На правах рукописи

Скворцова Мария Ивановна

МАТЕМАТИЧЕСКИЕ МОДЕЛИ И АЛГОРИТМЫ В ИССЛЕДОВАНИЯХ СВЯЗИ МЕЖДУ СТРУКТУРОЙ И СВОЙСТВАМИ ОРГАНИЧЕСКИХ СОЕДИНЕНИЙ

05.13.18 – математическое моделирование,
численные методы и комплексы программ

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
доктора физико-математических наук

Москва – 2007

Работа выполнена
в Московской государственной академии тонкой химической технологии (МИТХТ)
им. М. В. Ломоносова

ОФИЦИАЛЬНЫЕ ОППОНЕНТЫ:

доктор физико-математических наук, профессор Сапоженко А.А.
доктор физико-математических наук, профессор Чернозатонский Л.А.
доктор химических наук, профессор Гюльмалиев А.М.

ВЕДУЩАЯ ОРГАНИЗАЦИЯ:

Институт математического моделирования РАН

Защита состоится «___» _____2007 г. в «___» час. на заседании диссертационного совета Д 501.001.43 при Московском государственном университете им. М.В. Ломоносова по адресу: 119992, г. Москва, ГСП-2, Ленинские горы, МГУ, факультет вычислительной математики и кибернетики, ауд. 685.

С диссертацией можно ознакомиться в библиотеке факультета вычислительной математики и кибернетики МГУ им. М.В.Ломоносова.

Автореферат разослан «____» _____2007 г.

Ученый секретарь
диссертационного совета, Захаров Е. В.
доктор физико-математических наук

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ.

1. АКТУАЛЬНОСТЬ ТЕМЫ. Проблема моделирования связи между структурой и свойствами органических соединений является одной из важнейших математических задач современной теоретической химии. Найденные закономерности позволяют, минуя эксперимент, прогнозировать свойства новых химических соединений непосредственно по их структуре и могут быть использованы для планирования целенаправленного поиска соединений с заданными свойствами.

К настоящему времени синтезировано огромное количество химических соединений (около 20 млн.), которые интенсивно вовлекаются в сферу практического использования. Однако экспериментальное определение различных свойств этих веществ (физико-химических, разных видов биологической активности) часто связано со значительными трудностями, возникающими, например, при получении достаточного количества вещества, его очисткой, возможной нестойкостью, токсичностью и т. д., и, кроме того, не всегда возможно. Такие исследования требуют значительных финансовых и временных затрат. В связи с этим разработка любых теоретических методов расчета свойств веществ по их структуре, минуя эксперимент, является актуальной научно-практической задачей. Следует отметить, что выявленные закономерности могут быть использованы и при разработке новых теорий о связи свойств веществ с их строением, а также при изучении механизмов действия биологически активных соединений.

Приведем краткую характеристику наиболее распространенного современного подхода к моделированию связи «структура-свойство». Имеется выборка соединений с известными численными значениями некоторого свойства этих соединений. Структура соединений описывается при помощи набора молекулярных параметров $x_1,\ldots,x_n$, в качестве которых используются топологические, электронные, геометрические характеристики молекул или значения каких-либо физико-химических свойств. Как правило, математическая модель связи «структура-свойство» в рамках этого подхода имеет вид уравнения, связывающего численные значения исследуемого свойства y и молекулярных параметров $x_1,\ldots,x_n$ при помощи некоторой функции f:

$$y=f(x_1,\ldots,x_n). \qquad (1)$$

Вид функции f предполагается известным, однако f зависит от ряда подгоночных параметров. Эти параметры подбираются по известным численным значениям рассматриваемого свойства соединений заданной выборки так, чтобы соотношение (1) выполнялось бы как можно более точно на этой выборке.

Модели связи «структура-свойство» могут иметь и другую форму, отличную от уравнения (1). Например, используются модели, определяемые заданием некоторой количественной меры молекулярного подобия $d(S_1,S_2)$ пары соединений S_1 и S_2, характеризующей количественно степень их сходства. Принцип расчета свойств соединений в рамках этого подхода базируется на постулате «близкие структуры имеют близкие свойства»: для оценки свойства какого-либо соединения S_0 в базе данных находят соединение S, ближайшее к S_0 по мере d, и полагают, что значения свойств этих соединений равны.

Важное место в вышеуказанных исследованиях занимают способы количественного описания структуры молекул, т.е. выбор параметров $x_1,\ldots,x_n$. От этого выбора значительно зависит эффективность модели. Параметры $x_1,\ldots,x_n$ могут быть как экспериментальными, так и расчетными. Использование расчетных параметров в моделях связи «структура-свойство» более предпочтительно, т. к. они могут быть вычислены даже для гипотетических структур. Для получения этих параметров в качестве основы используется классическая структурная формула молекулы, которую можно рассматривать как меченый граф. По структурной формуле могут быть построены другие меченые графы. Вершины таких графов, называемых молекулярными, обычно соответствуют атомам (или фрагментам), а ребра – химическим связям молекулы. Метки вершин кодируют атомы различной химической природы, а метки ребер – связи разного типа.

Метки типа буквенных символов характеризуют атомы и связи качественно, а числовые метки (веса) – количественно. Веса вершин и ребер могут быть взяты как из литературы (например, заряды ядер или ковалентные радиусы атомов), так и рассчитаны при помощи специальных стандартных программ, позволяющих определить электронные и геометрические характеристики молекул (например, могут быть найдены потенциалы ионизации, межатомные расстояния или рассчитаны заряды на атомах). На рис.1 в качестве примера приведена структурная формула 1,3-дихлорфенола и соответствующий ей меченый граф, в котором вершины соответствуют атомам углерода, а их метки A, B, C кодируют атомы углерода, в зависимости от присоединенных к ним фрагментов H, Cl или OH.

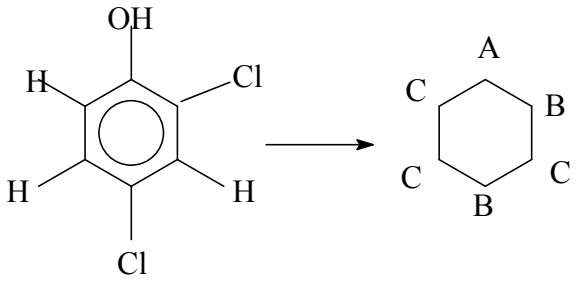

Рис.1.

Таким образом, каждой молекулярной структуре могут быть сопоставлены различные инварианты $x_1,\ldots,x_n$ соответствующего молекулярного графа (т.е. числа, вычисляемые по графу, не зависящие от способа нумерации его вершин). Инварианты графов, для построения которых использовалась лишь информация о топологии молекулы и, возможно, литературные данные о количественных характеристиках атомов и связях разного типа, в теоретической химии обычно называют топологическими индексами. Инварианты графов, связанных с пространственными моделями молекул, называют геометрическими дескрипторами. Если же для вычисления весов графа использовались квантово-химические методы, то соответствующие инварианты называют квантово-химическими дескрипторами. При построении молекулярного графа возможна и комбинация этих подходов. Отметим, что все вышеуказанные молекулярные параметры, имеющие различную химическую интерпретацию и различные способы их построения, имеют единую математическую основу – это инварианты меченых графов.

В последние десятилетия опубликовано большое число работ, посвященных моделированию связи «структура-свойство». В подавляющем большинстве случаев для описания молекулярной структуры используются разнообразные топологические индексы, что связано с относительной простотой их вычисления. Область научных исследований, связанная с математическим моделированим связи «структура-свойство», возникла на стыке органической химии, дискретной математики, регрессионного анализа, программирования и ее иногда рассматривают как часть математической химии или химической информатики. Многочисленные работы, посвященные этой тематике, публикуются в таких международных журналах, как *Journal of Chemical Information and Computer Science, Journal of Computational Chemistry, Journal of Mathematical Chemistry, Computers and Chemistry* и. т. д. Интенсивное развитие данного направления связано, прежде всего, с широким внедрением ЭВМ в химические исследования, созданием баз данных по структурам и свойствам соединений, а также доступностью вычислительной техники для химиков. Все это делает возможным проводить статистический анализ накопленной информации с целью выявления различных скрытых закономерностей. Наличие многочисленных примеров успешного применения вышеуказанного подхода для моделирования связи «структура-свойство» как для физико-химических свойств, так и для разных видов биологической активности, показывающих эффективность применяемого метода, также способствует развитию данного направления.

Однако, несмотря на наличие большого числа отдельных, частных результатов в этой области, полученных для конкретных свойств и выборок соединений, общие, универсальные, теоретически обоснованные методы построения и исследования моделей вида (1) в настоящее время не разработаны. В задачах такого типа заранее неизвестно, от каких именно структурных особенностей зависит данное свойство, и каким образом. Поиск моделей обычно происходит путем перебора некоторого ограниченного числа стандартных вариантов, причем, как правило, обнаруживаются корреляции между различными молекулярными параметрами. Поэтому разработка и обоснование общих подходов к моделированию связи «структура-свойство», применимых к произвольным свойствам и классам органических соединений и допускающих компьютерную реализацию и автоматизацию, является актуальной задачей. Основой для разработки таких подходов может служить общая математическая природа изучаемых химических объектов (структурных формул органических соединений) – все они представляют собой *меченые графы*, а также возможность рассматривать различные наблюдаемые свойства соединений (в предположении, что они зависят лишь от структуры молекул) как некоторые *инварианты* соответствующих графов.

2. ЦЕЛИ РАБОТЫ. При моделировании связи «структура-свойство» вышеописанным методом возникают следующие ***проблемы:***

*1) Выбор весов вершин и ребер молекулярного графа в конкретной задаче***.** Для решения этой проблемы нет определенных, обоснованных методов;

2) Выбор функции f (или меры молекулярного подобия d) и инвариантов $x_1,\ldots,x_n$ для описания структуры молекул в конкретной задаче. Отметим, что число инвариантов графов бесконечно даже для одного, фиксированного способа взвешивания графа. Как правило, большинство инвариантов, используемых в теоретической химии, получают при помощи формальных математических операций с графами, поэтому им трудно дать достаточно ясную физико-химическую или структурную интерпретацию. Заранее не известно, от каких именно структурных особенностей зависит данное свойство, и каким образом. Поэтому никаких четких правил выбора молекулярных параметров $x_1,\ldots,x_n$ и аппроксимирующей функции f (или меры d) для построения модели не существует*;*

3) Оценка области применимости модели связи «структура-свойство». Очевидно, что любая математическая модель, построенная по ограниченному набору данных, имеет свою область применимости. В связи с этим возникает задача определения области применимости модели связи «структура-свойство», т. е. определения того класса химических соединений, свойства которых могут быть рассчитаны при помощи построенной модели с заданной точностью. Прогнозирование свойств соединений без учета области применимости модели может привести к неверным результатам;

4) Разработка методов компьютерной генерации химических структур, обладающих заданной величиной свойства, на основе модели типа (1) (обратная задача в проблеме связи «структура-свойство»). Как отмечалось выше, основная цель построения моделей типа (1) - прогнозировать численные значения свойств других соединений из некоторого заданного набора, минуя эксперимент, и находить среди них соединения с требуемыми свойствами. Однако могут существовать соединения (возможно, еще не синтезированные), не входящие в этот набор, которые имеют требуемое значение рассматриваемого свойства. Такие новые, перспективные соединения не будут обнаружены при вышеописанном подходе. В связи с этим в рамках исследований связи «структура-свойство» естественно сформулировать так называемую *обратную задачу*, заключающуюся в исчерпывающей генерации структур, обладающих заданным значением свойства y_0. При наличии модели типа (1), где $x_1,\ldots,x_n$ - инварианты графов, эта проблема может быть сведена к математической задаче исчерпывающей генерации графов (возможно, определенного класса) с заданным значением инварианта $f(x_1,\ldots,x_n)$ и решена теоретико-графовыми методами. Однако уравнения типа (1) могут иметь разный вид, зависящий

от функции f и инвариантов $x_1,\ldots,x_n$. Отдельные методы решения обратных задач для конкретных случаев уравнения (1), учитывающие их специфику, не применимы к другим случаям. В связи с этим необходима разработка алгоритмов решения таких задач для наиболее типичных или общих случаев уравнения (1).

Цели работы связаны с указанными выше проблемами. Они таковы:

1) Разработать и теоретически обосновать ряд общих детерминированных методов построения теоретико-графовых моделей связи «структура-свойство» вида (1), применимых к различным свойствам и классам соединений, для случая, когда их структуры представлены *произвольно мечеными графами*. Провести тестирование предложенных методов моделирования связи «структура-свойство».

2) Разработать систему автоматической генерации инвариантов графов разнообразной природы, моделирующую логику действий человека, конструирующего инварианты для вышеуказанных задач (систему «искусственного интеллекта»), и исследовать ее возможности. Разработать на основе этой системы подход к моделированию связи «структура-свойство», альтернативный указанному выше детерминированному подходу, когда подходящий набор параметров для модели отбирается из конечного, достаточно большого числа инвариантов, сгенерированных автоматически с использованием процедуры случайного выбора. Провести тестирование предложенного метода построения моделей связи «структура-свойство».

3) Разработать обоснованные подходы для конструктивного определения областей применимости моделей вида (1) некоторых специальных типов и провести их тестирование.

4) Разработать алгоритмы решения обратных задач в проблеме связи «структура-свойство» на основе уравнений (1) различных видов и провести их тестирование .

5) Разработать методы построения моделей связи «структура-свойство» и прогнозирования свойств химических соединений на основе концепции молекулярного подобия и провести их тестирование.

6) Разработать ряд комбинаторных алгоритмов на графах, применяемых в компьютерной химии и химической информатике (алгоритмы поиска канонической нумерации вершин графа, установления изоморфизма графов, поиска группы симметрии графа, нахождения всех заданных подграфов в графе).

3. НАУЧНАЯ НОВИЗНА И ПРАКТИЧЕСКАЯ ЗНАЧИМОСТЬ РАБОТЫ.

Диссертационная работа посвящена разработке и обоснованию математических методов решения основных задач, возникающих при моделировании связи «структура-свойство» органических соединений: построения моделей, определения их областей применимости, конструирования химических соединений с заданными свойствами на основе построенных уравнений. В качестве исходных данных для такого моделирования используются базы данных по структурам и свойствам химических соединений. Обработка этих данных позволяет выявить скрытые закономерности между структурой и свойствами органических соединений. В качестве математических моделей химических соединений используются произвольно меченые графы. В диссертации:

1) Разработан и обоснован ряд новых *методов* построения моделей связи «структура-свойство» в терминах инвариантов молекулярных графов. Эти методы носят общий характер, применимы к произвольным свойствам и к произвольным выборкам химических соединений, представленных произвольно мечеными графами. Методы строго детерминированы и допускают компьютерную реализацию. Проведено *тестирование* предложенных подходов для моделирования связи «структура-свойство» для разнообразных свойств (физико-химические, биологическая активность, вычисляемые молекулярные параметры) и классов соединений, показавшее их практическую применимость и эффективность.

2) Разработана *интеллектуальная система,* предназначенная для автоматического конструирования произвольных наборов инвариантов графов различной природы для построения

корреляций «структура-свойство». В этой системе реализовано моделирование действий человека, конструирующего инварианты графа для вышеуказанной задачи. Предполагается, что выбор варианта действий в этом алгоритме в процессе конструирования происходит случайным образом. Использование случайного выбора позволяет освободиться от элементов субъективизма и выйти за рамки стандартного мышления в процессе такой деятельности. Проведено исследование возможностей этой системы. Показано, что основные, известные из литературы инварианты молекулярных графов (называемые в теоретической химии топологическими индексами) могут быть получены в рамках разработанной схемы. В то же соответствующий алгоритм позволяет получить принципиально новые пути построения инвариантов графов, в том числе и такие, которые практически не могут быть разработаны человеком «вручную». Предложенная схема, позволяет строить автоматически сколь угодно много инвариантов графов разного типа. Эти инварианты могут быть использованы при решении различных задач химической информатики, математической и компьютерной химии, в том числе при моделировании связи «структура-свойство». Следует отметить, что аналогов предложенной системы нет.

3) На основе разработанной схемы конструирования инвариантов графов предложен новый *метод* построения моделей связи «структура-свойство», а также проведено его *тестирование* для построения корреляций «структура-свойство» для физико-химических свойств и биологической активности органических соединений различных классов, показавшее его практическую применимость и эффективность.

4) Проведено исследование задачи *определения области применимости* модели связи «структура-свойство» для заданной допустимой погрешности расчета свойств соединений, а также предложен ряд *методов* ее решения. Проведено *тестирование* этих методов, показавшее, что использование областей применимости моделей при прогнозировании свойств соединений, определенных в соответствии с разработанными подходами, позволяет сократить долю ошибочных прогнозов.

5) Разработаны алгоритмизированные *методы* решения различных *обратных задач* в исследованиях связи «структура-свойство». Эти методы позволяют провести *исчерпывающую* генерацию химических структур определенного класса, имеющих заданное значение y_0 рассматриваемого свойства (или заданный интервал (y_1, y_2) значений свойства), на основе предварительно построенной модели вида $y=f(x_1,...,x_N)$, связывающей значения рассматриваемого свойства y и некоторые инварианты молекулярных графов $x_1,...,x_N$. Рассмотрены базовые корреляционные уравнения, содержащие различные инварианты, широко используемые при моделировании связи «структура-свойство» и допускающие определенную структурную интерпретацию. Проведено *тестирование* предложенных методов.

6) Предложены *модели* связи «структура-свойство» нового типа, которые отражают широко распространенный в химии постулат «близкие структуры имеют близкие свойства», позволяющие в ряде случаев оценивать свойство соединения на основе его сходства с другим соединением, для которого значение изучаемого свойства известно. Эти модели имеют следующий вид: $|y_i-y_j|=d(G_i,G_j)$, где y_i, y_j – значения свойств i–ого и j–ого соединений, представленных графами G_i и G_j, а $d(G_i,G_j)$ - некоторая симметричная функция двух аргументов G_i и G_j, значения которой количественно характеризуют степень подобия G_i и G_j. Предложен *метод* оптимального подбора меры $d(G_i,G_j)$ в этом соотношении, а также способ оценки свойств соединений на основе такой модели. Проведено *тестирование* метода, а также его *сравнение* с двумя другими методами, использующими другие меры подобия. Это сравнение показывает, что предложенный в работе метод дает более точный результат, чем остальные методы.

7) Разработан *алгоритм* оптимального подбора меры подобия при прогнозировании свойств соединений по методу «ближайшего соседа». Предлагаемый подход позволяет построить меру подобия, дающую наилучший результат при вышеуказанном способе прогнозирования свойств соединений, по крайней мере, для исходной выборки соединений.

Проведено *тестирование* метода и его *сравнение* с другими методами оценки свойств соединений, основанными на других мерах подобия. Это сравнение показывает, что предложенный в работе подход дает более точный результат, чем остальные методы.

8) Разработаны новые *комбинаторные алгоритмы* на графах, используемые при решении различных задач теоретической, компьютерной и математической химии, связанных с кодированием, идентификацией и анализом структурных особенностей графов. Эти алгоритмы позволяют строить каноническую нумерацию вершин графа, находить группу симметрии графа, устанавливать изоморфизм пары графов, находить все подграфы графа, изоморфные заданному подграфу. Алгоритмы математически обоснованы и применимы к графам произвольного вида, имеющим любые веса вершин и ребер.

9) Определены три *новых класса прикладных задач* в теории графов, имеющих практическое применение в области химии, а также предложены *методы* их решения или исследования. Полученные теоретико-графовые результаты являются основой алгоритмов моделирования связи «структура-свойство», разработанных в диссертации.

Первый класс задач связан с восстановлением аналитического вида инварианта меченых графов некоторого множества по всем или некоторым его значениям на графах этого множества. Для решения или исследования задач такого типа в работе предложена новая стратегия, основанная на введении и использовании понятия *базиса инвариантов меченых графов*. Предложены три определения базиса инвариантов графов, доказан ряд теорем о свойствах базисов, дана химическая интерпретация полученных математических результатов, предложены варианты наборов базисных инвариантов.

Второй класс задач связан с проблемой определения такого набора подграфов меченого графа (названных *базисными подграфами*), по которому граф восстанавливается однозначно. Предложена стратегия решения этой задачи, основанная на использовании ряда результатов спектральной теории графов. Получены теоретические результаты, позволяющие выявить один из возможных наборов таких подграфов.

Третий класс задач связан с нахождением аналитического вида произвольной симметричной меры подобия меченых графов. Выведена аналитическая формула для такой меры, из которой получен ряд важных следствий. Найденная формула позволяет строить меры подобия, удовлетворяющие определенным условиям и адаптировать их к конкретным химическим задачам.

10) Предложена формализация постулата *«близкие структуры имеют близкие свойства»,* являющегося основой некоторых методов прогнозирования свойств соединений, и проведено теоретическое исследование его справедливости. Указаны общие случаи, когда вышеуказанное утверждение будет заведомо верным или заведомо неверным. Актуальность таких исследований связана с широким внедрением компьютеров в химические исследования, что приводит к необходимости формализаций различных понятий и эмпирических правил, разработанных в химии. Кроме того, анализ этого постулата важен для обоснования методов прогнозирования свойств соединений, которые на нем основаны.

Таким образом, в работе предложен ряд новых математических моделей и алгоритмов в рамках исследований связи между структурой и свойствами органических соединений для случая, когда структура молекул представлена произвольно мечеными графами. Проведено тестирование предложенных методов, показавшее их практическую применимость и эффективность. Предложенные алгоритмы могут быть реализованы в виде компьютерных программ. Эти программы могут использоваться как самостоятельно, так и в составе уже имеющихся комплексов программ, предназначенных для исследования связи «структура-свойство». Следует отметить, что для решения одной и той же задачи (например, построения модели связи «структура-свойство», определения области ее применимости) в работе предлагается сразу *несколько* методов. Их совместное использование позволит повысить достоверность получаемых результатов.

Разработанные методы имеют большое практическое значение для моделирования связи между структурой и свойствами органических веществ, прогнозирования свойств соединений по их структуре, целенаправленного поиска соединений с заданными свойствами в области медицины, сельского хозяйства, промышленности, техники и т. д. Предложенные методы могут быть рекомендованы к внедрению в научно-исследовательских институтах, лабораториях и других организациях, занимающихся поиском соединений с определенным набором свойств разного профиля.

Полученные результаты могут быть включены в спецкурсы по математическому моделированию в химии, медицинской химии, теории графов, прикладной математике. Ряд приведеных в работе результатов был использован автором при чтении спецкурса по дисциплине «Теория графов» в МИТХТ им. М. В. Ломоносова.

4. ЛИЧНЫЙ ВКЛАД АВТОРА. Постановки задач, рассматриваемых в Главах 1-5, методы их решения, а также алгоритмы на графах из §6.2, §6.4 Главы 6 разработаны автором. Алгоритм из §6.3 Главы 6 разработан совместно с д.х.н. Трачом С. С. Теоретические результаты (определения, теоремы 1.1-1.12, 5.1-5.3) получены лично автором. Тестирование предложенных методов и алгоритмов в ряде случаев выполнено автором самостоятельно, а в ряде – совместно с соавторами публикаций по теме диссертации. Проведение компьютерно-статистических экспериментов по проверке гипотез о свойствах графов, описанных в §1.3-1.5, выполнено совместно с Федяевым К.С. В разработке компьютерных программ участвовали: Баскин И.И., Словохотова О.Л., Федяев К.С., Пасюков А.В., Дозор И.Н., Трач С.С., Гальперн Е.Г.

5. АПРОБАЦИЯ РАБОТЫ. Основные результаты диссертации были представлены на следующих конференциях и симпозиумах: Всесоюзной конференции «Использование вычислительных машин в химических исследованиях и спектроскопии молекул» (Рига, 1986); Всесоюзной школе-семинаре по автоматизации химических исследований (Тбилиси, 1988); Межреспубликанской научно-практической конференции «Синтез, фармакология и клинические аспекты новых психотропных и сердечно-сосудистых средств» (Волгоград, 1989);VIII - ой Всесоюзной конференции «Использование вычислительных машин в спектроскопии молекул и химических исследованиях» (Новосибирск, 1989); Межвузовских конференциях «Молекулярные графы в химических исследованиях» (Одесса, 1987; Калинин,1990); I-ой Всесоюзной конференции по теоретической органической химии (ВАТОХ) (Волгоград,1991); Symposium “QSAR and Molecular Modeling: Concepts, Computational Tools and Biological Applications” (Spain, Barcelona, 1995); 11-th European Symposium on Quantitative Structure - Activity Relationships: Computer-Assisted Lead Finding and Optimization, (France, Lausanne, 1996); International Conference on Inverse and Ill- Posed Problems (IIPP-96), (Russia, Moscow, 1996); International Symposium CACR - 96, (Russia, Moscow, 1996); IV-ом Российском научном конгрессе «Человек и лекарство» (Москва, 1997); I-ой, II-ой, III-ей, IV-ой Всероссийских конференциях «Молекулярное моделирование» (Москва, 1998г, 2001 г., 2003 г., 2005); Ninth International Workshop on Quantitative Structure-Activity Relationships in Environmental Sciences, (Bulgaria, Bourgas, 2000); International School-Seminar on Computer Automatization and Information, (Russia, Moscow, 2000); II-ом Международном симпозиуме «Компьютерное обеспечение химических исследований», (Москва, 2001); Memorial International Symposium “Modern Trends in Organometallic and Catalitic Chemistry. Mark Vol’pin (1923-1996)” (Russia, Moscow, 2003); Fourth Indo-US Workshop on Mathematical Chemistry (With Application to Drug Discovery, Environmental Toxicology, Chemoinformatics and Bioinformatics), (Pune, Maharashtra, India, 2005); 11-ой Международной конференции «Математические модели физических процессов» (Россия, Таганрог, 2005); XIX Международной научной конференции «Математические методы в технике и технологиях» (Россия, Воронеж, 2006).

Научные исследования по теме диссертации были поддержаны следующими грантами: INTAS-93-32-33 («Development of New Technique for Quantitative Structure-Activity Relationships and Molecular Design»); INTAS-00-03-63 («Virtual Computational Chemistry Laboratory – CCLAB»); РФФИ - №95-03-09696а («Разработка новых нейросетевых методов исследования связи между структурой и свойствами органических соединений. Компьютерное конструирование и синтез соединений с заданными свойствами»); РФФИ - № 98-03-32955а («Разработка новых методов компьютерного дизайна органических соединений с заданными свойствами на основе искусственных нейросетей. Конструирование и синтез перспективных структур»); РФФИ- №96-03-33003а («Математические модели, алгоритмы и программы решения задач дизайна органических реакций»).

6. ПУБЛИКАЦИИ. По теме диссертации опубликовано 73 работы, среди которых 35 статей в журналах и сборниках (в том числе 24 статьи в журналах, рекомендованных ВАК), 34 тезиса докладов на конференциях, 2 главы в монографиях, 2 учебно-методических пособия.

7. СТРУКТУРА И ОБЪЕМ ДИССЕРТАЦИИ. Диссертация состоит из введения, шести глав, выводов, списка цитированной литературы (210 наименований), списка публикаций автора по теме диссертации (73 наименования) и Приложения. Работа изложена на 272 стр., содержит 35 таблиц, 49 рисунков. Каждая глава посвящена отдельной тематике, рассматриваемой в рамках общей задачи исследования связи «структура-свойство», и имеет логическую завершенность. В Главе 1 разработан ряд детерминированных методов построе-ния моделей связи «структура-свойство» на основе базисных инвариантов и базисных подграфов молекулярных графов. В Главе 2 описана система автоматической генерации инвариантов графов для моделирования связи «структура-свойство», использующая элементы случайного выбора. В Главе 3 рассматриваются различные методы определения областей применимости моделей связи «структура-свойство». Глава 4 посвящена алгоритмам решения обратных задач в исследованиях связи «структура-свойство» на основе различных базовых моделей связи «структура-свойство». В Главе 5 предложены модели, связывающие степень близости свойств и степень сходства химических соединений, отражающие постулат «близкие структуры имеют близкие свойства». Глава 6 посвящена описанию ряда алгоритмов на графах, используемых для их кодирования, идентификации и исследования структурных особенностей. Приложение содержит краткие описания некоторых из компьютерных программ, использованных для тестирования разработанных методов.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ.

ГЛАВА 1. Методы построения моделей связи «структура-свойство» на основе базисных инвариантов и базисных подграфов молекулярных графов.

Постановки химических задач и их теоретико-графовые формулировки.

Рассматривается следующая общая ***проблема моделирования связи «структура-свойство»***: по заданной выборке органических соединений $\{S_i\}$ $(i=1,...,k)$, представленных классическими структурными формулами с известными численными значениями некоторого свойства $\{y_i\}$, построить уравнение вида $y=f(S)$, связывающее значения изучаемого свойства y и структуры S данных соединений при помощи некоторой функции f. Основная цель построения модели - оценить значения свойств y_i других соединений S_i, не включенных в исходную выборку. Следовательно, на этапе применения модели возникает задача определения ее области применимости, т. е. выделения такого подмножества структур в некотором заданном множестве $\{S_i\}$ $(i=k+1,...,N)$, свойства которых могут быть рассчитаны при помощи уравнения $y=f(S)$ с заданной допустимой погрешностью $\varepsilon \geq 0$.

Пусть ***математической моделью*** химического соединения S является произвольно ***меченый граф*** G, вершины и ребра которого соответствуют атомам и связям молекулы, а метки

вершин и ребер кодируют атомы и связи различной химической природы. Метки могут быть как числами, так и произвольными символами. Способ выбора меток и их интерпретация для дальнейших исследований не важны. Если отождествить структуру *S* с соответствующим молекулярным графом *G*, то свойство *y* (функцию от структуры) можно рассматривать как инвариант графа *y=f(G)* (т.е. число, определяемое по графу, значение которого не зависит от способа нумерации его вершин).

Для этого способа представления химических структур впервые предложены ***теоретико-графовые формулировки*** вышеуказанных общих задач, возникающих при моделировании связи «структура-свойство» и прогнозировании свойств соединений:

- задача построения уравнения типа *y=f(S)* равносильна задаче восстановления аналитического вида некоторого инварианта *y=f(G)* графа *G* по набору его значений $y_i=f(G_i)$ *(i=1,…,k)* на исходной выборке графов (возможно, с заданной погрешностью ε);

- задача определения области применимости построенной модели равносильна определению условий на граф *G* из некоторого множества $\{G_i\}$ *(i=k+1,…,N),* при которых значения инварианта *y=f(G)* на этом графе однозначно определяются по его значениям на заданных графах $\{G_i\}$ *(i=1,…,k)* (возможно, с заданной погрешностью ε).

Эти формулировки позволяют: а) ***определить новый класс прикладных задач*** в теории графов, имеющих практическое применение в области химии, а также разработать методы решения таких задач; б) ***применить аппарат теории графов*** для разработки и обоснования новых методов исследования связи «структура-свойство».

Исследование теоретико-графовых задач, связанных с проблемой моделирования связи «структура-свойство». Для решения или исследования вышеуказанных задач теории графов предложены две стратегии. Первая стратегия основана на использовании понятия *базиса инвариантов графов* заданного множества меченых графов, введенном в диссертации. *Базисом инвариантов графов* заданного множества в общем случае естественно назвать такой набор инвариантов, через который может быть выражен (при помощи некоторых функциональных соотношений) любой инвариант графов этого множества (возможно, неоднозначно). Вторая стратегия основана на использовании понятия *базисных подграфов* меченого графа, введенном в диссертации. *Базисными подграфами* меченого графа назван такой набор подграфов этого графа, по которому он восстанавливается однозначно.

1) Первая стратеия: поиск базисных инвариантов графов.

Базис инвариантов графов может быть определен разными способами. В Главе 1 введены три определения базиса, доказан ряд теорем о свойствах базисов, предложены возможные наборы базисных инвариантов, на основе полученных теоретических результатов разработаны общие методы построения моделей связи «структура-свойство».

● ***Определение 1 базиса инвариантов графов.*** Набор инвариантов $\{g_j\}$ *(j=1,…,M)* графов множества $\{G_i\}$ *(i=1,…,N)* назовем *базисным,* если любой инвариант *f(G)* графов этого множества однозначно представляется в виде линейной функции от них, т.е.:

$$f(G)=\sum_{j=1}^{M} a_j g_j(G), \quad (G \in \{G_i\}, j=1,\ldots,N),$$

где a_j *(j=1,…,M)* - некоторые константы, не зависящие от *G*, а зависящие только от *f*.

Сформулированы и доказаны ***теоремы*** о свойствах базиса в смысле *определения 1*.

ТЕОРЕМА 1.1 *(необходимые и достаточные условия на набор инвариантов, при которых они образуют базис).* Набор инвариантов $\{g_j\}$ *(j=1,…,M)* образует базис множества инвариантов графов $\{G_i\}$ *(i=1,…,N)* в смысле определения 1 тогда и только тогда, когда *M=N* и $detB \neq 0$, где $B=(b_{ij})$ - матрица с элементами $b_{ij}=g_j(G_i)$, *i, j=1,…,N*.

ТЕОРЕМА 1.2 *(описание множества всех базисов инвариантов).* Пусть $\{g_j\}$ $(j=1,\ldots,N)$ – некоторый базис инвариантов графов множества $\{G_i\}$ $(i=1,\ldots,N)$ в смысле определения 1, A – произвольная невырожденная квадратная матрица размера N. Построим набор инвариантов $\{h_j\}$ $(j=1,\ldots,N)$ по формуле:

$$\boldsymbol{h}=A\boldsymbol{g}\ , \qquad (2)$$

где $\boldsymbol{g}=(g_1,\ldots,g_N)$, $\boldsymbol{h}=(h_1,\ldots,h_N)$ - вектора – столбцы. Тогда:

1) Инварианты $\{h_j\}$ $(j=1,\ldots,N)$ также являются базисом инвариантов графов в смысле определения 1; 2) Любые два базиса $\boldsymbol{h}$ и $\boldsymbol{g}$ связаны между собой при помощи формулы (2) с некоторой невырожденной матрицей A.

ТЕОРЕМА 1.3 *(о существовании базиса инвариантов, равных числам вхождения в граф определенных подграфов)*. Рассмотрим множество графов $\{G_i\}$ $(i=1,\ldots,N)$. Тогда инварианты $g_j(G)$, равные числам вхождения подграфа $H_j=G_j$ $(j=1,\ldots,N)$ в граф G, образуют базис инвариантов графов заданного множества.

ТЕОРЕМА 1.4 *(о существовании базиса инвариантов, часть которых постоянна на выделенном подмножестве графов).* Пусть в множестве графов $\{G_i\}$ $(i=1,\ldots,N)$ выделено подмножество $\{G_i\}$ $(i=1,\ldots,k;\ k\le N)$. Тогда существует базис $\{f_p\}$ $(p=1,\ldots,N)$ инвариантов графов множества $\{G_i\}$ $(i=1,\ldots,N)$, такой, что его $N-k+1$ элемент постоянен на подмножестве $\{G_i\}$ $(i=1,\ldots,k)$. При этом $N-k+1$ - максимальное число базисных инвариантов, обладающих вышеуказанным свойством.

ТЕОРЕМА 1.5 *(характеристическое свойство графов выделенного подмножества графов).* Пусть в множестве графов $\{G_i\}$ $(i=1,\ldots,N)$ выделено подмножество $\{G_i\}$ $(i=1,\ldots,k;\ k\le N)$, а $\{f_p\}$ $(p=1,\ldots,N-k+1)$ - базис инвариантов, постоянных на подмножестве графов $\{G_i\}$ $(i=1,\ldots,k)$, т. е. $f_p(G_i)=c_p$, где c_p - некоторые константы, зависящие только от индекса p $(p=1,\ldots,N-k+1)$ (см. теорему 1.4). Тогда не существует графа G_i $(i=k+1,\ldots,N)$, такого, что $f_p(G_i)=c_p$ $(p=1,\ldots,N-k+1)$.

ТЕОРЕМА 1.6 *(Об общем виде произвольного инварианта на выделенном подмножестве графов).* Пусть в множестве графов $\{G_i\}$ $(i=1,\ldots,N)$ выделено подмножество $\{G_i\}$ $(i=1,\ldots,k;\ k\le N)$, а инварианты $\{f_p\}$ $(p=N-k+2,\ldots,N)$ и константы c_p $(p=1,\ldots,N-k+1)$ те же, что и в теоремах 1.4 и 1.5. Тогда на любом графе $G=G_i$ $(i=1,\ldots,k)$ инвариант f представляется в виде:

$$f(G)=a_0+\sum_{p=N-k+2}^{N} a_p f_p(G)\ , \qquad (a_0=\sum_{p=1}^{N-k+1} a_p c_p=const), \qquad (3)$$

причем коэффициенты $\boldsymbol{a}=(a_0,a_{N-k+2},\ldots,a_N)$ однозначно определяются по значениям $f(G_i)$ $(i=1,\ldots,k)$.

ТЕОРЕМА 1.7.*(необходимое и достаточное условие для восстановления значения инварианта графа по набору значений этого инварианта для других графов).* Пусть в множестве графов $\{G_i\}$ $(i=1,\ldots,N)$ выделено подмножество $\{G_i\}$ $(i=1,\ldots,k;\ k\le N)$. Значение инварианта $f(G)$ для графа $G\ne G_i$ $(i=1,\ldots,k)$ определяется по уравнению (3) тогда и только тогда, когда инвариант f и граф G удовлетворяют условию:

$$\sum_{p=1}^{N-k+1} a_p f_p(G)=a_0\ . \qquad (4)$$

Следствие из теоремы 1.7.

Из теоремы 1.7 следует, что для проверки возможности вычисления $f(G)$ $(G\ne G_i,\ i=1,\ldots,k)$ по $f(G_i)$ $(i=1,\ldots,k)$ необходимо знать значения a_p $(p=1,\ldots,N-k+1)$ (значения $f_p(G)$ и a_0 - известны). Однако их невозможно определить по исходным данным. Следовательно, без дополнительных предположений относительно инварианта f и графа G *в принципе невозможно* решить вышеуказанный вопрос. Однако можно указать следующие *достаточные условия* на f и G, при которых выполнено условие (4). Предположим, что инвариант f такой, что $a_p=0$ при некоторых значениях p $(1\le p\le N-k+1)$ (причем хотя бы для одного значения p), а граф G из множества $\{G_i\}$

(i=k+1,...,N) такой, что $f_p(G)=c_p$ для остальных значений *p, 1≤p≤N-k+1*. Легко видеть, что в этом случае выполнено условие (4).

Поставим следующий вопрос: можно ли вообще не накладывать вышеуказанные ограничения на инвариант *f*, а ввести ограничения только на граф *G*? Предположим, что $f_p(G)=c_p$ для любого *p, 1≤p≤N-k+1*. Однако, как было доказано ранее, такого графа *G* вообще не существует, и эти ограничения становятся бессмысленными.

ТЕОРЕМА 1.8 *(обобщение теоремы 1.7)*. Предположим, что задана допустимая точность *ε≥0* расчета значения инварианта *f(G), G=G_i (i=1,...,N)* и для графов *G=G_i (i=1,...,k)* получено приближенное уравнение вида

$$f'(G)=\sum_{p\in S_1} a_p' f_p(G)+a_0' \ , \qquad (5)$$

где S_1 -некоторое подмножество множества $S=\{N-k+2,...,N\}$ и $|f(G)-f'(G)|\le\varepsilon$, а *f(G)* определено по формуле (3). Обозначим $S_2=\{1,...,N-k+1\}$. Значение *f(G)* для графа *G=G_i (i=k+1,...,N))* вычисляется с точностью *ε* по уравнению (5) (т.е. $|f(G)-f'(G)|\le\varepsilon$) тогда и только тогда, когда

$$\left|\sum_{p\in S_1} f_p(G)(a_p-a_p')+\sum_{p\in S\setminus S_1} f_p(G)a_p+\sum_{p\in S_2} fp(G)a_p-a_0'\right|\le\varepsilon. \qquad (6)$$

Следствие из теоремы 1.8.

Сформулируем достаточные условия, при которых *f(G)* определяется по уравнению (5). Как и в случае теоремы 1.7, предположим, что *f* и *G* таковы, что при $p\in S_2$ либо $a_p=0$, либо $f_p(G)=c_p$.

Тогда $\sum_{p\in S_2} a_p f_p=a_0$,

а условие (6) примет вид:

$$\left|\sum_{S_1} f_p(G)(a_p-a_p')+\sum_{S\setminus S_1} f_p(G)a_p+a_0-a_0'\right|\le\varepsilon. \qquad (7)$$

Все величины, входящие в это неравенство, *определяются по начальным данным*, поэтому его можно использовать на практике.

ТЕОРЕМА 1.9 (*аналог теоремы 1.8*). Предположим, что задана допустимая точность *ε≥0* расчета значения инварианта *f(G), G=G_i (i=1,...,N)* и для графов *G=G_i (i=1,...,k)* получено точное уравнение (3), а из него - приближенное уравнение путем замены некоторых инвариантов f_p *(p=N-k+2,...,N)* на их средние значения

$$b_p=(1/k)\sum_{i=1}^{k} f_p(G_i)$$

на подмножестве графов G_i *(i=1,...,k)*. Обозначим: $S=\{N-k+2,...,N\}$, $S_2=\{1,...,N-k+1\}$, S_1 - множество номеров базисных инвариантов, оставшихся в приближенном уравнении. Таким образом, приближенное уравнение будет иметь следующий вид:

$$f'(G)=\sum_{p\in S_1} a_p f_p(G)+A_0 \qquad (A_0=a_0+\sum_{p\in S\setminus S_1} a_p b_p) \ , \qquad (8)$$

причем $|f(G_i)-f'(G_i)|\le\varepsilon$ *(i=1,...,k)*. Значение *f(G)* для графа $G\ne G_i$ *(i=1,..,k)* вычисляется с точностью *ε* по уравнению (8) (т.е. $|f(G)-f'(G)|\le\varepsilon$) тогда и только тогда, когда

$$\left|\sum_{p\in S\setminus S_1} (b_p-f_p(G))a_p+\sum_{p\in S_2} (c_p-f_p(G))a_p\right|\le\varepsilon \ . \qquad (9)$$

Следствие из теоремы 1.9.

Сформулируем достаточные условия, при которых *f(G)* определяется по уравнению (8). Как и в случае следствия из теоремы 1.7, предположим, что *f* и *G* таковы, что при $p\in S_2$ либо $a_p=0$, либо $f_p(G)=c_p$. Тогда $\sum_{p\in S_2} a_p f_p=a_0$,

а условие (9) примет вид

$$\Big/\sum_{S\backslash S_1}(b_p\text{-}f_p(G))a_p\Big/\leq\varepsilon. \qquad (10)$$

Все величины, входящие в это неравенство, *определяются по начальным данным,* поэтому его можно использовать на практике.

Методологические выводы из ТЕОРЕМ 1.1-1.9 и их интерпретация:

1) Из теорем 1.1-1.3 следует, что для *любой* выборки химических структур и *любого* свойства *всегда* можно построить бесконечно много точных линейных моделей связи «структура-свойство», используя базисные инварианты. При этом *всегда* в качестве базисных инвариантов можно взять числа вхождения в структуру определенных фрагментов (подграфов). В качестве таких подграфов могут быть использованы сами графы заданной выборки. На основании точных моделей можно строить приближенные, отбрасывая несущественные параметры. Таким образом, теоремы 1.1-1.3 являются основой новой общей, математически обоснованной методологии построения моделей связи «структура-свойство». Кроме того, эти результаты можно рассматривать как обоснование довольно распространенного в исследованиях связи «структура-свойство» фрагментного подхода, когда предполагается, что величина некоторого свойства представляется в виде суммы вкладов отдельных структурных фрагментов.

2) Теорема 1.4 позволяет описать множество всех инвариантов, каждый из которых принимает одно и то же значение на всех графах заданной выборки, т. е. найти *все общее* у заданных графов в терминах их инвариантов. Эта задача теории графов тесно связана с проблемой определения молекулярного сходства. Полученные результаты важны для корректного определения области применимости модели связи «структура-свойство», которая, в свою очередь, также связана с этим понятием. Обычно «сходство» соединений определяется путем визуального выявления некоторых общих имеющихся или отсутствующих фрагментов у структур выборки. Это равносильно тому, что рассматриваются следующие инварианты, связанные с определенными фрагментами: если данный фрагмент присутствует в структуре, то значение инварианта полагается равным «1», если нет, то значение инварианта равно «0». Таким образом, сходными объявляются те структуры, для которых эти инварианты принимают одинаковые значения. При этом выбор таких фрагментов происходит субъективным образом, и некоторые из них могут быть не обнаружены. Теорема 1.4 позволяет дать описание множества всех таких инвариантов, выявляя тем самым скрытые общие черты заданной выборки структур.

3) Однако, как следует из теоремы 1.5, в практических задачах *нельзя* использовать для определения сходства некоторой структуры и структур заданной выборки *все то общее* (в терминах инвариантов графов), что обнаружено у этих структур: никакая новая структура не будет иметь этих характеристик.

4) Теорема 1.7 связана с возможностью экстраполяции найденной зависимости «структура-свойство» на новые соединения. В ней даны необходимые и достаточные условия на исходную выборку соединений, на новое соединение, для которого осуществляется прогноз, на исследуемое свойство, при которых это возможно. Из этих условий, в частности следует, что: а) на основе исходных данных *в принципе невозможно* определить, принадлежит ли данный граф области применимости построенной модели; б) можно предложить достаточные условия на свойство и граф, при которых эта задача разрешима: свойство не должно зависеть от некоторых структурных особенностей (что можно только предполагать и нельзя получить из исходных данных), а граф должен обладать определенным сходством с графами исходной выборки; в) чем меньше структурных факторов влияет на рассматриваемое свойство, тем меньше ограничений требуется на новые структуры и тем шире область применимости построенной модели. Теоремы 1.8, 1.9 обобщают теорему 1.7 на случай, когда вычисление значений рассматриваемого свойства допускается с определенной погрешностью ε, а для вычислений используется приближенное уравнение. Таким образом, теоремы 1.4-1.9, могут служить основой для разработки новых,

математически обоснованных методов определения областей применимости моделей связи «структура-свойство».

Метод построения моделей связи «структура-свойство» и его тестирование. На основании полученных теоретических результатов предложен общий алгоритмизированный ***метод №1*** построения приближенной модели связи «структура-свойство» по набору N молекулярных графов. Метод заключается в следующем: для описания структуры графов рассматриваются N инвариантов, равных числам вхождения в произвольный граф графов этой выборки, а затем из них отбирается относительно небольшое число параметров, дающих модель удовлетворительной точности. Метод универсален: он позволяет построить точную модель связи «структура-свойство» для *любой выборки* химических соединений, представленных *любыми* мечеными графами и *любого* свойства химических соединений (физико-химического, биологической активности) или какого-либо вычисляемого молекулярного параметра. Таким образом, метод основан на определенном, строго детерминированном и теоретически обоснованном способе выбора инвариантов графов и аппроксимирующей функции в модели связи «структура-свойство». Число параметров, исключаемых из точной модели для получения приближенной модели заданной точности $\varepsilon \geq 0$, зависит от состава выборки, рассматриваемого свойства, числа ε, а также от способа представления химических соединений молекулярными графами. Метод может быть модифицирован следующим образом: наряду с вышеуказанными подграфами рассматриваются также подграфы самого «маленького» по числу вершин графа, и наилучший набор параметров отбирается из соответствующего объединенного набора.

Проведено ***тестирование*** предложенного метода на основе баз данных по разнообразным свойствам и классам соединений. Рассматривались: **1)-3)** алканы с известными значениями температуры кипения $t_{кип.}$, критической температуры $t_{кр.}$, критического давления $P_{кр.}$; **4)** сульфиды с известными значениям температуры кипения $t_{кип.}$; ***5)*** спирты с известными значениями параметра $y=-logX$, где X – растворимость соединения в воде; ***6)*** амины с известными значениями температуры кипения $t_{кип.}$; ***7)*** эфиры с известными значениями токсичного действия (на мышей) $y=-lgC$ (C - концентрация вещества, вызывающая заданный биологический эффект). Для оценки качества модели в соответствии с принятыми критериями использовались коэффициент корреляции R и среднеквадратичное отклонение s для регрессии, построенной для расчетных и экспериментальных значений свойства как для обучающей, так и для контрольной выборки соединений; рассматривались также коэффициент корреляции R_{cv} и среднеквадратичное отклонение s_{cv} для регрессии, полученной в процедуре «скользящего контроля» (“cross-validation”) в случае отсутствия контрольной выборки. Построенные модели обладают достаточно высокой точностью и имеют хорошую прогностическую способность, что свидетельствует об эффективности предложенного метода.

● Определение 2 базиса инвариантов графов.

Назовем набор инвариантов $\{g_i\}$ $(i=1,2,...)$ меченых графов некоторого множества $\{G_i\}$ $(i=1,2,...;\ G_{i1}\neq G_{i2}\ ,\ i1\neq i2)$ *базисным,* если: 1) для любых графов G_{i1} и G_{i2} и $(i1\neq i2)$ из этого множества вектора $\boldsymbol{g(G_{i1})}=(g_1(G_{i1}),g_2(G_{i1}),...)$ и $\boldsymbol{g(G_{i2})}=(g_1(G_{i2}),g_2(G_{i2}),...)$ различны; 2) любой инвариант $f(G)$ графов любого конечного подмножества графов исходного множества $\{G_i\}$ $(i=1,2,...)$ может быть представлен в виде некоторой функции h от $g_1, g_2,...$, т.е. $f(G)=h(g_1(G),g_2(G),...)$, причем h не зависит от G, а зависит от инварианта f и выбранного подмножества графов.

Отметим, что в *определении 2,* в отличие от *определения 1,* не требуется, чтобы: а) рассматриваемое множество графов было бы конечным; б) любой инвариант графа представлялся бы в виде линейной функции от базисных инвариантов; в) любой инвариант однозначно выражался бы через базисные инварианты.

Далее введены два набора инвариантов простых графов и проведено их исследование на базисность в смысле *определения 2*.

Для построения первого набора инвариантов рассматриваются все графы F_k с $k\geq 1$ вершинами, состоящие из объединения нескольких несвязных компонент, каждая из которых является либо цепью, либо циклом, или циклом, к некоторым вершинам которого присоединено еще по одной вершине. В случае $k=1$ граф F_1 состоит из одной вершины. Все такие графы для одного фиксированного k нумеруются произвольным образом и обозначаются через $F_{k,m}$ $(m=1,2,...)$. На рис. 2 приведены все такие графы при $k=5$. Пусть $x_{k,m}$ - инвариант, равный числу вхождения в некоторый граф G подграфа $F_{k,m}$.

Второй набор инвариантов строится на основе первого следующим образом. Нумеруются все вхождения $F_{k,m}$ в граф G и j-ое вхождение обозначается через $F_{k,m,j}$. Каждому $F_{k,m,j}$ сопоставляется число

$$\mu_{k,m,j} = \sum \frac{1}{\sqrt[n_i]{v_1 v_2 \ldots v_{n_i}}},$$

где суммирование проводится по всем компонентам связности $F_{k,m,j}$, n_i - число вершин в i - ой компоненте, v_p (p=1,2,…) - степени вершин $F_{k,m,j}$ в G. Инвариант $\varphi_{k,m}$ определяется так:

$$\varphi_{k,m} = \sum_j \mu_{k,m,j} \quad .$$

Для исследования наборов инвариантов $\{x_{k,m}\}$ и $\{\varphi_{k,m}\}$ на базисность в смысле *определения 2* использованы разные методы исследования: 1) строгое математическое доказательство соответствующих утверждений для графов определенных классов; 2) выявление на основе некоторых теоретических результатов тех наборов графов, для которых могут нарушаться условия базисности; нахождение таких графов в разных классах графах с последующей непосредственной проверкой соответствующих утверждений для них; 3) проведение компьютерно-статистического эксперимента, в ходе которого случайным образом генерируются различные выборки графов и для них проверяется выдвигаемая гипотеза.

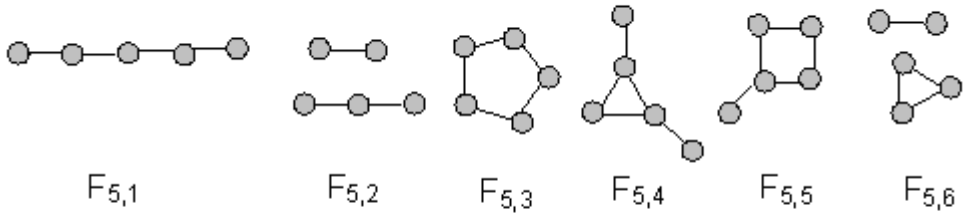

Рис. 2. Все графы $F_{k,m}$ при $k=5$.

ТЕОРЕМА 1.10. Набор инвариантов $x_{k,m}$ является базисным (в смысле определения 2) для множества графов $\{G_i\}$ $(i=1,2,..,)$, состоящего из графов типа $F_{k,m}$.

Аналогичные результаты получены и для набора $\{\varphi_{k,m}\}$.

При исследовании базисности введенных инвариантов методами 2) и 3) во всех рассмотренных случаях выдвигаемая гипотеза оказалась верна; при этом было установлено, что в качестве функции h можно взять полином степени не более двух.

На основании результатов проведенных комбинированных исследований наборы инвариантов $\{x_{k,m}\}$ и $\{\varphi_{k,m}\}$ были названы *базисными (*в смысле *определения 2).*

Метод построения моделей связи «структура-свойство» и его тестирование.

Предложен общий алгоритмизированный ***метод №2*** построения моделей связи «структура-свойство», основанный на введенных выше инвариантах $\{\varphi_{k,m}\}$. Согласно этому

методу, для данной выборки молекулярных графов строятся все возможные инварианты $\{\varphi_{k,m}\}$, а также их квадраты и попарные произведения, и затем из этого набора параметров отбирается небольшое число параметров, дающих удовлетворительную линейную модель.

Проведено ***тестирование*** предложенного метода на основе баз данных по углеводородам различных классов с различными свойствами. Рассматривались: **(1-4)** температура кипения, критическая температура, молярная рефракция, молярный объем алканов; **(5)** температура кипения циклосодержащих углеводородов; **(6)** полная π-электронной энергия бензоидных углеводородов. Полученные модели обладают достаточно высокой точностью.

● Определение 3 базиса инвариантов графов.

Назовем семейство инвариантов $\{Z_1(\alpha_1), Z_2(\alpha_2),\ldots\}$ произвольного набора графов $\{G_i\}$, зависящих от параметров $\alpha_1, \alpha_2,\ldots$, *базисным*, если для любого инварианта f любой выборки графов $\{G_1,\ldots,G_N\}$ найдется N инвариантов $Z_{j1}(\alpha_1),\ldots, Z_{jN}(\alpha_N)$ из этого множества, и N чисел $\alpha_1',\ldots, \alpha_N'$, таких, что рассматриваемый инвариант f однозначно представляется в виде линейной комбинации $Z_1(\alpha_1'),\ldots,Z_N(\alpha_N')$:

$$f=\sum_{i=1}^{N} c_i Z_{ji}(\alpha_i').$$

Далее введено семейство инвариантов $\{\psi_{k,m}(\alpha)=\varphi_{k,m}/n^{\alpha}\ (k,m\geq 1);\ \psi_{0,0}(\alpha)=n^{\alpha}\}$, где n - число вершин графа G, $\alpha=\alpha(k,m)\geq 0$ - произвольный параметр, который для каждой пары (k,m) может принимать любые значения. Это семейство инвариантов является обобщением рассмотренного ранее набора инвариантов $\{\varphi_{k,m}\}$.

Проведено исследование семейства инвариантов $\{\psi_{k,m}(\alpha)\}$ на базисность в смысле *определения 3*. Для этой цели использованы два различных метода: 1) строгое математическое доказательство выдвигаемой гипотезы для определенных классов графов; 2) проверка гипотезы в ходе компьютерно-статистического эксперимента.

ТЕОРЕМА 1.11. Инварианты $\{\psi_{k,m}(\alpha)\}$ являются базисными в смысле *определения 3* для любого множества графов $\{G_i\}$, $i=1,\ldots,N$, удовлетворяющего одному из следующих условий: а) все графы данного множества имеют различное число вершин $n_1, n_2,\ldots, n_N$; б) каждый граф из данного множества является графом типа $F_{k,m}$ при некоторых (k,m).

Для графов произвольного множества базисность соответствующих инвариантов проверялась при помощи компьютерно-статистического эксперимента, описанного выше. Во всех рассмотренных случаях выдвинутая гипотеза оказалась справедливой. На основании полученных результатов введенные параметры были названы *базисными (*в смысле *определения 3)*.

Метод построения моделей связи «структура-свойство» и его тестирование.

Разработан общий алгоритмизированный ***метод №3*** построения моделей связи «структура-свойство». Метод заключается в следующем: 1) задается конечный набор M значений параметра α: $\alpha_1=0, \alpha_2,\ldots,\alpha_M$ с фиксированным значением шага h и заданным максимальным значением α_M; 2) строятся инварианты $\{\psi_{k,m}(\alpha)\}$ для всех фрагментах $F_{k,m}$, которые присутствуют в заданном множестве структур, при всех выбранных значениях параметра α; 3) из этого множества инвариантов отбираются наилучшие для построения линейной модели. Если полученный результат является неудовлетворительным (по каким-либо критериям), то процедура повторяется для других значений α_M или h.

Проведено ***тестирование*** предложенного метода. Для этой цели было использовано несколько баз данных по физико-химическим свойствам углеводородов различных классов и значениям некоторых широко известных топологических индексов. Рассматривались следующие свойства: **1)** температура кипения; **2)** критическая температура; **3)** молярная рефракция; **4)** теплота образования; **5)** теплота сгорания; **6)** критическое давление; **7)** молярный объем; **8)**

теплота испарения; **9)** поверхностное натяжение; **10)** плотность; **11)** энтальпия образования; **12)** температура плавления; **13)** энергия Гиббса; **14)** удельная теплоемкость; **15)** показатель преломления. В качестве топологических индексов были взяты индексы Винера, Хосойя, молекулярной связности, индексы молекулярной формы Кира, полная π-электронная энергия. Рассмотренные базы разбивались на обучающую и контрольную выборки так, чтобы число структур в последней составляло примерно 10% от общего числа структур базы. По обучающей выборке строилось уравнение связи «структура – свойство»; затем оно использовалось для расчета свойств соединений контрольной выборки. Было построено 27 моделей, для каждой из которых определялись коэффициент корреляции и среднеквадратичное отклонение как для обучающей выборки, так и для контрольной. В этих примерах были использованы значения $h=0.1$, $\alpha_M=3, 4, 5, 6$.

Полученные результаты свидетельствуют об эффективности предложенного метода: построенные модели обладают высокой точностью и имеют хорошую прогнозирующую способность. Таким образом, разработанный метод позволяет *единообразно* описывать *различные* свойства *разнообразных* классов углеводородов.

2) Вторая стратегия: поиск базисных подграфов графа. Рассматривается задача поиска такого набора подграфов взвешенного графа G, по которому граф G может быть восстановлен однозначно (т. е. *базисных подграфов*). При этом желательно, чтобы среди этих подграфов были бы подграфы с относительно небольшим числом вершин.

Идея поиска таких подграфов основана на следующих известных результатах спектральной теории графов: 1) собственные числа взвешенного графа с n вершинами однозначно определяются по набору его подграфов на $k=1,2,...,n$ вершинах, состоящих из объединения изолированных вершин, ребер и циклов; 2) граф однозначно определяется по набору его собственных чисел и соответствующих линейно независимых собственных векторов; однако в общем случае граф не определяется однозначно по набору собственных чисел. В связи с этим возникает следующая задача: найти подграфы, определяющие однозначно и собственные вектора графа. Отметим, что вышеуказанная проблема для собственных векторов более сложная, чем для собственных чисел, так как: 1) собственные вектора зависят от собственных чисел; 2) в общем случае может быть несколько линейно-независимых собственных векторов, соответствующих одному и тому же собственному числу; 3) компоненты собственных векторов зависят от нумерации вершин графа.

В этом разделе Главы 1 дано решение вышеуказанной проблемы: выведены формулы, связывающие собственные вектора графа и его некоторые подграфы. Полученные результаты сформулированы в виде ***теоремы 1.12***. На их основе выделен объединенный набор подграфов, который используется для определения как собственных чисел, так и собственных векторов графа. Эти подграфы названы *базисными.*

Метод построения моделей связи «структура-свойство» и его тестирование. На основе полученных теоретических результатов, связанных с базисными подграфами, предложен общий алгоритмизированный ***метод №4*** построения моделей связи «структура-свойство». Согласно этому методу, для описания структуры молекулярных графов рекомендуется использовать инварианты, равные числам вхождения в граф введенных в работе базисных подграфов, а в качестве аппроксимирующей функции в модели следует использовать многочлен нескольких переменных от этих параметров. Предложено две методики построения этого многочлена.

Проведено ***тестирование*** предложенного метода на основе баз данных по биологической активности разнообразных классов соединений, а также его сравнение с другими методами моделирования связи «структура-свойство» на используемых данных. Рассматривались: ***1)*** галоидпроизводные метана и этана с известными значениями их наркотической активности $lnAD_{50}$ (AD_{50} - концентрация вещества, вызывающая анестезию у половины подопытных

животных); ***2)*** нитробензолы и нитротолуолы с известными значениями мутагенной активности $ln\mu$ (на *Salmonella typhimurium*, μ - количество ревертантов на наномоль); ***3)*** хлорзамещенные анилины с известными значениями токсичности $logEC_{50}^{-1}$, где EC_{50} - концентрация вещества, вызывающая уменьшение интенсивности люминесценции в 2 раза у морских бактерий *Photobacterium phosphoreum*. Построенные модели обладают достаточно высокой точностью, что свидетельствует об эффективности предложенного метода.

Таким образом, в Главе 1 разработаны и обоснованы четыре новых *метода* построения моделей связи «структура-свойство» в терминах инвариантов молекулярных графов. Методы носят *общий* характер, применимы к *произвольным* свойствам и *произвольным* выборкам химических соединений. Два из них позволяют учесть метки соответствующих молекулярных графов, которые могут быть произвольными символами; два других используют представления структур в виде простых графов. Методы строго детерминированы и допускают компьютерную реализацию. Проведено тестирование предложенных подходов для моделирования связи «структура-свойство» для *разнообразных* свойств (физико-химических, биологической активности), вычисляемых молекулярных параметров и классов соединений, показавшее их широкую практическую применимость и эффективность. Кроме того, получен ряд новых *теоретических результатов* в области теории графов, являющихся основой для разработки соответствующих алгоритмов.

ГЛАВА 2. Система автоматической генерации инвариантов графов для моделирования связи «структура-свойство».

Постановка задачи: ***разработать алгоритм конструирования инвариантов графов***: 1) моделирующий действия человека, строящего инварианты для использования их в корреляциях «структура-свойство»; 2) в котором выбор элементарных шагов в процессе конструирования инвариантов происходит случайным образом; 3) позволяющий генерировать как известные, так и новые инварианты графов. Цель разработки такой системы – получать произвольное количество разнообразных инвариантов графов для построения на их основе моделей связи «структура-свойство».

Целесообразность создания вышеуказанной системы обусловлена тем, что не всегда удается построить достаточно хорошие корреляции «структура-свойство», используя для этих целей даже достаточно большие наборы вполне определенных параметров, построенных «вручную». Это связано с тем, что: а) инвариантов графов в принципе существует бесконечно много, и использование какого-либо одного и того же конечного, фиксированного набора инвариантов для всех случаев не всегда приводит к требуемому результату; б) как правило, в процессе построения конкретной модели обнаруживаются корреляции между различными инвариантами. Последнее можно объяснить, в частности, тем, что при конструировании инвариантов «вручную» часто происходит применение одних и тех же приемов построения и действий «по аналогии».

Система автоматической генерации инвариантов графов. В Главе 2 детально описана ***интеллектуальная система***, предназначенная для автоматического (компьютерного) конструирования инвариантов графов для построения корреляций «структура-свойство», удовлетворяющая вышеперечисленным требованиям. Для создания такого алгоритма было проанализировано около сотни известных из литературы инвариантов графов, нашедших успешное применение при построении корреляций «структура-свойство». На основании проведенного анализа выделено несколько достаточно простых процедур, допускающих формальное описание. Установлено, что из этих процедур конструируются алгоритмы построения известных инвариантов путем их определенного сочетания, в том числе и размещения одной процедуры внутри другой. При этом в процессе выполнения каждой такой процедуры необходимо произвести выбор одного варианта из нескольких возможных. В связи с

отсутствием теоретического обоснования (как с точки зрения математики, так и с точки зрения теоретической химии) принятия того или иного решения, в разработанном алгоритме предложено любой выбор проводить случайным образом. Однако выбор может быть сделан и исследователем. В этом случае процесс конструирования инвариантов будет управляемым.

Алгоритм описан в терминах блок-схем и состоит из двух последовательных этапов: 1) Создание Базы Матриц (БМ) графа; 2) Построение инвариантов графа по матрицам из БМ или по другим инвариантам. В связи с необходимостью выбора одного варианта из нескольких возможных на разных этапах алгоритма неотъемлемой частью структуры алгоритма являются предварительно составленные Списки возможных вариантов действий. Эти Списки можно как сокращать, так и расширять, добавляя в них новые варианты.

Далее в качестве примера на рис.3 приведена блок-схема 1-ого этапа. В качестве входных данных на этом этапе используется матрица смежности (или весов) $A_0=(a_{ij})$ графа. Результатом работы алгоритма на этом этапе является База Матриц (БМ) введенного графа, полученных из A_0 по разным правилам. Матрица A_0 также заносится в БМ. На 1-ом этапе задаются Списки 1-5, содержащие варианты преобразования A_0**.** Например, в Списке 1 приведены варианты начальных весов вершин графа, в Списке 2 - варианты начальных весов пар вершин; Списки 3 и 4 содержат варианты преобразований весов вершин или весов пар вершин. Так как некоторые варианты в Списках 1-5 предполагают использование каких-либо функций или определенных подграфов, то также вводятся дополнительные Списки 6-9 (перечни функций f одной переменной, симметричных функций F многих переменных, симметричных функций g двух векторных аргументов; перечень специальных подграфов).

Исследование возможностей системы генерации инвариантов графов. Показано, что основные, известные из литературы инварианты молекулярных графов (называемые в теоретической химии топологическими индексами) могут быть получены в рамках разработанной схемы. Рассмотрено 42 топологических индекса различного типа, причем некоторые из них в действительности представляют собой целые семейства инвариантов. Примерами таких являются индексы связности порядка $h\geq 1$, для вычисления которых рассматриваются все цепи фиксированной длины $h\geq 1$ в графе, или информационные индексы порядка $k\geq 1$, где k - номер координационной сферы атома.

В то же время при анализе структуры алгоритма и содержания списков возможных вариантов, заложенных в него, выявляются принципиально новые пути построения инвариантов графов, которые могут оказаться полезными в корреляциях «структура-свойство». При реализации алгоритма можно получить довольно сложные и громоздкие по конструкции инварианты, которые практически не могут быть построены человеком «вручную», но также могут оказаться полезными в вышеуказанных задачах.

Используя предложенную схему, которая является, по сути, *алгоритмом генерации алгоритмов генерации инвариантов*, можно строить автоматически сколь угодно много инвариантов разного типа при помощи компьютера.

Метод построения моделей связи «структура-свойсво» на основе системы генерации инвариантов графов и его тестирование. Предложен следующий ***метод*** построения моделей связи «структура-свойство». Сначала генерируется некоторое множество инвариантов, затем из них выбирается небольшое число наилучших каким-либо стандартным образом (например, при помощи пошаговой линейной регрессии). Если результат оказался неудовлетворительным (с точки зрения какого-либо критерия), то можно расширить или заменить исходное множество инвариантов, используя генератор инвариантов повторно. Кроме того, можно построить много разных моделей для одних и тех же данных, и использовать для оценки свойств соединений все эти модели, усредняя получаемые результаты.

Проведено ***тестирование*** предлагаемого подхода для построения корреляций «структура-свойство» для физико-химических свойств и биологической активности органических соединений различных классов. Рассматривались: ***1)-5)*** энтальпия образования,

температура кипения, критическая температура, критическое давление, 3D-индекс Винера ${}^3W$ алканов C_2-C_8; ***6)-7)*** ингибирование микросомального пара-гидроксилирования анилина цитохромом P_{450} (степень ингибирования характеризуется величиной pIC_{50},=-$lgIC_{50}$, где IC_{50} - концентрация вещества, приводящая к 50% ингибированию гидроксилирования анилина), а также температура кипения $t_{кип.}$ алифатических спиртов; ***8), 9)*** параметр гидрофобности $logP$ (P – коэффициент распределения соединения между водой и н-октанолом), а также токсичность, характеризуемая величиной $logEC_{50}^{-1}$ (EC_{50} – концентрация вещества, вызывающая 50% уменьшение биолюминисценции морских бактерий *Photobacterium phosphoreum* в течение 30 мин.) хлорзамещенных фенолов. Полученные результаты свидетельствуют об эффективности предложенного подхода.

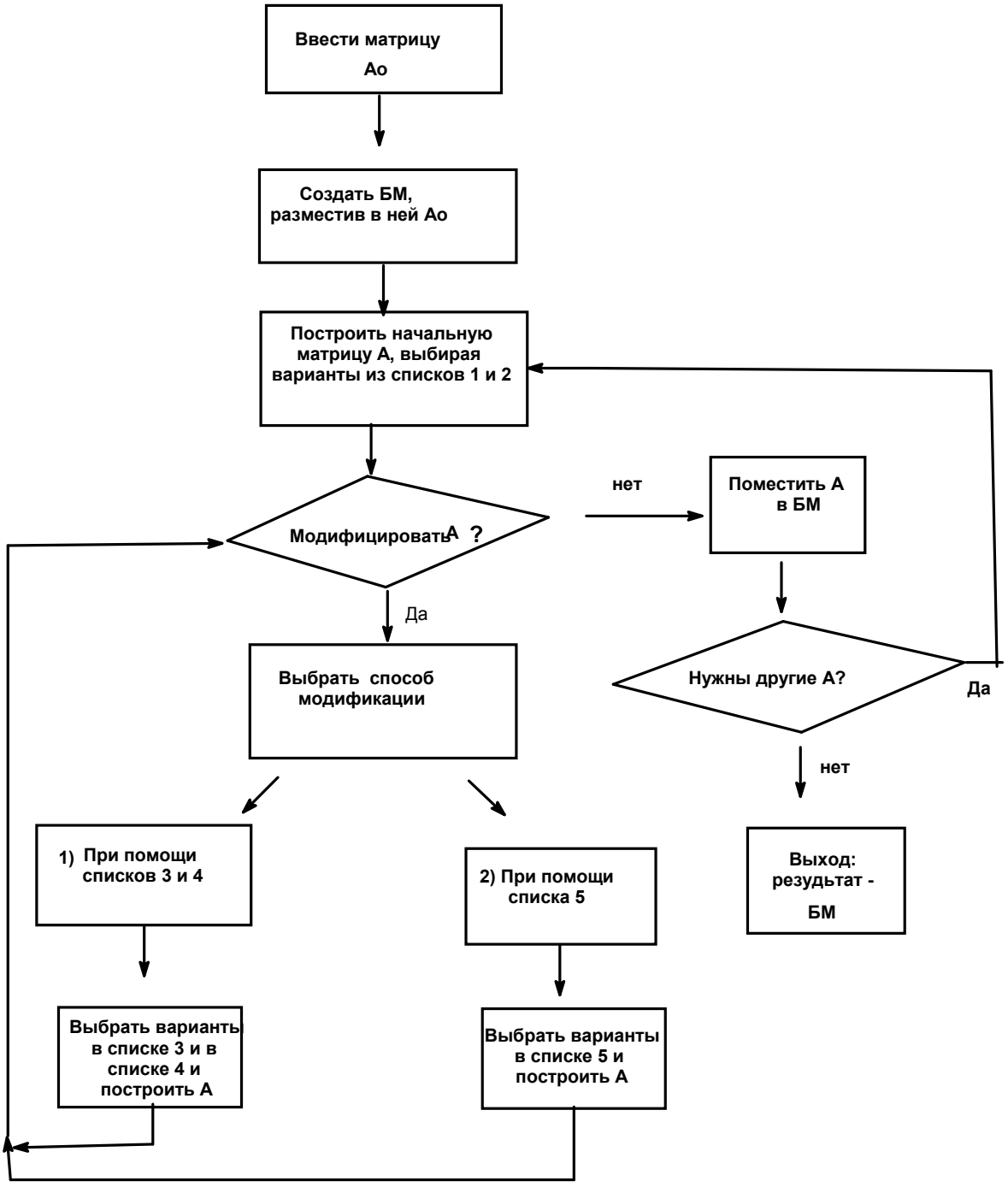

Рис.3. Блок-схема 1-ого этапа алгоритма генерации инвариантов графа.

Таким образом, в Главе 2 разработана *система автоматичекой генерации инвариантов графов* различной структуры (топологических индексов) и в любом заданном количестве. В ней используются элементы случайного выбора возможных элементарных шагов в процессе конструирования инвариантов. Система позволяет получать как основные известные инварианты графов (топологические индексы), так и новые, которые вряд ли могут быть построены «вручную». На основе разработанного алгоритма предложен новый *метод* построения моделей связи «структура-свойство», а также приведены примеры его применения для различных физико-химических свойств соединений и видов биологической активности. Следует отметить, что аналогов предложенной системы нет.

ГЛАВА 3. Методы определения областей применимости моделей связи «структура-свойство».

Постановка задачи: ***определить область применимости (ОП)*** построенной модели связи «структура-свойство», т. е. то множество химических соединений, свойства которых могут быть рассчитаны по соответствующему уравнению с заданной погрешностью ε. Эта задача возникает на этапе прогнозирования свойств соединений при помощи построенной модели. Очевидно, что использование любой математической модели без учета ее ОП может дать неверный результат.

При исследовании проблемы конструктивного определения ОП по исходным данным прежде всего возникает вопрос о *принципиальной возможности* ее решения. В Главе 1 было теоретически доказано, что на основе исходных данных *в принципе невозможно* определить, принадлежит ли данный граф (т. е. химическая структура) области применимости построенной модели, т. е. исходных данных *недостаточно* для детерминированного решения этой проблемы. В то же время эти результаты позволяют выявить вид дополнительной информации, необходимой для решения поставленной задачи. Однако априорное выделение класса соединений, близкого в каком-либо смысле к реальной ОП (РОП), является важным при прогнозировании свойств соединений, так как его использование при прогнозировании свойств способствует сокращению доли ошибочных предсказаний. Такой класс соединений естественно назвать теоретической областью применимости модели (ТОП). Можно ожидать, что введение ТОП приведет к «пропуску» некоторых искомых соединений. Однако с практической точки зрения более важно уменьшить число ошибочных прогнозов, которые повлекут за собой неоправданные финансовые и временные затраты, чем «пропустить» перспективное соединение.

В связи с отмеченными выше особенностями поставленной задачи можно предложить два принципиально разных подхода к определению ТОП моделей связи «структура-свойство». Один из них базируется на выдвижении ряда гипотез относительно рассматриваемого свойства, которые, по сути, позволяют увеличить объем исходной информации. Другой подход носит вероятностный характер. Однако и в этом случае используется ряд гипотез, в частности, предположения о характере распределения некоторых случайных величин.

В данной главе описаны два общих, конструктивных метода априорного определения ТОП уравнений связи «структура-свойство» при заданной погрешности расчета свойств ε. Приведены обоснования предложенных методов, а также результаты их тестирования.

Вероятностный метод определения ОП. Предложен ***вероятностный подход*** к определению области применимости линейной модели связи «структура-свойство» следующего вида:

$$y=a_1x_1+\ldots+a_mx_m ,$$

в которой параметры $a_1,\ldots,a_m$ определяются по исходной выборке k соединений методом наименьших квадратов, а $x_1,\ldots,x_m$ - любые молекулярные параметры. Пусть $y_{расч}$ - величина свойства, рассчитанная по вышеприведенному уравнению, y – экспериментальное значение

свойства, M – множество, структуры которого требуется разделить на два класса: принадлежащие и не принадлежащие ТОП соответствующего уравнения. Согласно определению, *РОП* вышеприведенного уравнения состоит из тех соединений, для которых $|y-y_{расч}|\le\varepsilon$. Так как невозможно учесть все факторы, влияющие на заданное свойство, его экспериментальное значение y можно рассматривать как случайную величину. Следовательно, выполнение условия $|y-y_{расч}|\le\varepsilon$ представляет собой случайное событие, и можно рассмотреть его вероятность $P(|y-y_{расч}|\le\varepsilon)$. Будем считать, что *ТОП* состоит из тех соединений, для которых $P(|y-y_{расч}|\le\varepsilon)\ge\alpha_{кр}$, где $\alpha_{кр}$ – некоторое пороговое (критическое) значение этой вероятности. Основная идея предложенного метода заключается в определении порога $\alpha_{кр}$ по заданному значению ε, и дальнейшей оценке вероятности $P(|y-y_{расч}|\le\varepsilon)$ для тестируемого соединения. Подход базируется на ряде гипотез, в частности, на предположении о том, что некоторые случайные величины, связанные с изучаемым свойством, распределены по нормальному закону. Критерий принадлежности некоторого соединения ТОП построенной модели заключается в выполнении для тестируемого соединения некоторого числового неравенства. Для его проверки необходимо знание значений параметров $x_1,\dots,x_m$ для тестируемого соединения и для соединений исходной выборки, значений свойств соединений исходной выборки, чисел $\alpha_{кр}$ и ε, а также значения $t(\alpha_{кр}, k-m)$- квантили уровня $\alpha_{кр}$ распределения Стьюдента с $k-m$ степенями свободы.

Проведено ***тестирование*** предложенного метода. При этом проверялись степень совпадения РОП и ТОП, степень сокращения доли ошибочных предсказаний и доля «пропущенных» соединений при использовании ТОП. В качестве множества M рассматривалось множество всех алканов C_2-C_8 (39 соединений) с известными значениями температуры кипения. Обучающая выборка состояла из $k=12$ соединений этого класса. По этим данным было построено линейное уравнение связи «структура-свойство», содержащее такие параметры как $ln(\chi+1)$, где χ - индекс Рандича, и n - число атомов углерода в молекуле. Рассматривался ряд значений ε $(0<\varepsilon\le5(^0C))$, и для этих значений определялось качество предложенного метода. Проведенные исследования показали, что описанный выше метод позволяет в 94-97% случаев отбросить соединения, не принадлежащие *РОП*, и в 80-85% случаев верно определить, принадлежит ли РОП данное соединение. Если проводить прогнозирование свойств всех соединений исходного множества, не выделяя ТОП, то доля верных предсказаний составляет 90%; если прогнозирование проводить только внутри ТОП, то доля верных предсказаний - 97%; доля «пропущенных» соединений -13%.

Аналогичные результаты были получены и для ряда других уравнений, построенных для тех же данных, и содержащих такие параметры, как индекс Рандича, индекс Винера, число атомов углерода в молекуле.

Метод определения ОП на основе базисных инвариантов. Предложен ***метод*** определения ТОП модели связи «структура-свойство» специального вида на основе базисных инвариантов (в смысле определения 1), рассмотренных в Главе 1, и ряда соответствующих теоретических результатов.

Рассматриваемые модели связи «структура-свойство» строятся следующим образом. Пусть задано множество соединений, представленных графами $\{G_i\}$ $(i=1,\dots,N)$, и выборка соединений из них $\{G_i\}$ $(i=1,\dots,k)$ с известными значениям некоторого свойства $\{y_i\}$ $(i=1,\dots,k)$. Пусть $\{f_j\}$ $(j=1,\dots,N)$ – базис инвариантов графов исходного множества, такой, что $N-k+1$ его элементов с номерами $k,\dots,N$ постоянны на графах $\{G_i\}$ $(i=1,\dots,k)$, т.е. $f_p(G_i)=c_p$, $i=1,\dots,k$. Предположим, что по исходным данным сначала построено точное уравнение связи «структура-свойство» следующего вида:

$$y=\sum_{p=1}^{k-1}a_p f_p(G)+a_0.$$

Пусть из него получено приближенное уравнение (с заданной погрешностью ε) путем замены некоторых инвариантов f_p (например, с номерами $p=m+1,...,k-1$) на константы b_p, равные их средним на выборке значениям:

$$y=\sum_{p=1}^{m} a_p f_p(G)+A_0 \quad (A_0=a_0+\sum_{p=m+1}^{k-1} a_p b_p).$$

В Главе 1 были даны некоторые *достаточные условия* на рассматриваемое свойство и молекулярный граф G (т.е. химическую структуру), при которых значение свойства этой структуры определяется по вышеуказанному уравнению с точностью ε (см. Теорему 1.9 и следствие из нее). Первое из них – это независимость рассматриваемого свойства для соединений исходного множества от некоторых базисных инвариантов f_p с номерами $p=k,...,N$ (что можно только предполагать и нельзя получить из исходных данных). Второе условие - это выполнение для графа G равенств вида $f_p(G)=c_p$ для остальных номеров $p=k,...,N$. Третье условие - это выполнение следующего неравенства:

$$|\sum_{p=m+1}^{k-1} a_p(f_p(G)-b_p)|\le\varepsilon.$$

Из этих условий следует, что число L_1 ограничений типа равенств на структуры графов из ТОП связаны с числом L_2 гипотез о независимости свойства от некоторых базисных параметров так: $L_1+L_2=N-k+1$. Таким образом, чем меньше факторов влияет на величину данного свойства, тем меньше структурных ограничений надо вводить на графы из ТОП.

На основании этих теоретических результатов предложен следующий ***метод*** определения ТОП вышеприведенного уравнения: 1) выдвигается ряд гипотез о независимости рассматриваемого свойства от некоторых структурных параметров, задаваемых инвариантами f_p; 2) для анализируемого графа G проверяется ряд соответствующих ограничений типа равенств и одно ограничение типа неравенства, приведенные выше; если все эти условия выполняются, то граф G считается принадлежащим ТОП.

Проведено ***тестирование*** предложенного метода. Проверялись степень совпадения РОП и ТОП, степень сокращения доли ошибочных предсказаний и доля «пропущенных» соединений при использовании ТОП. Рассмотрено множество всех алканов C_2-C_7 ($N=21$), с известными значениями температуры кипения y. В качестве обучающей выборки использовано множество всех алканов C_2-C_5 ($k=7$), а $\varepsilon=5(^0C)$. Выдвигаемые гипотезы основаны на представлении о том, что температура кипения зависит, в основном, от размера и степени разветвленности молекул, а числа вхождения в граф некоторых специальных подграфов могут служить количественной мерой этих структурных особенностей. Проведенные исследования показали, что при классификации исходных соединений на «принадлежащие/не принадлежащие» РОП при помощи ТОП была сделана лишь одна ошибка, т.е. правильная классификация соединений была проведена в 95% случаев. Если проводить прогноз свойств всех соединений исходного множества, не выделяя ТОП, то доля верных прогнозов составляет 43%; если прогнозирование проводить внутри ТОП, то доля верных прогнозов – 100%; доля «пропущенных» соединений - 5%.

Таким образом, в Главе 3 рассмотрена задача определения ОП модели связи «структура-свойство», построенной в результате анализа ограниченного набора данных (при заданной допустимой погрешности расчета свойств ε, зависящей от конкретной задачи). Доказано, что данная задача в принципе не может быть решена на основе анализа исходных данных. При этом указан вид дополнительной информации, необходимой для ее решения. Предложены два общих *метода* определения теоретической области применимости моделей связи «структура-свойство» специального вида, учитывающие заданную погрешность ε. Один из них использует аппарат теории вероятности и базируется на гипотезе о том, что некоторые величины, связанные с рассматриваемым свойством, являются случайными величинами, распределенными по

нормальному закону. Второй подход опирается на понятие базисных инвариантов и их свойства и используется для моделей определенного типа. В этом подходе также необходимо выдвижение некоторых гипотез относительно рассматриваемого свойства. Проведено тестирование предложенных методов, показавшее, что учет теоретической области применимости при прогнозировании свойств соединений позволяет снизить долю ошибочных прогнозов.

ГЛАВА 4. Обратные задачи в исследованиях связи «структура-свойство»: теоретико - графовый подход.

Постановка задачи. ***Обратная задача (ОЗ)*** в исследованиях связи «структура-свойство» - это задача *исчерпывающей* генерации химических структур определенного класса, имеющих заданное значение y_0 рассматриваемого свойства (или заданный интервал (y_1, y_2) значений свойства), на основе предварительно построенной базовой модели связи «структура-свойство» следующего вида:

$$y=f(x_1,...,x_N),$$

где y - значение рассматриваемого свойства, $x_1,...,x_N$ - какие-либо молекулярные параметры, f - некоторая функция. Если в качестве параметров $x_1,...,x_N$ использованы инварианты соответствующих молекулярных графов, то ОЗ сводится к исчерпывающей генерации молекулярных графов по заданному значению их инварианта, задаваемому выражением вида $f(x_1,...,x_N)$.

Метод ОЗ важен для целенаправленного поиска соединений с заданными свойствами. По сравнению с традиционным подходом к поиску таких соединений, когда при помощи базовой модели «структура-свойство» последовательно тестируется определенный набор соединений и затем из него отбираются подходящие соединения, метод ОЗ имеет явное преимущество: он позволяет дать *исчерпывающее* (с математической точки зрения) решение поставленной задачи. Такая особенность этого метода позволяет выявить структуры *новых* соединений (возможно, еще не синтезированных), которые, согласно прогнозу, должны обладать требуемым свойством.

В Главе 4 описаны алгоритмы решения ОЗ для некоторых наиболее популярных инвариантов графов, используемых в теоретической химии при построении корреляций «структура-свойство» и ставших в определенном смысле «классическими». Проведено тестирование предложенных алгоритмов.

Типы рассмотренных базовых моделей связи «структура-свойство».

Рассматриваются ***модели связи «структура-свойство»*** следующих видов:

1) а) Уравнение содержит только один молекулярный параметр χ, называемый *индексом Рандича:*

$$\chi=\Sigma(v_i\, v_j)^{-1/2}$$

(v_i и v_j - степени вершин i и j, суммирование проводится по всем ребрам (i,j) молекулярного графа). Предполагается, что χ может быть выражен однозначно из этого уравнения; рассматривается как случай произвольных графов так и случай молекулярных графов, соответствующих ката-конденсированным бензоидным углеводородам; **б)** корреляционное уравнение, наряду с индексом χ содержит и ряд других целочисленных параметров, ограниченных на рассматриваемом классе графов.

2) Уравнение содержит *индекс Винера* W и рассматривается для ациклических молекулярных графов:

$$W=\sum_{i<j}d_{ij}$$

(d_{ij} - расстояние между вершинами i и j, суммирование проводится по всем парам вершин (i,j), $i<j$.).

3) Уравнение содержит *«каппа»-индексы Кира* ${}^i\kappa$ (i=0, 1, 2, 3), предложенные для количественной характеристики различных особенностей «формы» молекулы, представленной простым графом. Эти молекулярные параметры определяются в терминах числа вершин графа n и числа путей ${}^iP$ длины i $(i=1, 2, 3)$ в графе по следующим формулам:

$^{1}\kappa=n(n-1)^{2}/{}^{1}P^{2}$, $^{2}\kappa=(n-1)(n-2)^{2}/{}^{2}P^{2}$,
$^{3}\kappa=(n-3)(n-2)^{2}/{}^{3}P^{2}$ (для четного $n>3$); $^{3}\kappa=(n-1)(n-3)^{2}/{}^{3}P^{2}$ (для нечетного $n>3$).

Индекс $^{0}\kappa$ определяется по формуле: $^{0}\kappa=-n\Sigma(n_i/n)\log_2(n_i/n)$, где n_i – число топологически эквивалентных вершин в i – ом классе эквивалентности. Разбиение вершин на классы происходит по каким-либо их топологическим характеристикам, причем самое «мелкое» разбиение соответствует орбитам группы симметрии графа.

4) Уравнение содержит индексы $^{i}\kappa$ ($i=0,1,2,3$), а также их обобщения $^{i}\kappa_{\alpha}$ ($i=1,2,3$), разработанные для учета гетероатомов и кратных связей в молекуле. Они вычисляются аналогично $^{i}\kappa$ ($i=1,2,3$), но в вышеприведенных формулах вместо n используется величина $n+\alpha$, а вместо ^{i}P - величина $^{i}P+\alpha$ при некотором параметре α, вычисляемом по взвешенному графу. Для вычисления α атомы молекулы классифицируют по химическим символам атомов и распределениям типов связей; для атома каждого типа определенным способом вычисляют параметр α_j, зависящий от ковалентного радиуса атома, затем α вычисляют по формуле $\alpha=\Sigma\alpha_j$.

3) Уравнение содержит *информационные топологические индексы* разных типов, но одного порядка k.

Предположим, что химические соединения представлены в виде классических структурных формул, т.е. в виде вершинно – и реберно-меченых графов. Пусть атомы в молекуле разбиты на классы эквивалентности по окрестностям k-ого порядка ($k\geq 0$). *Информационными топологическими индексами,* соответствующими такой классификации атомов, являются следующие инварианты:
$IC_k=-\Sigma\, n_i/n_i\log_2 n_i/n_i$ (*Information Content*), $SIC_k=IC_k/\log_2 n$ (*Structural Information Content*),
$CIC_k=\log_2 n$ (*Complement Information Content*), $BIC_k=IC_k/\log_2 q$ (*Boundary Information Content*),
$TIC_k=n\cdot IC_k$ (*Total Information Content)*, (q – общее число связей в молекуле).

Аналогичные инварианты можно построить и для произвольно меченого графа.

4) Уравнение содержит *индекс Хосойя Z,* а также такие параметры как общее число вершин графа n и числа n_i вершин графа степени i=1, 2, 3, 4. Инвариант Z определяется по формуле:

$$Z=\sum_{k=0}^{\left[\frac{n}{2}\right]} p_k\,,$$

где p_k - число подграфов, состоящих из k несмежных ребер граф, $p_0=1$, n – число вершин графа. Отметим, что для ациклических графов индекс Хосойя равен сумме модулей коэффициентов характеристического полинома графа. Рассматриваются простые графы, степени вершин которых не превосходят четырех. Кроме того, предполагается, что индекс Z может быть выражен однозначно из вышеуказанного уравнения.

Алгоритмы решения обратных задач и их тестирование. Приведены ***алгоритмы*** решения ОЗ для вышеуказанных корреляционных уравнений. Проведено их ***тестирование*** для конструирования химических соединений с заданными интервалами значений определенных свойств. Для этой цели предварительно были построены разнообразные модели связи «структура-свойство» вышеописанного вида.

Рассматривались: (1) температура кипения алканов; (2) температура кипения циклосодержащих углеводородов; (3) токсичность простых эфиров; (4) теплота парообразования алканов; (5) растворимость спиртов в воде; (6) параметр гидрофобности *logP,* где P - коэффициент распределения соединения в системе октанол-вода для кислородсодержащих соединений (кетонов, ненасыщенных и насыщенных спиртов, карбоновых кислот); (7) температура кипения аминов; (8) температура кипения сульфидов. Во всех рассмотренных случаях имеется хорошее соответствие между экспериментальными данными и результатами компьютерной генерации соединений с заданными свойствами.

Рассмотрим следующий *пример решения ОЗ*. По базе данных, содержащих предельные спирты (N=50) с известными значениями физико-химического свойства: -logX (X - растворимость спиртов в воде (в мольных долях)), построено уравнение вида:

$$-logX=-0.8+1.186lnZ \quad (R=0.976,\ s=0.21).$$

Поставим задачу: найти все соединения этого класса, для которых $2.6 \leq -\log X \leq 3.0$. Построено 20 структур, изображенных на рис. 4. Для соединений №№ 1-11 значения свойства известны. При этом для 9 структур экспериментальные значения свойства действительно лежат в заданном интервале; для 2 структур - незначительно выходят за пределы интервала (для №3 - 2.542; для №8 – 2.588). Для соединений №№ 12-20 экспериментальные значения рассматриваемого свойства неизвестны.

Таким образом, в Главе 4 рассматривается ряд алгоритмов решения ОЗ в исследованиях связи «структура-свойство» на основе предварительно построенных базовых моделей, содержащих различные инварианты графов (топологические индексы). Рассмотренные топологические индексы находят широкое применение в корреляциях «структура-свойство» и допускают определенную структурную интерпретацию (например, как количественная мера ветвления, компактности, симметрии, «формы», неоднородности молекулы и т. д.). Базовые корреляционные уравнения могут содержать как один, так и несколько различных инвариантов. Уравнения, содержащие какие-либо другие инварианты, не рассматриваемые в данной главе, в ряде случаев можно свести к уравнениям, содержащим уже рассмотренные инварианты, используя корреляционные соотношения между различными инвариантами. Применение алгоритмов и их эффективность продемонстрированы на конкретных примерах.

ГЛАВА 5. Построение моделей связи «структура-свойство» и прогнозирование свойств химических соединений на основе концепции молекулярного подобия.

Постановка задачи. В Главе 5 рассматривается один из широко распространенных подходов к построению моделей связи «структура-свойство», основанный на постулате *«близкие структуры имеют близкие свойства»*. Для реализации этого метода необходимо: 1) иметь базу данных, содержащую структуры соединений *{S}* и значения их свойств; 2) выбрать способ математического описания структуры молекул, при котором структуре S соответствует объект M; 3) на множестве выбранных математических объектов *{M}* задать количественную меру подобия этих объектов: $d(M_1,M_2)\geq 0$.

Для прогнозирования свойства y_0 соединения S_0 в рамках этого подхода используются различные методы, суть которых заключается в следующем: 1) для S_0 следует найти соединение S, «ближайшее» к нему в базе данных (или несколько «ближайших») и положить $y_0=y$ (или y_0 равно среднему арифметическому свойств «ближайших» соединений). Метод такого типа целесообразно использовать, в частности, тогда, когда исходная база данных очень разнородна по своему составу, и не удается построить удовлетворительную модель вида $y=f(S)$. Однако разбиение базы на части структурно-близких соединений приводит к малоинформативным выборкам небольшого размера.

Следует отметить, что меры подобия, обычно используемые для прогнозирования свойств в рамках этого подхода, зависят лишь от структур сравниваемых соединений и *не зависят* ни от исходной выборки, ни от рассматриваемого свойства. Имеются примеры, показывающие, что в то же время результат выбора «ближайшего» соседа (следовательно, и разультат прогнозирования) зависит от использованной меры подобия. Кроме того, различных мер подобия существует бесконечно много, а правил выбора меры в конкретной задаче – нет. В связи с этим основная задача, рассматриваемая в данной главе, такова: *разработать алгоритмы подбора меры подобия, дающей наилучший результат при прогнозировании свойств соединений в рамках вышеуказанного метода, в предположении, что структуры соединений представлены графами.*

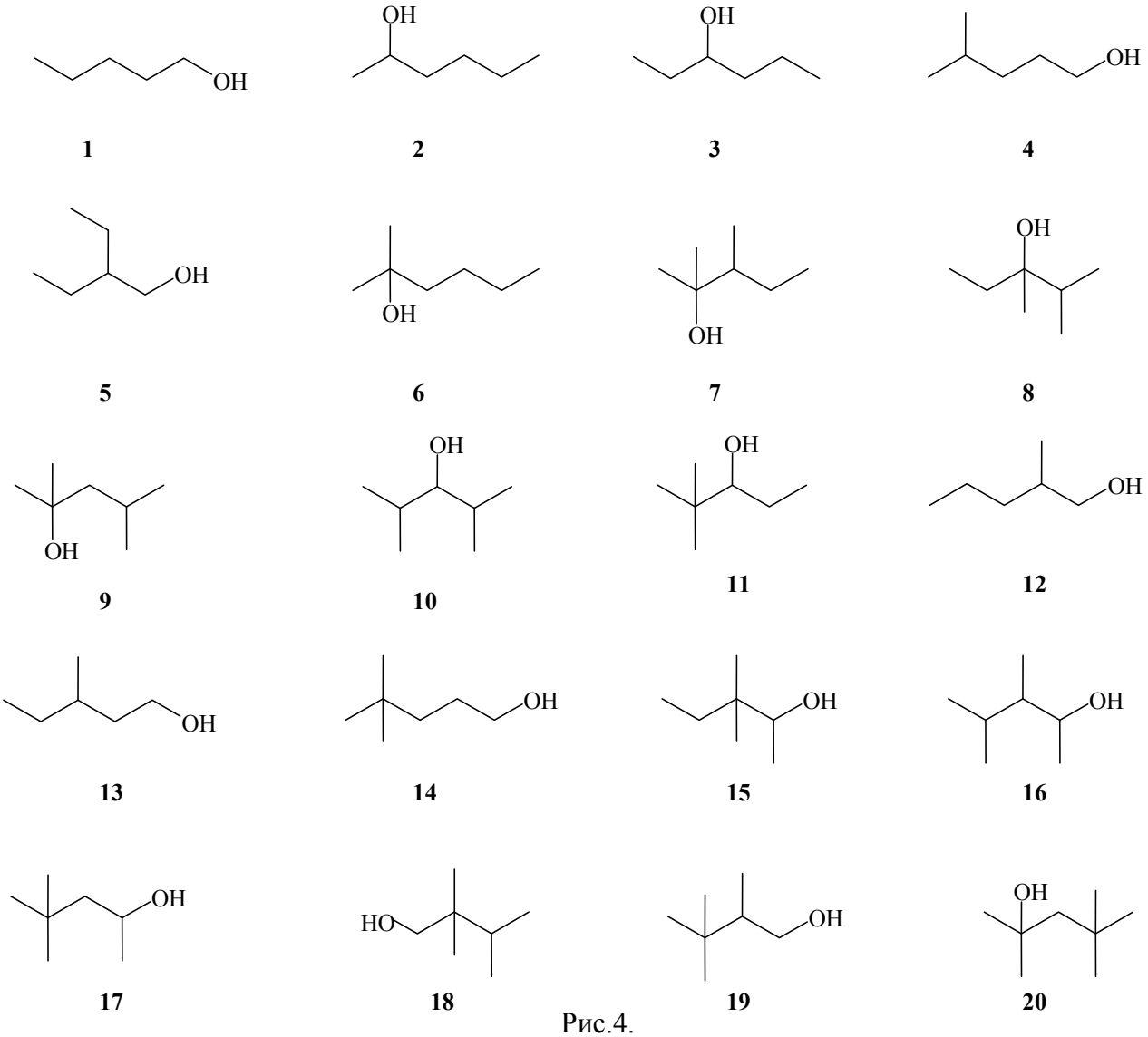

Рис.4.

<u>Общая аналитическая формула для произвольной меры подобия молекулярных графов.</u> Выведена ***общая аналитическая формула*** для произвольной симметричной меры подобия $d(G_k,G_l)$, заданной на произвольном множестве графов $\{G_i\}$, $i=1,...,N$. Доказана ***теорема,*** соласно которой существует единственная симметричная квадратная матрица $M=(m_{ij})$ $(i,j=1,...,N-1)$ такая, что мера $d(G_k,G_1)$ представляется в следующем виде:

$$d(G_k,G_l)=M(\boldsymbol{f_k}-\boldsymbol{f_l})\cdot(\boldsymbol{f_k}-\boldsymbol{f_l})$$

где $\boldsymbol{f_k}=(f_1(G_k),...,f_{N-1}(G_k))$, $\boldsymbol{f_l}=(f_1(G_l),...,f_{N-1}(G_l))$ - вектора-столбцы, компоненты которых – это значения некоторых базисных инвариантов исходного множества графов (в смысле определения 1, рассмотренного в Главе 1) на графах G_k и G_l, соответственно, $M(\boldsymbol{f_k}-\boldsymbol{f_l})$ - произведение M и $\boldsymbol{f_k}-\boldsymbol{f_l}$, символ « · » обозначает скалярное произведение соответствующих векторов.

Из этой теоремы следует, что: **1)** Варьируя матрицу M, можно получить меру подобия, которая принимает *любые заданные значения* для каждой пары графов из рассматриваемого

множества графов; ***2*)** Полученная формула позволяет строить *бесконечно много* новых мер подобия, варьируя матрицу M, и адаптировать их к конкретной задаче.

Метод построения моделей связи «структура - свойство», основанный на оптимальном подборе меры подобия. Предложен ***метод*** построения моделей связи «структура-свойство» и прогнозирования свойств химических соединений, основанный на приведенных выше теоретических результатах. Для разработки этого метода используется аппроксимация вышеуказанной точной формулы для меры подобия. Предполагается, что структура соединений исходной выборки описана при помощи некоторых векторов $\boldsymbol{X}$ длины $k<N-1$, мера подобия $d(G_i,G_j)$ задается формулой, аналогичной вышеуказанной формуле: $d(G_i,G_j)=M(\boldsymbol{X^{(i)}-X^{(j)}})\cdot(\boldsymbol{X^{(i)}-X^{(j)}})$, где матрица M с неопределенными элементами имеет порядок k. Элементы матрицы M подбираются так, чтобы $|y_i-y_j|=d(G_i,G_j),\ i,j=1,...,N;\ i>j$. В предлагаемом подходе мера подобия подбирается некоторым оптимальным образом по исходным данным. Заключительный этап - прогнозирование свойства y_0 нового соединения G_0 - тоже изменен. Вместо метода m «ближайших соседей» (где всегда остается вопрос о выборе числа m) предлагается другой подход, в котором для вычисления y_0 используются все исходные данные. Однако для прогнозирования свойств может быть использован и метод «ближайших соседей».

Проведено ***тестирование*** разработанного метода и его ***сравнение*** с другими аналогичными методами. Рассмотрена база данных, состоящая из *76* соединений различных химических классов (спирты, фенолы, кетоны, карбоновые кислоты, простые и сложные эфиры, амины, амиды, нитрилы, галогенпроизводные, гетероциклические соединения и.т.д.) с известными значениями параметра $log\ P$, где P - коэффициент распределения соединения между водой и *н*-октанолом. Качество построенной модели оценивалось по коэффициенту корреляции R и среднеквадратичному отклонению s, найденным для корреляции между расчетными и экспериментальными значениями свойства. Приведены результаты сравнения построенной модели и двух других моделей, полученных другими авторами для тех же данных, где для оценки степени подобия использовались две другие меры подобия. Из этих результатов следует, что предлагаемый метод дает наилучшую модель из этих трех.

Оптимальный подбор меры подобия при прогнозировании свойств по методу «ближайшего соседа». Рассмотрена задача построения оптимальной меры подобия молекулярных графов при прогнозировании свойств соединений по методу одного «ближайшего соседа». Предполагается, что задана некоторая выборка молекулярных графов $\{G_i\}$ с известными значениями некоторого свойства y_i, причем все эти значения - различны.

Предложен ***метод*** построения меры подобия в вышеуказанной задаче, использующий известные значения свойств соединений исходной выборки. Метод позволяет построить меру подобия, дающую *наилучший результат* при вышеуказанном способе прогнозирования свойств соединений, по крайней мере, для исходной выборки (т. е. «ближайший сосед» каждого соединения имеет значение свойства, ближайшее к значению свойства исследуемого соединения). Метод основан на использовании общей аналитической формулы для произвольной меры подобия, полученной в данной главе.

Проведено ***тестирование*** предложенного метода и его ***сравнение*** с шестью аналогичными методами, использующими другие меры подобия, зависящие от различных особенностей молекулярного строения и не зависящие от исследуемого свойства. Рассмотрена база данных, содержащая структурные формулы нитрозаминов с известными значениями мутагенности $y=ln\mu$ (на *Salmonella typhimurium*, μ - число ревертантов на наномоль). Установлено, что предложенный метод дает более точный результат, чем остальные методы.

Формализация и интерпретация постулата «близкие структуры имеют близкие свойства». Впервые рассмотрен вопрос о возможной формализации постулата *«близкие структуры имеют близкие свойства»* и проведено исследование его справедливости. Актуальность таких исследований связана с широким внедрением компьютеров в химические исследования, что приводит к необходимости формализаций различных понятий и

эмпирических правил, разработанных в химии. Кроме того, анализ этого постулата важен для обоснования методов прогнозирования свойств соединений, которые на нем основаны.

Для проведения теоретического исследования справедливости этого утверждения рассмотрен *общий случай*, когда химические структуры представлены в виде некоторых математических объектов M_i $(i=1,...,N)$, и на множестве этих объектов задана некоторая симметричная функция $d(M_i,M_j)$ - мера подобия этих объектов. Предполагается, что $d(M_i,M_j)=0$ тогда и только тогда, когда $M_i=M_j$. Пусть заданы числа y_i $(i=1,...,N)$ – значения некоторого свойства соответствующих соединений. Естественно считать мерой близости свойств величину $|y_i-y_j|$. Предположим, что заранее указаны численные критерии подобия свойств и структур, т.е. такие числа $\varepsilon\geq 0$ и $\delta\geq 0$, что если $d(M_i,M_j)\leq\delta$, то структуры M_i, M_j считаются «близкими», и если $|y_i-y_j|\leq\varepsilon$, то значения свойств считаются «близкими». Очевидно, что число ε задается исследователем и зависит от конкретной задачи, а варьируемыми характеристиками являются $d(M_i,M_j)$ и δ. Вышеуказанный постулат в этом случае можно сформулировать так: *если для любых структур* M_i *и* M_j $d(M_i,M_j)<\delta$, *то* $|y_i-y_j|<\varepsilon$. Легко видеть, что это утверждение является аналогом определения равномерной непрерывности функции $f(x)$ одной переменной на заданном числовом промежутке X.

Приведенная формулировка этого постулата позволяет провести теоретическое исследование его справедливости в общем виде. Предполагается, что мера подобия такова, что $d(M_i,M_j)=0$ тогда и только тогда, когда $M_i=M_j$. Доказано, что для любой выборки структур, представленных в виде некоторых математических объектов M_i $(i=1,...,N)$, любого свойства y, любой меры подобия $d(G_i,G_j)$ верны следующие оценки:

$$a\cdot d(M_i,M_j)\leq|y_i-y_j|\leq b\cdot d(M_i,M_j),$$

где a и b - константы, зависящие от меры, свойства, и выборки структур. Этот результат сформулирован в виде ***теоремы.***

Из полученного результата сделан ряд выводов: **1)** Постулат будет *всегда справедливым*, если выбрать $\delta=\varepsilon/b$; **2)** Предположим, что для данной выборки не все значения свойств близки, т.е. найдется пара M_i и M_j, что $|y_i-y_j|>\varepsilon$. Тогда, если $\delta=maxd(M_i,M_j)$, то постулат *не будет справедлив* на данной выборке; **3)** Если выбрать ε очень большим, то постулат *будет справедлив* при любых δ и $d(M_i,M_j)$; **4)** Из полученных неравенств следует *качественный вывод*: чем меньше величина $d(M_i,M_j)$, тем меньше величина $|y_i-y_j|$, так что для «очень близких» структур их свойства также «очень близки». Этот качественный вывод, следующий из строгих математических рассуждений, по сути, и есть утверждение неформализованного постулата, обычно используемого в теоретической химии для предсказания свойств соединений.

ГЛАВА 6. Алгоритмы на графах, используемые для их кодирования, идентификации и исследования структурных особенностей.

Постановка задачи: ***разработать и обосновть ряд алгоритмов*** для произвольно меченых графов: канонизации графа, установления изоморфизма пары графов, нахождения группы симметрии графа, нахождения заданных подграфов в графе. Эти алгоритмы могут быть использованы как для решения ряда прикладных задач компьютерной и теоретической химии и химической информатики (например, при создании информационно-поисковых систем, анализе связи «структура-свойство» с помощью ЭВМ, компьютерном синтезе, масс-спектрометрии и т. д.), так и представляют самостоятельный интерес в теории графов.

Разработаны следующие алгоритмы на графах: 1) поиска канонической нумерации вершин взвешенного графа и его группы автоморфизмов, основанного на использовании ряда спектральных характеристик графа (даны примеры реализации алгоритма и некоторые результаты его тестирования на быстродействие при программной реализации); 2) установления изоморфизма графов G_1 и G_2 и поиска группы симметрии $AutG$ графа G (приведены некоторые результаты тестирования алгоритма на быстродействие при его программной реализации); 3) поиска всех подграфов, изоморфных заданному подграфу, в произвольно взвешенном графе

(прилагается акт о внедрении соответствующей компьютерной программы в ИОХФ им. А. Е. Арбузова в исследования по планированию органического синтеза).

* * *

ВЫВОДЫ

1) Разработан и обоснован ряд новых *методов* построения моделей связи «структура-свойство» в терминах инвариантов молекулярных графов. Эти методы носят общий характер, применимы к произвольным свойствам и к произвольным выборкам химических соединений, представленных произвольно мечеными графами. Методы строго детерминированы и допускают компьютерную реализацию. Проведено *тестирование* предложенных подходов для моделирования связи «структура-свойство» для разнообразных свойств (физико-химические, биологическая активность, вычисляемые молекулярные параметры) и классов соединений, показавшее их практическую применимость и эффективность.

2) Разработана *интеллектуальная система,* предназначенная для автоматического конструирования произвольных наборов инвариантов графов различной природы для построения корреляций «структура-свойство». В ней реализовано моделирование последовательности действий человека, конструирующего инварианты графа для вышеуказанной задачи. Проведено исследование возможностей этой системы. Получаемые таким образом инварианты могут быть использованы при решении различных задач химической информатики, математической и компьютерной химии, в том числе при моделировании связи «структура-свойство».

3) На основе разработанной схемы конструирования инвариантов графов предложен новый *метод* построения моделей связи «структура-свойство». Проведено *тестирование* предлагаемого подхода для построения корреляций «структура-свойство» для физико-химических свойств и биологической активности органических соединений различных классов, показавшее его эффективность.

4) Проведено исследование задачи *определения области применимости* модели связи «структура-свойство» для заданной допустимой погрешности расчета свойств соединений, а также предложено два *метода* ее решения. Проведено тестирование этих методов.

5) Разработаны *методы* решения различных *обратных задач* в исследованиях связи «структура-свойство». Эти методы позволяют провести *исчерпывающую* генерацию химических структур определенного класса, имеющих заданное значение y_0 рассматриваемого свойства (или заданный интервал (y_1, y_2) значений свойства), на основе предварительно построенной модели вида $y=f(x_1,...,x_N)$, связывающей значения y изучаемого свойства и некоторые инварианты молекулярных графов $x_1,...,x_N$. Рассмотрены базовые модели, содержащие различные инварианты (топологические индексы), широко используемые при моделировании связи «структура-свойство» и допускающие определенную структурную интерпретацию. Проведено *тестирование* разработанных методов, показавшее хорошее соответствие получаемых результатов и экспериментальных данных.

6) Предложены *модели* связи «структура-свойство» нового типа, отражающие широко распространенный в химии постулат «близкие структуры имеют близкие свойства». Эти модели имеют следующий вид: $|y_i-y_j|=d(G_i,G_j)$, где y_i, y_j – численные значения свойств i-ого и j-ого соединений, представленных графами G_i и G_j, а $d(G_i,G_j)$ - некоторая симметричная функция двух аргументов G_i и G_j, значения которой количественно характеризуют степень подобия G_i и G_j. Предложен *метод* оптимального подбора меры подобия $d(G_i,G_j)$ в этом соотношении, а также способ оценки свойств соединений на основе такой модели. Проведено *тестирование* этого метода, а также его сравнение с двумя другими методами, использующими другие меры подобия.

7) Предложен *алгоритм* оптимального подбора меры подобия при прогнозировании свойств соединений по методу «ближайшего соседа». Подход позволяет построить меру подобия, дающую наилучший результат при вышеуказанном способе прогнозирования свойств соединений, по крайней мере, для исходной выборки соединений. Проведено *тестирование*

предложенного метода и его сравнение с шестью другими методами оценки свойств соединений, разработанных на основе других мер подобия.

8) Разработаны новые *комбинаторные алгоритмы* на графах, используемые при решении различных задач теоретической, компьютерной и математической химии, связанных с кодированием, идентификацией и анализом структурных особенностей графов. Эти алгоритмы позволяют строить каноническую нумерацию вершин графа, находить группу симметрии графа, устанавливать изоморфизм пары графов, находить все подграфы графа, изоморфные заданному подграфу. Алгоритмы математически обоснованы и применимы к графам произвольного вида, имеющим любые веса вершин и ребер.

9) Определены и исследованы *три новых класса прикладных задач* теории графов, имеющих практическое применение в области химии. Первый класс задач связан с проблемой восстановления аналитического вида инварианта меченых графов некоторого множества по всем или некоторым его значениям на графах этого множества. Второй класс задач связан с проблемой определения такого набора подграфов меченого графа (названных базисными подграфами), по которому граф восстанавливается однозначно. Третий класс задач связан с задачей аналитического представления произвольной симметричной меры подобия меченых графов произвольного конечного множества. Введен ряд новых определений, а также сформулирован и доказан ряд новых теорем в теории графов. Полученные теоретические результаты являются основой алгоритмов моделирования связи «структура-свойство», разработанных в диссертации.

10) Предложена формализация постулата *«близкие структуры имеют близкие свойства»,* являющегося основой некоторых методов прогнозирования свойств соединений, и проведено теоретическое исследование его справедливости. Указаны общие случаи, когда вышеуказанное утверждение будет заведомо верным или заведомо неверным.

***

Автор глубоко признателен академику Н. С. Зефирову за предоставленную возможность работать в области математической химии, помощь в организации научной работы и обсуждение научных результатов, находящихся на стыке математики и химии.

Автор выражает искреннюю благодарность заслуженному деятелю науки РФ, д.ф.-м.н., профессору Карташову Э. М. за внимание к настоящей работе, ценные замечания и полезное обсуждение рукописи диссертации.

***

СПИСОК ПУБЛИКАЦИЙ ПО ТЕМЕ ДИССЕРТАЦИИ.

1. Зефиров Н.С., Трач С.С., Станкевич (Скворцова) М. И. Алгоритм установления изоморфизма молекулярных графов // Тезисы докладов Всесоюзной конференции «Использование вычислительных машин в химических исследованиях и спектроскопии молекул», Рига, 1986, с.190-192.

2. Станкевич (Скворцова) М. И., Баскин И. И., Зефиров Н.С. Автоматизированный поиск структурных фрагментов. Алгоритм и программа // Журнал структурной химии, 1987, т.28, № 6, с. 136-137.

3. Станкевич (Скворцова) М.И., Баскин И.И., Зефиров Н.С. Комбинаторные модели и алгоритмы в химии. Поиск структурных фрагментов // Деп. ВИНИТИ АН СССР 11.06 1986, № 4288-В86, 27 стр.

4. Станкевич (Скворцова) М.И., Станкевич И.В., Зефиров Н.С. Топологические индексы в органической химии // Успехи химии, 1988, т 57, № 3, с. 337-366.

5. Девдариани Р.О., Станкевич (Скворцова) М.И., Палюлин В.А., Зефиров Н.С. Оценка с помощью ЭВМ температур плавления для некоторых классов органических соединений // Тезисы докладов Всесоюзной школы-семинара по автоматизации химических исследований, Тбилиси, 1988, с. 39.

6. Баскин И.И., Станкевич (Скворцова) М.И., Девдариани Р.О., Зефиров Н.С. Комплекс программ для нахождения корреляций «структура - свойство» на основе топологических индексов // Журнал структурной химии, 1989, т. 30, № 6, с.145-147.
7. Гордеева Е.В., Баскин И.И., Девдариани Р.О., Зефиров Н.С., Станкевич (Скворцова) М.И. Методология решения обратной задачи в проблеме связи «структура-свойство» для случая топологических индексов // ДАН СССР, 1989, т. 307, № 3, с. 613-616.
8. Станкевич (Скворцова) М. И. , Баскин И. И., Зефиров Н.С., Гордеева Е. В., Палюлин В. А., Девдариани Р. О. О проблеме восстановления химических структур по заданным значениям топологических индексов // Тезисы докладов межреспубликанской научно-практической конференции «Синтез, фармакология и клинические аспекты новых психотропных и сердечно-сосудистых средств», Волгоград, 1989, С. 28.
9. Зефиров Н.С., Скворцова М. И., Станкевич И. В., Томилин О. Б. Об одном способе нумерации вершин молекулярных графов // Тезисы докладов VIII-ой Всесоюзной конференции «Использование вычислительных машин в спектроскопии молекул и химических исследованиях», Новосибирск, 1989, с. 176-177.
10. Скворцова М.И., Баскин И.И., Девдариани Р.О., Палюлин В.А., Зефиров Н.С. О проблеме генерации структур органических соединений с заданными свойствами // Тезисы докладов VIII-ой Всесоюзной конференции «Использование вычислительных машин в спектроскопии молекул и химических исследованиях», Новосибирск, 1989, с. 250.
11. Девдариани Р.О., Палюлин В. А., Скворцова М. И., Баскин И. И., Зефиров Н.С. Прогнозирование температур плавления ароматических соединений некоторых классов на основе использования взвешенных топологических индексов // Тезисы докладов VIII-ой Всесоюзной конференции «Использование вычислительных машин в спектроскопии молекул и химических исследованиях», Новосибирск, 1989, с. 251.
12. Зефиров Н. С., Скворцова М. И., Станкевич И. В. Генерация структур поликонденсированных бензоидных углеводородов по индексу Рандича // Тезисы докладов VIII-ой Всесоюзной конференции «Использование вычислительных машин в спектроскопии молекул и химических исследованиях», Новосибирск, 1989, с. 252-253.
13. Stankevitch (Skvortsova) M. I., Tratch S. S., Zefirov N. S. Combinatorial Models and Algorithms in Chemistry. Search for Isomorphisms and Automorphisms of Molecular Graphs // J. Comput. Chem.,1988, v.9, N 4, p. 303-314.
14. Скворцова М. И., Станкевич И. В.. Зефиров Н. С. Топологические свойства ката-конденсированных бензоидных углеводородов: индекс Рандича и его связь с химическим строением // Тезисы докладов Межвузовской конференции «Молекулярные графы в химических исследованиях», Калинин, 1990, с. 84.
15. Скворцова М. И., Станкевич И. В., Томилин О. Б., Зефиров Н. С. Проекционные операторы и каноническая нумерация вершин молекулярных графов // Тезисы докладов Межвузовской конференции «Молекулярные графы в химических исследованиях», Калинин, 1990, с. 85-86.
16. Скворцова М. И., Словохотова О. Л., Палюлин В. А., Зефиров Н. С. Решение обратной задачи в проблеме связи «структура-свойство» для топологических индексов, характеризующих молекулярную форму // Тезисы докладов I-ой Всесоюзной конференции по теоретической органической химии (ВАТОХ), Волгоград, 1991, с. 551.
17. Станкевич И. В., Скворцова М. И. Обобщенный индекс Рандича как функционал от π - электронной плотности // Тезисы докладов I-ой Всесоюзной конференции по теоретической органической химии (ВАТОХ), Волгоград, 1991, с. 97.
18. Скворцова М. И., Станкевич И. В., Зефиров Н. С. Генерация молекулярных структур поликонденсированных бензоидных углеводородов по индексу Рандича // Журнал структурной химии, 1992, т. 33, № 3, с. 99-104.

19. Станкевич И. В., Скворцова М. И., Томилин О. Б., Зефиров Н. С. Использование проекционных операторов для нумерации атомов и исследования свойств симметрии молекулярных структур // Журнал структурной химии, 1992, т. 33, № 3, с. 93-98.
20. Скворцова М. И., Баскин И. И., Словохотова О. Л., Палюлин В. А., Зефиров Н. С. Обратная задача в QSAR/QSPR-анализе для случая топологических индексов, харак-теризующих молекулярную форму (индексов Кира) // ДАН, 1992, т. 324, № 2, с. 344-348.
21. Станкевич И. В., Скворцова М. И., Зефиров Н.С. Топологические свойства сопряженных углеводородов: обобщенный индекс Рандича как функционал от π-электронной плотности // ДАН, 1992, т.324, № 1, с.133-137.
22. Skvortsova M. I., Baskin I. I., Slovokhotova O. L., Palyulin V. A., Zefirov N. S. The Inverse Problem in QSAR/QSPR Studies for the Case of Topological Indices Characterizing Molecular Shape (Kier Indices) // J. Chem. Inform.Comput.Sci., 1993, v.33, N 4, p. 630-634.
23. Stankevich I.V., Galpern E. G., Chistyakov A. L., Baskin I. I., Skvortsova M. I., Zefirov N. S., Tomilin O. B. Spectral Theory of Graphs in Chemistry.1. Projection Operators and Canonical Numeration of Graph Vertices // J. Chem. Inform.Comput.Sci. 1994, v. 34, N 5, p. 1105-1108.
24. Скворцова М. И., Баскин И. И., Словохотова О. Л., Зефиров Н. С. Методология построения общей модели связи «структура-свойство» на топологическом уровне // ДАН, 1994, т. 336, N 4, с. 496-499.
25. Баскин И. И., Скворцова М. И., Станкевич И. В., Зефиров Н. С. О базисе инвариантов помеченных молекулярных графов // ДАН, 1994, т. 339, N 3, с. 346-350.
26. Stankevich I. V., Skvortsova M. I., Kolmykov V. A., Subbotin V. F., Mnukhin V. B. Spectral Graph Theory in Chemistry // In: Mathematical Methods in Contemporary Chemistry. (Ed. Kuchanov S. I.; Gordon and Breath Publishers, Amsterdam), 1996, p.101-141.
27. Skvortsova M. I., Baskin I. I., Palyulin V. A., Slovokhotova O. L., Zefirov N. S. Structural Design. Inverse Problems for Topological Indices in QSAR/QSPR Studies // In: AIP Conference Proceedings 330, E.C.C.C.1, Computational Chemistry, F.E.C.S. Conference, Nancy, France, May 1994, Eds.: F. Bernardy, J.-L. Rivail; p. 486-499.
28. Baskin I. I., Skvortsova M. I., Stankevich I. V., Zefirov N. S. On the Basis of Invariants of Labeled Molecular Graphs // J. Chem. Inform. Comput. Sci., 1995, v. 35, N. 3, p. 527-531.
29. Stankevich I. V., Skvortsova M. I., Zefirov N. S. On a Quantum-Chemical Interpretation of Molecular Connectivity Indices for Conjugated Hydrocarbons // J. Mol. Strut. (THEOCHEM), 1995, v. 342, p.173-179.
30. Zefirov N. S., Palyulin V. A., Skvortsova M. I., Baskin I. I. Inverse Problem in QSAR // In: QSAR and Molecular Modeling: Concepts, Computational Tools and Biological Applications; Barcelona, Prous Science Publishers, 1995. p. 40.
31. Скворцова М. И., Баскин И. И., Словохотова О. Л., Палюлин В. А., Зефиров Н. С. Обратная задача в проблеме связи «структура-свойство» для случая корреляционного уравнения, содержащего произвольные топологические дескрипторы // ДАН, 1996, т. 346, N 4, с. 497-500.
32. Skvortsova M. I., Stankevich I. V., Baskin I. I., Palyulin V. A., Zefirov N. S. Analytical Description of a Set of Similarity Measures Defined on Molecular Graphs // In: Proceedings of the 11th European Symposium on Quantitative Structure-Activity Relationships: Computer-Assisted Lead Finding and Jptimization; Lausanne, September 1-6, 1996, p. 13.B.
33. Skvortsova M. I., Baskin I. I., Slovokchotova O. L., Palyulin V. A., Zefirov N. S. Inverse Problems in Quantitative Structure-Property Relationships Studies: Molecular Graph Reconstruction Using Graph Invariants // In: Proceedings of International Conference on Inverse and Ill-Posed Problems (IIPP-96), September 9-14, 1996, Moscow, Russia, p.169.
34. Baskin I. I., Skvortsova M. I., Stankevich I. V., Zefirov N. S. The Basis of Invariants of Labeled Molecular Graphs and Its Applications to Molecular Properties Prediction // In: Book of Abstracts. International Symposium CACR-96; December 17-18; 1996, Moscow, Russia; p. 39.

35. Skvortsova M. I., Baskin I. I., Stankevich I. V., Zefirov N. S. New Method for Constructing Linear "Structure-Property" Equations // In: Book of Abstracts. International Symposium CACR-96; December 17-18; 1996, Moscow, Russia; p. 60.
36. Skvortsova M. I., Baskin I. I., Stankevich I. V., Palyulin V. A., Zefirov N. S. Molecular Similarity in Structure-Property Relationships Studies. Analitical Description of the Complete Set of Graph Similarity Measures // In: Book of Abstracts; International Symposium CACR-96; December 17-18; 1996, Moscow, Russia; p. 67.
37. Skvortsova M.I., Baskin I.I., Stankevich I.V., Zefirov N. S. A New Approach to to the Problem of Defining Applicability Range of QSAR/QSPR Models // In: Book of Abstracts. International Symposium CACR-96; December 17-18; 1996, Moscow, Russia; p. 67-68.
38. Baskin I. I., Skvortsova M. I., Palyulin V. A., Zefirov N. S. Quantitative Chemical Structure-Property/Activity Studies Using Artifical Neural Networks // Foundations of Computing and Decision Sciences. 1997, v. 22, N 2, p.107-116.
39. Скворцова М. И., Баскин И. И., Станкевич И. В., Зефиров Н. С. Об одном способе построения линейных уравнений связи «структура-свойство» // ДАН, 1996, т.351, № 1, с. 78-80.
40. Скворцова М. И., Станкевич И. В., Баскин И.И., Палюлин В. А., Зефиров Н. С. Аналитическое описание множества мер подобия молекулярных графов // ДАН, 1996, т.350, № 6, с. 786-788.
41. Зефиров Н. С., Палюлин В. В., Молчанова М. С., Скворцова М. И., Баскин И. И. Структурная генерация и QSAR // Тезисы докладов IV – ого Российского научного конгресса «Человек и лекарство»; Москва, 8-12 апреля 1997г.; с. 261.
42. Скворцова М. И., Словохотова О. Л., Баскин И. И., Палюлин В. А., Зефиров Н. С. Обратная задача в проблеме связи «структура-свойство» для случая информационных топологических индексов // ДАН, 1997, т. 357, № 1, с. 72-74.
43. Skvortsova M. I., Baskin I. I., Stankevich I. V., Palyulin V. A., Zefirov N. S. Molecular Similarity. 1. Analitycal Description of Graph Similarity Measures // J. Chem. Inform. Comput. Sci. 1998, v.38, N 5, p. 785-790.
44. Skvortsova M. I., Baskin I. I., Skvortsov L. A., Palyulin V. A., Zefirov N. S., Stankevich I. V. Chemical Graphs and Their Basis Invariants//J. Mol. Struct. (THEOHEM), 1999, v. 466, p.211-217.
45. Скворцова М. И., Баскин И. И., Станкевич И. В., Палюлин В. А., Зефиров Н. С. Новый метод прогнозирования свойств химических соединений на основе оптимизации меры молекулярного подобия//Тезисы докладов I-ой Всероссийской конференции «Молекулярное моделирование», 14-16 апреля 1998г., (РАН, отделение общей и технической химии, Институт геохимии и аналитической химии им. В. И. Вернадского, Москва, 1998); С. 66.
46. Скворцова М. И., Баскин И. И., Словохотова О. Л., Палюлин В. А., Зефиров Н. С. Обратная задача в проблеме связи «структура-свойство» для случая топологических индексов // Тезисы докладов I-ой Всероссийской конференции «Молекулярное моделирование», 14-16 апреля 1998г., (РАН, отделение общей и технической химии, Институт геохимии и аналитической химии им. В. И. Вернадского, Москва, 1998); С. 67.
47. Станкевич И. В., Чистяков А. Л., Скворцова М. И. Исследование структуры и свойств некоторых эндоэндральных кластеров и обобщение понятия молекулярной топологической формы // Известия РАН, сер. химическая, 1999, № 3, с. 436-440.
48. Скворцова М. И., Станкевич И. В. Теория графов в структурной химии. Молекулярные графы. Часть I. - Изд-во МИТХТ. – 1998. - 88 с.
49. Artemenko N. V., Baskin I. I., Halberstam N. M., Skvortsova M. I., Palyulin V. A., Zefirov N. S. Combination of Substructural Approach and Neural Networks as a Universal Tool for Predicting Physico-Chemical Properties of Organic Compounds // In: Book of Abstracts. "QSAR 2000. Crosroads to the XXI Century. Ninth International Workshop on Quantitative Structure-Activity Relationships in Environmental Sciences». September 16-20, 2000, Bourgas, Bulgaria, p.66-67.

50. Скворцова М. И., Федяев К. С., Палюлин В. А., Зефиров Н. С. О вероятностном подходе к определению области применимости уравнений связи «структура-свойство» // ДАН, 2000, т. 375, № 1, с. 46-49.
51. Пасюков А. В., Скворцова М. И., Палюлин В. А., Зефиров Н. С. Метод прогнозирования свойств химических соединений, основанный на оптимальном подборе меры молекулярного подобия // ДАН, 2000, т. 374, № 6, с.786-789.
52. Skvortsova M. I., Fedyaev K. S., Palyulin V. A., Zefirov N. S. An Automatic Search for Chemical Structures with Given Properties Correlating Hosoya Index // In: Book of Abstracts of International School-Seminar on Computer Automatization and Information, Moscow, 2000 (Russian Academy of Sciences, Moscow State University, Russian Research Center "Kurchatov Institute", MC RAS-MS "ELICS", ACS'2000); p.47-48.
53. Skvortsova M. I., Fedyaev K. S., Palyulin V. A., Zefirov N. S. A Probability Technique for the Construction of the Applicability Range for "Structure-Property" Equation // In: Book of Abstracts of International School-Seminar on Computer Automatization and Information, Moscow, 2000 (Russian Academy of Sciences, Moscow State University, Russian Research Center "Kurchatov Institute", MC RAS-MS "ELICS", ACS'2000); p. 45-46.
54. Скворцова М. И., Федяев К. С., Палюлин В. А., Зефиров Н. С. О вероятностном подходе к определению области применимости в QSAR // Тезисы докладов II-ой Всероссийской конференции «Молекулярное моделирование», 24-26 апреля 2001 г., Москва, 2001 (РАН, Отделение Общей и технической химии, Институт геохимии и аналитической химии им. В. И. Вернадского, МГУ им. М. В. Ломоносова); с. 36.
55. Скворцова М. И., Федяев К. С., Палюлин В. А., Зефиров Н. С. Обратная задача в проблеме связи «структура-свойство» для индекса Хосойя // Тезисы докладов II-ой Всероссийской конференции «Молекулярное моделирование», 24-26 апреля 2001 г., Москва, 2001 (РАН, Отделение Общей и технической химии, Институт геохимии и аналитической химии им. В. И. Вернадского, МГУ им. М. В. Ломоносова); с. 99.
56. Скворцова М. И., Станкевич И. В. Теория графов в структурной химии. Спектры графов и их применение в теории сопряженных молекул. Часть II. – Москва. - МИТХТ им. М. В. Ломоносова. – 2001. - 64 с.
57. Скворцова М. И., Федяев К. С., Баскин И. И., Палюлин В. А., Зефиров Н. С. Структурно-вероятностный подход к определению области применимости линейной модели связи «структура-свойство» // Тезисы докладов II-ого международного симпозиума «Компьютерное обеспечение химических исследований» (Москва, 22-23 мая 2001 г.) и III-ей Всероссийской школы-конференции по квантовой и вычислительной химии им. В. А. Фока (Великий Новгород, 21-25 мая 2001 г.); с. 126.
58. Скворцова М. И., Федяев К. С., Баскин И. И., Палюлин В. А., Зефиров Н. С. Об одном способе кодирования химических структур в задачах построения математических моделей связи "структура-свойство" // Тезисы докладов II-ого международного симпозиума «Компьютерное обеспечение химических исследований» (Москва, 22-23 мая 2001 г.) и III-ей Всероссийской школы-конференции по квантовой и вычислительной химии им. В. А. Фока (Великий Новгород, 21-25 мая 2001 г.); с. 127.
59. Скворцова М. И., Федяев К. С., Палюлин В. А., Зефиров Н. С. Обратная задача в проблеме связи «структура-свойство» для случая корреляционного уравнения, содержащего индекс Хосойя // ДАН, 2001, т. 379, № 2, с. 209-213.
60. Скворцова М. И., Федяев К. С., Баскин И.И., Палюлин В. А., Зефиров Н.С. Новый способ кодирования химических структур на основе базисных фрагментов // ДАН, 2002, т. 382, № 5, с. 645-648.
61. Скворцова М. И., Федяев К. С., Палюлин В. А., Зефиров Н. С. Обратная задача в проблеме связи «структура-свойство» для случая корреляционных уравнений, содержащих базисные топологические дескрипторы//Тезисы докладов III-ей Всероссийской конференции

«Молекулярное моделирование», 15-17 апреля 2003 г. (Москва, РАН, Отделение химии и наук о материалах, Ин-т геохимии и аналитической химии им. В. И. Вернадского, МГУ им. М. В. Ломоносова); с. 110.
62. Скворцова М. И., Федяев К. С., Палюлин В. А., Зефиров Н. С. Базисные топологические дескрипторы и их применение для построения корреляций «структура-свойство» // Тезисы докладов III-ей Всероссийской конференции «Молекулярное моделирование», 15-17 апреля 2003 г. (Москва, РАН, Отделение химии и наук о материалах, Ин-т геохимии и аналитической химии им. В. И. Вернадского, МГУ им. М. В. Ломоносова); с. 109.
63. Skvortsova M. I., Fedyaev K. S., Palyulin V. A., Zefirov N. S. Molecular Design of Chemical Compounds with Prescribed Properties from QSAR Models Containing the Hosoya Index // Internet Electron. J. Mol. Des. 2003, N 2, p.70-95; http://www.biochempress.com.
64. Skvortsova M. I., Fedyaev K. S., Palyulin V. A., Zefirov N. S. Inverse Problem in Quantitatve Structure-Property Relationships Studies for Correlations Constructed by Basic Topological Molecular Descriptors // In: Book of Abstracts "Modern Trends in Organometallic and Catalitic Chemistry. Mark Vol'pin (1923-1996) Memorial International Symposium, Moscow, May 18-23, 2003", p.181.
65. Скворцова М. И., Федяев К. С., Палюлин В. А., Зефиров Н. С. Моделирование связи между структурой и свойствами углеводородов на основе базисных топологических дескрипторов // Известия АН (сер. химическая), 2004, № 8, с. 1527-1535.
66. Скворцова М. И., Станкевич И. В. О связи между собственными векторами взвешенных графов и их подграфами // Дискретная математика, 2004, т.16, вып. 4, с. 32-40.
67. Скворцова М. И., Палюлин В. А., Зефиров Н. С. Компьютерный дизайн топологических индексов в органической химии // Тезисы докладов IV-ой Всероссийской конференции «Молекулярное моделирование", 12-15 апреля 2005 г., Москва (РАН – Отделение наук о Земле, ГЕОХИ им. В.И. Вернадского, МГУ им. М.В. Ломоносова); с.93.
68. Скворцова М. И., Палюлин В. А., Зефиров Н. С. Построение моделей связи «структура-свойство» на основе концепции молекулярного подобия путем оптимального подбора меры подобия и молекулярных дескрипторов//Тезисы докладов IV-ой Всероссийской конференции «Молекулярное моделирование", 12-15 апреля 2005 г., Москва (РАН - Отделение наук о Земле, ГЕОХИ им. В. И. Вернадского, МГУ им. М. В. Ломоносова), с. 94.
69. Skvortsova M. I., Stankevich I. V. Eigenvectors of Weighted Graphs: Supplement to Sachs' Theorem // J. Mol. Struct. (THEOCHEM), 2005, v.719, p. 213-223.
70. Palyulin V. A., Skvortsova M. I., Zotov A. Yu., Zefirov N.S. Computed-Aided Design of Topological Indices // In: Book of Abstracts. Fourth Indo-US Workshop on Mathematical Chemistry (With Applications to Drug Discovery, Environmental Toxicology, Chemoinformatics and Bioinformatics), Yanuary 8-12, 2005, Pune, Maharashtra, India; p.16.
71. Скворцова М. И., Федяев К. С., Палюлин В. А., Зефиров Н. С. Теоретико-графовый подход к моделированию связи между строением и свойствами углеводородов // Сборник научных трудов 11-ой Международной конференции «Математические модели физических процессов», т. 1 (29-30 июня 2005, Россия, Таганрог, ТГПИ); с. 254-259.
72. Скворцова М. И., Станкевич И. В. Система искусственного интеллекта для конструирования инвариантов графов в органической химии // Сборник трудов XIX Международной научной конференции «Математические методы в технике и технологиях», т. 6 (30 мая-2 июня 2006, Россия, Воронеж, ВГТА); с. 62-64.
73. Скворцова М. И., Станкевич И. В., Палюлин В. А., Зефиров Н. С. Концепция молекулярного подобия и ее применение для прогнозирования свойств органических соединений // Успехи химии, 2006, т.75, № 11, 1074-1093.