–17] положено предположение о том, что численную величину некоторого свойства соединений заданного ряда можно представить в виде суммы вкладов отдельных структурных фрагментов. Разработан[6,7,18] фрагментный метод расчета физико-химических свойств, в частности энтальпий образования предельных углеводородов. Он основан на топологическом графовом подходе к рассмотрению структур. В соответствии с общими принципами теории графов структурная формула любого химического соединения может быть описана в терминах теории графов: в качестве вершин молекулярного графа рассматривают атомы, в качестве ребер — химические связи. В качестве фрагментов предложено использовать цепочки определенной длины, подразделяя атомы углерода на первичные, вторичные, третичные и четвертичные. Вклад цепочек, содержащих по две или три вершины (атома), рассматривали как поправку на влияние окружения первого и второго порядка. Описанный подход получил дальнейшее развитие в рамках программного комплекса ЭММА,[19]

© 2003 «Известия Академии наук. Серия химическая», Российская академия наук, Отделение химии и наук о материалах Российской академии наук, Институт органической химии им. Н. Д. Зелинского Российской академии наук

в котором в качестве статистического метода используют множественную линейную регрессию (МЛР). В рамках этого комплекса создана программа для расчета фрагментных дескрипторов Fragment 1, позволяющая осуществлять генерацию линейных (от одного до девяти атомов), циклических (трех—шестичленных) и трех типов разветвленных фрагментов[20,21]. Модифицированная версия этой программы[22], которая использована и в настоящей работе, включена в программный комплекс NASAWIN.[23] Ранее разработан программный комплекс BIBIGON,[24] позволяющий строить линейные зависимости «структура—свойство» с использованием фрагментных дескрипторов. Для характеристики структуры в этом подходе используют цепочки связанных атомов $A_1...A_k$. Каждому атому приписывают набор меток, кодирующих тип данного атома (химический символ, число неводородных соседей, кратность химических связей, положение атома в цикле или в цепи).

Фрагментные подходы успешно использовали в работах по поиску взаимосвязи биологической активности молекул и их химического строения. Так, в методе Фри—Вильсона[25—27] биологическую активность (A) описывали как сумму вкладов заместителей при определенном общем фрагменте.

Следующим шагом в развитии вычислительных методов с использованием фрагментов послужило введение достаточно простого языка для описания химических структур (ФКСП — фрагментный код суперпозиции подструктур), основанного на концепции дескрипторных центров[28—31]. В рамках этого подхода в химической структуре выделяют так называемые дескрипторные центры (атомы или их группы), которые могут являться либо центрами взаимодействия между биологически активной молекулой и биомишенью, либо реакционными центрами. В несколько модифицированном виде ФКСП положен в основу программного комплекса PASS.[32] Данная программа в первом варианте позволяла оценивать вероятность проявления разнообразными органическими соединениями 114 видов биологической активности. При дальнейшем развитии программы PASS было предложено использовать в качестве подструктурных дескрипторов «многоуровневые окружения атомов» (MNA — Multilevel neighborhoods of atoms)[33,34]. Разработан[31] так называемый логико-структурный подход (ЛСП), предназначенный для выявления связей «структура—активность». В рамках этого метода активность химического соединения прогнозируют на основании присутствия в нем определенных структурных фрагментов, называемых фармакофорами (способствующими проявлению активности) и фармакофобами (препятствующими проявлению активности). Активность структур также оценивали[35] при помощи логических функций, представляющих собой конъюнкции и дизъюнкции предикатов, являющихся индикаторами наличия определенных фрагментов в структуре.

Фрагментный подход реализован в компьютерной программе CASE (Computer automated structure evaluation)[36—38]. В этом подходе молекулярное свойство получают простым суммированием всех локальных значений для атомов или линейных фрагментов, входящих в данную структуру. Однако индивидуальный вклад каждого атома определяется не только его собственной природой, но и природой окружения. Для успешной работы программы сформированы основные группы структурных параметров: наличие тяжелых атомов различных типов гибридизации и функциональных групп. Последний вариант программ CASE и MultiCASE интересен еще и тем, что стало возможным автоматически находить структурные фрагменты, оказывающие заметное положительное или отрицательное влияние на биологическую активность, так называемые биофоры и биофобы[38]. Структурные фрагменты использовали[39,40] для прогнозирования мутагенной активности при помощи искусственных нейронных сетей (ИНС). Число вхождений простейших одноатомных фрагментов в сочетании с ИНС применяли[41] для прогнозирования физико-химических свойств и биологической активности ряда органических соединений.

Предложена концепция молекулярной голограммы[42], по существу представляющей собой вектор, который описывает вхождение разнообразных фрагментов (задаваемых в явном виде при помощи линейной нотации) в структуру химического соединения. В рамках этого подхода поиск зависимости между структурой и свойством (биологической активностью) химического соединения осуществляется при помощи метода частичных наименьших квадратов (PLS — Partial least squares)[43], что позволяет работать с большим количеством скоррелированных дескрипторов. Этот подход также дает возможность интерпретировать получаемые модели при помощи цветового кодирования изображения молекулярной структуры. В результате удается выявить части молекул, являющиеся благоприятными или неблагоприятными для проявления биологической активности данного вида.

Однако большинство описанных методов имеют ряд ограничений: 1) слишком высокий уровень обобщения в классификации атомов; 2) отсутствие гибкости при выборе фрагментов; 3) использование во многих случаях линейного статистического метода, что ухудшает качество модели и требует введения слишком большого количества поправочных коэффициентов.

В предыдущей работе[44] нами рассмотрены методологические аспекты применения фрагментных дескрипторов на примере построения линейных моделей «структура—свойство». Однако во многих случаях зависимость физико-химических свойств от дескрипторов носит существенно нелинейный характер, ее общий вид обычно заранее неизвестен. Применение в таких случаях аппарата ИНС позволяет успешно решать задачу прогнозирования свойств органических соединений[45—47].

Под искусственными нейронными сетями понимают совокупность методов вычислительной математики, объединенных общей идеей имитации функ-

ционирования головного мозга человека при обработке информации[48]. Обычно полагают, что искусственные нейронные сети («нейросети») состоят из набора простых вычислительных устройств, называемых «нейронами», и совокупности «синапсов», связывающих их. Каждый синапс характеризуется числом, называемым «синаптическим весом». Нейросеть обладает способностью к обучению. Под обучением обычно понимают подстройку значений синаптических весов, которая минимизирует определенный функционал ошибки, зависящий от задачи. При решении регрессионных задач в результате обучения нейросеть подстраивает свои синаптические веса таким образом, чтобы минимизировать ошибку прогнозирования значений выходного вектора чисел на основе значения входного вектора. В частности, при анализе зависимостей «структура—свойство» нейронная сеть учится на основе входного вектора значений дескрипторов прогнозировать выходной вектор свойств химических соединений[46]. Наиболее распространенной архитектурой нейросетей является многослойная сеть прямого распространения с обратным распространением ошибки. В рамках этой архитектуры входные значения дескрипторов задают значения активации нейронов входного слоя. Спрогнозированные значения свойств снимают с нейронов выходного слоя, а для промежуточных вычислений используют также нейроны «скрытого» слоя, количество которых определяет сложность моделируемой зависимости.

В настоящей работе для оценки эффективности предлагаемого фрагментного подхода в сочетании с аппаратом ИНС при моделировании физико-химических свойств органических соединений проведено как линейно-регрессионное, так и нейросетевое моделирование плотности (для жидких веществ), вязкости и давления насыщенного пара органических соединений. Выбор именно этих свойств обусловлен их практической важностью. В частности, предсказание плотности необходимо для разработки высокоэнергоемких соединений. Знание вязкости важно для оптимизации нефтехимических процессов. Прогнозирование давления насыщенного пара позволяет оценивать скорость испарения и абсорбции, а также максимально возможные концентрации в воздухе веществ, потенциально являющихся факторами загрязнения окружающей среды.

Методика исследования

Построение и анализ моделей «структура—свойство» проводили по следующей схеме. На первом этапе для всех соединений из базы данных, включающей информацию о структурах химических соединений и их свойствах, рассчитывали фрагментные дескрипторы (число вхождений структурных фрагментов в химическую структуру)[22], причем максимальный размер фрагментов варьировали от одного до десяти атомов. При расчете исключали фрагменты, встречающиеся в выборке для ≤1 % соединений, а также статистически идентичные. Для каждого дескриптора D_i рассчитывали нелинейные модификации: квадрат (D_i^2), квадратный корень ($D_i^{1/2}$), десятичный логарифм ($\lg D_i$, вычисляемый только для фрагментов, содержащихся во всех структурах базы данных), отношение значения дескриптора к числу неводородных атомов в молекуле (D_i/N_a). На следующем этапе часть дескрипторов отбрасывали таким образом, чтобы все парные коэффициенты корреляции r между оставшимися дескрипторами не превышали 0.97.

Использование наряду с фрагментными дескрипторами их нелинейных модификаций вполне оправдано. Для исследования этого вопроса предварительно был проведен сравнительный анализ линейно-регрессионных и нейросетевых моделей (методику их построения см. ниже) для четырех наборов дескрипторов: как содержащих, так и не содержащих перечисленные выше модификации, с максимальным числом атомов, равным единице и двум. Проведенный анализ показал, что статистические характеристики построенных моделей с дескрипторами и их нелинейными модификациями значительно лучше аналогичных параметров для моделей, построенных без включения нелинейных модификаций дескрипторов.

После этого базу данных разбивали на три выборки — обучающую (80% соединений), контрольную (10% соединений) и выборку для оценки предсказательной способности модели (10% соединений). Разбивку проводили десятью разными способами таким образом, чтобы каждое соединение из базы данных присутствовало по одному разу в каждой из двух последних выборок. Затем для каждого из 10 или 13 (для разных баз данных) первоначальных наборов дескрипторов (различающихся максимальным размером фрагментов) и каждой разбивки базы данных проводили отбор дескрипторов при помощи процедуры пошаговой множественной линейной регрессии, в которой включение каждого последующего нового дескриптора определялось уменьшением ошибки прогноза для контрольной выборки. После этого из 10 или 13 первоначальных наборов для каждого варианта разбивки базы данных отбирали оптимальный дескриптор в соответствии со средней ошибкой прогноза для контрольных выборок. Полученные из него 10 наборов дескрипторов были далее использованы в исследовании при помощи многослойных нейронных сетей с обратным распространением ошибок[46,48].

На следующем этапе для каждой разбивки базы данных строили по пять нейросетевых моделей для каждого числа скрытых нейронов, которое варьировали от двух до восьми. Обучение проводили при помощи «обобщенного дельта-правила» (параметр скорости 0.25, момент 0.9) до достижения минимальной ошибки прогноза для контрольной выборки. После этого определяли оптимальное число скрытых нейронов, обеспечивающее наименьшие ошибки для контрольных выборок, и результаты прогнозирования при помощи половины наилучших (т.е. дающих наименьшую ошибку для контрольной выборки) моделей для всех соединений усредняли. В результате для каждого свойства получили следующие четыре параметра: коэффициент корреляции, усредненный по всем обучающим выборкам (R_{av}), а также среднеквадратичные значения ошибок для трех типов выборок. Поскольку данные из третьей выборки не участвовали ни в построении моделей, ни в их отборе, то именно среднеквадратичная ошибка для этого типа выборок и служит объективной оценкой прогнозирующей способности построенных моделей.

Обсуждение полученных результатов

Ранее прогнозирование плотности, вязкости и давления насыщенного пара осуществляли преимущественно без использования фрагментного подхода.

3-2-4691

Например, для прогнозирования давления насыщенных паров алканов и алкенов применяли[49] автокорреляционный метод. Компоненты автокорреляционных векторов (дескрипторы при проведении анализа) рассчитывали из площади поверхности 186 соединений по методу Бонди[50]. Ранее для моделирования данного свойства в качестве дескрипторов использовали топоструктурные, топохимические и геометрические параметры[51] или квантово-химические дескрипторы[52]. При прогнозировании плотности применяли[53] дескрипторы, основанные на соотношении молекулярных масс и объемов молекул. Разработаны[54] нейросетевые модели описания физико-химических свойств с помощью теоретико-графовых дескрипторов. Для прогнозирования плотности алкенов использовали топологические дескрипторы[55,56]. Вязкость многие исследователи также рассчитывали либо на основе фрагментных подходов[57–62], либо с помощью методов изучения молекулярных свойств (молекулярной рефракции, дипольном моменте, критической температуре, молярной магнитной восприимчивости, энергии когезии), применяемых в качестве дескрипторов[63,64].

Рассмотрим применение описанной выше методики на примерах нейросетевого моделирования вязкости, плотности и давления насыщенного пара. Для моделирования вязкости органических соединений использована база данных[61] для 367 органических соединений различных классов: линейных, разветвленных, моно- и бициклических алканов, алкенов и алкинов, аренов, спиртов, простых и сложных эфиров, кетонов, альдегидов, карбоновых кислот, нитрилов, иминов, аминов, амидов, галоген- и серосодержащих соединений и нитросоединений. Из выборки, приведенной в опубликованной ранее работе[61], были исключены два соединения, для которых авторами приведены одинаковые названия, но разные значения вязкости (соединения 266 и 267). Базу данных разбивали десятью разными способами на три выборки: обучающую (293 соединения), контрольную (37 соединений) и для оценки прогнозирующей способности (37 соединений). Согласно описанной выше схеме с помощью процедуры пошаговой линейной регрессии из рассчитанного множества дескрипторов проводили их отбор для десяти различных вариантов разбивки базы данных. В процессе построения каждой линейно-регрессионной модели последовательно вводили дескрипторы до достижения наилучшей прогнозирующей способности для контрольной выборки. Результаты полученных линейно-регрессионных моделей для 13 наборов дескрипторов с различным максимальным размером фрагментов (130 моделей) представлены в табл. 1 и на рис. 1.

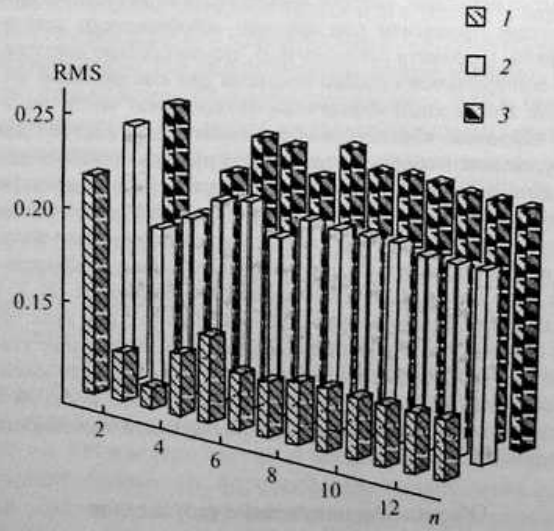

Рис. 1. Зависимость усредненной среднеквадратичной ошибки (RMS) от максимального размера (n) фрагментных дескрипторов для вязкости для обучающей (*1*), контрольной (*2*) и тестовой выборки (*3*).

Таблица 1. Усредненные статистические характеристики линейно-регрессионных моделей при варьировании максимального размера дескрипторов для моделирования вязкости органических соединений*

Число атомов	n	n_d	МЛР			
			R_{av}	RMS_{tr}	RMS_{val}	RMS_{test}
1	146	38±20	0.9204	0.2172	0.2366	0.2407
2	531	53±12	0.9740	0.1260	0.1857	0.1853
3	1757	46±16	0.9794	0.1113	0.1950	0.2119
4	1974	42±22	0.9593	0.1336	0.2079	0.2341
5	2183	34±21	0.9531	0.1470	0.2113	0.2330
6	2413	36±21	0.9681	0.1307	0.1960	0.2207
7	2566	33±19	0.9662	0.1302	0.2088	0.2392
8	2649	35±22	0.9656	0.1337	0.2075	0.2305
9	2703	33±20	0.9652	0.1348	0.2077	0.2322
10	2732	35±22	0.9658	0.1330	0.2081	0.2316
11	2945	35±22	0.9657	0.1331	0.2044	0.2297
12	2759	35±22	0.9657	0.1331	0.2044	0.2297
13	2770	35±22	0.9657	0.1331	0.2044	0.2297

* Обозначения: n — общее число дескрипторов, n_d — среднее число отобранных дескрипторов, МЛР — множественная линейная регрессия; R_{av} — коэффициент корреляции; RMS_{tr}, RMS_{val} и RMS_{test} — среднеквадратичная ошибка для обучающей, контрольной и тестовой выборок соответственно.

Как видно из рис. 1, использование фрагментных дескрипторов с числом атомов более двух—трех не приводит к улучшению качества регрессионных моделей. При построении нейросетевых моделей наилучшие статистические характеристики получены для множества дескрипторов с максимальным размером фрагментов, равным трем. Оптимальный набор дескрипторов выбирали по значению среднеквадратичной ошибки для контрольной выборки, поскольку некорректно ориентироваться как на минимум ошибки для обучающей выборки (во избежание построения переопределенных моделей), так и на минимум ошибки выборки для оценки предсказательной способности (поскольку данные для этой выборки следует использовать только для оценки предсказательной способности, а не для построения моделей).

Само по себе наличие оптимального значения максимального размера для генерируемых фрагментов, которое обеспечивает наилучшую прогнозирующую способность моделей, не очевидно и поэтому заслуживает отдельного рассмотрения. Связано это с тем, что при увеличении размеров фрагментов число их типов, а следовательно, и число фрагментных дескрипторов резко возрастает. В то же время при прочих равных условиях (т.е. при одинаковой ошибке для обучающей выборки и одинаковом числе отобранных дескрипторов), как следует из целого ряда математических теорий (см. ниже), прогнозирующая способность статистической модели ухудшается с увеличением первоначального числа дескрипторов, из которых производят отбор. Действительно, согласно статистической теории прогнозирования Вапника—Червоненкиса[65] минимальный размер выборки соединений, необходимый для достижения заданного качества прогнозирования, зависит как от числа отобранных дескрипторов, так и от их первоначального числа. В последнем случае для бинарных дескрипторов (так называемых признаков) показан логарифмический характер зависимости минимального размера выборки от логарифма числа первоначальных дескрипторов. Следовательно, при фиксированном размере выборки качество модели ухудшается по мере увеличения первоначального числа дескрипторов. Таким образом, эффективное число дескрипторов в статистической модели («размерность Вапника—Червоненкиса») в общем случае не равно числу отобранных дескрипторов. Оно зависит также от первоначального числа дескрипторов, из которых производят отбор. К аналогичным выводам приводит и теория индуктивных выводов[48]. Как известно[66], ожидаемая ошибка статистической модели на данных, не входящих в обучающую выборку, определяется степенью сжатия информации с помощью этой модели. Чем меньше суммарная длина описания данных с помощью модели и описания самой модели, тем меньше ошибка предсказаний при помощи этой модели. Длина описания модели M равна количеству информации, необходимой для выбора этой модели из множества с априорным распределением вероятностей $P(M)$, что можно аппроксимировать как $-\lg P(M)$. Ясно, что чем больше первоначальное число, из которого отбирают дескрипторы, тем меньше априорная вероятность получаемой модели и, следовательно, тем больше длина описания модели и ожидаемая ошибка прогноза. Отсюда следует крайне важный вывод для построения моделей «структура—свойство»: нельзя неограниченно увеличивать набор дескрипторов, подаваемых на вход статистических процедур, надеясь, что нужные дескрипторы все равно будут автоматически отобраны, поскольку вероятность ошибочного выбора при этом возрастает. Следовательно, при построении моделей с наилучшей прогнозирующей способностью следует оптимизировать не только наборы отбираемых дескрипторов, но и первоначальные наборы дескрипторов, подаваемых на вход автоматических процедур отбора, что и было продемонстрировано в настоящей работе.

Вторым интересным моментом, отраженным в табл. 1, является наличие максимумов у зависимости среднего числа отбираемых при построении моделей дескрипторов от максимального размера (т.е. числа атомов во фрагменте) генерируемых дескрипторов. Положение максимумов на этой зависимости в точности совпадает с положением минимумов ошибок прогноза. Для объяснения этого факта можно воспользоваться следующей легко наблюдаемой, по крайней мере для физико-химических свойств, закономерностью: с увеличением размера фрагментов доля «ценных» для построения моделей фрагментных дескрипторов уменьшается. Точное число «ценных» дескрипторов определить трудно. Кроме того, они могут коррелировать друг с другом. В связи с этим отмеченную тенденцию можно оценить приближенно с помощью зависимости среднего числа отбираемых дескрипторов от общего числа генерируемых фрагментов с определенным максимальным размером (рис. 2). При малом размере фрагментов выигрыш за счет добавления «ценных» дескрипторов превышает проигрыш за счет возрастания вероятности ошибочного выбора. Поэтому число отбираемых дескрипторов возрастает, а ошибка прогнозирования уменьшается. С увеличением размеров фрагментов доля «ценных» дескрипторов уменьшается параллельно с возрастанием их общего числа, в результате чего проигрыш за счет возрастания вероятности ошибочного выбора начинает превышать выигрыш за счет добавления «ценных» дескрипторов. Вследствие этого качество моделей, выражаемое их прогнозирующей способностью, ухудшается, т.е. возрастают ошибки прогноза. Поскольку в использованном нами варианте пошагового отбора дескрипторы включаются в модель по достижении наилучшей прогнозирующей способности, оцениваемой по контрольной выборке, то параллельно с уменьшением числа включаемых в модель дескрипторов ухудшается прогнозирующая способность. Иными словами, с возрастанием размера фрагментов часть «ценных» дескрипторов оказывается «зашумленными» за счет возрастания общего числа дескрипторов, и они уже не могут быть

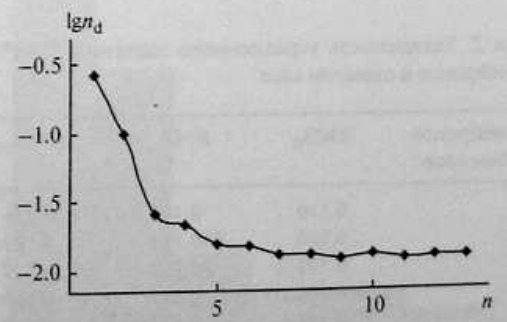

Рис. 2. Зависимость логарифма доли отбираемых в модели дескрипторов (n_d) от их общего числа (n) при построении моделей для вязкости.

автоматически отобраны, что и приводит к уменьшению числа отбираемых дескрипторов.

Рассмотрим теперь отобранные дескрипторы. При моделировании вязкости наиболее важным (в настоящей работе меру важности мы определяем по количеству моделей, в которые вошел данный дескриптор) оказывается вклад общего числа неводородных атомов в молекуле (N_a), отношение числа Me-групп, связанных с атомом С, к числу неводородных атомов $N(Me—C)/N_a$, а также отношение числа *н*-пропильных групп к числу неводородных атомов $N(Pr)/N_a$. Кроме того, следует отметить важность таких дескрипторов, как число аминогрупп $N(NH_2)$ и атомов N при двойной связи $N(=N)/N_a$, цепочек, содержащих гидроксильные группы $N(C_{sp^3}—C_{sp^3}—C_{sp^3}—OH)/N_a$, атомов галогенов и амидных групп. Можно предположить, что первые три дескриптора описывают ван-дер-ваальсово взаимодействие между молекулами, а остальные — электростатическое (включая образование водородных связей). При построении модели не учитывали дескрипторы, описывающие ароматические связи, поскольку с их учетом снижается качество прогноза.

После построения ряда нейросетевых моделей (350 моделей) с варьированием числа нейронов в скрытом слое от двух до восьми было выбрано оптимальное число скрытых нейронов, равное семи (табл. 2), хотя практически при любом количестве скрытых нейронов статистические параметры моделей отличались несущественно.

Для 50 моделей, построенных с семью скрытыми нейронами, на рис. 3 представлена зависимость между среднеквадратичной ошибкой выборки для оценки предсказательной способности (тестовой) RMS_{test} и среднеквадратичной ошибкой для контрольной выборки RMS_{val}. Тангенс угла наклона прямой, отвечающей линейной аппроксимации зависимости (см. рис. 3) среднеквадратичных ошибок для данной выборки, положителен. Подобные изменения наблюдаются для всех ансамблей моделей с различным числом скрытых нейронов. В отдельных случаях может также встречаться и отрицательное значение тангенса угла (ансамбли моделей плотности жидких органических соединений при некоторых других, отличных от оптимального, количествах скрытых нейронов). Можно предположить, что знак угла наклона указанной линии характеризует недоопределенность или переопределенность моделей. В результате при наличии подобного угла наклона появляется возможность улучшить качество построенных моделей, т.е. их предсказательную способность. Таким образом, из 50 построенных моделей с данным количеством скрытых нейронов выбирается половина моделей с низшими значениями среднеквадратичной ошибки для контрольной выборки RMS_{val}. В таблице 3 представлены статистические параметры моделирования на каждой стадии.

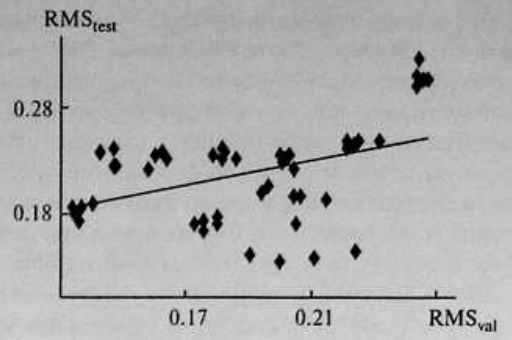

Рис. 3. Зависимость усредненного значения RMS_{test} от RMS_{val} для вязкости.

Окончательные результаты (см. табл. 3) значительно лучше усредненных по линейно-регрессионным моделям и являются статистически более обоснованными, поскольку они учитывают больший массив моделей. Корреляция усредненных по всему массиву моделей расчетных данных для всех выборок с экспериментальными значениями представлена на рис. 4.

На рисунке 5 приведено распределение ошибок прогноза для вязкости. Следует отметить, что для всей выборки среднеквадратичная ошибка RMS_{test} для спрогнозированных значений рассматриваемого свойства составляет 0.141 логарифмической единицы. Среди соединений, для которых значения вязкости предсказываются с большими ошибками, оказались в основном сильно полярные соединения (неко-

Таблица 2. Зависимость усредненного значения RMS* от числа нейронов в скрытом слое

Число нейронов в скрытом слое	RMS_{tr}	RMS_{val}	RMS_{test}
2	0.110	0.193	0.226
3	0.106	0.191	0.222
4	0.108	0.192	0.220
5	0.106	0.192	0.219
6	0.105	0.191	0.219
7	0.105	0.189	0.219
8	0.105	0.191	0.220

* Обозначения см. в табл. 1.

Таблица 3. Статистические параметры* моделирования на различных этапах исследования с использованием нейросетевых моделей

Этап	*R*	RMS_{tr}	RMS_{val}	RMS_{test}
МЛР	0.9794	0.111	0.195	0.212
Усреднение по 50 моделям	0.9815	0.105	0.189	0.219
Расчет по индивидуальным вкладам для 50 моделей	0.9904	0.078	0.177	0.208
Усреднение по 25 лучшим моделям	0.9814	0.106	0.161	0.212
Расчет по индивидуальным вкладам для 25 лучших моделей	0.9885	0.084	0.104	0.141

* Обозначения см. в табл. 1.

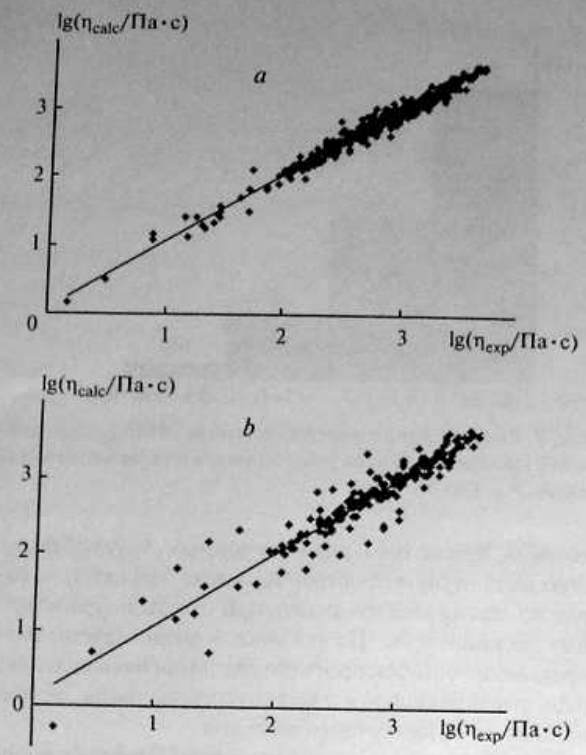

Рис. 4. Результаты моделирования вязкости (lg(η/Па·с)): обучающая выборка (*a*), выборка для оценки предсказательной способности (*b*).

торые из них могут образовывать водородные связи): 2-метилпентан-2,4-диол (Δ = 0.76), тринитрат глицерина (Δ = –0.72), муравьиная кислота (Δ = –0.58), дибутил-*о*-фталат (Δ = –0.60), циклогексанол (Δ = –0.64), 2-метоксиэтанол (Δ = 0.58), акриловая кислота (Δ = 0.54), трифторуксусная кислота (Δ = 0.54), 4-гидрокси-4-метилпентан-2-он (0.52), этоксибензол (Δ = 0.49), 2-метилбутан-2-ол (Δ = –0.48), бутан-1,3-диол (Δ = –0.48), дибутилмалеат (Δ = 0.45), глицерин (Δ = 0.43) и *N*,*N*-диметилформамид (Δ = 0.42).

По аналогичной схеме было выполнено моделирование плотности[67] жидких органических соединений (на основе базы данных, включающей 803 соединения — алканы, алкены, алкины, арены, аллены, спирты, простые и сложные эфиры, нитросоединения, альдегиды, карбоновые кислоты, кетоны, нитрилы, амины, имины, амиды, соединения, содержащие гетероатомы, моно-, би- и трициклические структуры) и давления насыщенного пара[68] (на основе базы данных из 352 углеводородов — линейных, разветвленных и циклических алканов, алкенов, аренов и их галогенпроизводных).

Важными для моделирования плотности органических соединений являются число sp^3- и sp^2-гибридных атомов С ($N(C_{sp^3})/N_a$ и $N(H_2C=)/N_a$) и относительное количество различных гетероатомов (в частности, галогенов, кислорода, азота, кремния, серы и т.д.), что можно связать с различной массой, ковалентными и ван-дер-ваальсовыми радиусами разных элементов. Разнообразные поправки описываются такими дескрипторами, как число тройных связей C≡C, и дескрипторами, характеризующими разветвленность. При построении модели не учитывали дескрипторы, описывающие ароматические связи, поскольку при их учете не улучшалось качество прогноза. Среднеквадратичная ошибка прогноза RMS_{test} для всей выборки составила 0.051 г·см$^{-3}$. Среди веществ, плотность которых (г·см$^{-3}$) предсказывается с большой ошибкой, оказались соединения, содержащие редко встречающиеся в базе данных фрагменты и гетероатомы: селенофенол (Δ = 0.47), иодметилтриметилсилан (Δ = –0.29), 1,4-дииодбутан (Δ = –0.27), *R*(–)-пропиленгликоль (Δ = 0.25), хлорид германия(IV) (Δ = 0.20), 1,2-дитиобис(триметилсилил)этан (Δ = –0.20) и некоторые другие. На рисунке 6 приведено распределение ошибок прогноза значений плотности для ряда жидких органических со-

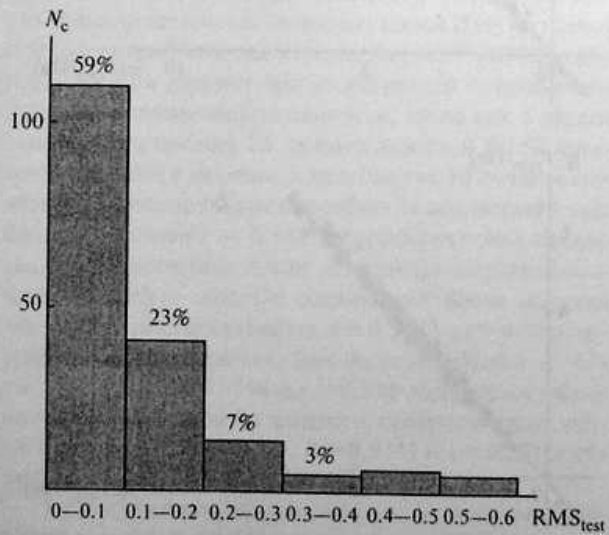

Рис. 5. Распределение ошибок прогноза (RMS_{test}) для вязкости органических соединений (lg(η/Па·с)); N_c — число соединений.

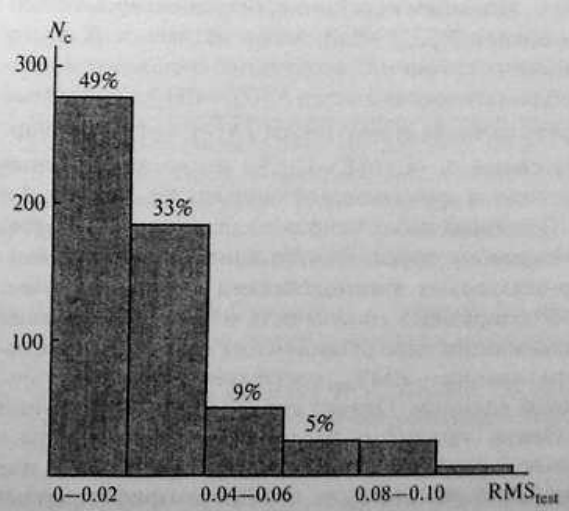

Рис. 6. Распределение ошибок прогноза (RMS_{test}) для плотности жидких органических соединений (d^{20}/г·см$^{-3}$); N_c — число соединений.

3–4–4691

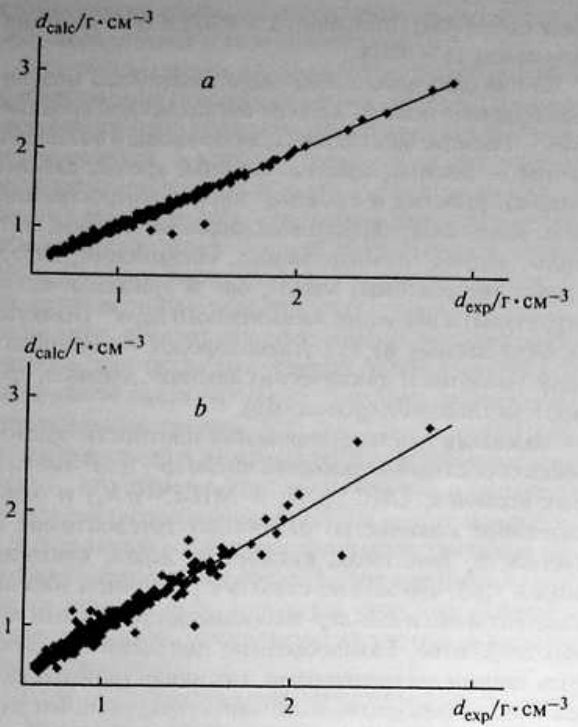

Рис. 7. Результаты моделирования плотности (d^{20}/г·см$^{-3}$): обучающая выборка (*a*), выборка для оценки предсказательной способности (*b*).

единений по полученному ансамблю из 25 лучших моделей. Корреляция усредненных по всему массиву моделей расчетных данных с экспериментальными значениями для плотности жидких органических соединений представлена на рис. 7.

При моделировании давления насыщенного пара органических соединений среди наиболее значимых дескрипторов, присутствующих практически во всех моделях, оказались квадрат числа углеродных атомов N^2(C); логарифм общего количества неводородных атомов $\lg N_a$; число атомов галогена, связанных с атомом C, входящим в состав шестичленных ароматических циклов N[C_{Ar}—Hal]; число метиленовых групп, связанных с атомом C, входящим в состав шестичленных ароматических циклов N[C_{Ar}—CH_2]; квадратный корень из числа атомов фтора $\sqrt{N[F]}$; число ординарных связей C—C N(C—C)/N_a; число двухатомных фрагментов ароматической системы N[—C_{Ar}—C_{Ar}] и др. Подобный набор наиболее важных дескрипторов, по-видимому, обусловлен доминирующей ролью вандер-ваальсовых взаимодействий в данном случае. Прогнозирующая способность модели для давления насыщенного пара органических соединений достаточно высока — RMS_{test} составляет 0.152 логарифмической единицы. Однако для отдельных соединений значения логарифма давления предсказываются с большой ошибкой: для иодбензола (Δ = –0.81), дибромдифторметана (Δ = –0.56), бромтрифторметана (Δ = 0.42) и 1,1,2-трифторэтана (Δ = –0.38). Среди таких соединений оказались в основном непредельные и полярные соединения, содержащие атомы галогенов. Кроме того, для отмеченных структур большую роль играют пространственные эффекты, которые не всегда удается учесть при помощи фрагментных дескрипторов. На рисунке 8 представлено распределение ошибок прогноза давления насыщенного пара углеводородов и галогенуглеводородов, усредненных по 25 наилучшим моделям.

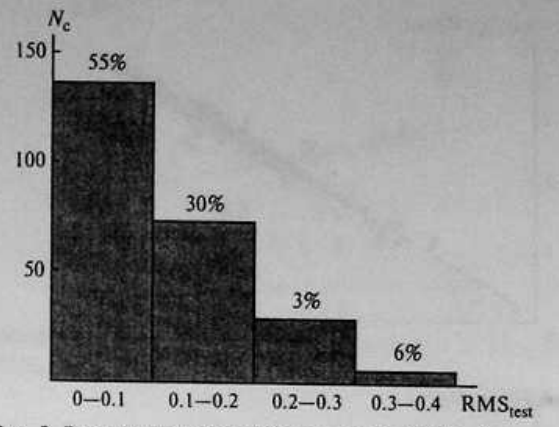

Рис. 8. Распределение ошибок прогноза (RMS_{test}) для давления насыщенного пара углеводородов и галогенуглеводородов (P_{sat}/Па).

Корреляция усредненных по всему массиву моделей расчетных данных для давления насыщенного пара по всем выборкам с экспериментальными значениями приведена на рис. 9.

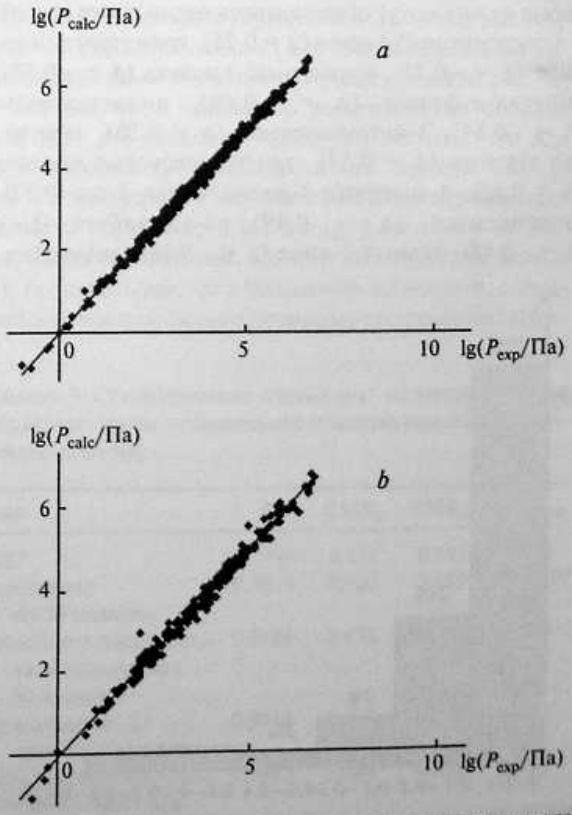

Рис. 9. Результаты моделирования давления насыщенного пара (P_{sat}/Па): обучающая выборка (*a*), выборка для оценки предсказательной способности (*b*).

Таблица 4. Параметры[a] нейросетевых и линейно-регрессионных моделей (числа в скобках) для ряда физико-химических свойств

Свойство	N_c[b]	n_d[c]	R_{av}	RMS_{tr}	RMS_{val}	RMS_{test}	Типы фрагментов[22]
Вязкость[58], lg(η/Па·с)	367	46±16	0.9885 (0.9794)	0.084 (0.111)	0.104 (0.195)	0.141 (0.212)	p1, p2, p3
Плотность[65], d^{20}/г·см$^{-3}$	803	69±21	0.9980 (0.9885)	0.021 (0.038)	0.046 (0.055)	0.051 (0.067)	p1, p2, p3, p4, p5, c4, c5, s4, s5
Давление насыщенного пара[66], lg(P_{sat}/Па)	352	56±9	0.9971 (0.9902)	0.090 (0.198)	0.122 (0.248)	0.152 (0.258)	p1, p2

[a] Обозначения см. в табл. 1. [b] Число соединений. [c] Среднее число отобранных дескрипторов.

В сводной таблице 4 представлены результирующие статистические параметры полученных линейно-регрессионных и нейросетевых моделей для указанных выше физико-химических свойств. Прогнозирующая способность нейросетевых моделей (которую наиболее корректно оценивать по значению RMS_{test}, т.е. по среднеквадратичной ошибке для выборки для оценки предсказательной способности) превосходит аналогичные показатели регрессионных моделей. Кроме того, построенные нейросетевые модели по ряду показателей превосходят лучшие из опубликованных моделей. В частности, результаты прогнозирования плотности жидкостей для соединений различных классов оказались близки к наилучшей из опубликованных моделей[69]: для 303 соединений $R = 0.9874$ и $s = 0.0458$. Однако наша модель построена по значительно более представительной выборке. Полученная нами модель для предсказания вязкости (Па·с) жидких органических соединений значительно превосходит по всем показателям наилучшие из опубликованных моделей[61,62]. В частности, модель Иванчиуча[61] для 337 соединений дает среднеквадратичную ошибку скользящего контроля 0.38 логарифмической единицы, а модель Катрицкого[62] для 361 соединения дает стандартное отклонение 0.22 логарифмической единицы (в нашей модели 0.14 для выборки для оценки прогноза). Точность предсказания давления насыщенных паров (Па) по нашей модели лучше, чем для модели Джурса[68] (RMS_{test} для выборки для оценки прогнозирующей способности 0.209 логарифмической единицы, тогда как в нашей модели при оценке 25 лучших моделей 0.152 логарифмической единицы, а при оценке 10 лучших моделей прогнозирующая способность возрастает в еще большей степени — 0.142 логарифмической единицы) и существенно лучше остальных опубликованных моделей — для 476 соединений Басак получил модель с характеристиками $R = 0.9182$ и $s = 0.29$ логарифмической единицы, для модели Лайанга из 479 соединений $R = 0.9798$ и $s = 0.534$ логарифмической единицы и в модели Катрицкого, построенной по данным для 411 соединений, $R = 0.9742$ и $s = 0.331$ логарифмической единицы[69].

Таким образом, использование фрагментных дескрипторов в сочетании с аппаратом ИНС позволяет получать высокоточные модели для прогнозирования ряда физико-химических свойств органических и элементоорганических соединений с учетом лишь топологических параметров.

Список литературы

1. L. Pogliani, *Chem. Rev.*, 2000, **100**, 3827.
2. H. J. Bernstein, *J. Chem. Phys.*, 1952, **20**, 263.
3. H. J. Bernstein, *Trans. Faraday Soc.*, 1962, **58**, 2285.
4. S. W. Benson, F. R. Cruickshank, D. M. Golden, G. R. Haugen, H. E. O'Neal, A. S. Rodgers, R. Shaw, and R. Walsh, *Chem. Rev.*, 1969, **69**, 279.
5. В. М. Татевский, *Теория физико-химических свойств молекул и веществ*, МГУ, Москва, 1987, 239 с.
6. Е. А. Смоленский, *Журн. физ. хим.*, 1964, **38**, 1288 [*J. Phys. Chem. USSR*, 1964, **38** (Engl. Transl.)].
7. Е. А. Смоленский, *Докл. АН СССР*, 1976, **230**, 373 [*Dokl. Chem.*, 1976 (Engl. Transl.)].
8. C. T. Zahn, *J. Chem. Phys.*, 1934, **2**, 671.
9. M. Sounders, Jr., C. S. Matthews, and C. O. Hurd, *Ind. Eng. Chem.*, 1949, **41**, 1037.
10. M. Sounders, Jr., C. S. Matthews, and C. O. Hurd, *Ind. Eng. Chem.*, 1949, **41**, 1048.
11. J. L. Franklin, *Ind. Eng. Chem.*, 1949, **41**, 1070.
12. J. L. Franklin, *J. Chem. Phys.*, 1953, **21**, 2029.
13. T. L. Allen, *J. Chem. Phys.*, 1959, **31**, 1039.
14. A. J. Kalb, A. L. H. Chung, and T. L. Allen, *J. Am. Chem. Soc.*, 1966, **88**, 2938.
15. В. М. Татевский, *Химическое строение углеводородов и закономерности в их физико-химических свойствах*, МГУ, Москва, 1953, 320 с.
16. В. М. Татевский, В. А. Бендерский, С. С. Яровой, *Закономерности и методы расчета физико-химических свойств парафиновых углеводородов*, МГУ, Москва, 1960, 114 с.
17. В. М. Татевский, *Классическая теория строения молекул и квантовая механика*, Химия, Москва, 1973, 516 с.
18. Е. А. Смоленский, Л. В. Кочарова, *Докл. АН СССР*, 1982, **264**, 112 [*Dokl. Chem.*, 1982 (Engl. Transl.)].
19. Д. Е. Петелин, В. А. Палюлин, Н. С. Зефиров, Дж. У. МакФарланд, *Докл. АН*, 1992, **327**, 508 [*Dokl. Chem.*, 1992 (Engl. Transl.)].
20. И. И. Баскин, В. А. Палюлин, Н. С. Зефиров, *Тез. докл. конф. «Молекулярные графы в химических исследованиях»*, Калинин, 1990, 5.
21. И. И. Баскин, В. А. Палюлин, Н. С. Зефиров, *Тез. докл. 1-й Всесоюз. конф. по теоретической органической химии*, Волгоград, 1991, 557.
22. Н. В. Артеменко, И. И. Баскин, В. А. Палюлин, Н. С. Зефиров, *Докл. АН*, 2001, **381**, 203 [*Dokl. Chem.*, 2001 (Engl. Transl.)].
23. И. И. Баскин, Н. М. Гальберштам, В. А. Палюлин, Н. С. Зефиров, *Тр. VII Всерос. конф. «Нейрокомпьютеры и их*

применение» НКП-2001 с международным участием, под ред. А. И. Галушкина, Ин-т проблем управления им. В. А. Трапезникова РАН, Москва, 2001, 419.
24. М. И. Кумсков, Л. А. Пономарева, Е. А. Смоленский, Д. Ф. Митюшев, Н. С. Зефиров, *Изв. АН, Сер. хим.*, 1994, 1391 [*Russ. Chem. Bull.*, 1994, **43**, 1317 (Engl. Transl.)].
25. S. M. Free and J. W. Wilson, *J. Med. Chem.*, 1964, **7**, 395.
26. A. Cammarata, *J. Med. Chem.*, 1972, **15**, 573.
27. T. Fujita and T. Ban, *J. Med. Chem*, 1971, **14**, 148.
28. В. В. Авидон, *Хим.-фарм. журн.*, 1974, **8**, 22 [*Pharm. Chem. J.*, 1974, **8** (Engl. Transl.)].
29. В. В. Авидон, В. С. Аролович, С. П. Козлова, Л. А. Пирузян, *Хим.-фарм. журн.*, 1978, **12**, 88 [*Pharm. Chem. J.*, 1978, **12** (Engl. Transl.)].
30. V. V. Avidon, I. A. Pomerantsev, A. B. Rozenblit, and V. E. Golender, *J. Chem. Inf. Comp. Sci.*, 1982, **22**, 207.
31. A. B. Rozenblit and V. E. Golender, *Logical Combinatorial Algorithms for Drug Design, Research Studies Press*, Wiley and Sons, New York—Chihester—Brisbane—Toronto, 1983, 352 pp.
32. Yu. V. Borodina, D. A. Filimonov, and V. V. Poroikov, *Pharm. Chem. J.*, 1996, **30**, 760.
33. D. Filimonov, V. Poroikov, Yu. Borodina, and T. Gloriozova, *J. Chem. Inf. Comput. Sci.*, 1999, **39**, 666.
34. V. V. Poroikov, D. A. Filimonov, Yu. V. Borodina, A. A. Lagunin, and A. Kos, *J. Chem. Inf. Comput. Sci.*, 2000, **40**, 1349.
35. С. К. Котовская, Л. А. Тюрина, Е. Ю. Чернова, Г. А. Мокрушина, О. Н. Чупахин, А. П. Новикова, В. И. Ильенко, *Хим.-фарм. журн.*, 1989, **22**, 310 [*Pharm. Chem. J.*, 1989, **22** (Engl. Transl.)].
36. G. Klopman, *J. Am. Chem. Soc.*, 1984, **106**, 7315.
37. G. Klopman and H. S. Rosenkranz, *Mutat. Res.*, 1994, **305**, 33.
38. A. R. Cunningham, G. Klopman, and H. S. Rosenkranz, *Mutat. Res.*, 1998, **405**, 9.
39. M. Brinn, P. T. Walsh, M. P. Payne, and B. Bott, *SAR QSAR Environ. Res.*, 1993, 1, 169.
40. M. W. Brinn, M. P. Payne, and P. T. Walsh, *Chem. Eng. Res. Des.*, 1993, **71**(A3), 337.
41. F. R. Burden, *Quant. Struct.-Act. Relat.*, 1996, **15**, 7.
42. T. Hurst and T. Heritage, *Thes. 213th ACS Natl. Meeting*, San Francisco, CA, 1997, CINF 019.
43. A. Hoskuldsson, *J. Chemometrics*, 1988, **2**, 211.
44. N. S. Zefirov and V. A. Palyulin, *J. Chem. Inf. Comput. Sci.*, 2002, **42**, 1112.
45. И. И. Баскин, А. О. Айт, Н. М. Гальберштам, В. А. Палюлин, М. В. Алфимов, Н. С. Зефиров, *Докл. АН*, 1997, **357**, 57 [*Dokl. Chem.*, 1997 (Engl. Transl.)].
46. J. Zupan and J. Gasteiger, *Neural Networks for Chemists. An Introduction*, Wiley—VCH Publishers, Weinheim—New York—Chichester—Brisbane—Singapore—Toronto, 1993, **1**, 244 pp.
47. И. И. Баскин, В. А. Палюлин, Н. С. Зефиров, *Докл. АН*, 1993, **332**, 713 [*Dokl. Chem.*, 1993 (Engl. Transl.)].
48. А. А. Ежов, С. А. Шумский, *Нейрокомпьютер и его применение в экономике*, МИФИ, Москва, 1998, 57.
49. M. Chastrette, D. Cretin, and F. Tiyal, *C. R. Acad. Sci.*, 1994, **318**, 1059.
50. A. Bondi, *J. Phys. Chem.*, 1964, **68**, 441.
51. S. C. Basak, B. D. Gute, and G. D. Grunwald, *J. Chem. Inf. Comput. Sci.*, 1997, **37**, 651.
52. C. Liang and D. A. Gallagher, *J. Chem. Inf. Comput. Sci.*, 1998, **38**, 321.
53. M. Karelson and A. Perkson, *Comput. Chem.*, 1999, **23**, 49.
54. A. A. Gakh, E. G. Gakh, B. G. Stumper, and D. W. Noid, *J. Chem. Inf. Comput. Sci.*, 1994, **34**, 832.
55. R. Zhang, S. Liu, M. Liu, and Z. Hu, *Comput. Chem.*, 1997, **21**, 335.
56. S. Liu, R. Zhang, M. Liu, and Z. Hu, *J. Chem. Inf. Comput. Sci.*, 1997, **37**, 1146.
57. K. G. Joback and R. C. Reid, *Chem. Eng. Commun.*, 1987, **57**, 233.
58. P. Škubla, *Collect. Czech. Chem. Commun.*, 1985, **50**, 1907.
59. C. Hansch and T. Fujita, *J. Am. Chem. Soc.*, 1964, **86**, 1616.
60. T. Fujita, J. Iwasa, and C. Hansch, *J. Am. Chem. Soc.*, 1964, **86**, 5175.
61. O. Ivanciuc, T. Ivanciuc, P. A. Filip, and D. Cabrol-Bass, *J. Chem. Inf. Comput. Sci.*, 1999, **39**, 515.
62. A. R. Katritzky, K. Chen, Y. L. Wang, M. Karelson, B. Lucic, N. Trinajstic, T. Suzuki, and G. Schuurmann, *J. Phys. Org. Chem.*, 2000, **13**, 80.
63. T. Suzuki, K. Ohtaguchi, and K. Koide, *Comput. Chem. Eng.*, 1996, **20**, 161.
64. T. Suzuki, R.-U. Ebert, and G. Schuurmann, *J. Chem. Inf. Comput. Sci.*, 1997, **37**, 1122.
65. В. Е. Вапник, А. Я. Червоненкис, *Теория распознавания образов*, Наука, Москва, 1979, 237 с.
66. J. Rissanen, in *Complexity, Entropy and the Physics of Information*, Ed. W. H. Zurek, Addison-Wesley, California, Redwood City, 1990, 117.
67. *Flukalog Database*, Fluka Chemie AG, 1995.
68. E. S. Goll and P. C. Jurs, *J. Chem. Inf. Comput. Sci.*, 1999, **39**, 1081.
69. A. R. Katritzky, U. Maran, V. S. Lobanov, and M. Karelson, *J. Chem. Inf. Comput. Sci.*, 2000, **40**, 1.

Поступила в редакцию 26 февраля 2002;
после доработки — 23 мая 2002