

Нейронные сети как метод поиска зависимостей структура – свойство органических соединений

Н.М.Гальберштам, И.И.Баскин, В.А.Палюлин, Н.С.Зефиров

Институт органической химии им. Н.Д.Зелинского Российской академии наук

119991 Москва, Ленинский просп., 47, факс (095) 135–5328

Московский государственный университет им. М.В.Ломоносова, Химический факультет

119992 Москва, Ленинские горы, факс (095) 939–3181

Рассмотрены работы, в которых нейросетевой подход применяется для моделирования зависимостей свойств органических соединений от их структуры. Наряду с основными принципами нейросетевого моделирования обсуждаются также особенности нейросетевого подхода, характерные для представления и классификации структурных химических данных. Приведена краткая информация о нейросетевых моделях спектральных характеристик, реакционной способности, физико-химических свойств и биологической активности органических соединений.

Библиография — 159 ссылок.

Оглавление

I. Введение	706
II. Основные принципы нейросетевого моделирования	707
III. Способы описания структур исследуемых соединений	710
IV. Применение нейросетевых алгоритмов для поиска зависимостей структура – свойство	725

I. Введение

Применение вычислительной техники практически во всех областях науки в последние десятилетия приобрело поистине массовый характер. Богатейший объем экспериментальных данных, накопленный к настоящему времени, позволяет уделять особое внимание методам обобщения и математической обработки характеристик уже исследованных веществ с целью компьютерного моделирования и предсказания свойств новых, еще не синтезированных либо не изученных соединений. Это в свою очередь открывает широкие возможности для решения одной из главных задач химической

науки — целенаправленного поиска новых веществ и материалов с заранее заданными свойствами.

Искусственные нейронные сети становятся одним из наиболее широко применяемых методов построения разнообразных количественных зависимостей. Благодаря возможности построения нелинейных моделей любого уровня сложности методология искусственных нейронных сетей нашла широкое применение в рамках поиска зависимостей между структурами органических соединений и их физико-химическими свойствами (QSPR) или биологической активностью (QSAR).

Среди известных архитектур нейросетей, используемых для поиска зависимостей структура – свойство (здесь и далее под термином «свойство» будем понимать как физико-химические свойства органических соединений, так и их биологическую активность), первое место по популярности занимает многослойная нейросеть прямого распространения, обучающаяся по методу обратного распространения ошибки. Ее привлекательность связана в первую очередь со способностью нейросетей этого типа обобщать и аппроксимировать данные с высокой точностью, а также с возможностью обрабатывать с их помощью большие массивы разрозненной информации.

Поиск нейросетевых зависимостей структура – свойство можно свести к последовательному выполнению следующих операций: 1) структуры исследуемых соединений описывают некими числовыми параметрами — дескрипторами; 2) выбирают оптимальный набор дескрипторов; 3) весь массив экспериментальных данных разбивают на обучающую и контрольную выборки; 4) выбирают наиболее подходящий тип, архитектуру и метод обучения нейросети; 5) выбирают статистические параметры для оценки качества обучения

Н.М.Гальберштам. Кандидат химических наук, научный сотрудник лаборатории математической химии и компьютерного синтеза ИОХ РАН. Телефон: (095)939–3557, e-mail: n.halb@g23.relcom.ru

И.И.Баскин. Кандидат химических наук, старший научный сотрудник лаборатории органического синтеза химического факультета МГУ. Телефон: (095)939–3557, e-mail: baskin@org.chem.msu.ru

В.А.Палюлин. Кандидат химических наук, ведущий научный сотрудник той же лаборатории. Телефон: (095)939–3557, e-mail: vap@org.chem.msu.ru

Н.С.Зефиров. Академик, заведующий кафедрой органической химии химического факультета МГУ. Телефон: (095)939–1620, e-mail: zefirov@org.chem.msu.ru

Область научных интересов авторов: органическая химия, поиск количественных соотношений структура – свойство и структура – активность органических соединений, искусственные нейронные сети.

Дата поступления 23 июля 2002 г.

нейросети; 6) производят обучение нейросети и оценивают результаты этого обучения; 7) полученную нейросетевую модель используют для прогноза свойств и активностей неизвестных соединений.

Первые работы, посвященные использованию нейронных сетей (или перцептронов) для решения химических задач, появились в конце 1980-х годов.¹ С тех пор интерес химиков к применению нового метода начал стремительно расти.

Нейронные сети, часто также называемые искусственными нейронными сетями или вычислительными нейронными сетями, представляют собой упрощенную математическую модель обработки информации головным мозгом человека. Однако большинство современных архитектур нейронных сетей не воспроизводит в точности функции мозга, и их следует рассматривать как разновидность параллельных алгоритмов.^{2,3}

Благодаря своей способности обучаться, объединять и обобщать данные нейросети начали успешно применяться в химии, особенно в тех случаях, когда неизвестен аналитический вид зависимости между структурой и свойствами соединений.⁴

II. Основные принципы нейросетевого моделирования

1. Описание архитектур и методов обучения нейросетей, используемых для нахождения зависимостей структура – свойство

Нейронные сети, как правило, состоят из набора нескольких слоев из относительно простых, но соединенных между собой большим количеством связей вычислительных элементов, называемых нейронами. Слои обычно подразделяют на три группы: входной, скрытый и выходной (рис. 1). Каждая связь между двумя нейронами характеризуется действительным числом, называемым весом связи. Совокупность весовых коэффициентов связей между нейронами определяет вычислительные свойства нейросети, и обучение нейросети сводится к соответствующей настройке весовых коэффициентов.

Внутри каждого i -ого нейрона n -ого слоя выполняются двухстадийные вычисления (см. рис. 1): сначала вычисляется

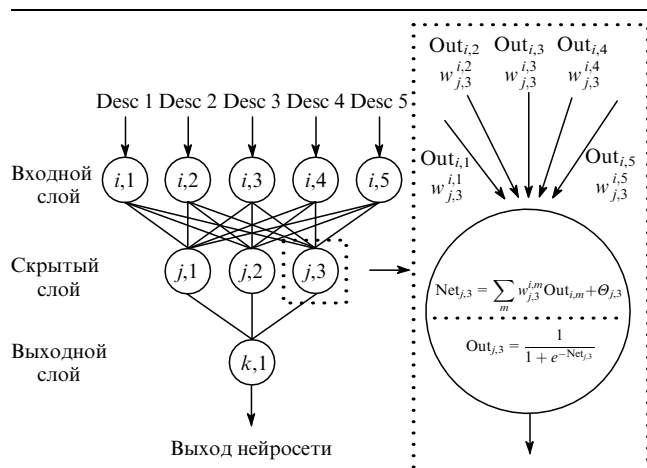


Рис. 1. Структура нейросети прямого распространения. В увеличенном фрагменте показаны основные вычислительные операции, выполняемые нейроном с сигмоидной передаточной функцией. Desc — дескриптор.

общий входной сигнал Net_j нейрона путем суммирования всех взвешенных входных сигналов, а затем с помощью передаточной функции вычисляется выходной сигнал Out_j этого нейрона.

$$Net_j = \sum_i w_{ij} Out_i + \theta_j,$$

$$Out_j = f(Net_j),$$

где Net_j — общий входной сигнал нейрона j , принадлежащего слою n ; Out_i — выходной сигнал нейрона i , принадлежащего слою $n - 1$; w_{ij} — весовой коэффициент связи между нейронами i и j ; θ_j — пороговая величина для нейрона j ; Out_j — выходной сигнал нейрона j слоя n .

В качестве передаточных чаще всего применяются сигмоидные функции (1); но, кроме того, могут использоваться пороговые (2), линейные (3), гиперболические тангенциальные (4) и ряд других функций.^{5,6}

$$Out_j = \frac{1}{1 + e^{-Net_j}}, \quad (1)$$

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

$$Out_j = k Net_j \quad (k = \text{const}), \quad (3)$$

$$Out_j = \frac{e^{Net_j} - e^{-Net_j}}{e^{Net_j} + e^{-Net_j}}. \quad (4)$$

Как правило, вместо использования пороговых величин θ в каждый слой нейросети (за исключением выходного слоя) добавляют так называемые псевдонейроны смещения² с постоянным выходным сигналом, равным 1.

Все веса связей перед началом обучения инициализируют случайными числами. Правильный выбор границ инициализации может сократить время обучения нейросети и улучшить качество получаемых нейросетевых моделей.^{7,8} В отдельных случаях, когда процесс обучения (т.е. настройки весов связей) не сходится, обучение повторяют при других параметрах обучения либо с другой инициализацией весов связей.

Архитектура нейронной сети определяется топологией соединения нейронов между собой. Для моделирования зависимостей структура – свойство, как правило, применяют нейросети, состоящие из нескольких слоев, число которых диктуется конкретной архитектурой нейросети.^{2,9} Для всех нейронов, принадлежащих одному слою, характерно одинаковое число входных связей, соединяющих нейрон с предыдущим слоем. Рассчитанными выходными значениями всей нейросети являются выходные сигналы последнего вычислительного слоя сети. При применении нейросетей для прогнозирования свойств химических соединений значения выходных сигналов входных нейронов устанавливаются по значениям рассчитанных дескрипторов (после нормировки, см. ниже), а с выходных нейронов снимаются сигналы, отождествляемые со значениями прогнозируемых свойств (тоже с учетом нормировки).

а. Многослойные нейросети прямого распространения

В большинстве публикаций, посвященных поиску зависимостей структура – свойство, используют многослойные нейросети прямого распространения. К основным достоинствам таких нейросетей можно отнести их способность находить многопараметровые нелинейные зависимости, характеризующиеся высокой точностью интерполяции даже в тех случаях, когда экспериментальные данные недостаточно репрезентативны или содержат шум.¹⁰

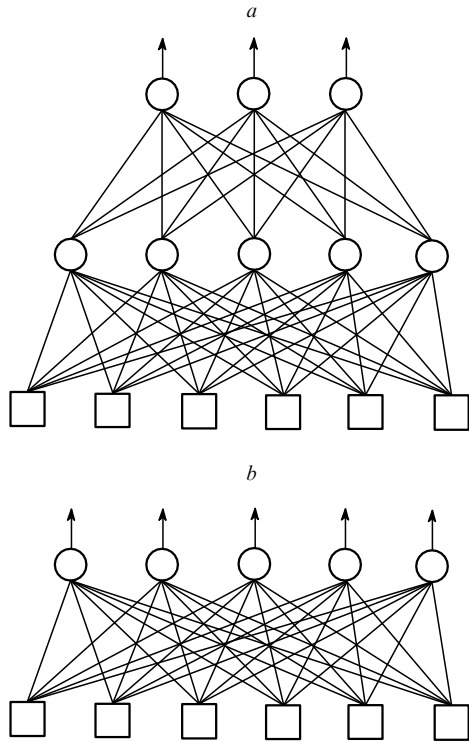


Рис. 2. Классические нейросети прямого распространения с одним скрытым слоем (а) и без скрытых слоев (б).

Для нейросетей прямого распространения характерна послойная передача сигнала от входа нейросети к ее выходу. Классические нейросети прямого распространения могут содержать несколько скрытых слоев (рис. 2,а) или состоять только из входного и выходного слоев (рис. 2,б).

При обучении многослойных нейросетей прямого распространения настройка весовых коэффициентов связей проводится последовательно, начиная со связей выходного слоя, поэтому методы обучения таких нейросетей носят название методов обратного распространения ошибки.^{11, 12} Ниже приведено краткое описание наиболее широко применяемых методов.

б. Обобщенное дельта-правило

Для каждого слоя n из N слоев нейросети обучение выполняется по правилу Уидроу – Хоффа или дельта-правилу^{13, 14}

$$\Delta w_{ij} = \eta \delta_j o_i,$$

где Δw_{ij} — изменение весового коэффициента связи между нейронами i и j ; η — эмпирическая константа, характеризующая скорость обучения; δ_j — ошибка вычислений для нейрона j , принадлежащего слою n ; o_i — выходной сигнал нейрона i , принадлежащего слою $n - 1$.

Для того чтобы избежать осцилляции, часто возникающей, если поверхность ошибок характеризуется очень узкой областью минимума, в случае обобщенного дельта-правила вводят дополнительную величину, называемую моментом обучения и рассчитываемую с учетом изменения весовых коэффициентов на предыдущей итерации^{13–15}

$$\Delta w_{ij}^{(k+1)} = \eta \delta_j o_i + \alpha \Delta w_{ij}^{(k)},$$

$$\delta_j = \begin{cases} f'(\text{Net}_j)(t_j - o_j), & n = N \\ f'(\text{Net}_j) \sum_m \delta_m w_{jm}, & n \neq N, \end{cases}$$

где t_j и o_j — экспериментальное и расчетное значения исследуемого свойства, соответствующие нейрону j ; α — константа, характеризующая влияние момента; k — номер итерации.

в. Метод гибкого распространения (RPROP)

При обучении по методу гибкого распространения для подстройки весовых коэффициентов используется только информация о знаках частных производных функции ошибки нейросети E (обычно сумма квадратов ошибок на выходных нейронах).^{16, 17} Величина, на которую изменяются весовые коэффициенты, вычисляется следующим образом:

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

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

Здесь $\frac{\partial E^{(t)}}{\partial w_{ij}}$ — значение частной производной функции ошибки нейросети по весу связи w_{ij} на i -ой итерации, η^+ и η^- — эмпирические константы, характеризующие увеличение или уменьшение скорости обучения; t — номер итерации.

г. Квази-ньютоновские методы обучения

Эта группа методов базируется на ньютоновском методе аппроксимации функций. На каждой итерации вычисляется вектор, компоненты которого равны частным производным функции ошибки нейросети по всем весовым коэффициентам. Наиболее удачным методом из этой группы является метод Бройдена – Флетчера – Гольдфарба – Шанно (BFGS).^{14, 18}

2. Другие архитектуры нейросетей

Для моделирования зависимостей структура – свойство с успехом применяют нейросети и с другими архитектурами. Ниже приведено краткое описание некоторых таких архитектур.

Для нейросетей прямого распространения, содержащих хотя бы один скрытый слой, возможно наличие дополнительных прямых связей, соединяющих нейроны входного и выходного слоя,^{19–21} а также рекурсивных связей между выходом нейрона и его входом.^{21, 22}

Для задач, связанных с классификацией соединений, Кваснича с соавт.^{23, 24} предложил использовать нейросеть в форме дерева, причем топология этой нейросети была мак-

симально приближена к структурам изучаемых молекул. Входным нейронам приписывались значения 0 или 1 таким образом, чтобы получать подграфы, изоморфные структурам изучаемых молекул. Позднее подобный подход был использован в компьютерной программе «ChemNet».²⁵

Предложен^{26,27} интересный нейросетевой способ описания исследуемых соединений, позволяющий получать зависимости их свойств от структуры без вычисления дескрип-

торов для молекул в целом, а используя лишь дескрипторы атомов и связей. Для представления структур вводится дополнительная нейросеть, состоящая из двух основных блоков: сенсорного поля, на вход которого подаются исходную структурную информацию, и «глаз», преобразующих полученные от сенсорного поля данные в сигнал, инвариантный к перенумерации атомов в молекулах исследуемых соединений. Обработанная таким образом структурная информация подается на вход обычной нейросети прямого распространения для получения зависимостей структура–свойство.

Нейросети Кохонена (см. работы^{28–32}) (рис. 3,а), широко используемые для кластерного анализа, позволяют получить такое отображение исходных данных, при котором близкие вектора входных значений отображаются в близко расположенные на решетке выходные нейроны.

Нейросети встречного распространения^{33,34} используют два различных вида слоев: скрытый слой Кохонена и выходной слой Гроссберга (рис. 3,б). К существенным достоинствам нейросетей, обучающихся по методу встречного распространения, можно отнести сравнительно небольшое (порядка нескольких сот) число итераций, необходимых для обучения нейросети,^{35,36} а также способность находить глобальный минимум функции ошибки при любых стартовых настройках весовых коэффициентов.^{37,38} К недостаткам таких нейросетей следует отнести худшую аппроксимирующую способность по сравнению с рядом других архитектур.

Нейросеть с линейной передаточной функцией для нейронов выходного слоя и скрытыми нейронами, выполняющими радиально-базисную нелинейную аппроксимацию, носит название радиально-базисной нейросети и также используется для нахождения зависимостей структура–свойство.^{39–41}

Работа нечетких ARTMAP-нейросетей (fuzzy adaptive resonance theory for mapping) основана на процедуре кластеризации (категоризации) векторов обучающей выборки на основе теории адаптивного резонанса (см. работу⁴² и ссылки в ней). Процедура категоризации векторов в рамках этого метода основана на сравнении очередного вектора с эталонными векторами, описывающими уже найденные ранее категории (кластеры). Если очередной вектор «похож» по определенному критерию близости на один из эталонных векторов, то он используется для его подстройки, в противном же случае он сам объявляется представителем новой категории данных и запоминается в виде нового эталонного вектора. Описанная процедура реализуется в виде нейросети, состоящей из слоя сравнения, оценивающего «сходство» векторов, слоя распознавания, каждый нейрон которого описывает свою категорию (кластер) данных, и нескольких дополнительных элементов. Подобная архитектура получила название ART-1 для категоризации бинарных векторов и ART-2 для категоризации векторов вещественных чисел. ARTMAP представляет собой модульную нейросеть, состоящую из двух сетей типа ART для категоризации векторов (в случае QSAR/QSPR эти векторы связаны с векторами, соответствующими дескрипторам и свойствам органических соединений), и модуля сравнения, в котором происходит запоминание «ассоциаций» между категориями дескрипторов и свойств.

Нейросеть функциональных связей⁴³ состоит из двух слоев: входного, образованного нейронами, выходы которых представляют собой результаты действия нелинейного преобразования (например, взятие квадрата, квадратного корня, синуса и т.д.) на значение какого-либо дескриптора, и выходного слоя, каждый нейрон которого соответствует какому-либо прогнозируемому свойству. Следует отметить, что эта категория нейросетей представляет собой нейросетевую имитацию линейного регрессионного анализа с использованием нелинейно-преобразованных дескрипторов и свойств.

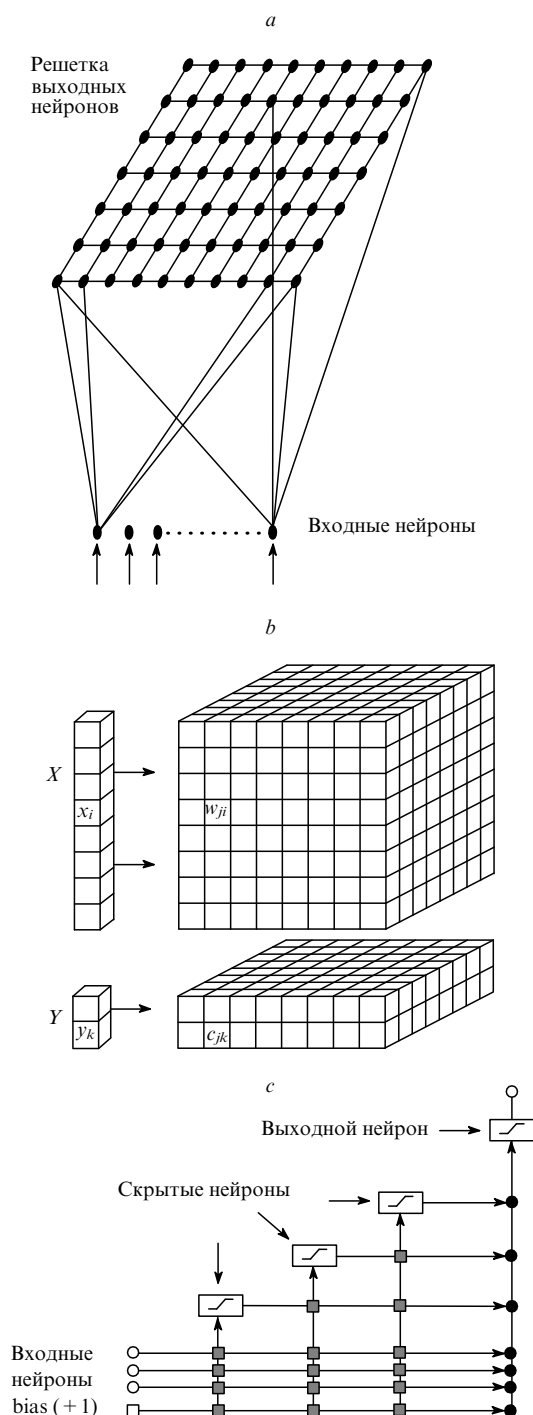


Рис. 3. Условное изображение архитектур некоторых нейросетей, используемых для моделирования зависимостей структура–свойство:

а — нейросеть Кохонена, б — нейросеть встречного распространения, с — каскадная нейросеть.

Иногда в ходе моделирования зависимостей структура – свойство несколько нейросетей различной архитектуры объединяют вместе, получая таким образом иерархическую систему нейросетей.^{21, 44}

Помимо нейросетей со статической архитектурой (ее необходимо выбрать до начала обучения) существуют так называемые каскадные нейросети (рис. 3,с), топология которых формируется в ходе обучения.^{45–47} Обучение начинается с нейросети, содержащей только входные и выходные нейроны. Если ошибка нейросети превышает пороговую величину, после каждого цикла обучения добавляется один скрытый нейрон, в который поступают сигналы от всех входных и ранее добавленных скрытых нейронов (поэтому можно считать, что каждый скрытый нейрон образует отдельный слой). При добавлении очередного скрытого нейрона веса идущих к нему связей определяются только один раз, исходя из требования максимизации ковариации значений выходного сигнала с ошибкой, и больше в процессе обучения не меняются. Что же касается выходного нейрона, то веса идущих к нему связей всякий раз подстраиваются для уменьшения функции ошибки нейросети.

III. Способы описания структур исследуемых соединений

Первый вопрос, который возникает при использовании нейронных сетей, особенно для решения химических задач, — каким способом описывать исходные данные для построения модели.

Совокупность параметров, подаваемых на вход нейросети, как правило, представляют в виде вектора. Поэтому необходимо каким-либо образом описать топологическую или трехмерную структуру молекул в виде одномерного массива без существенных потерь значимой информации. При выборе описания необходимо выполнить следующие условия:³⁶

- 1) количество и тип используемых дескрипторов одинаковы для всех описываемых структур;
- 2) каждому дескриптору соответствует определенная компонента в векторе параметров;
- 3) различным молекулам соответствуют различные вектора параметров.

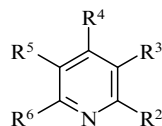
Как правило, при нейросетевом моделировании зависимостей структура – свойство для описания структур соединений используют набор неких чисел — дескрипторов,⁴⁸ в которые тем или иным способом заложена необходимая информация. По типу описываемой информации дескрипторы можно подразделить на фрагментные,⁴⁹ топологические,^{50, 51} квантово-химические⁵² и физико-химические. Для всех этих групп дескрипторов требуется проведение дополнительных вычислительных процедур, предшествующих этапу построения нейросетевой модели. Кроме того, необходимо иметь четкое представление о природе изучаемой зависимости, чтобы выбрать группу и тип дескрипторов, которые целесообразно использовать в каждом конкретном случае.

В большинстве работ, посвященных нейросетевому моделированию зависимостей структура – свойство, для вычисления дескрипторов используют различные программные комплексы (например, ADAPT^{53–55}); в результате обычно получается избыточный набор дескрипторов, из которого при помощи различных процедур отбирают наиболее весомые. Объяснить физический смысл отобранных дескрипторов и проследить их связь с исследуемым свойством часто бывает достаточно трудно.

Попытки применить хорошо известный способ представления химической структуры в виде модифицированной матрицы связности, содержащей на диагоналях заряды ядер соответствующих атомов, а в качестве недиагональных эле-

ментов — формальные порядки связей, были предприняты Элродом с соавт.⁵⁶ Авторы предложили формировать входной вектор для нейросети путем объединения диагонали матрицы связности с поддиагональными элементами, взятыми построчно. Для описания химической реакции авторы предложили применять две матрицы связности — для реагентов и продуктов реакции.⁵⁷ Очевидно, что из-за большого числа элементов в матрице связности такой подход может быть применен только к соединениям, состоящим из небольшого числа атомов. Кроме того, данное представление структур не является инвариантным относительно перенумерации атомов.

Классический подход к построению количественных зависимостей структура – свойство используют для описания серий структурных аналогов с помощью параметров заместителей, таких, например, как σ -константы Гаммета, E_s -константы Тафта, липофильные константы π , параметры L , B_1 , B_2 (STERIMOL) и т.д. Однако если структуры содержат топологически эквивалентные позиции замещения, применение этих параметров становится не вполне корректным. Например, в молекуле пиридина существуют две пары топологически эквивалентных расположений заместителей — в α -положении (заместители R^2 и R^6) и в β -положении (R^3 и R^5), в результате чего одному и тому же соединению будет соответствовать разный набор дескрипторов, основанных на использовании параметров заместителей, что может привести к разным прогнозам.



Для решения этой проблемы был предложен подход,⁵⁸ при котором в исходную выборку добавляют копии соединений с различным расположением заместителей в топологически эквивалентных позициях, обладающие той же активностью, что и исходное соединение. Благодаря такому расширению исходной выборки удается добиться независимости прогноза в случае неоднозначности отнесения заместителей по топологически эквивалентным положениям.

В работе⁵⁹ предложен интересный метод, позволяющий сократить количество входных параметров нейросети практически без потери значимой информации. Подход основан на вычислении главных компонент,⁶⁰ являющихся линейными комбинациями исходных дескрипторов. Полученные таким образом величины используют далее в качестве входных параметров нейросети.

1. Нормировка входных параметров нейросети

Практически во всех работах, посвященных применению нейросетей для решения химических задач, приводят сведения о нормировке входных данных. Как правило, входные параметры нормируют таким образом, чтобы их значения лежали в интервале от 0 до 1 либо характеризовались единичной дисперсией.

2. Выбор наиболее значимых дескрипторов

Не менее важно иметь возможность выбрать из всех рассчитанных параметров только те, которые вносят существенный вклад в изучаемое свойство. Такой выбор можно производить предварительно или в ходе обучения нейросети.

Предварительный отбор параметров может осуществляться как с применением независимых методов, например пошагового множественного регрессионного анализа,^{61, 62} так и с помощью нейросетевых алгоритмов.⁶³ В последнем

случае тип нейросети, используемой для отбора дескрипторов, как правило, не совпадает с типом основной нейросети, применяемой для построения модели.

Одним из примеров может служить нейросеть, использующая «генетический» алгоритм обучения.^{64–66} «Генетический» алгоритм — это стохастический метод оптимизации, который основан на эволюционных принципах. В ходе применения этого алгоритма из разнообразных вариантов описания объектов на протяжении нескольких поколений выбирают то описание, которое включает в себя наиболее значимые параметры. Значимость параметров, т.е. удачность генетического сочетания, определяется с помощью оценочной функции, вычисляемой нейросетью как остаточная погрешность для исследуемого свойства. В литературе описаны разнообразные методы порождения нового поколения в рамках «генетического» алгоритма. Наиболее известные из них — метод аппроксимации генетической функции⁶⁷ и эволюционное программирование.⁶⁸

Бринн с соавт.⁶⁹ предложил использовать так называемое эволюционное приближение для того, чтобы отбросить малозначимые дескрипторы. Из множества рассчитанных дескрипторов случайным образом выбирают некоторую часть. Их подают на вход нейросети, а на выходе нейросети рассчитывают среднеквадратичную ошибку. Затем один из входных нейронов исключается из нейросети, и рассчитывается новая среднеквадратичная ошибка; процедуру повторяют для всех нейронов. Дескрипторы ранжируют по величине изменения ошибки на выходе нейросети. Затем, начиная с самых малозначимых дескрипторов, производят удаление нейронов из нейросети до тех пор, пока ошибка на выходе не начнет существенно изменяться. Все отброшенные дескрипторы заменяются на новые из набора рассчитанных дескрипторов, и вычисления повторяют.

В ходе обучения нейросети малозначимые дескрипторы также могут быть отброшены. Для этого используются так называемые отсекающие алгоритмы (pruning), суть которых заключается в упрощении архитектуры нейросети путем отсева связей с малозначимыми весовыми коэффициентами. Если все связи, исходящие из входного нейрона, оказываются малозначимыми, то такой нейрон может быть отброшен.^{70, 71} Упрощающие методы можно подразделить на две группы:⁷² методы чувствительности, основанные на вычислении значимости или чувствительности всех связей с последующим отсевом связей с наименьшей чувствительностью, и методы штрафного компонента, основанные на введении так называемых штрафных функций, уменьшающих до нулевого значения весовые коэффициенты в ходе обучения.

В работе Тетко с соавт.⁷³ приводится подробное описание пяти методов, используемых для вычисления чувствительности связей. В работах^{74, 75} описан метод отбора наиболее значимых дескрипторов на основе расчета чувствительностей связей и вклада каждого дескриптора в общую ошибку вычислений нейросети.

Приведено описание реконструкционного метода, основанного на процессе забывания, при котором абсолютные величины всех весовых коэффициентов пошагово уменьшаются.⁷⁶ Повторение циклов обучение–забывание приводит к усилению значимых связей и ослаблению малозначимых.

Еще одним примером метода, использующего штрафные функции, может служить алгоритм вырождения весовых коэффициентов.^{13, 20} На каждом шаге обучения рассчитанный весовой коэффициент связи уменьшается на некоторую часть его старой величины так, что в ходе обучения малозначимые весовые коэффициенты уменьшаются до 0.

3. Выбор оптимальной обучающей выборки

Следующей очень важной задачей, возникающей при использовании нейросетей в химии, является составление обучающей выборки. Химические соединения, образующие обучающую выборку, должны равномерно распределяться по всему диапазону возможных значений входных параметров.

Большинство статей содержат данные о случайном разбиении на обучающую и контрольную выборки всего массива исследуемых соединений, и только в нескольких из них приводится описание специальных процедур для подбора оптимальной обучающей выборки. Так, в работах^{35, 77, 78} описано использование нейросетей Кохонена, а в работе⁷⁹ — нейросети на базе теории адаптивного резонанса (ART 2-A, см. выше) для предварительной классификации всего массива экспериментальных данных. Важное значение имеет также подбор правильного размера обучающей выборки.

4. Оптимизация числа скрытых нейронов

Количество скрытых нейронов определяет количество настраиваемых параметров нейросети, т.е. весовых коэффициентов. Слишком большое количество скрытых нейронов, при котором число настраиваемых параметров сопоставимо с числом структур в обучающей выборке, приводит к тому, что нейросеть пытается «запомнить» обучающую выборку (эффект переопределенности модели); нейросеть с недостаточным числом скрытых нейронов будет не в состоянии использовать всю входную информацию для обобщения и классификации.^{80, 81}

В ряде работ^{82–84} была показана целесообразность использования специального параметра r , позволяющего оценить оптимальное количество скрытых нейронов

$$r = \frac{N}{P},$$

где N — количество молекул в обучающей выборке; P — общее число связей в нейросети.

Оптимальные значения этого параметра находятся в интервале $1.8 < r < 2.2$. Рассчитанное таким образом количество скрытых нейронов позволяет построить нейросеть с наилучшей прогнозирующей способностью.⁸⁰

5. Эффект «переучивания» нейросетей

Эффект переопределенности модели (см. выше) приводит к возникновению так называемого «переучивания» нейросети.^{85–87}

Процесс обучения нейросетей можно подразделить на две фазы: обобщение информации и ее запоминание. Для первой фазы характерно уменьшение средней ошибки вычисления исследуемого свойства как для соединений из обучающей выборки, так и для соединений из контрольной выборки. Однако на следующей фазе — фазе запоминания — ошибка, вычисляемая для контрольной выборки, начинает возрастать на фоне продолжающей уменьшаться ошибки для обучающей выборки. Вследствие этого полностью обученная нейросеть хорошо предсказывает свойства лишь для соединений, входящих в обучающую выборку; для остальных соединений качество прогноза резко ухудшается. Этот эффект особенно заметен, когда первая фаза обучения нейросети непродолжительна или вообще отсутствует. «Переучивание» удастся сгладить при уменьшении реального (за счет уменьшения числа входных параметров нейросети) или эффективного (с использованием методов «регуляризации», например байесовой регуляризации^{88, 89}) количества подстраиваемых параметров модели.⁸⁶

Широко распространена также практика преждевременного прерывания обучения нейросети в момент смены фаз обучения. В большинстве опубликованных работ оценку прогнозирующей способности нейросети производили на базе той же контрольной выборки, которая служила для определения момента начала «переучивания». Однако в этом случае ошибка оказывается заниженной, что было продемонстрировано в работе⁹⁰.

6. Анализ активностей скрытых нейронов

Анализ результатов ряда работ, полученных с помощью трехслойной нейросети при моделировании зависимостей структура–свойство, показал, что скрытые нейроны можно использовать для кластерного анализа исходных соединений.

Квасничка с соавт.⁹¹ использовал рекуррентный кластерный анализ для классификации функциональных групп. Каждая функциональная группа представлялась в виде вектора выходных значений скрытых нейронов, который был нормализован для всех исследуемых соединений с помощью стандартных статистических процедур.

В работе³¹ приводится анализ активностей скрытых нейронов трехслойной ассоциативной нейросети для классификации данных из обучающей выборки на примере образцов оливковых масел из девяти различных районов Италии на основе данных о содержании в них восьми жирных кислот. Авторам удалось получить более четкую классификацию по районам производства масел, чем та, что была найдена с помощью нейросети Кохонена.

7. Анализ полученных нейросетевых зависимостей

Чрезвычайно важным аспектом применения практически любого статистического метода (включая нейронные сети) для построения моделей структура–свойство в химии (особенно при использовании физико-химических и квантово-химических дескрипторов) является необходимость содержательной интерпретации получаемых моделей, согласующейся с имеющимися данными из соответствующих областей химии, физики или биологии. Очень часто именно наличие подобной интерпретации является наиболее весомым аргументом в пользу правильности построенной модели. Невозможность выявления четких нейросетевых зависимостей исследуемых свойств от входных параметров, т.е. представление нейросети в виде «черного ящика», до сих пор служит основным аргументом против применения нейронных сетей для решения задач в химии.

В большинстве работ полученные нейросетевые модели никак не интерпретируются авторами, однако возможность анализа нейросетевых зависимостей показана в работах^{92,93}. Была также продемонстрирована^{94–96} возможность оценки вкладов входных параметров в нейросетевые модели для нейросетей с одним выходным нейроном с использованием аналитических формул для вычисления частных производных нейросетевых функций. Для этой цели предложили дополнить стандартную сигмоидную передаточную функцию параметром b , выражающим степень смещения линейных и нелинейных операций.

$$\text{Out}_j = \frac{\beta}{(1 + e^{-\alpha \text{Net}_j})} + (1 - b)\text{Net}_j,$$

где Out_j — выходной сигнал нейрона j ; Net_j — суммарный входной сигнал нейрона j ; α и β — коэффициенты.

В работе⁹⁷ предложено использовать специально рассчитываемые статистические параметры, характеризующие работу обученной нейронной сети. Эти параметры позво-

ляют интерпретировать нейросетевые модели привычным способом, аналогичным способам интерпретации линейных регрессионных моделей. Действительно, значимость того или иного параметра, входящего в линейное регрессионное уравнение, определяется величиной и знаком соответствующего этому параметру числового коэффициента, который в свою очередь равен величине частной производной, взятой по данному параметру от регрессионной функции, а коэффициенты вычисляются в ходе регрессионного анализа на множестве всех представленных экспериментальных данных. Таким образом, чтобы получить аналогичную интерпретацию нейросетевой модели, необходимо вычислить значения частных производных исследуемых свойств последовательно по значениям всех исходных дескрипторов на множестве примеров из обучающей выборки. Таким способом вычисляется набор статистических функций, позволяющий оценить вклад всех исходных дескрипторов в построенные нейросетевые модели.

8. Выбор функции для оценки погрешности нейросетевых расчетов

Для оценки получаемых нейросетевых моделей используют, как правило, стандартные статистические функции. Для этого вычисляют, например, среднеквадратичные ошибки,^{61,98} коэффициенты корреляции,^{36,99} коэффициенты распознавания^{43,100} и т.п.

Для оценки прогнозирующей способности нейросетевых моделей часто используют процедуру скользящего контроля (см., например, работы^{35,82,101}).

9. Сравнение нейросетевого алгоритма со стандартными статистическими методами обработки информации

Во многих случаях, для того чтобы сделать выводы об адекватности использования нейросетевого моделирования, сравнивают модели, полученные с помощью нейросетевых алгоритмов, с моделями, полученными другими методами.

Так, результаты нейросетевой классификации соединений сравнивают с результатами иерархического кластерного анализа, анализа главных компонент,³² с результатами PLS (Partial Least Squares) в рамках метода CoMFA (Comparative Molecular Field Analysis),⁹ с химическими экспертными системами⁵⁶ и т.д.

Нейросетевые модели зависимостей структура–свойство чаще всего сравнивают с уравнениями, полученными с помощью множественной линейной регрессии.^{77,102,103} В большинстве работ сделан вывод о том, что с помощью нейросетевых методов можно получить лучшие результаты как при описании зависимостей структура–свойство, так и для прогнозирования свойств новых соединений. Однако в работе¹⁰⁴ сделан вывод о слабой прогнозирующей способности нейросетей по сравнению с множественной линейной регрессией, что может быть объяснено неудачным выбором параметров нейросетевой модели.

Как было отмечено в работах^{95,105}, существует связь между нейросетевым моделированием и множественным регрессионным анализом. Показано, что результаты, полученные с помощью нейросети с одним скрытым слоем, которая использует линейную передаточную функцию для всех нейронов, идентичны результатам множественной линейной регрессии.

Если нейроны скрытого слоя трехслойной нейросети характеризуются нелинейной передаточной функцией (чаще всего сигмоидной), а нейрон выходного слоя — линейной, то полученная нейросеть, называемая MR-нейросетью, будет работать подобно нелинейной множественной регрессии.^{96,106,107}

Таблица 1. Характеристики нейросетевых моделей, полученных при моделировании зависимостей структура – спектральные свойства и структура – индексы удерживания в хроматографии.

Моделируемое свойство	Исследуемые соединения	Количество соединений	Дескрипторы	Архитектура нейросетей		Результаты обучения нейросети	Статистические характеристики обучения нейросети ^a	Ссылки
				количество нейронов	тип нейросети ^b			
ИК-спектры	Монозамещенные бензолы	185	Топологические	32 – (10 × 10) – 128	A	Для разных соединений	$R = 0.42 - 0.94$	99
	Алифатические соединения, содержащие от 4 до 100 атомов углерода	200	Рассчитанные ИК-спектры (метод РМЗ)	254 – (от 76 до 650) – 1	B	Уменьшение ошибки метода РМЗ для разных соединений от ~10% до ~1.25%	—	108
Спектры ЯМР ¹³ C	Ациклические алканы	21	Фрагментные	12 – 2 – 4 – 6 – 2 – 1	C	Рассчитанные величины химических сдвигов близки к экспериментальным	—	24
	Монозамещенные бензолы	20	Топологические и физико-химические	11 – 6 – 4	D	Удовлетворительная классификация и прогнозирование химических сдвигов	—	22
	То же	20	То же	11 – 6 – 4	E	То же	—	22
	Разнородная выборка	40 000	Фрагментные	360 – (5, 10 или 20) – 1	D	Усредненные результаты обучения трех нейросетей	$MD_t = 1.79, SD_t = 2.10, R_t = 0.979; MD_p = 1.75, SD_p = 1.97, R_p = 0.981$ (м.д.)	109
	Пиранозы и пиранозиды	55	Топологические, электронные, геометрические	11 – 5 – 1	D	—	$RMS_t = 1.03, RMS_{cv} = 0.766, RMS_p = 1.11$ (м.д.)	110
	Фуранозы и фуранозиды	56	То же	8 – 5 – 1	D	—	$RMS_t = 1.58, RMS_{cv} = 0.995, RMS_p = 0.898$ (м.д.)	110
	Рибонуклеозиды	17	Квантово-химические	4 – 2 – 1	D	—	$RMS_t = 0.69, RMS_{cv} = 0.47, RMS_p = 0.39$ (м.д.)	111
	Ациклические алкены	130	Топологические	12 – 4 – 1	D	Четыре сети с различными передаточными функциями	$SD_t = 0.59 - 0.63, SD_p = 0.89 - 1.07$ (м.д.)	112
Спектры ЯМР	Аминокислоты	123	Физико-химические	Иерархические сети: 52 – 4 – 3, 39 – 4 – 3	F	—	$C_t = 77\%$ для ¹³ C _α $C_t = 75.3\%$ для ¹ H _α $C_t = 78.3\%$ для ¹ HN	44
Индексы удерживания в газовой хроматографии	Алканы, алкены, спирты, эфиры, кетоны	216	Типы связей	29 – 2 – 1	D	—	$RMS_t = 14.7, RMS_p = 23.6$	102
Индексы удерживания в хроматографии на обращенной фазе	Стероиды	85	Квантово-химические	11 – 3 – 1	D	—	$SD_p = 0.08, SD_{cv} = 0.09$	113

Таблица 1 (окончание).

Моделируемое свойство	Исследуемые соединения	Количество соединений	Дескрипторы	Архитектура нейросетей		Результаты обучения нейросети	Статистические характеристики обучения нейросети ^a	Ссылки
				количество нейронов	тип нейросети ^b			
Индексы Ковача	Замещенные фенолы	43	Молярная рефракция, дипольный момент	2–3–6–1	A	—	$RMS_p = 1.1$ для носителя SE-30 $RMS_p = 2.1$ для носителя OV-225 $RMS_p = 0.8$ для носителя NGA	37
Индексы удерживания в тонкослойной хроматографии	Замещенные бензойные кислоты	22	Квантово-химические	7–4–1	D	—	$R = 0.973$	103

Примечание. Здесь и в табл. 2–4 архитектуры нейросетей описываются количеством нейронов для каждого слоя нейросети, начиная с входного и заканчивая выходным.

^a Параметры для оценки результатов обучения нейросетей: R — коэффициент корреляции; MD — среднее отклонение; RE — относительная ошибка; SD — стандартное отклонение; RMS , MSE , S — среднеквадратичная ошибка; C — коэффициент распознавания; SE — стандартная ошибка; MAE , AAE — средняя абсолютная ошибка; RSD — относительное стандартное отклонение; ME — средняя ошибка прогноза; ADD — среднее абсолютное отклонение; RV — остаточное изменение; SSO — сумма квадратов разниц между вычисленной и экспериментальной величиной. Для подстрочных индексов использованы следующие буквенные обозначения: t — обучающая выборка; p , v — контрольная выборка; o — для всех соединений; cv — скользящий контроль.

^b A — нейросеть встречного распространения; B — полносвязанная нейросеть прямого распространения; C — нейросеть, воспроизводящая топологию исследуемых молекул; D — полносвязанная нейросеть прямого распространения, обучающаяся по дельта-правилу; E — рекуррентная нейросеть; F — полносвязанная нейросеть прямого распространения, обучающаяся по методу устойчивого распространения ошибки; G — нейросеть Кохонена; H — волновая нейросеть; I — полносвязанная нейросеть прямого распространения, обучающаяся по методу Бройдена – Флетчера – Гольдфарба – Шанно; J — радикально-базисная нейросеть; K — нечеткая ARTMAP-нейросеть; L — нейросеть прямого распространения с дополнительными прямыми связями между нейронами входного и выходного слоя, обучающаяся по дельта-правилу; M — нейросеть прямого распространения с дополнительными прямыми связями между нейронами входного и выходного слоя, обучающаяся по методу Бройдена – Флетчера – Гольдфарба – Шанно; N — полносвязанная нейросеть прямого распространения, обучающаяся по методу масштабированного сопряженного градиента; O — полносвязанная нейросеть прямого распространения, обучающаяся с использованием бейесовой регуляризации; P — нейросеть Кохонена с нелинейным распределением; Q — полносвязанная нейросеть прямого распространения с функциональными связями.

Таблица 2. Характеристики нейросетевых моделей, полученных при моделировании зависимостей структура – реакция способность органических соединений.

Моделируемое свойство	Исследуемые соединения	Количество соединений	Дескрипторы	Архитектура нейросетей		Результаты обучения нейросети	Статистические характеристики обучения нейросети ^a	Ссылки
				количество нейронов	тип нейросети ^a			
Разрыв связи С–С между атомами в α - и β -положениях по отношению к карбонильной группе	Карбонильные соединения	32	Фрагментные	8–(9 × 9)	G	Четкое разделение исходного массива соединений на четыре кластера по типам разрыва связей	—	28
Гетеролитический разрыв связей	Алифатические соединения	29	Физико-химические	7–(11 × 11)	G	Правильная классификация реакций по типу разрыва связей как для обучающей, так и для контрольной выборок	—	2
	То же	29	То же	7–3–1	D			
Соотношение изомеров при электрофильном замещении	Монозамещенные бензолы	45	Квантово-химические	6–10–2	D	—	$RMS_t = 5.2\%$, $RMS_p = 19.8\%$	2, 56
	То же	45	Таблицы связей заместителей	25–5–2	D	—	$RMS_t = 0.3\%$, $RMS_p = 12.1\%$	2, 56
Выход <i>m</i> -изомеров при нитровании	»	31	Фрагментные	9–4–4–1	D	Хорошее согласие с экспериментальными данными для обучающей выборки; для контрольной выборки результат на «полуколичественном уровне»	—	23
	»	31	»	9–5–5–1	I			
	»	31	»	9–6–6–1	I			
Присоединение галогеноводородов по правилу Марковникова	Алкены	25	Матрица связей	36–8–8–28	D	Правильное распознавание позиции присоединения для обучающей и контрольной выборок	—	57
Прямые и ретро-реакции Дильса – Альдера	Циклоалкены	36	То же	120–36–36–105	D	Правильная классификация на прямые и ретро-реакции и определение основного продукта при прямой реакции для обучающей и контрольной выборок	—	57
Элиминирование галогеноводородов по правилу Зайцева	Алкены	83	»	66–24–24–55	D	Правильное определение основного продукта реакции для обучающей и контрольной выборок	—	57
Константы комплексообразования с α -циклодекстрином	Замещенные бензолы	24	Физико-химические	3–3–3–3–1	D	—	$R = 0.97$, $SD_t = 0.22$ ($\ln K_a$)	114
Константы комплексообразования с β -циклодекстрином	Моно- и 1,4-дизамещенные бензолы	40	То же	6–6–1	H	—	$R = 0.992$, $SD_t = 0.089$ ($\ln K_a$)	115
Теплота образования ароматических солей	Дигидропиридины и родственные гетероциклические соединения	71	»	22–4–1	D	—	$R = 0.92$, $SD_t = 0.298$, $SD_p = 1.80$ ккал·моль ^{–1}	116

Таблица 2 (окончание).

Моделируемое свойство	Исследуемые соединения	Количество соединений	Дескрипторы	Архитектура нейросетей		Результаты обучения нейросети	Статистические характеристики обучения нейросети ^a	Ссылки
				количество нейронов	тип нейросети ^a			
Константы скорости присоединения метильного радикала	Ненасыщенные соединения различных типов	191	Квантово-химические	6–4–1	D	—	$RMS_t = 0.381$, $RMS_p = 0.496$	53
	То же	191	То же	7–3–1	D	—	$RMS_t = 0.424$, $RMS_p = 0.409$ ($\lg k$)	53
Реакция с тетразолиевым синим, состав мультикомпонентной смеси	Смеси кортизона и гидрокортизона	40	То же	6–4–2	D	—	$SE_t = 3.47\%$, $SE_p = 4.12\%$ для кортизона; $SE_t = 5.12\%$, $SE_p = 5.25\%$ для гидрокортизона	117

^a Обозначения см. в табл. 1.

Таблица 3. Характеристики нейросетевых моделей, полученных при моделировании зависимостей структура – физико-химические свойства органических соединений.

Моделируемое свойство	Исследуемые соединения	Количество соединений	Дескрипторы	Архитектура нейросетей		Статистические характеристики обучения нейросети ^a	Ссылки
				количество нейронов	тип нейросети ^a		
Температура кипения	Галогеналканы C ₁ –C ₄	171	Топологические	7–14–1	D	$R = 0.995$	118
	Ациклические простые эфиры, пероксиды, ацетали и их сернистые аналоги	185	Фрагментные	20–5–1	D	$R = 0.998$, $SE_t = 2.9^\circ\text{C}$, $SE_p = 5.1^\circ\text{C}$	82
	То же	185	Топологические	3–20–1	J	$R = 0.990$, $SE_t = 4.9^\circ\text{C}$, $SE_p = 5.9^\circ\text{C}$	39
	Тетрагидрофураны, тиофены, фураны, пираны	299	Топологические и электронные	16–3–1	D	$RMS_{cv} = 8.49^\circ\text{C}$	119
	Пиридины	291	То же	7–3–1	D	$RMS_{cv} = 15.8^\circ\text{C}$	119
	Разнородная выборка	298	Электротопологические	19–5–1	D	$R = 0.9975$, $MAE_t = 3.86^\circ\text{C}$, $MAE_p = 4.57^\circ\text{C}$	120
	Фторзамещенные этаны и пропаны	31	Топологические	8–6–1	D	$MAE_t = 4.7^\circ\text{C}$	121
	Алкены	82	Топологические	5–5–1	D	$RSD_t = 4.88\%$, $RSD_{cv} = 3.5\%$	122
	Алканы	140	Топологические и физико-химические	7–4–1	D	$AAE_t = 1.65^\circ\text{C}$, $AAE_p = 1.73^\circ\text{C}$, $AAE_o = 1.54^\circ\text{C}$	42
	»	140	То же	7–4–1	K	$AAE_t = 0.81^\circ\text{C}$, $AAE_p = 1.30^\circ\text{C}$	42
	Алкены	144	»	7–10–1	D	$AAE_t = 6.79^\circ\text{C}$, $AAE_p = 6.45^\circ\text{C}$, $AAE_o = 4.42^\circ\text{C}$	42
	»	144	»	7–10–1	K	$AAE_t = 0.73^\circ\text{C}$, $AAE_o = 0.95^\circ\text{C}$	42
	Алканы, алкены и алкины	327	»	7–9–1	D	$AAE_t = 6.09^\circ\text{C}$, $AAE_p = 4.68^\circ\text{C}$, $AAE_o = 4.85^\circ\text{C}$	42
	То же	327	»	7–9–1	K	$AAE_t = 1.15^\circ\text{C}$, $AAE_o = 1.35^\circ\text{C}$	42
	Разнородная выборка	298	Топологические, электронные и смешанные	8–3–1	I	$RMS_t = 7.75^\circ\text{C}$, $RMS_{cv} = 7.17^\circ\text{C}$, $RMS_p = 8.69^\circ\text{C}$	123
	То же	400	Топологические	26–36–2	J	$AAE_t = 11^\circ\text{C}$, $AAE_p = 14^\circ\text{C}$	40
	Углеводороды	134	То же	7–8–6	D	$ME_p = 1.19\%$	101
	Разнородная выборка	421	Наличие функциональных групп	36–3–4	L	$ADD_p = 2.9\%$	19
Критическая температура	То же	165	Электротопологические	19–4–1	D	$R = 0.9965$, $MAE_t = 4.39^\circ\text{C}$, $MAE_p = 5.59^\circ\text{C}$	120
	Фторзамещенные этаны и пропаны	38	Топологические и температура кипения	9–4–1	D	$MAE_t = 5.9^\circ\text{C}$	121
	Разнородная выборка	421	Наличие функциональных групп	36–3–4	L	$ADD_p = 3.1\%$	19
	Алканы	69	Топологические	12–5–6	D	$R = 0.994$, $S_t = 3.80$, $S_v = 3.94$	124
	»	69	Фрагментные	14–6–6	D	$R = 0.995$, $S_t = 3.37$, $S_v = 3.58$	124
Температура самовоспламенения	Ациклические углеводороды с низкой температурой самовоспламенения	47	Топологические и квантово-химические	5–3–1	I	$RMS_t = 8.77^\circ\text{C}$, $RMS_{cv} = 6.88^\circ\text{C}$, $RMS_p = 5.11^\circ\text{C}$	125
	Ациклические углеводороды с высокой температурой самовоспламенения	51	То же	6–2–1	I	$RMS_t = 18.5^\circ\text{C}$, $RMS_{cv} = 17.0^\circ\text{C}$, $RMS_p = 15.7^\circ\text{C}$	125

Таблица 3 (продолжение).

Моделируемое свойство	Исследуемые соединения	Количество соединений	Дескрипторы	Архитектура нейросетей		Статистические характеристики обучения нейросети ^a	Ссылки
				количество нейронов	тип нейросети ^a		
Температура самовоспламенения	Азотсодержащие органические соединения	40	Топологические и квантово-химические	6–2–1	I	$RMS_t = 34.9^{\circ}\text{C}$, $RMS_{cv} = 23.6^{\circ}\text{C}$, $RMS_p = 28.2^{\circ}\text{C}$	125
	Органические соединения, содержащие атомы кислорода или серы	132	То же	7–5–1	I	$RMS_t = 30.8^{\circ}\text{C}$, $RMS_{cv} = 29.7^{\circ}\text{C}$, $RMS_p = 32.5^{\circ}\text{C}$	125
Температура вспышки	Разнородная выборка	400	Топологические	26–36–2	J	$AAE_t = 10^{\circ}\text{C}$, $AAE_p = 12^{\circ}\text{C}$	40
Теплота испарения	Фторзамещенные этаны и пропаны	38	Топологические	8–4–1	D	$MAE_t = 1.1 \text{ кДж} \cdot \text{моль}^{-1}$	121
	Алканы	69	»	12–5–6	D	$R = 0.994$, $S_t = 0.44$, $S_v = 0.51$	124
	То же	69	Фрагментные	14–6–6	D	$R = 0.996$, $S_t = 0.44$, $S_v = 0.56$	124
Теплоемкость	Углеводороды	134	Топологические	7–8–6	D	$ME_p = 0.87\%$	101
Плотность	То же	134	»	7–8–6	D	$ME_p = 0.60\%$	101
	Фторзамещенные этаны и пропаны	38	»	8–4–1	D	$MAE_t = 0.03 \text{ г} \cdot \text{см}^{-3}$	121
	Алкены	82	»	5–5–1	D	$RSD_t = 0.43\%$, $RSD_{cv} = 0.4\%$	122
Показатель преломления	Углеводороды	134	»	7–8–6	D	$ME_p = 0.19\%$	101
	Алкены	82	»	5–5–1	D	$RSD_t = 0.13\%$, $RSD_{cv} = 0.14\%$	122
Энергия Гиббса	Углеводороды	134	»	7–8–6	D	$ME_p = 1.36\%$	101
Энтальпия	То же	134	»	7–8–6	D	$ME_p = 1.42\%$	101
Критический объем	Разнородная выборка	421	Наличие функциональных групп	36–3–4	L	$ADD_p = 3.4\%$	19
Ацентрический фактор	То же	421	То же	36–3–4	L	$ADD_p = 9.1\%$	19
Молярный объем	Алканы	69	Топологические	12–5–6	D	$R = 0.999$, $S_t = 0.84$, $S_v = 0.89$	124
	То же	69	Фрагментные	14–6–6	D	$R = 0.999$, $S_t = 0.88$, $S_v = 1.10$	124
Молярная рефракция	»	69	Топологические	12–5–6	D	$R = 1.000$, $S_t = 0.15$, $S_v = 0.18$	124
	»	69	Фрагментные	14–6–6	D	$R = 0.999$, $S_t = 0.20$, $S_v = 0.18$	124
Критическое давление	»	69	Топологические	12–5–6	D	$R = 0.984$, $S_t = 0.46$, $S_v = 0.39$	124
	»	69	Фрагментные	14–6–6	D	$R = 0.986$, $S_t = 0.44$, $S_v = 0.23$	124
	Разнородная выборка	421	Температура кипения, критическая температура, критический объем	3–3–1	L	$ADD_p = 6.4\%$	19
Поверхностное натяжение	Алканы	69	Топологические	12–5–6	D	$R = 0.996$, $S_t = 0.18$, $S_v = 0.28$	124
	»	69	Фрагментные	14–6–6	D	$R = 0.996$, $S_t = 0.17$, $S_v = 0.17$	124

Таблица 3 (продолжение).

Моделируемое свойство	Исследуемые соединения	Количество соединений	Дескрипторы	Архитектура нейросетей		Статистические характеристики обучения нейросети ^a	Ссылки
				количество нейронов	тип нейросети ^a		
Давление насыщенных паров [lg <i>V</i> <i>P</i> (Па)]	Соединения с гетероатомами	420	Топологические	8–3–1	I	$RMS_t = 0.26, RMS_p = 0.37, RMS_{cv} = 0.29$	54
	То же	420	»	10–4–1	I	$RMS_t = 0.19, RMS_p = 0.33, RMS_{cv} = 0.24$	54
	Углеводороды и галогенуглеводороды	352	Топологические, геометрические и электронные	7–3–1	I	$RMS_t = 0.163, RMS_{cv} = 0.163, RMS_p = 0.209$	126
Вязкость [lg <i>M</i> (мПа·с)]	Разнородная выборка	361	Экспериментальные физико-химические величины	9–3–1	D	$R = 0.958, RMS_t = 0.118, RMS_p = 0.161$	127
Растворимость в воде [lg <i>S</i> (моль·л ^{–1})]	То же	350	Квантово-химические	17–18–1	D	$SD_t = 0.23, SD_p = 0.43$	128
	То же	332	Топологические, электронные и геометрические	9–6–1	I	$RMS_t = 0.394, RMS_{cv} = 0.358, RMS_p = 0.343$	129
	Разнородная выборка лекарственных препаратов	211	Электротопологические и топологические	23–5–1	D	$R_t^2 = 0.90, SE_t = 0.46, R_p^2 = 0.86, SE_p = 0.53$	130
	Разнородная выборка	136	Топологические, электронные, геометрические и смешанные	9–3–1	I	$RMS_t = 0.145, RMS_{cv} = 0.151, RMS_p = 0.166$	131
Растворимость в диоксиде углерода, находящемся в сверхкритическом состоянии [lg <i>S</i> (моль·л ^{–1})]	То же	58	Топологические и квантово-химические	7–2–1	I	$RMS_t = 0.65, RMS_{cv} = 0.68, RMS_p = 0.64$	132
Коэффициент активности при бесконечном разведении (lg γ^∞)	»	325	То же	12–6–1	I	$RMS_t = 0.472, RMS_{cv} = 0.538, RMS_p = 0.484$	133
Липофильность (log <i>P</i>) ^b	»	250	Квантово-химические	8–5–1	D	$R = 0.923, SE_p = 0.379$	134
	»	250	То же	13–6–1	D	$R = 0.952, SE_p = 0.300$	134
	»	1870	Молекулярная масса и электротопологические	39–5–1	D	$R_t^2 = 0.90, RMS_t = 0.46, R_p^2 = 0.94, RMS_p = 0.4$	135
	»	323	Квантово-химические	6–7–1	I	$SE_p = 0.30$	136
	»	323	То же	11–3–1	D	$SD_t = 0.31, SD_p = 0.29, SD_{cv} = 0.32$	113
	Соединения с гетероатомами	7719	Тополого-физико-химические	35–32–1	D	$R_t = 0.97, RMS_t = 0.37, R_p = 0.98, RMS_p = 0.39$	137
	Соединения с гетероатомами	7719	Тополого-физико-химические	35–32–1	D	$R_t = 0.97, RMS_t = 0.37, R_p = 0.98, RMS_p = 0.39$	137
Температура фазового перехода (°C)	Нематические жидкокристаллические соединения	17383	Фрагментные	205–100–1	D	$SD_t = 3.8, SD_p = 16.4$	138, 139
	То же	6304	»	205–10–1	D	$SD_p = 18.8$	138, 139
Сольватохромный параметр полярности – поляризуемости	Разнородная выборка	333	Топологические, электростатические, квантово-химические, структурные	16–7–1	D	$R^2 = 0.980$	77
Молярная адсорбция (lg ϵ)	Асимметричные фосфобис-азопроизводные	43	Топологические	4–4–3–1	I	$RMS_t = 0.05669, RMS_p = 0.09621$	61

Таблица 3 (окончание).

Моделируемое свойство	Исследуемые соединения	Количество соединений	Дескрипторы	Архитектура нейросетей		Статистические характеристики обучения нейросети ^a	Ссылки
				количество нейронов	тип нейросети ^a		
Чувствительность к удару ($\lg H_{50\%}$) ^c	Нитросоединения	204	Топологические, квантово-химические	11–2–1	D	$R = 0.941, SE_p = 0.154$	62
	»	204	То же	13–2–1	D	$R = 0.937, SE_p = 0.159$	62
	»	204	»	13–3–1	D	$R = 0.95, SE_p = 0.13$	140
Положение длинноволновой полосы поглощения	Симметричные цианиновые красители	398	Квантово-химические, индикаторы наличия заместителей	8–10–1	D	$R_s = 0.9928, S_t = 10.6 \text{ нм}, S_v = 7.0 \text{ нм}$	141
Межатомные расстояния	Соединения, содержащие два гетероатома	2615	Геометрические	78–8–1	D	$C_t = 80\%, C_p = 80\%$ для удаленных атомов; $C_t = 66\%$ для сближенных атомов	100
Запах	Алифатические спирты	99	Топологические	6–3–1	D	$C_t = 100\%, C_p = 85\%$	142
Общая энергия π -электронов	Незамещенные полициклические углеводороды	265	»	3–3–1	D	$SD_t = 0.06, SD_{cv} = 0.27$	143
Индуктивные и резонансные константы заместителей	Одновалентные функциональные группы ^d	37	»	14–8–2	D	—	144

^a Обозначения см. в табл. 1.^b $\log P$ — десятичный логарифм коэффициента распределения в системе *n*-октанол–вода.^c $H_{50\%}$ — высота, при которой происходит взрыв с вероятностью 50%.^d Для данных исследуемых соединений результат обучения сети — правильная классификация заместителей и хорошая оценка величин констант заместителей как для обучающей, так и для контрольной выборки.

Таблица 4. Характеристики нейросетевых моделей, полученных при моделировании зависимостей структура – биологическая активность органических соединений.

Моделируемое свойство	Количество соединений	Исследуемые соединения	Дескрипторы	Архитектура нейросетей		Результаты обучения	Статистические характеристики обучения нейросети ^a	Ссылки
				количество нейронов	тип нейросети ^a			
Сродство к транспортным белкам (глобулинам)	31	Кортикостероиды и тестостероны	Физико-химические	32 – (10 × 10)	G	Четкое разделение стероидов на три кластера (сильное, среднее и слабое сродство)	—	99
		То же	Карта атомных координат	9 – (7 × 7)	G	Разделение стероидов на два кластера (сильное и слабое сродство)	—	29
		»	Карта атомных координат и заряды	9 – (7 × 7)	G	Четкое разделение по типу рецепторов для обучающей и контрольной выборок	—	30
Сродство к рецепторам эстрогена	22	Дизамещенные производные гексэстрола	Топологические и физико-химические	2 – 6 – 1	D	—	$RMS_t = 0.09131$ ($\lg RBA$) ^b	145
	16	Простые эфиры гексэстрола	Топологические	1 – 4 – 2 – 1	D	—	$RMS_t = 0.018219$ ($\lg RBA$)	145
	16	Дезоксигексэстроны	То же	1 – 4 – 2 – 1	D	—	$RMS_t = 0.018219$ ($\lg RBA$)	145
Сродство к рецепторам прогестерона	55	Производные андрост-4-ен-3-она	Физико-химические	6 – 2 – 1	M	—	$R = 0.96$, $RMS_t = 0.30$, $RMS_p = 1.32$ ($\lg IC_{50}$) ^c	20
Сродство к рецепторам бензодиазепина GABA _A	57	1,4-Бензодиазепин-3-оны	То же	10 – 3 – 1	N	—	$R_t = 0.954$, $SD_t = 0.007$, $R_{cv} = 0.901$ ($\lg IC_{50}$)	146
				6 – 2 – 1	N	—	$SD_{cv} = 0.009$, $R_{cv} = 0.938$	146
Ингибирование обратной транскриптазы вируса человеческого иммунодефицита	44	Производные AZT ^d , TIVO ^e и подобных веществ	Топологические	4 – 3 – 2	D	—	$C_t = 100\%$, $C_p = 86\%$	74, 75
	107	Производные 1-[(2-гидроксиэтокс)метил]-6-(фенилтио)тимина	Квантово-химические	6 – 6 – 1	D	—	$MSE_t = 0.073$, $MSE_p = 0.372$ ($\lg 1/C$) ^f	147
Ингибирование p56 ^{lck} тирозинкиназы белка	105	Замещенные флавоноиды	Физико-химические константы заместителей	6 – (10 × 10) – 1	A	—	$R^2 = 0.92$, $SD_p = 0.55$ ($\lg 1/IC_{50}$)	35
			Квантово-химические	3 – (10 × 10) – 1	A	—	$R^2 = 0.96$, $SD_p = 0.40$ ($\lg 1/IC_{50}$)	35
Ингибирование 5-липоксигеназы <i>in vitro</i>	68	Арилгидроксамовые кислоты	Физико-химические	1 – 4 – 1	D	—	$RMS_t = 0.09043$ ($\lg 1/K$) ^g	145
Ингибирование дигидрофолатредуктазы	157	5-Фенил-3,4-диамино-6,6-диметилдигидротриазины	То же	6 – 4 – 1	D	—	$R = 0.922$, $SD_t = 0.374$ ($\lg IC_{50}$)	80
	61	Симметричные 1-арил-2,2-диметил-4,6-диамино-1,2-дигидротриазины	»	3 – 1 – 1	M	—	$R = 0.87$, $SD_t = 0.27$	20
	68	2,4-Диамино-5-(<i>R</i> -бензил)пиримидины	Физико-химические	4 – 7 – 1	D	—	$R = 0.97$, $RV_t = 0.0126$, $RV_p = 0.323$	83

Таблица 4 (продолжение).

Исследуемые соединения	Количество соединений	Моделируемое свойство	Дескрипторы	Архитектура нейросетей		Результаты обучения	Статистические характеристики обучения нейросети ^a	Ссылки
				количество нейронов	тип нейросети ^a			
Ингибирование ацил-холестерин- <i>O</i> -ацил-трансферазы	157	Аминосульфонил-, гидроксисульфонил- и (2,6-диизопропилфенокси)-карбонаты, аминосульфонил- и гидроксисульфонилмочевины	Топологические, геометрические и электронные	8–3–1	I	—	$RMS_t = 0.226$, $RMS_{cv} = 0.208$, $RMS_p = 0.242$ ($\lg IC_{50}$)	55
Ингибирование связывания диазепама ³ H	245	Бензодиазепины	Топологические	50–4–1	O	—	$R^2 = 0.63$, $SE_p = 0.14$ ($\lg IC_{50}$)	88, 89
Противогрибковая активность	103	3-(<i>R</i>)-1-(3,5-Дихлорфенил)-пирролидин-2,5-дионы	Физико-химические параметры заместителей	5–(5×5)	O	Четкое разделение исходных веществ на группы	—	78
Антигельминтная активность	31	2-Гидроксифениламидамы	Физико-химические	3–3–1	D	—	$R = 0.919$, $RMS_t = 0.322$	64
		То же	То же	53–(8×8)	G	Четкое разделение на группы с различной активностью	—	32
Гипотензивная активность	24	3-(<i>R</i>)-Фенилтиопропил-замещенные гетероциклы	»	6–3–1	D	—	$SD_{cv} = 11.6\%$	148
	29	Арилакрилоилпиперазины	»	7–14–4	D	—	$C_t = 19/21$, $C_p = 6/8$	149
Противосудорожная активность	60	Производные бензодиазепаина	Физико-химические	14–15–57–1	A	—	$RMS_t = 0.16$	38
			То же	14–28–1	D	—	$SD_t = 0.33$	10, 38
			»	5–6–1	D	—	$R = 0.887$	2, 43
			»	5–1	Q	—	$R = 0.876$, $RMS_t = 0.369$	43
Противораковая активность (минимальная эффективная доза при постоянных инъекциях)	39	2,5-Бис(1-азиридирил)- <i>n</i> -бензохиноны	Физико-химические	6–7–37–1	A	—	$RMS_t = 0.08$	38
			То же	12–13–37–1	A	—	$RMS_t = 0.06$	38
			»	7–12–1	D	—	$RMS_t = 0.21$	2, 10, 38
			»	13–26–1	D	—	$RMS_t = 0.21$	10, 38
Противораковая активность (минимальная эффективная доза при одиночной инъекции)	39		Физико-химические	6–7–37–1	A	—	$RMS_t = 0.04$	38
			То же	12–13–37–1	A	—	$RMS_t = 0.20$	38
			»	7–12–1	D	—	$RMS_t = 0.25$	2, 10, 38
			»	13–26–1	D	—	$RMS_t = 0.24$	2, 10, 38
Противораковая активность (оптимальная доза при постоянных инъекциях)	39		Физико-химические	6–7–37–1	A	—	$RMS_t = 0.02$	38
			То же	12–13–37–1	A	—	$RMS_t = 0.05$	38
			»	7–12–1	D	—	$RMS_t = 0.19$	2, 10, 38
			»	13–26–1	D	—	$RMS_t = 0.18$	2, 10, 38
			»	4–13–1	D	—	$R = 0.960$	43
			»	4–1	Q	—	$R = 0.951$, $RMS_t = 0.169$	43
Противораковая активность (оптимальная доза при одиночной инъекции)	39		Физико-химические	6–7–37–1	A	—	$RMS_t = 0.08$	38
			То же	12–13–37–1	A	—	$RMS_t = 0.05$	38
			»	7–12–1	D	—	$RMS_t = 0.20$	2, 10, 38
			»	13–26–1	D	—	$RMS_t = 0.17$	2, 10, 38

Таблица 4 (продолжение).

Исследуемые соединения	Количество соединений	Моделируемое свойство	Дескрипторы	Архитектура нейросетей		Результаты обучения	Статистические характеристики обучения нейросети ^a	Ссылки
				количество нейронов	тип нейросети ^a			
Ингибирование раковых клеток, чувствительных к метотрексату	61	3-Замещенные триазины	Физико-химические	3–3–1	M	—	$R = 0.86, SD_t = 0.25$	20
Ингибирование раковых клеток, устойчивых к метотрексату	62		То же	3–1	M	—	$R = 0.58, SD_t = 0.53$	20
Механизмы противораковой активности	141	Лекарственные препараты	Активности по отношению к различным раковым клеткам	60–7–6	D	—	$C_p = 129/141$	150
Токсичность	91	Бензотиазольевые соли	Фрагментные	30–2–1	D	—	$SSO_t = 0.017, SSO_p = 0.850$	151
Антиканцерогенная активность	16	Карбохиноны	Физико-химические	7–12–5 7–4–5	D D	— —	$C_t = 100\%, C_p = 3/5$ $C_t = 100\%, C_p = 3/5$	84, 94, 149 84
Канцерогенная активность	45	Моно- и полизамещенные бензолы	Топологические	90–100–100–1	A	—	$R_t = 0.95, R_p = 0.70$	36
			Топологические и квантово-химические	90–100–100–1	A	—	$R_t = 0.98, R_p = 0.72$	36
			Квантово-химические	90–100–100–1	A	—	$R_t = 1.00, R_p = 0.63$	36
	11	Полициклические ароматические углеводороды	Спектр ЯМР ^{13}C	10–3–1	D	Четкое разделение на активные и неактивные соединения	—	76
	81	Ароматические соединения с азотсодержащими заместителями	Квантово-химические	6–12–2	D	—	$C_t = 84.6\%, C_p \approx 80\%$	152
	104		Топологические, физико-химические и квантово-химические	12–4–1	D	—	$R_{cv}^2 = 0.691, MSE_{cv} = 0.0416$ $[\lg(MW \cdot 1000/ID_{50})]^{h,i}$	153
Мутагенная активность	197	Ароматические и гетероароматические соединения	Физико-химические	2–4–2–1	D	—	$RMS_t = 0.14331 (\lg TA_{98})^j$	145
			То же	4–8–4–1	D	—	$RMS_t = 0.11515, RMS_p = 0.1216$	145
			»	4–4–1	D	—	$R_t = 0.919, SD_t = 0.789, R_p = 0.853, SD_p = 1.049 (\lg TA_{98})$	154
	488	Разнородная выборка	Фрагментные	100–10–2	D	Хорошее разделение на кластеры по величине и активности	—	69, 155
	54	Гетероциклические аналоги пирена и фенантрена, флуорены, дифенилы	Энергия НСМО, липофильность, наличие нитрогруппы в <i>n</i> -положении	3–2–1	D	—	$R = 0.87, S_t = 1.03, S_v = 1.47$ $(\ln N_{his+})^k$	156

Таблица 4 (окончание).

Исследуемые соединения	Количество соединений	Моделируемое свойство	Дескрипторы	Архитектура нейросетей		Результаты обучения	Статистические характеристики обучения нейросети ^a	Ссылки
				количество нейронов	тип нейросети ^a			
Ингибирование внутриклеточного ионного обмена Na^+/H^+	113	Производные бензоилгуанидина	Топологические, геометрические, электронные и комбинированные	5–4–1	D	—	$R_t^2 = 0.812$, $RMS_t = 0.278$, $R_p^2 = 0.697$, $RMS_p = 0.362$ ($\lg IC_{50}$)	157
Галлюциногенная активность	64	Фенилалкиламины	Фрагментные	15–(33 × 33)	G	Четкое разделение соединений на кластеры по величине активности	—	28
	35	Фенилалкиламины	Константы заместителей, индикаторы наличия заместителей	7–2–1	D	—	$R = 0.932$, $S_t = 0.55$, $S_v = 0.47$ ($\ln MU$) ^{l,m}	58
							$R = 0.820$, $S_t = 0.89$, $S_v = 0.54$ ($\ln MU$) ⁿ	58
Блокирование каналов Ca^{2+}	46	1,4-Дигидропиридины	Константы заместителей	5–2–1	D	—	$R = 0.832$, $S_t = 0.79$, $S_v = 0.71$ ($\lg 1/EC_{50}$) ^{m,o}	58
					D	—	$R = 0.870$, $S_t = 0.70$, $S_v = 1.59$ ($\lg 1/EC_{50}$) ⁿ	58
Адсорбция слизистой оболочкой кишечника человека	86	Лекарственные препараты	Квантово-химические и топологические	6–4–1	D	—	$RMS_t = 9.4\%$, $RMS_{cv} = 19.7\%$, $RMS_p = 16.0\%$	158
Антибактериальная активность ^p $MIC \leq 0.05$ $MIC \leq 0.10$ $MIC \leq 0.20$	111	Фторхинолоны	Топологические	62–2–1	D	—		
							$C_t = 100\%$, $C_p = 93.88\%$	159
							$C_t = 97.28\%$, $C_p = 82.36\%$ $C_t = 99.53\%$, $C_p = 87.99\%$	

^a Обозначения см. в табл. 1; ^b RBA — сродство к рецептору; ^c IC_{50} — концентрация, приводящая к 50%-ному ингибированию; ^d AZT — азидотимидин; ^e TIBO — 4,5,6,7-тетрагидроимидазо[4,5,1-*j,k*][1,4]бензодиазепин-2(1*H*)-он; ^f C — концентрация вещества, вызывающая определенный биологический отклик; ^g K — константа ингибирования; ^h MW — молекулярная масса; ⁱ ID_{50} — доза, вызывающая 50%-ное ингибирование; ^j TA_{98} — число ревертантов в тесте Эймса TA_{98} ; ^k N_{his+} — число ревертантов в тесте Эймса TA_{1538} ; ^l MU — мускариновые единицы; ^m в исходную базу добавляли топологически эквивалентные структуры; ⁿ без размножения базы; ^o EC_{50} — эффективная концентрация; ^p MIC — минимальная ингибирующая концентрация в $\text{мг} \cdot \text{мл}^{-1}$.

IV. Применение нейросетевых алгоритмов для поиска зависимостей структура – свойство

Нейросетевой подход может быть с успехом применен для нахождения как качественных, так и количественных зависимостей структура – свойство. В табл. 1–4 приводится краткая информация о нейросетевых моделях, используемых для моделирования спектральных характеристик, реакционной способности, физико-химических свойств и биологической активности органических соединений.

Информация, содержащаяся в ИК-спектрах^{99, 108} и спектрах ЯМР,^{22, 24, 44, 109–112} а также в хроматографических индексах удерживания,^{37, 102, 103, 113} тесно связана со структурой органических молекул. В цитированных в табл. 1 работах нейросетевой алгоритм обработки подобной информации был с успехом использован в решении разнообразных задач, связанных с поиском зависимости между структурой молекулы и спектральными свойствами, классификацией и предсказанием структурных данных, а также с уточнением структуры молекулы на основании спектральной информации.

Одними из первых появились работы, в которых нейросетевые методы были применены для классификации реакций органических соединений: по типу разрыва связей,^{2, 28} распознаванию позиции присоединения,⁵⁷ соотношению получающихся в ходе реакции изомеров,^{2, 23, 56} определению основного продукта реакции⁵⁷ и т.д. (табл. 2). Входной информацией для нейросетевых моделей служили матрицы связности органических соединений, использование которых не требовало никаких дополнительных вычислений, предшествующих этапу моделирования. С помощью нейросетевых методов можно находить также количественные оценки реакционной способности органических соединений и моделировать сложные кинетические процессы в ходе химических превращений.

Весьма велико число публикаций, в которых с помощью нейронных сетей анализируются различные физико-химические свойства химических соединений, причем к наиболее популярным свойствам можно отнести температуру кипения^{19, 39, 40, 42, 82, 101, 118–123} (в основном с использованием топологических индексов), плотность,^{101, 121, 122} растворимость в воде^{128–131} и липофильность^{134–137} органических соединений и др. (табл. 3).

Моделирование биологической активности органических соединений на сегодняшний день представляет большой практический интерес, и подавляющее большинство публикаций, связанных с использованием методологии искусственных нейронных сетей для поиска зависимостей структура – свойство, относится именно к этой области моделирования (табл. 4). В ряде работ приведены результаты успешного количественного нейросетевого моделирования сродства органических соединений к различным рецепторам,^{20, 145, 146} транспортным белкам,^{29, 30, 99} моделирования констант ингибирования ферментов,^{20, 35, 55, 74, 75, 80, 83, 145, 147} противораковой активности,^{20, 38, 43, 105, 150} канцерогенности,^{36, 76, 152, 153} мутагенности,^{69, 145, 154–156} антибактериальной активности¹⁵⁹ и т.д.

* * *

Искусственные нейронные сети представляют собой мощный и многогранный вычислительный метод решения задач, возникающих при обработке химической информации, а именно, при классификации и моделировании разрозненных экспериментальных данных. Количество публикаций в данной области увеличивается из года в год, что может свидетельствовать о неослабевающем интересе химиков-исследователей к моделированию зависимостей структура – свой-

ство с использованием аппарата искусственных нейронных сетей.

Литература

1. J.Zupan, J.Gasteiger. *Anal. Chim. Acta*, **248**, 1 (1991)
2. J.Gasteiger, J.Zupan. *Angew. Chem., Int. Ed. Engl.*, **105**, 503 (1993)
3. T.Aoyama, H.Ichikawa. *Chem. Pharm. Bull.*, **39**, 358 (1991)
4. J.A.Burns, G.M.Whitesides. *Chem. Rev.*, **93**, 2583 (1993)
5. J.Zupan, J.Gasteiger. *Neural Networks in Chemistry*. Wiley-VCH, Weinheim, 1999
6. И.В.Заенцев. *Нейронные сети: основные модели*. Воронеж. гос. ун-т, Воронеж, 1999
7. M.Lehtokangas, J.Saarinen. *Neurocomputing*, **20**, 265 (1998)
8. J.Y.F.Yam, T.W.S.Chow. *Neurocomputing*, **30**, 219 (2000)
9. I.Kovesdi, M.F.Dominguez-Rodriguez, L.Orfi, G.Naray-Szabo, A.Varro, J.G.Papp, P.Matyus. *Med. Res. Rev.*, **19**, 249 (1999)
10. *Neural Networks in QSAR and Drug Design*. (Ed. J.Devillers). Academic Press, London, 1996
11. А.Н.Горбань. *Обучение нейронных сетей*. ПараГраф, Москва, 1990
12. D.E.Rumelhart, G.E.Hinton, R.J.Williams. In *Parallel Data Processing. Vol. 1*. (Eds D.E.Rumelhart, J.L.McClelland). M.I.T. Press, Cambridge, MA, 1986. P. 318
13. *Stuttgart Neural Network Simulator*. User Manual, Version 4.0, 1995
14. *MatLab*. User Manual, Version 6.0, 2000
15. T.P.Vogl, J.K.Mangis, A.K.Rigler, W.T.Zink, D.L.Alkon. *Biol. Cybern.*, **59**, 256 (1988)
16. L.M.Patnaik, K.Rajan. *Neurocomputing*, **35**, 123 (2000)
17. M.Riedmiller, H.Braun. In *Proceedings of the IEEE International Conference on Neural Networks*. San Francisco, CA, 1993. P. 586
18. J.Dennis, R.B.Schnabel. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Prentice-Hall, Englewood Cliffs, NJ, 1983
19. M.J.Lee, J.T.Chen. *Ind. Eng. Chem. Res.*, **32**, 995 (1993)
20. Ajay. *J. Med. Chem.*, **36**, 3565 (1993)
21. L.Dolmatova, V.Tchistiakov, K.Ruckebusch, N.Dupuy, J.-P.Huvenne, P.Legrand. *J. Chem. Inf. Comput. Sci.*, **39**, 1027 (1999)
22. V.Kvasnicka, S.Sklenak, J.Pospichal. *J. Chem. Inf. Comput. Sci.*, **32**, 742 (1992)
23. V.Kvasnicka, J.Pospichal. *J. Mol. Struct. (THEOCHEM)*, **235**, 227 (1991)
24. V.Kvasnicka. *J. Math. Chem.*, **6**, 63 (1991)
25. D.B.Kireev. *J. Chem. Inf. Comput. Sci.*, **35**, 175 (1995)
26. И.И.Баскин, В.А.Палюлин, Н.С.Зефирова. *Докл. АН*, **333**, 176 (1993)
27. И.И.Баскин, В.А.Палюлин, Н.С.Зефирова. *J. Chem. Inf. Comput. Sci.*, **37**, 715 (1997)
28. B.Bienfait. *J. Chem. Inf. Comput. Sci.*, **34**, 890 (1994)
29. J.Polanski. *J. Chem. Inf. Comput. Sci.*, **36**, 694 (1996)
30. J.Polanski. *J. Chem. Inf. Comput. Sci.*, **37**, 553 (1997)
31. R.Kocjancic, J.Zupan. *J. Chem. Inf. Comput. Sci.*, **37**, 985 (1997)
32. V.S.Rose, I.F.Croall, H.J.MacFie. *Quantum Struct.-Act. Relat.*, **10**, 6 (1991)
33. R.Hecht-Nielsen. *Appl. Opt.*, **26**, 4979 (1987)
34. R.Hecht-Nielsen. *Neural Netw.*, **1**, 131 (1988)
35. M.Novic, Z.Nikolovska-Coleska, T.Solmajer. *J. Chem. Inf. Comput. Sci.*, **37**, 990 (1997)
36. M.Vracko. *J. Chem. Inf. Comput. Sci.*, **37**, 1037 (1997)
37. K.L.Peterson. *Anal. Chem.*, **64**, 379 (1992)
38. K.L.Peterson. *J. Chem. Inf. Comput. Sci.*, **35**, 896 (1995)
39. H.Lohninger. *J. Chem. Inf. Comput. Sci.*, **33**, 736 (1993)
40. J.Tetteh, T.Suzuki, E.Metcalf, S.Howells. *J. Chem. Inf. Comput. Sci.*, **39**, 491 (1999)
41. C.H.Wan, P.de B.Harrington. *J. Chem. Inf. Comput. Sci.*, **39**, 1049 (1999)
42. G.Espinosa, D.Yaffe, Y.Cohen, A.Arenas, F.Giralt. *J. Chem. Inf. Comput. Sci.*, **40**, 859 (2000)
43. Q.Liu, S.Hirono, I.Moriguchi. *Chem. Pharm. Bull.*, **40**, 2962 (1992)

44. W.Y.Choy, B.C.Sanctuary, G.Zhu. *J. Chem. Inf. Comput. Sci.*, **37**, 1086 (1997)
45. C.Cai, P.de B.Harrington. *J. Chem. Inf. Comput. Sci.*, **39**, 874 (1999)
46. V.V.Kovalishyn, I.V.Tetko, A.I.Luik, V.V.Kholodovych, A.E.P.Villa, D.J.Livingstone. *J. Chem. Inf. Comput. Sci.*, **38**, 651 (1998)
47. A.Micheli, A.Sperduti, A.Starita, A.M.Bianucci. *J. Chem. Inf. Comput. Sci.*, **41**, 202 (2001)
48. R.Todeschini, V.Consonni. *Handbook of Molecular Descriptors*. Wiley-VCH, Weinheim, 2000
49. Н.В.Артемченко, И.И.Баскин, В.А.Палюлин, Н.С.Зефиров. *Докл. АН*, **381**, 203 (2001)
50. М.И.Станкевич, И.В.Станкевич, Н.С.Зефиров. *Успехи химии*, **57**, 337 (1988)
51. L.Pogliani. *Chem. Rev.*, **100**, 3827 (2000)
52. M.Karelson, V.S.Lobanov, A.R.Katritzky. *Chem. Rev.*, **96**, 1027 (1996)
53. G.A.Bakken, P.C.Jurs. *J. Chem. Inf. Comput. Sci.*, **39**, 508 (1999)
54. H.E.McClelland, P.C.Jurs. *J. Chem. Inf. Comput. Sci.*, **40**, 967 (2000)
55. S.J.Patankar, P.C.Jurs. *J. Chem. Inf. Comput. Sci.*, **40**, 706 (2000)
56. D.W.Elrod, G.M.Maggiora, R.G.Trenary. *J. Chem. Inf. Comput. Sci.*, **30**, 477 (1990)
57. D.W.Elrod, G.M.Maggiora, R.G.Trenary. *Tetrahedron Comput. Methodol.*, **3**, 163 (1990)
58. I.I.Baskin, N.M.Halberstam, T.V.Mukhina, V.A.Palyulin, N.S.Zefirov. *SAR QSAR Environ. Res.*, **12**, 401 (2001)
59. V.N.Viswanadhan, G.A.Mueller, S.C.Basak, J.N.Weinstein. *J. Chem. Inf. Comput. Sci.*, **41**, 505 (2001)
60. A.M.Kshirsagar. *Multivariate Analysis*. Marcel Dekker, New York, 1972
61. M.Guo, L.Xu, H.Li, C.-Y.Hu. *Anal. Sci.*, **12**, 291 (1996)
62. H.Nefati, J.-M.Cense, J.-J.Legendre. *J. Chem. Inf. Comput. Sci.*, **36**, 804 (1996)
63. G.Castellano, A.M.Fanelli. *Neurocomputing*, **31**, 1 (2000)
64. S.-S.So, M.Karplus. *J. Med. Chem.*, **39**, 1521 (1996)
65. S.-S.So, M.Karplus. *J. Med. Chem.*, **40**, 4347 (1997)
66. S.-S.So, M.Karplus. *J. Med. Chem.*, **40**, 4360 (1997)
67. D.Rogers, A.J.Hopfinger. *J. Chem. Inf. Comput. Sci.*, **34**, 854 (1994)
68. B.T.Luke. *J. Chem. Inf. Comput. Sci.*, **34**, 1279 (1994)
69. M.W.Brinn, M.P.Payne, P.T.Walsh. *Chem. Eng. Res. Des.*, **71**, 337 (1993)
70. E.D.Karnin. *IEEE Trans. Neural Netw.*, **1**, 239 (1990)
71. J.H.Wikel, E.R.Dow. *Bioorg. Med. Chem. Lett.*, **3**, 645 (1993)
72. R.Reed. *IEEE Trans. Neural Netw.*, **4**, 740 (1993)
73. I.V.Tetko, A.E.P.Villa, D.J.Livingstone. *J. Chem. Inf. Comput. Sci.*, **36**, 794 (1996)
74. I.V.Tetko, V.Yu.Tanchuk, A.I.Luik. In *Proceedings 330 of AIP Conference. (E.C.C.C. 1 Computational Chemistry)*. (F.E.C.S. Conference). Nancy, 1995. P. 761
75. I.V.Tetko, V.Yu.Tanchuk, N.P.Chentsova, S.V.Antonenko, G.I.Poda, V.P.Kukhar, A.I.Luik. *J. Med. Chem.*, **37**, 2520 (1994)
76. Y.Isu, U.Nagashima, T.Aoyama, H.Hosoya. *J. Chem. Inf. Comput. Sci.*, **36**, 286 (1996)
77. D.Svozil, J.G.K.Sevcik, V.Kvasnicka. *J. Chem. Inf. Comput. Sci.*, **37**, 338 (1997)
78. D.Domine, J.Devillers, D.Wienke, L.Buydens. *Quantum Struct.-Act. Relat.*, **15**, 395 (1996)
79. D.Domine, J.Devillers, D.Wienke, L.Buydens. *J. Chem. Inf. Comput. Sci.*, **37**, 10 (1997)
80. T.A.Andrea, H.Kalayeh. *J. Med. Chem.*, **34**, 2824 (1991)
81. X.-H.Song, Z.Chen, R.-Q.Yu. *Sci. China, Ser. B*, **36**, 1443 (1993)
82. D.Cherqaoui, D.Villemin, A.Mesbah, J.M.Cense, V.Kvasnicka. In *Proceedings 330 of AIP Conference. (E.C.C.C. 1 Computational Chemistry)*. (F.E.C.S. Conference). Nancy, 1995. P. 767
83. S.-S.So, W.G.Richards. *J. Med. Chem.*, **35**, 3201 (1992)
84. I.V.Tetko, A.I.Luik, G.I.Poda. *J. Med. Chem.*, **36**, 811 (1993)
85. B.Curry, D.E.Rumelhart. *Tetrahedron Comput. Meth.*, **3**, 213 (1990)
86. A.S.Weigend, B.A.Huberman, D.E.Rumelhart. *Int. J. Neural Syst.*, **1**, 193 (1990)
87. D.T.Manallack, D.J.Livingstone. *Eur. J. Med. Chem.*, **34**, 195 (1999)
88. F.R.Burden, D.A.Winkler. *J. Med. Chem.*, **42**, 3183 (1999)
89. F.R.Burden, M.G.Ford, D.C.Whitley, D.A.Winkler. *J. Chem. Inf. Comput. Sci.*, **40**, 1423 (2000)
90. I.I.Baskin, M.I.Skvortsova, V.A.Palyulin, N.S.Zefirov. *Found. Comput. Decis. Sci.*, **22**, 107 (1997)
91. V.Kvasnicka, S.Sklenak, J.Pospichal. *J. Am. Chem. Soc.*, **115**, 1495 (1993)
92. J.Y.Choi, C.-H.Choi. *IEEE Trans. Neural Netw.*, **3**, 101 (1992)
93. S.-H.Oh, Y.Lee. *IEEE Trans. Neural Netw.*, **6**, 1005 (1995)
94. T.Aoyama, H.Ichikawa. *J. Chem. Inf. Comput. Sci.*, **32**, 492 (1992)
95. T.Aoyama, H.Ichikawa. *Chem. Pharm. Bull.*, **39**, 372 (1991)
96. H.Ichikawa, T.Aoyama. *SAR QSAR Environ. Res.*, **1**, 115 (1993)
97. I.I.Baskin, A.O.Ait, N.M.Halberstam, V.A.Palyulin, N.S.Zefirov. *SAR QSAR Environ. Res.*, **13**, 35 (2002)
98. G.M.Maggiora, D.W.Elrod, R.G.Trenary. *J. Chem. Inf. Comput. Sci.*, **32**, 732 (1992)
99. J.H.Schuur, P.Selzer, J.Gasteiger. *J. Chem. Inf. Comput. Sci.*, **36**, 334 (1996)
100. S.N.Jordan, A.R.Leach, J.Bradshaw. In *Proceedings 330 of AIP Conference. (E.C.C.C. 1 Computational Chemistry)*. (F.E.C.S. Conference). Nancy, 1995. P. 550
101. A. A.Gakh, E.G.Gakh, B.G.Sumpster, D.W.Noid. *J. Chem. Inf. Comput. Sci.*, **34**, 832 (1994)
102. A.Bruchmann, P.Zinn, C. M. Haffer. *Anal. Chim. Acta*, **283**, 869 (1993)
103. R.C.Glen, V.S.Rose, J.C.Lindon, R.J.Ruane, I.D.Wilson, J.K.Nicholson. *J. Planar Chromatogr.*, **4**, 432 (1991)
104. D.T.Manallack, D.D.Ellis, D.J.Livingstone. *J. Med. Chem.*, **37**, 3758 (1994)
105. T.Aoyama, Y.Suzuki, H.Ichikawa. *J. Med. Chem.*, **33**, 2583 (1990)
106. B.Lucić, N.Trinajstić. *J. Chem. Inf. Comput. Sci.*, **39**, 121 (1999)
107. B.Lucić, D.Amić, N.Trinajstić. *J. Chem. Inf. Comput. Sci.*, **40**, 403 (2000)
108. C.J.Barden, M.K.Boysworth, F.A.Palocsay. *J. Chem. Inf. Comput. Sci.*, **38**, 483 (1998)
109. J.Meiler, R.Meusinger, M.Will. *J. Chem. Inf. Comput. Sci.*, **40**, 1169 (2000)
110. B.E.Mitchell, P.C.Jurs. *J. Chem. Inf. Comput. Sci.*, **36**, 58 (1996)
111. D.L.Clouser, P.C.Jurs. *J. Chem. Inf. Comput. Sci.*, **36**, 168 (1996)
112. O.Ivanciuc, J.-P.Rabine, D.Cabrol-Bass, A.Panaye, J.P.Doucet. *J. Chem. Inf. Comput. Sci.*, **36**, 644 (1996)
113. J.Grunenberg, R.Herges. *J. Chem. Inf. Comput. Sci.*, **35**, 905 (1995)
114. Q.X.Guo, S.H.Luo, H.Wang, M.S.Zhang, Y.C.Liu. *J. Chem. Res. (S)*, 38 (1996)
115. L.Liu, Q.-X.Guo. *J. Chem. Inf. Comput. Sci.*, **39**, 133 (1999)
116. M.E.Brewster, M.-J.Huang, A.Harget, N.Bodor. *Tetrahedron*, **48**, 3463 (1992)
117. S.Ventura, M.Silva, D.Pérez-Bendito, C.Hervás. *J. Chem. Inf. Comput. Sci.*, **37**, 287 (1997)
118. A.T.Balaban, S.C.Basak, T.Colburn, G.D.Grunwald. *J. Chem. Inf. Comput. Sci.*, **34**, 1118 (1994)
119. L.M.Egolf, P.C.Jurs. *J. Chem. Inf. Comput. Sci.*, **33**, 616 (1993)
120. L.H.Hall, C.T.Story. *J. Chem. Inf. Comput. Sci.*, **36**, 1004 (1996)
121. A.A.Gakh, E.G.Gakh, B.G.Sumpster, D.W.Noid, L.D.Trowbridge. *J. Fluorine Chem.*, **73**, 107 (1995)
122. S.Liu, R.Zhang, M. Liu, Z.Hu. *J. Chem. Inf. Comput. Sci.*, **37**, 1146 (1997)
123. E.S.Goll, P.C.Jurs. *J. Chem. Inf. Comput. Sci.*, **39**, 974 (1999)
124. И.И.Баскин, В.А.Палюлин, Н.С.Зефиров. *Докл. АН*, **332**, 713 (1993)
125. B.E.Mitchel, P.C.Jurs. *J. Chem. Inf. Comput. Sci.*, **37**, 538 (1997)
126. E.S.Goll, P.C.Jurs. *J. Chem. Inf. Comput. Sci.*, **39**, 1081 (1999)
127. T.Suzuki, R.-U.Ebert, G.Schürmann. *J. Chem. Inf. Comput. Sci.*, **37**, 1122 (1997)
128. N.Bodor, A.Harget, M.-J.Huang. *J. Am. Chem. Soc.*, **113**, 9480 (1991)
129. B.E.Mitchell, P.C.Jurs. *J. Chem. Inf. Comput. Sci.*, **38**, 489 (1998)
130. J.Huuskonen, M.Salo, J.Taskinen. *J. Chem. Inf. Comput. Sci.*, **38**, 450 (1998)

131. J.M.Sutter, P.C.Jurs. *J. Chem. Inf. Comput. Sci.*, **36**, 100 (1996)
132. H.L.Engelhardt, P.C.Jurs. *J. Chem. Inf. Comput. Sci.*, **37**, 478 (1997)
133. B.E.Mitchell, P.C.Jurs. *J. Chem. Inf. Comput. Sci.*, **38**, 200 (1998)
134. J.M.Cense, B.Diawara, J.J.Legendre, G. Rouillet. In *Proceedings 330 of AIP Conference. (E.C.C.C. 1 Computational Chemistry). (F.E.C.S. Conference)*. Nancy, 1995. P. 556
135. J.J.Huuskonen, D.J.Livingstone, I.V.Tetko. *J. Chem. Inf. Comput. Sci.*, **40**, 947 (2000)
136. A.F.Duprat, T.Huynh, G.Dreyfus. *J. Chem. Inf. Comput. Sci.*, **38**, 586 (1998)
137. J.Devillers, D.Domine, C.Guillon, W.Karcher. *J. Pharm. Sci.*, **87**, 1086 (1998)
138. H.Kränz, V.Vill, B.Meyer. *J. Chem. Inf. Comput. Sci.*, **36**, 1173 (1996)
139. R.Schröder, H.Kränz, V.Vill, B.Meyer. *J. Chem. Soc., Perkin Trans. 2*, 1685 (1996)
140. H.Nefati, J.-J.Legendre, C.Michot. In *Proceedings 330 of AIP Conference. (E.C.C.C. 1 Computational Chemistry). (F.E.C.S. Conference)*. Nancy, 1995. P. 562
141. И.И.Баскин, А.О.Айт, Н.М.Гальберштам, В.А.Палюлин, М.В.Алфимов, Н.С.Зефилов. *Докл. АН*, **357**, 57 (1997)
142. M.Chastrette, D.Cretin, C.El Aidi. *J. Chem. Inf. Comput. Sci.*, **36**, 108 (1996)
143. O.Ivanciuc. *Rev. Roum. Chim.*, **40**, 567 (1995)
144. V.Kvasnicka, S.Sklenak, J.Pospichal. *Theor. Chim. Acta*, **86**, 257 (1993)
145. N.Ghoshal, S.N.Mukhopadhyay, T.K.Ghoshal, B.Achhari. *Indian J. Chem., Sect. B*, **32**, 1045 (1993)
146. S.-S.So, M.Karplus. *J. Med. Chem.*, **39**, 5246 (1996)
147. M.Jalali-Heravi, F.Parastar. *J. Chem. Inf. Comput. Sci.*, **40**, 147 (2000)
148. M.Wiese. *Quantum Struct.-Act. Relat.*, **10**, 369 (1991)
149. T.Aoyama, Y.Suzuki, H.Ichikawa. *J. Med. Chem.*, **33**, 905 (1990)
150. J.N.Weinstein, K.W.Kohn, M.R.Greuer, V.K.Viswanadhan, L.V.Rubinstein, A.P.Monks, D.A.Scudiero, L.Welch, A.D.Koutsokos, A.J.Chiausa, K.D.Paull. *Science*, **258**, 447 (1992)
151. S.Hatrik, P.Zahradnik. *J. Chem. Inf. Comput. Sci.*, **36**, 992 (1996)
152. R.Vendrame, R.S.Braga, Y.Takahata, D.S.Galvao. *J. Chem. Inf. Comput. Sci.*, **39**, 1094 (1999)
153. G.Gini, M.Lorenzini. *J. Chem. Inf. Comput. Sci.*, **39**, 1076 (1999)
154. N.Ghoshal, S.N.Mukhopadhyay, T.K.Ghoshal, B.Achhari. *Bioorg. Med. Chem. Lett.*, **3**, 329 (1993)
155. M.Brinn, P.T.Walsh, M.P.Payne, B.Bott. *SAR QSAR Environ. Res.*, **1**, 169 (1993)
156. И.К.Любимова, С.К.Абилев, Н.М.Гальберштам, И.И.Баскин, В.А.Палюлин, Н.С.Зефилов. *Изв. АН. Сер. биол.*, 180 (2001)
157. G.W.Kauffman, P.C.Jurs. *J. Chem. Inf. Comput. Sci.*, **40**, 753 (2000)
158. M.D.Wessel, P.C.Jurs, J.W.Tolan, S.M.Muskal. *J. Chem. Inf. Comput. Sci.*, **38**, 726 (1998)
159. J.Jaén-Oltra, T.Salabert-Salvador, F.J.Garcia-March, F.Pérez-Giménez, F.Tomás-Vert. *J. Med. Chem.*, **43**, 1143 (2000)

NEURAL NETWORKS AS A METHOD FOR ELUCIDATING STRUCTURE-PROPERTY RELATIONSHIPS FOR ORGANIC COMPOUNDS

N.M.Halberstam, I.I.Baskin, V.A.Palyulin, N.S.Zefirov

*N.D.Zelinsky Institute of Organic Chemistry, Russian Academy of Sciences
47, Leninsky prosp., 119991 Moscow, Russian Federation, Fax +7(095)135-5328
Department of Chemistry, M.V.Lomonosov Moscow State University
Leninskie Gory, 119992 Moscow, Russian Federation, Fax +7(095)939-3181*

Publications devoted to the use of neuronet approach to the modelling of structure-property relationships for organic compounds are surveyed. The main principles of the neuronet modelling are described together with the aspects related to representation and modelling of structural chemical data. Brief information on the neuronet models of spectral characteristics, reactivity, physicochemical properties and biological activity of organic compounds is presented.
Bibliography — 159 references.

Received 23rd July 2002