Федеральное государственное бюджетное учреждение
«Научно-исследовательский институт биомедицинской химии
имени В. Н. Ореховича»
Российской академии медицинских наук

На правах рукописи

Ромеро Рейес Илякай Владиславовна

ОЦЕНКА АФФИННОСТИ КОМПЛЕКСОВ БЕЛОК-ЛИГАНД С ПРИМЕНЕНИЕМ НЕЙРОННЫХ СЕТЕЙ

Специальность: 05.13.18 – математическое моделирование,
численные методы и комплексы программ

Диссертация на соискание ученой степени
кандидата физико-математических наук

Научный руководитель: кандидат физико-математических наук
Филимонов Дмитрий Алексеевич

Москва – 2014

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ 4

ГЛАВА 1. ОБЗОР ЛИТЕРАТУРЫ 9

1.1 Методы компьютерного конструирования лекарств 9

1.1.1 Конструирование лекарств на основе структуры лигандов 9

1.1.2 Конструирование лекарств на основе структуры белка-мишени 14

1.2 Искусственные нейронные сети 20

1.2.1 Модель нейрона 21

1.2.2 Персептроны 24

1.2.3 Алгоритмы обучения многослойного персептрона 27

1.2.4 Переобучение и переподгонка данных 33

1.3 Методы снижения размерности 35

1.3.1 Линейные методы 36

1.3.2 Глобально нелинейные методы 37

1.3.3 Локально линейные методы 41

1.3.4 Расширение вложения для новых точек 44

1.4 Параллельные вычисления с использованием технологии CUDA 46

ГЛАВА 2. РАЗРАБОТКА МЕТОДА ОЦЕНКИ АФФИННОСТИ КОМПЛЕКСОВ БЕЛОК-ЛИГАНД 52

2.1 Объекты исследования 52

2.2 Молекулярное моделирование 54

2.3 Численный метод оценки аффинности 55

2.3.1 Входные параметры и выходные значения 55

2.3.2 Предварительная обработка данных 56

2.3.3 Структура ИНС и ее оптимизация 58

2.3.4 Параметры оценки моделей 60

2.3.5 Результаты построения моделей с использованием метода главных компонент 61

2.3.6 Результаты построения моделей с использованием нелинейных методов снижения размерности 67

ГЛАВА 3. ТЕСТИРОВАНИЕ РАЗРАБОТАННОГО МЕТОДА ОЦЕНКИ АФФИННОСТИ 74

3.1 Объекты исследования 74

3.2 Модели оценки аффинности 75

3.2.1 3D QSAR модели 76

3.2.2 Модели на основе молекулярного моделирования 76

3.2.3 Результаты построения моделей 78

3.3 Тестирование моделей оценки аффинности 79

ГЛАВА 4. ПРОГРАММНАЯ РЕАЛИЗАЦИЯ РАЗРАБОТАННОГО МЕТОДА 81

4.1 Итоговая модель 81

4.2. Параллельная реализация 82

ВЫВОДЫ 88

СПИСОК СОКРАЩЕНИЙ 89

СПИСОК ЛИТЕРАТУРЫ 91

ВВЕДЕНИЕ

Актуальность темы исследования. В настоящее время при создании и поиске новых лекарственных соединений активно применяют различные компьютерные методы поиска, молекулярного моделирования и конструирования фармакологически перспективных соединений *de novo* [1]. Применение данных методов позволяет существенно ускорить процессы разработки и внедрения и снизить их стоимость.

Одной из важных задач является компьютерный расчет аффинности предполагаемых лигандов к макромолекуле-мишени. Наиболее распространены два подхода: расчет изменения свободной энергии при связывании лиганда с макромолекулой-мишенью методами молекулярного моделирования и подбор для каждой мишени оценочной функции. Однако компьютерные оценки изменения свободной энергии сопряжены с фундаментальными ограничениями и имеют низкую точность, так как при данном подходе часто сложно учесть энтропийную составляющую энергии взаимодействия. Второй подход основан на использовании выборок соединений с соответствующими экспериментально измеренными величинами для подбора параметров оценочных функций. Для этих методов основное ограничение заключается в том, что такие модели дают хорошие предсказания только для лигандов того же химического класса, что и соединения из обучающей выборки. При этом дескрипторы для лигандов каждого класса необходимо подбирать индивидуально. Поэтому возникает необходимость создания новых численных методов оценки аффинности с учетом и минимизацией недостатков известных подходов.

В данной работе предложен комбинированный метод на основе универсальных дескрипторов, не зависящих от химического класса лигандов и получаемых путем объединения результатов молекулярного моделирования комплексов белок-лиганд и подхода на основе лигандов с известными свойствами. Для решения задачи построения оценочной функции в данной работе

применяются искусственные нейронные сети (ИНС), широко используемые в последние годы в задачах прогнозирования физико-химических свойств органических соединений и их биологической активности. Для снижения числа подаваемых на вход ИНС параметров и ускорения процедуры вычисления параметров ИНС были использованы методы снижения размерности.

Цель работы: разработка численных методов для оценки аффинности комплексов белок–лиганд с применением нейронных сетей и их реализация в виде комплекса программ на графических процессорах.

Задачи исследования:

1. Подготовить выборки данных по белкам, лигандам и комплексам белок–лиганд и выполнить молекулярное моделирование отобранных комплексов и расчет энергетических параметров их взаимодействия по результатам молекулярной динамики.
2. Разработать численный метод для оценки аффинности лигандов к ядерным рецепторам стероидных гормонов на основе комплексного подхода, совмещающего методы молекулярного моделирования и искусственных нейронных сетей и провести тестирование на независимой выборке.
3. Разработать высокопроизводительную программную реализацию метода с применением параллельных вычислений с использованием графических процессоров.

Научная новизна. Впервые предложен метод оценки аффинности нестероидных лигандов к рецепторам глюкокортикоидов и прогестерона с использованием методов нелинейного снижения размерности и ИНС, и показана возможность применения метода расширения вложения для новых точек [2] в задаче оценки параметров взаимодействия комплексов белок–лиганд. Разработанные методы реализованы в виде программ, поддерживающих параллельные вычисления на основе графических процессоров.

Практическая значимость работы. Результаты диссертации могут быть использованы для оценки аффинности лигандов к ядерным рецепторам стероидных гормонов на основе физико-химических дескрипторов лигандов и составляющих изменения энергии взаимодействия комплексов белок–лиганд. Предлагаемый метод позволяет получить статистически значимые модели оценки аффинности для рассматриваемых рецепторов (коэффициент детерминации $\overline{R^2} = 0{,}94$) в отличие от моделей по оценочным функциям молекулярного моделирования (коэффициент детерминации $\overline{R^2} < 0{,}1$).

Разработанный автором метод был применен в лаборатории структурной биоинформатики ФГБУ «ИБМХ» РАМН для оценки аффинности стероидных лигандов ко всем ядерным рецепторам стероидных гормонов и показал хорошую предсказательную способность [3].

Также совместно с вычислительным экспериментом для ряда лигандов-пентаранов рецептора прогестерона был проведен синтез в Институте органической химии имени Н. Д. Зелинского РАН и тестирование *in vitro* Московском государственном университете имени М. В. Ломоносова. Предсказанные значения аффинности комплексов рецептор–лиганд хорошо согласуются с результатами экспериментальной проверки.

Положения, выносимые на защиту.

- Новый метод оценки аффинности лигандов к внутриклеточным рецепторам глюкокортикоидов и прогестерона с использованием нелинейного снижения размерности и искусственных нейронных сетей на основе физико-химических параметров лиганда и составляющих изменения энергии взаимодействия комплексов белок–лиганд.
- Применение метода расширения вложения для новых точек в задаче оценки аффинности комплексов белок–лиганд.
- Параллельная программная реализация разработанных методов с использованием графических процессоров.

Достоверность результатов работы.

- На этапе вычисления дескрипторов применяются общепринятые, стандартно используемые программные пакеты (Dock 6.5 [4], SYBYL 8.1 [5], AMBER 9 [6]).
- Для процедур подготовки обучающей выборки и обучения ИНС разработана MATLAB-реализация с использованием встроенного пакета для ИНС Neural Network Toolbox [7], реализация на C++, а также параллельная C++\CUDA C [8] реализация для гибридных вычислительных систем с графическими процессорами. Тестовые расчеты подтверждают совпадение результатов работы упомянутых компьютерных реализаций.
- На этапе снижения размерности обучающей выборки проведены расчеты на основе различных известных подходов, обеспечивающие согласующиеся между собой результаты.
- Сопоставление расчетных данных с измерениями, полученными в результате синтеза и исследования *in vitro* для ряда соединений прегна-D´-пентаранов, лигандов внутриклеточного рецептора прогестерона, в лаборатории химии стероидных соединений Института органической химии имени Н.Д. Зелинского РАН и Московском государственном университете имени М.В. Ломоносова подтверждает эффективность разработанного подхода и достоверность получаемых результатов по оценке аффинности комплексов белок-лиганд.

Апробация результатов работы. Основные положения и результаты диссертационной работы докладывались на:

- XVI Российском национальном конгрессе «Человек и лекарство», Москва, Россия, 2010;
- IV сессии научной школы-практикума молодых ученых и специалистов в рамках VIII Всероссийской межвузовской конференции молодых ученых, Санкт-Петербург, Россия, 2011;

- XVI Международной конференции по нейрокибернетике, Ростов-на-Дону, 2012;
- Международной суперкомпьютерной конференции «Научный сервис в сети Интернет: поиск новых решений», Новороссийск, Россия, 2012;
- XVII научной конференции молодых ученых и специалистов (ОМУС-2013) к 100-летию В. П. Джелепова, Дубна, Россия, 2013.

ГЛАВА 1. ОБЗОР ЛИТЕРАТУРЫ

1.1 Методы компьютерного конструирования лекарств

Методы компьютерного конструирования лекарств позволяют решать следующий ряд задач [9, 10]:

1. поиск новых биологически активных соединений, которые взаимодействуют с требуемым рецептором и оказывают минимальное влияние на другие белки;
2. моделирование взаимодействия лиганда с макромолекулой-мишенью;
3. поиск зависимостей между структурой лигандов и целевой биологической активностью (свойством);
4. предсказание биологической активности для новых соединений;
5. конструирование новых биологически активных молекул.

Задача компьютерной оценки аффинности (сродства) лиганда к макромолекуле-мишени представляет собой частный случай общей задачи поиска зависимостей структура-свойство [11]. Для решения таких задач применяют методы, которые можно разделить на две группы: на основе структур лигандов (Ligand-Based Drug Design, LBDD) и на основе структуры белка-мишени (Structure-Based Drug Design, SBDD) [10].

1.1.1 Конструирование лекарств на основе структуры лигандов

Первая группа методов используется в случае неизвестной пространственной структуры мишени. К ним относят построение фармакофорных моделей [12], моделей «псевдорецептора» [11, 13], методы поиска количественных соотношений «структура-активность» (Quantitative Structure-Activity Relationship, QSAR) [14].

Методы QSAR базируются на описании лигандов с помощью молекулярно-структурных параметров (дескрипторов) и построении функциональной

зависимости (модели) между значениями дескрипторов и величиной заданной биологической активности. При таком подходе главную роль играет набор выбранных дескрипторов, описывающий все особенности структуры лиганда, от которых может зависеть значение активности. Построенная на выбранных дескрипторах модель позволяет предсказывать значение активности для новых молекул из того же узкого химического класса соединений, на которых проводилось построение.

Классификация методов QSAR базируется на размерности способа представления дескрипторов или описания структуры соединения [15, 16]. В таблице 1.1 представлены основные виды методов QSAR и соответствующие используемые дескрипторы.

В методах 1D QSAR для поиска соотношений «структура-активность» используют физико-химические характеристики молекулы, такие как pK_i, logP и другие.

Таблица 1.1 – Классификация методов QSAR

Метод QSAR	Дескрипторы
1D QSAR	Молекулярные характеристики pK_i, logP и другие [17-19]
2D QSAR	Топологические характеристики структур: индексы связности, 2D фармакофоры и другие [20-22]
3D QSAR	Потенциалы молекулярных полей с учетом пространственной структуры молекулы [23-26]
4D QSAR	Дескрипторы 3D QSAR+ учет конформаций лигандов [27, 28]
5D QSAR	Дескрипторы 4D QSAR+учет изменения конформации лиганда при связывании с рецептором [29, 30]
6D QSAR	Дескрипторы 5D QSAR + учет эффекта растворителя [31]

В методах 2D QSAR модели строят на основе плоской структурной

формулы, где учитывается топология молекул, в неявном виде отражающая возможные конформации исследуемого соединения.

В методах 3D QSAR и выше пространственную структуру учитывают в явном виде. В таких подходах содержится информация о составе, топологии и пространственной форме молекулы.

Методы 3D QSAR

Традиционные методы 3D QSAR [25, 32] моделируют зависимость биологической активности лигандов с использованием потенциалов молекулярных полей. Первоначально, производят выравнивание рассматриваемых лигандов и, предполагая, что лиганды лежат в центре связывания мишени сходным образом, формируют общую для всех лигандов воображаемую трехмерную решетку. Далее для оценки потенциалов различных видов молекулярных полей вычисляют энергии взаимодействия между атомами выравненных структур и тестирующими атомами, которые помещают в узлы трехмерной решетки. Тестирующими атомами являются:

- атомы углерода (по умолчанию, но возможно использование других, например, атомов кислорода) для оценки потенциалов стерического поля, описываемого потенциалом Леннарда-Джонса

$$E_{vwW} = \sum_{i=1}^{n} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right), \tag{1.1}$$

где E_{vwW} – энергия взаимодействия Ван-дер-Ваальса,

r_{ij} – расстояние между i-ым атомом молекулы и j-ой точкой решетки, куда помещен тестирующий атом,

A_{ij} и B_{ij} – константы потенциала Леннарда-Джонса.

- протоны для расчета потенциалов электростатического поля по закону

Кулона:

$$E_{ele} = \sum_{i=1}^{n} \frac{q_i q_j}{D r_{ij}}, \quad (1.2)$$

где E_{ele} – энергия кулоновского взаимодействия,

q_i – частичный заряд i-ого атома молекулы,

q_j – заряд тестирующего атома,

D – диэлектрическая постоянная,

r_{ij} – расстояние между i-ым атомом молекулы и j-ой точкой решетки, куда помещен тестирующий атом.

- молекулы воды для определения гидрофобных областей и вероятности образования водородных связей.

Затем формируют матрицу результатов вычислений и с помощью статистического анализа получают регрессионное уравнение количественных соотношений между значениями энергий взаимодействия и целевыми значениями биологической активности.

Метод CoMFA. Одним из наиболее популярных методов описания взаимодействия, применяемых в 3D QSAR, является метод сравнительного анализа молекулярных полей CoMFA (Comparative Molecular Field Analysis), который был разработан Крамером в 1988 году [33].

Описание нековалентных взаимодействий между лигандом и белком в классическом CoMFA осуществляют с помощью потенциалов электростатических и стерических полей [33]. В непосредственной близости к поверхности атомов оба потенциала имеют очень быстрое изменение, которому соответствуют значения бесконечности, если позиции атомов двух молекул перекрываются. Чтобы избежать этого, определяют отсечение потенциалов $E_{cut-off}$ на некотором радиусе отсечения $r_{cut-off}$ и используют все значения $E(r)$ при $r > r_{cut-off}$, которые больше (выходят за пределы) значений отсечения (рисунок 1.1).

Количественные соотношений между значениями энергий взаимодействия и целевыми значениями биологической активности вычисляют с помощью метода частичных наименьших квадратов [34]. Качество построенных моделей отслеживают посредством перекрестного контроля, например, с исключением по одному (Leave-one-out Cross-validation, LOO) [35, 36]. Значение коэффициента детерминации предсказаний Q^2 по LOO для надежных предсказательных моделей должно быть не менее 0,5 [23, 37, 38].

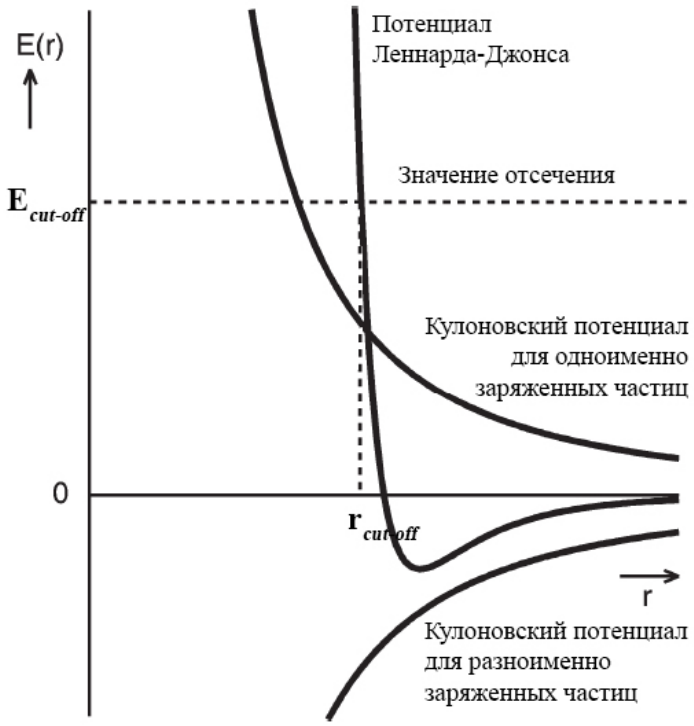


Рисунок 1.1 – Потенциалы электростатического и стерических полей в методе CoMFA, соответствующие потенциалам Кулона и Леннарда-Джонса [39].

Метод CoMSIA. Метод сравнительного анализа индексов молекулярного подобия CoMSIA (Comparative Molecular Similarity Indices Analysis) [40] является развитием метода CoMFA. В этой модификации тестирующие атомы используют для вычисления индексов молекулярного подобия с исследуемыми молекулами в различных точках трехмерной решетки. С помощью этих индексов описывают электростатические, стерические, гидрофобные поля, а также поля водородных

связей. По этим полям строят корреляционные зависимости с целевыми значениями биологической активности, как в CoMFA. Главное преимущество методов CoMSIA состоит в описании потенциалов в непосредственной близости к поверхности атомов с помощью гауссиана (рисунок 1.2), что не требует введения отсечения потенциалов и позволяет избежать резкого изменения в их зависимости.

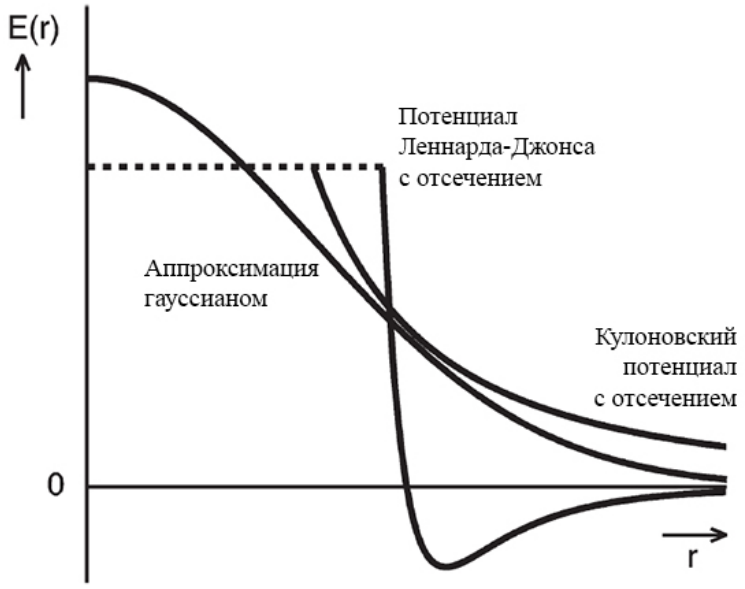


Рисунок 1.2 – Потенциалы электростатического и стерических полей в методе CoMSIA, на основе потенциалов Кулона и Леннарда-Джонса, соответственно [39].

Для методов на основе структуры лигандов главное ограничение заключается в том, что модели на их основе дают хорошие предсказания только для лигандов того же химического класса, что и соединения из обучающей выборки.

1.1.2 Конструирование лекарств на основе структуры белка-мишени

Вторая группа методов требует знания структуры белка-мишени с местом связывания рассматриваемых лигандов. Для этого необходимы данные о трехмерной

структуре рецептора и/или места связывания с лигандом, которые исследуют методами рентгеноструктурного анализа, ядерного магнитного резонанса, компьютерным моделированием на основе гомологии [41]. Если эта информация доступна, то наиболее распространенным подходом поиска потенциальных биологически активных веществ являются методы докинга и *de novo*. Основной плюс их применения состоит в быстром получении результатов расчетов.

Методы молекулярного докинга и молекулярной механики

Метод молекулярного докинга [42] позволяет определить наиболее достоверную ориентацию и конформацию лиганда в месте связывания белка-мишени. На выходе этот метод дает набор гипотез вероятных положений лигандов. Наиболее выгодную ориентацию и положение лиганда выявляют по оценочной функции энергии взаимодействия образования комплекса. Основные сложности применения метода молекулярного докинга связаны с учетом гибкости рецептора, которые частично могут быть решены методами молекулярной механики и динамики. Также к одной из трудностей докинга относят способ построения оценочной функции [43], что приводит к «отдаленности» оценок энергии взаимодействия комплексов от реальных величин [44].

Методы молекулярной механики и динамики позволяют получить более точную оценку энергии взаимодействия образования комплекса по сравнению со встроенными оценочными функциями молекулярного докинга, так как позволяют учитывать конформационные изменения лиганда и места связывания белка-мишени, а также их кинематические и термодинамические свойства. В зависимости от необходимой степени детализации используют два вида приближений:

- упрощенные модели на основе молекулярной механики, которые позволяют рассчитать функцию потенциальной энергии систем, называемую силовым полем,
- полноатомное приближение с помощью методов молекулярной динамики, которые позволяют на основе функции силового поля, полученной по

принципам молекулярной механики, рассчитать значения действующих на все атомы сил, используемых, в дальнейшем, при численном интегрировании уравнений движений.

Для уточнения результатов докирования в большинстве случае достаточно детализации на уровне молекулярной механики.

В основе молекулярной механики лежит идея рассмотрения молекулярной системы комплекса белок-лиганд как микроскопической механической системы, согласно которой все атомы соединены механическими пружинами, контролирующими длину ковалентных связей, углы ковалентных связей, вращение вокруг связей и т.д., а взаимодействие атомов осуществляется по классическим невалентным потенциалам [45] (рисунок 1.3).

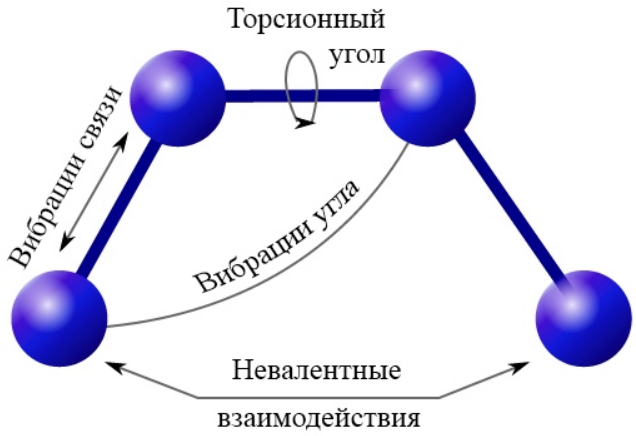


Рисунок 1.3 – Схематическое изображение молекулярной системы в рамках подхода молекулярной механики.

По функции потенциальной энергии вычисляют силы взаимодействия между атомами, которые подставляют в уравнения движения. В рамках молекулярной механики используют уравнения движения Ньютона:

$$m_i \frac{d\vec{r}_i(t)}{dt^2} = \vec{F}_i, i = 1,2, \dots n\,, \tag{1.3}$$

где m_i – масса i-го атома,

$\{\vec{r}_1, \vec{r}_2, \dots, \vec{r}_n\}$ – совокупность координат,

$\vec{F}_i$ – результирующая сила, действующая на i-ый атом,

n – количество атомов в системе.

В силу аддитивности полной потенциальной энергии системы комплекса белок-лиганд силовое поле задают в виде суммы ковалентных (U_{bonded}) и нековалентных ($U_{non-bonded}$) взаимодействий [46]:

$$U_{MM} = U_{bonded} + U_{non-bonded} \,. \tag{1.4}$$

U_{bonded} представляет собой сумму вкладов гармонических колебаний ковалентных связей, углов между смежными ковалентными связями и торсионных взаимодействий для правильных и неправильных торсионных углов:

$$\begin{aligned} U_{bonded} = & \sum_{bonds} K_r (r - r_0)^2 + \sum_{angles} K_\theta (\theta - \theta_0)^2 + \\ & + \sum_{dihedral} \frac{U_n}{2} (1 + \cos(n\phi - \phi_0)) \,, \end{aligned} \tag{1.5}$$

где r – межатомное расстояние,

θ - валентный угол между смежными ковалентными связями,

ϕ – торсионный угол поворота,

n – периодичность вращения,

U_n – высота барьера вращения,

r_0, θ_0, ϕ_0 – равновесные значения соответствующих величин.

$U_{non-bonded}$ представляет собой сумму потенциала электростатических взаимодействий, потенциала Леннарда-Джонса и потенциала водородных связей:

$$U_{non-bonded} = \sum_{i<j} \frac{q_i q_j}{\varepsilon r_{ij}} + \sum_{i<j} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right) + \sum_{H-bonds} \left(\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}} \right), \quad (1.6)$$

где q_i – частичный заряд i-ого атома,

r_{ij} – расстояния между i-ым и j-ым атомами,

$\varepsilon = \varepsilon(r_{ij})$ – функция диэлектрической проницаемости среды как функция межатомного расстояния,

A_{ij} и B_{ij} – константы потенциала Леннарда-Джонса,

C_{ij} и D_{ij} – эмпирические параметры потенциала водородных связей.

Поскольку биологические молекулы находятся в водном окружении, необходимо учитывать изменение сольватации лиганда и мишени при образовании комплекса. Наиболее распространенным и точным методом является явное добавление молекул воды в систему, которые оказывают влияние на все составляющие нековалентного взаимодействия.

В области молекулярного моделирования методами молекулярной механики широко распространен метод MM-PBSA/MM-GBSA [47], который позволяет оценить изменение свободной энергии ΔG комплекса с учетом влияния растворителя. Термодинамический цикл для вычисления ΔG образования комплекса представлен на рисунке 1.4. $\Delta G^{\text{связывания в}}_{\text{газовой среде}}$ представляет собой изменение энергии взаимодействия между лигандом и рецептором в газовой среде, а $\Delta G^{\text{сольватации}}_{\text{рецептора}}$, $\Delta G^{\text{сольватации}}_{\text{лиганда}}$ и $\Delta G^{\text{сольватации}}_{\text{комплекса}}$ – вклады сольватации в изменение свободной энергии рецептора, лиганда и комплекса, соответственно.

Изменение свободной энергии связывания лиганда с рецептором оценивают как разность изменений свободной энергии комплекса, рецептора и лиганда:

$$\Delta G_{\text{связвания лиганда с рецептором}} = \Delta G_{\text{комплекса}} - \\ -\left(\Delta G_{\text{рецептора}} + \Delta G_{\text{лиганда}}\right), \quad (1.7)$$

где $\Delta G = \bar{U}_{MM} + \bar{G}_{PBSA} - TS_{MM}$. Электростатическую составляющую свободной энергии сольватации $\bar{G}_{PBSA}$ вычисляют по уравнению Пуассона-Больцмана [48] или по обобщенной модели Борна – $\bar{G}_{GBSA}$. [49]. Энтропийную составляющую $-TS_{MM}$ оценивают посредством анализа нормальных мод траектории или квазигармонического анализа траектории [50]. Гидрофобную составляющую свободной энергии сольватации рассчитывают по доступной для растворителя площади поверхности.

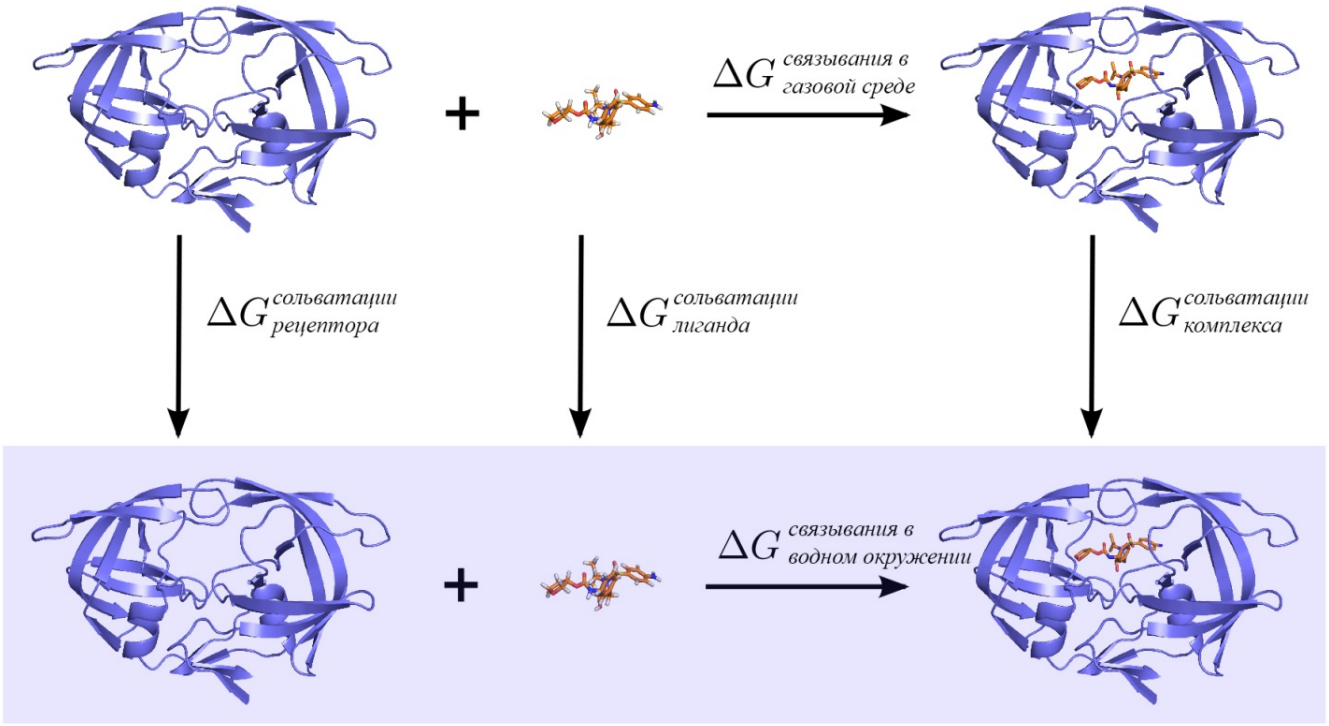


Рисунок 1.4 – Термодинамический цикл для вычисления свободной энергии образования комплекса

Основной недостаток методов молекулярной механики заключается в длительности расчетов и недостаточного учета энтропийного вклада энергии взаимодействия.

Метод *de novo*

Вторым подходом работки низкомолекулярных лигандов, структура которых максимально соответствует строению места связывания целевого рецептора, является метод конструирования структуры *de novo*. В этом случае структуру лигандов воссоздают на базе небольших фрагментов, помещаемых в место связывания рецептора с минимизацией стерических затруднений и максимизацией энергий связывания [51]. Но недостатки такого подхода заключаются в низкой надежности оценки связывания лиганда с рецептором, а также в довольно общих принципах моделирования молекул, из-за чего большую часть сгенерированных соединений невозможно синтезировать на практике.

1.2 Искусственные нейронные сети

Одно из перспективных направлений QSAR связано с применением искусственных нейронных сетей [52, 53]. Первое применение ИНС в этой области было отмечено в начале 1970-х годов. В 1971 году Гиллер и соавт. [54] опубликовали результаты исследования классификации 1,3-диоксанов на активные и неактивные с точки зрения их физиологической активности с использованием персептрона Розенблатта [55], единственного известного на тот момент вида нейронных сетей. Спустя 20 лет в 1990 году с работ Аояма и соавт. [56, 57] стартовал следующий этап развития этого направления. За последние годы этот подход к моделированию количественных соотношений «структура-активность» сильно развился и превратился в установившуюся область науки с успешным практическим применением.

ИНС могут быть использованы для решения любой разрешимой задачи, начиная от простого сложения двоичных чисел до доказательства теорем. Разрешению таких задач посвящен специальный раздел теории вычислительных систем (computer science), нейроматематика [58]. На практике, как правило, ИНС используют для решения некорректных задач, для которых могут быть

предложены альтернативные численные решения, таких, как аппроксимация функций, распознавание образов, кластеризация и снижение размерности. Эти задачи в точности соответствуют тем задачам, которые решают в большинстве исследований QSAR с использованием нейронных сетей [57, 59-62].

1.2.1 Модель нейрона

При создании метода искусственных нейронных сетей в их основе лежала идеология воспроизведения нервной системы и мозга человека, состоящих из нейронов, соединенных между собой нервными волокнами, по которым происходит передача импульсов. **Биологический нейрон** представляет собой специальную клетку, которая имеет ядро и два вида отростков (рисунок 1.5): группа дендритов, принимающих импульс, и единственный аксон, передающий импульс.

Связь биологических нейронов между собой осуществляется через синапсы между аксонами и дендритами нейронов. Каждую связь характеризуют силой синаптической связи, преобразующей электрохимический импульс при передаче соседнему нейрону.

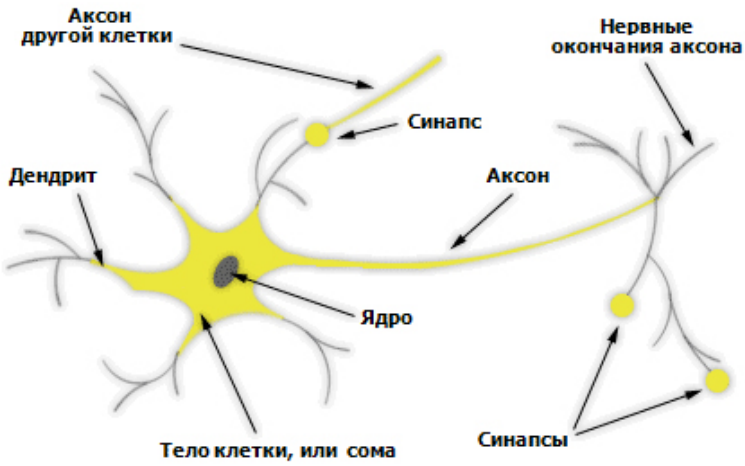


Рисунок 1.5 – Модель биологической нейронной сети [63]

Искусственный нейрон представляет собой упрощенную математическую модель биологического нейрона. Его структура представлена на рисунке 1.6.

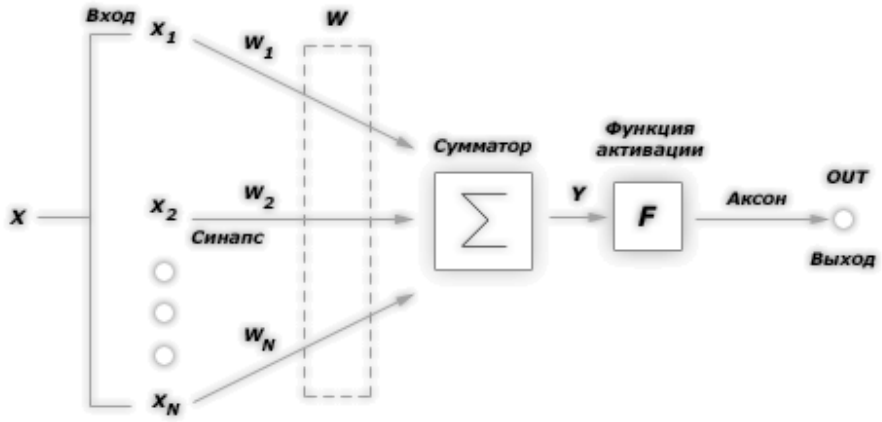


Рисунок 1.6 – Структура искусственного нейрона [64]

У искусственного нейрона также есть группа «синапсов», характеризующихся отдельными весами и осуществляющих связь между соседними нейронами, умножая вектор входных сигналов $X = [x_1, x_2, \dots, x_N]$ на вектор весов $W = [w_1, w_2, \dots, w_N]$, характеризующий силу синаптической связи. Аналогом тела клетки нейрона является сумматор, выполняющий сложение $Y = w_0 + \sum_{i=1}^{N} x_i w_i$ входных сигналов, взвешенных относительно соответствующих синапсов нейрона, и порогового значения w_0 активации нейрона.

Функция активации (возбуждения) $F(Y)$ определяет выходной результирующий уровень данного нейрона, с которым сигнал возбуждения поступает по выходной связи (аксону) на синапсы соседних нейронов.

Среди функций активации выделяют три основных вида:

1. Пороговая функция – функция Хэвисайда:

$$F(Y) = \begin{cases} 1, \text{если } Y \geq 0 \\ 0, \text{если } Y < 0 \end{cases} \quad (1.8)$$

Такой вид функции активации был использован в первой математической модели искусственного нейрона, созданной МакКаллоком и Питтсом [65].

2. Кусочно-линейная функция:

$$F(Y) = \begin{cases} 1, \text{если } Y \geq \frac{1}{2} \\ |Y|, \text{если } -\frac{1}{2} < Y < \frac{1}{2} \\ 0, \text{если } Y \leq -\frac{1}{2} \end{cases} \quad (1.9)$$

3. Сигмоидальная функция
   а. Логистическая функция

$$F(\alpha Y) = \frac{1}{1 + \exp(-\alpha Y)}, \quad (1.10)$$

где α – угол наклона, область значений функции: $(0,1)$.

   б. Гиперболический тангенс

$$F(Y) = \tanh(\alpha Y) = \frac{\exp(\alpha Y) - \exp(-\alpha Y)}{\exp(\alpha Y) + \exp(-\alpha Y)}, \quad (1.11)$$

где α – угол наклона, область значений функции: $(-1,1)$.

Обе эти функции являются монотонно возрастающими и всюду дифференцируемыми. При этом они усиливают малые сигналы и предотвращают насыщение от больших сигналов.

1.2.2 Персептроны

Архитектуру ИНС определяют по структуре связей искусственных нейронов между собой. Одной из первых ИНС появился ***персептрон Розенблатта*** [55] (рисунок 1.7).

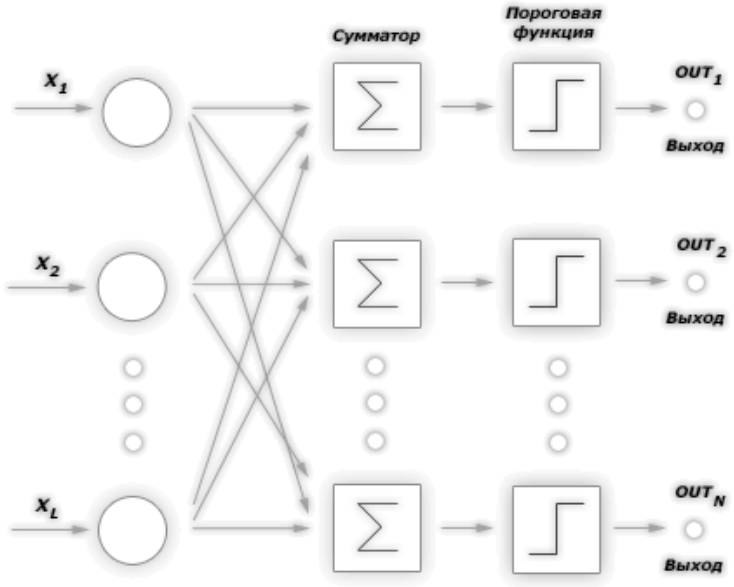


Рисунок 1.7 – Персептрон Розенблатта [66]

Данная ИНС была однослойной, и в качестве искусственного нейрона был использован нейрон МакКаллока-Питтса. Розенблатт первоначально предположил, что предложенная им нейронная сеть сможет воспроизвести любую логическую функцию. Однако Минский и Пейперт [67] выявили принципиальные неустранимые ограничения однослойных персептронов, и в дальнейшем, к середине 80-ых годов, распространение получил многослойный вариант персептрона, в котором имеются несколько слоев нейронов и другой вид функции активации.

К этому времени было выяснено, что замена пороговой функции активации на непрерывно дифференцируемую, например, сигмоидальную функцию,

позволяет устранить недостатки персептрона Розенблатта. Многослойные персептроны (рисунок 1.8) включают в себя множество входных узлов, которые образуют входной слой; один или несколько скрытых слоев нейронов; один выходной слой нейронов. В силу того, что в ИНС такого типа сигнал передается в прямом направлении, от входного слоя к выходному, их также называют нейронными сетями прямого распространения.

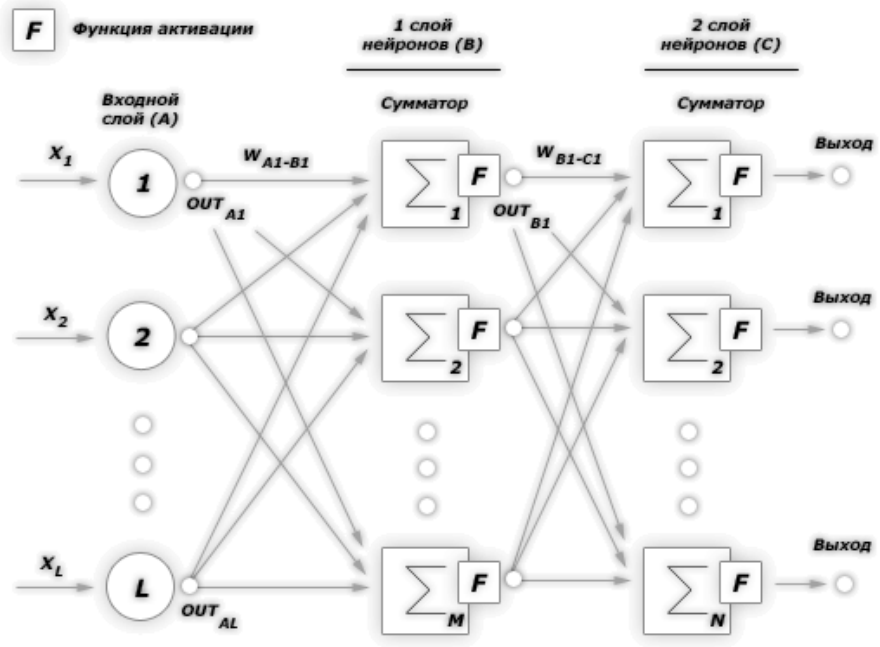


Рисунок 1.8 – Трехслойный персептрон [68]

Согласно теореме Колмогорова (1957 год) любая непрерывная функция n аргументов представима в виде суперпозиции непрерывных функций одного аргумента и суммирования [69].

Теорема. *Любая непрерывная функция* n *аргументов на единичном кубе* $[0,1]^n$ *представима в виде*

$$F(x_1, x_2, \dots, x_n) = \sum_{k=1}^{2n+1} h_k \left(\sum_{i=1}^{n} \varphi_{ik}(x_i) \right), \qquad (1.12)$$

где h_k, φ_{ik} *– непрерывные функции одной переменной, причем* φ_{ik} *не зависят от выбора* F.

В этой теореме и многочисленных ее развитиях [70, 71] представлено утверждение об универсальных аппроксимационных возможностях произвольной нелинейности: с помощью линейных операций и единственного нелинейного элемента φ можно спроектировать систему, вычисляющую любую непрерывную функцию с любой желаемой точностью. При этом указанная система имеет структуру нейронной сети с одним скрытым слоем, поэтому ИНС являются универсальными аппроксиматорами непрерывных функций.

В 1989 году в работе Гиросси и Пожжио [72] отмечено, что теорема Колмогорова требует выбирать нефиксированные функции φ_{ik} для каждого исследуемого случая, в то время как в многослойном персептроне функции фиксированы и параметризованы. Также теорема Колмогорова не содержит условия непрерывной дифференцируемости функций активации φ_{ik}, необходимого для обеспечения обобщающей способности ИНС. Однако в 1992 году в работе Курковой [73] было показано, что, поскольку ИНС является только аппроксимацией, указанные разногласия с теоремой Колмогорова являются не существенными. Куркова переформулировала указанную теорему в терминах последовательности сигмоидальных функций и доказала способность трехслойного персептрона с фиксированными функциями активации (например, сигмоидальными в скрытом слое) аппроксимировать любую непрерывную функцию многих переменных.

К обязательному этапу настройки ИНС относят процедуру обучения – алгоритм для последующих расчетов при максимально возможной точности, допустимой для поставленной задачи, в результате которого устанавливаются значения весов синаптических связей [71]. Обучение возможно «с учителем» , когда известны целевые значения в некоторых или всех точках исходного множества. В этом случае установка параметров ИНС осуществляется путем минимизации ошибки ее выходных значений на этом множестве. Такой вид

обучения позволяет строить модели аппроксимации, регрессии и классификации. Также известно обучение «без учителя», в ходе которого по мере сходства элементов между собой определяют окрестности схожих точек, строя новое представление исходных данных или «карту». К наиболее популярным видам таких ИНС относят самоорганизующееся карты или карты Кохонена [74].

1.2.3 Алгоритмы обучения многослойного персептрона

Одним из алгоритмов обучения с учителем для многослойных персептронов является **алгоритм обратного распространения ошибки** [75, 76]. Обучение по этому алгоритму осуществляется двумя проходами по всем слоям ИНС: прямым и обратным. Во время прямого прохода вектор обучающего множества подается на входной слой ИНС с фиксированными весами и распространяется до выходного слоя. В ходе обратного прохода все веса настраиваются согласно правилу коррекции ошибок: сначала вычисляют сигнал ошибки, равный разности выходного сигнала ИНС и целевого значения, затем этот сигнал, по которому корректируют веса, распространяется в направлении, обратном направлению распространения входного сигнала. Операции прямого и обратного прохода повторяют для каждого вектора обучающего множества.

Пусть выходной сигнал z многослойного персептрона задается следующим образом:

$$z_i = f(a_i), \quad a_i = \sum_j w_{ij}\, h_j + \theta_i\,, \qquad h_j = \begin{cases} x_j \text{ для входного слоя} \\ g(a_j) \text{ для скрытого слоя} \end{cases}, \tag{1.13}$$

где f – линейная функция выходного слоя,

w_{ij} – веса синаптических связей,

$\theta_i = w_{i0}$ – смещение для формирования порога чувствительности функции активации [77],

x_j – входные данные,

g – сигмоидальная функция активации скрытого слоя.

Процедура обучения заключается в минимизации функционала ошибки E в пространстве весов w [78]:

$$E = \frac{1}{2}\sum_p \left(y^{(p)} - t^{(p)}\right)^2 \to \min_w, \tag{1.14}$$

где t – целевые значения выходного сигнала,

$p = 1, \dots, N_{train}$, N_{train} – число обучающих примеров.

При выборе наиболее часто используемой нормы L_2, алгоритм обучения ИНС сводится к методу наименьших квадратов.

Методы первого порядка

В алгоритме обратного распространения ошибки изменение весов Δw_n на шаге обучения n по расширенному варианту дельта-правила [75] задается следующим выражением:

$$\Delta w_n = -\eta \nabla E_n + \alpha \Delta w_{n-1},$$
$$\frac{\partial E}{\partial w_{ij}} = d_i h_j, \quad d_i = \begin{cases} \sum_k d_k w_{ki}\, g'(a_i) \text{ для скрытого слоя} \\ (y_i - t_i) f'(a_i) \text{ для выходного слоя} \end{cases} \tag{1.15}$$

Этот алгоритм отличается от алгоритма наискорейшего спуска тем, что ∇E_n вычисляется на небольшом поднаборе обучающего множества, и наличием параметра момента α, $0 \le \alpha \le 1$. Производительность этого алгоритма зависит от параметров сети, среди которых главную роль играет параметр скорости обучения

η, $0 \leq \eta \leq 1$. Однако выбор параметров основан на эмпирических правилах и на методе «проб и ошибок». Стохастический режим алгоритма обучения (в котором образцы из обучающего множества последовательно представляются на вход сети в случайном порядке) обычно показывает лучшую сходимость, но требует существенно больше числа шагов и может дать менее точный результат [76] по сравнению с пакетным режимом алгоритмом обучения (в котором на вход разом подаются все образцы из обучающего множества).

Алгоритм обучения (1.16) имеет следующие недостатки:

- параметры обучения выбираются эмпирически;
- метод является неустойчивым, когда параметры подобраны не достаточно оптимально;
- метод не обеспечивает точного приближения к минимуму функционала ошибки;
- трудно определить точку остановки обучения.

Метод сопряженных градиентов [79] используется для повышения эффективности градиентных методов:

$$\Delta w_n = \eta(-\nabla E_n + \beta \Delta w_{n-1}) , \tag{1.16}$$

где параметр скорости обучения η выбирается на каждом шаге обучения путем линейного поиска вдоль выбранного направления. Разработаны несколько методов вычисления коэффициента β, наиболее популярными [80, 81] являются

$$\beta = \begin{cases} \nabla E_n(\nabla E_n - \nabla E_{n-1})/\|\nabla E_{n-1}\|^2, \text{по методу Полака} - \text{Рибьеры} \\ \|\nabla E_n\|^2/\|\nabla E_{n-1}\|^2 , \text{по методу Флетчера} - \text{Ривса} \end{cases} \tag{1.17}$$

Ввиду затрат на вычисление β, этот метод требует больше вычислительного времени, чем метод наискорейшего спуска.

Методы второго порядка

Методы второго порядка используются в большинстве задач минимизации в силу их лучшей сходимости. В **методе Ньютона** [82] веса изменяются по формуле

$$\Delta w = -\eta(\nabla^2 E)^{-1}\nabla E\,, \tag{1.18}$$

где матрица Гессе $\nabla^2 E$ задается как

$$\nabla^2 E = \frac{\partial^2 E}{\partial w_{ij}\partial w_{mn}} = d_i\frac{\partial h_j}{\partial w_{mn}} + \frac{\partial d_i}{\partial w_{mn}}h_j\,. \tag{1.19}$$

Данный метод не всегда приводит к положительному результату, поскольку $\nabla^2 E$ может быть знакопеременной и близкой к вырожденной в окрестности минимума.

Метод Ньютона с регуляризацией [83]

$$\begin{gathered}\Delta w = -\eta H^{-1}A^*\nabla E \\ H = A^*A + \alpha\|\nabla E\|^2, A = \nabla^2 E\end{gathered} \tag{1.20}$$

стабильно сходится при больших значениях α (α~0,1 − 1), но не лучше метода наискорейшего спуска. Матрица H плохо обусловлена при регуляционном члене, стремящимся к нулю.

Поскольку использование матрицы Гессе не позволяет достичь приемлемого результата, используют следующий **модифицированный метод Ньютона** [84, 85]:

$$\Delta w = -\eta H^{-1}\nabla E\,, \tag{1.21}$$

где H являются положительно определенной и хорошо обусловленной. Когда $\|(\nabla^2 E)^{-1}\| > A$ в окрестности минимума, где A – максимально допустимое значение, то при достижении минимума $H \nrightarrow \nabla^2 E$.

$$H = \nabla^2 E + \mu I, \tag{1.22}$$

где μ отвечает за что, что собственные значения оператора H больше или равны $\delta > 0$. При выполнении этого условия H^{-1} называют оператором сжатия.

Параметр скорости обучения η можно вычислить также путем линейного поиска вдоль выбранного направления. Данная процедура незначительно увеличивает время вычислений, так как большую часть времени занимает вычисление матрицы Гессе. Эмпирически выявлено, что когда решение близко к минимуму, значение η стабилизируется вблизи 1. В этом случае линейный поиск дает результат за несколько итераций. Также удобно вычислять η по следующему выражению [86]:

$$\eta_{n+1} = \eta_n \frac{\|\nabla E_{n-1}\|}{\|\nabla E_n\|}, \quad 0 < \eta_{n+1} \leq 1 \tag{1.23}$$

где η_0 должен быть существенно мал.

Этот метод имеет следующие преимущества:

- параметры обучения выбираются автоматически;
- метод обеспечивает вычисление минимизации с более высокой точностью, по сравнению с расширенным вариантом дельта-правила;
- в методе необходимо меньше итераций расчетов для получения сходимости.

Главным недостатком всех вариантов метода Ньютона является процессорное время и требуемая память, необходимые на одну итерацию (эпоху) обучения. Поэтому модифицированный метод Ньютона применим в задачах с

небольшими ИНС и если скорость обучения не является решающим фактором.

Поскольку вычисление матрицы Гессе требует больших временных ресурсов, в **квазиньютоновских методах** [82] применяют различные варианты ее аппроксимации. В этих методах обратную матрицу Гессе заменяют рекурсивным вычислением:

$$\Delta w = -\eta H_n \nabla E, \;\; H_n = f(H_{n-1}), \tag{1.24}$$

где параметр скорости обучения η выбирается с помощью линейного поиска, а $H_n = f(H_{n-1})$ вычисляют по ряду методов [87], наиболее распространенными из которых являются методы Дэвида-Флетчера-Пауэлла и Бройдена-Флетчера-Гольдфарба-Шанно.

Но квазиньютоновские методы менее стабильны по сравнению с методом Ньютона и более подвержены переобучению из-за наличия больших флуктуации при процедуре обучения.

Поскольку в данных методах нет необходимости вычислять матрицу Гессе, наибольшая часть времени вычислений уходит на линейный поиск, поэтому полезно использовать более экономичные по времени методы поиска.

Одним из методов с специальным способом аппроксимации матрицы Гессе является **метод Гаусса-Ньютона** [82], где

$$\Delta w = -(J^T J)^{-1} \nabla E \,, \tag{1.25}$$

где $\left[J_{pm}\right] = \partial y^p / \partial w_m$ – матрица Якоби.

Функцию передачи $y(x)$ рассматривают линейной, т.е. при вычислении матрицы Гессе не учитываются вторые производные $y(x)$ по весам. Аппроксимация $H = J^T J$ является положительно определенной матрицей.

Поскольку H плохо обусловлена на первых итерациях, в **методе Левенберга-Марквардта** [88] используют

$$H = J^T J + \mu I \ , \qquad (1.26)$$

где μ – фактор демпинга, который динамически изменяется на каждой эпохе обучения, $\mu \geq 0$, I – единичная матрица. При $\mu \to 0$ этот метод вырождается в метод Гаусса-Ньютона, при $\mu \gg 0$ – в метод наискорейшего спуска. По сравнению с методом наискорейшего спуска метод Гаусса-Ньютона обладает большей точностью и скоростью сходимости в окрестности локального минимума, поэтому задача минимизации функционала ошибки методом Левенберга-Марквардта сводится к переходу к методу Гаусса-Ньютона. По этой причине при уменьшении функционала ошибки на текущей итерации по сравнению с предыдущей параметр μ уменьшают, в случае возрастания ошибки – параметр μ увеличивают. Сходимость этого метода немного хуже, чем в случае метода Гаусса-Ньютона [76, 82, 89]. И в отличие от квазиньютоновских методов, методы Гаусса-Ньютона и Левенберга-Марквардта не дают существенного выигрыша в скорости вычислений.

По совокупности процессорного времени, необходимого для обучения, (малого числа итераций обучения) и получаемой точности метод Левенберга-Марквардта относят к наиболее эффективным методам для обучения ИНС с обратным распространением ошибки [76, 82, 90].

1.2.4 Переобучение и переподгонка данных

В ходе обучения ИНС на вход сети подается обучающая выборка, на основе которой с помощью алгоритма обучения устанавливаются значения весов синаптических связей. При этом необходимо получить ИНС с хорошей обобщающей способностью: корректное отображение входного множества примеров, не использованных в обучающей выборке, на выходное множество. Если ИНС спроектирована с учетом хорошего обобщения, то она будет давать корректное отображение, даже если входное множество примеров будет

отличаться от примеров из обучающей выборки. Но в случае чрезмерно тонкой подгонки в результате обучения ИНС теряет способность к обобщению и только запоминает данные обучения. Такое явление называют переобучением (overtraining) [91]. Также для ИНС существует проблема переподгонки данных (overfitting), когда с увеличением числа нейронов в скрытых слоях ИНС также теряет хорошую обобщающую способность [92, 93].

Для предотвращения переобучения существуют три основных метода [94]:

- уменьшение числа параметров ИНС путем уменьшения количества входных и скрытых нейронов, но такой подход приводит к созданию слишком упрощенных моделей, получаемых на небольших выборках, поэтому для сохранения максимальной информативности входного множества и получения полноценных моделей можно использовать различные методы снижения размерности;
- на основе принципа минимизации структурного риска ввод регуляризационного члена в функцию риска, минимизируемую в ходе обучения ИНС, но недостаток этого способа заключается в необходимости проведения многократного обучения ИНС для разных значений относительного веса регуляризатора;
- остановка процедуры обучения при достижении минимума ошибки на тестовой выборке (рисунок 1.9), такой метод представляет собой одну из форм регуляризации [95] и требует значительно меньше вычислительных ресурсов.

При использовании последнего метода предотвращения переобучения ИНС дополнительно применяют трехвыборочный подход [92, 94]. В этом подходе исходный набор данных делят на три выборки: обучающую, по которой проводят настройку весов; контрольную, по которой оценивают прогнозирующую способность; тестовую, по которой отслеживают переобучение.

Во избежание проблемы переподгонки данных в работе Тетко и соавт. [92] для оценки предсказательной способности ИНС было предложено использовать

параметры модели по методу перекрёстного контроля с исключением по одному. Суть этого метода заключается в исключении одного элемента из исходного набора, оставшуюся часть набора используют для настройки параметров сети, а исключенную точку используют для тестирования. Повторяя такую процедуру для всех точек исходного набора, получают оценку предсказательной способности ИНС.

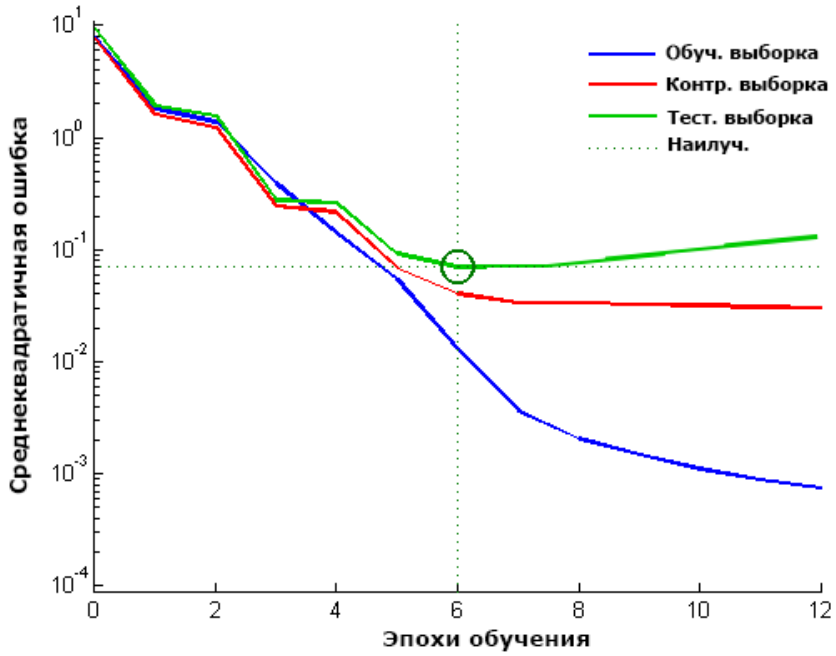


Рисунок 1.9 – Эффект переобучения ИНС

1.3 Методы снижения размерности

Как было указано ранее, одним из подходов предотвращения переобучения ИНС является уменьшения числа ее параметров. Число входных нейронов можно уменьшить посредством снижения размерности входного множества. Задача снижения размерности может быть определена следующим образом.

Пусть есть матрица X размерности $n \times D$, состоящая из n векторов x_i размерности D, и этот набор данных имеет внутреннюю размерность d (где $d < D$, часто $d \ll D$). В терминах геометрии внутренняя размерность означает, что

точки набора данных X лежат на или в окрестности многообразия размерности d, которое вложено в -мерное пространство. Метод снижения размерности преобразует набор данных X в новый набор данных Y размерности d с сохранением геометрической структуры, насколько это возможно. Обычно геометрия вложенного многообразия и внутренняя размерность d набора данных X не известны. Таким образом, задача снижения размерности является некорректно поставленной задачей, которую можно решить только при соблюдении некоторых свойств данных.

Методы снижения размерности можно разделить на три группы [96]: 1) линейные методы; 2) глобально нелинейные методы; 3) локально линейные методы.

1.3.1 Линейные методы

С начала XX века в статистике известны два линейных метода снижения размерности: метод главных компонент (первоначально известный как преобразование Карунена-Лоэва) и линейный дискриминантный анализ (первоначально известный как преобразование Фишера).

Метод главных компонент

Метод главных компонент, Principal Components Analysis (PCA) [97-99], позволяет построить низкоразмерное представление данных, которое описывает значительную часть дисперсии данных без существенных потерь информативности. Такое построение осуществляется нахождением ортогонального базиса меньшей размерности, в котором дисперсия данных максимальна. Векторы главных компонент представляют собой ортонормированный набор d собственных векторов $t_1, \dots, t_d$ ковариационной матрицы, расположенных в порядке убывания ее собственных значений λ: $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d \geq 0$. Таким образом, матрица перехода T к ортонормированному

базису, гиперплоскости проекции, строится из векторов главных компонент $t_1, \dots, t_d$. Низкоразмерное представление y_i данных x_i вычисляется с помощью отображения:

$$Y = (X - \bar{X})T \tag{1.27}$$

Однако в полученной плоскости базис может быть выбран произвольным образом. А в случае наличия одинаковых собственных векторов гиперплоскость поверхности определяется не однозначно.

Метод дискриминантного анализа

Метод дискриминантного анализа, Linear Discriminant Analysis (LDA), [100] позволяет найти линейную комбинацию признаков, которая разделяет два или более класса объектов наилучшим образом. Полученную комбинацию можно использовать как линейный классификатор или для снижения размерности пространства признаков перед последующей классификацией. В отличие от большинства других методов снижения размерности LDA является методом классификации с обучением.

С помощью LDA находится такое линейное преобразование M, которое максимизирует межклассовое и минимизирует внутриклассовое расстояние в пространстве признаков.

1.3.2 Глобально нелинейные методы

Глобально нелинейные методы снижения размерности, подобно PCA и LDA, максимально сохраняют глобальные свойства данных. В этом подразделе представлено описание двух основных глобально нелинейных методов: многомерного шкалирования и изометрического отображения. Также к этой

группе методов относят вложение стохастической близости (Stochastic Proximity Embedding) [101], быструю универсальную развертку по максимуму невязки (Fast Maximum Variance Unfolding) [102], ядерный метод главных компонент (Kernel PCA) [103], ядерный дискриминантный анализ (Generalized Discriminant Analysis) [104], метод диффузионных карт (Diffusion maps) [105, 106], вложение стохастической близости соседей (Stochastic Neighbor Embedding) [107] и многослойные системы автокодирования (multilayer autoencoders) [108].

Многомерное шкалирование

Многомерное шкалирование, Multidimensional scaling (MDS) [109, 110], представляет собой совокупность нелинейных методов, которые преобразуют высокоразмерное представление данных X в низкоразмерное Y с сохранением попарных расстояний между точками. Качество преобразования выражается с помощью функции стресса, которая представляет собой величину ошибки между попарными расстояниями в низкоразмерном и высокоразмерном представлениях данных. В качестве метрики можно использовать метрику евклидова пространства:

$$\begin{aligned} \|x_i - x_j\|^2 &= \sum_{k=1}^{D} \left(x_i^k - x_j^k\right)^2, \\ \|y_i - y_j\|^2 &= \sum_{k=1}^{d} \left(y_i^k - y_j^k\right)^2 \end{aligned} \tag{1.28}$$

Требование о максимально возможном приближении расстояний между объектами в X и Y можно достичь через минимизацию функции стресса, задаваемую в виде:

$$S(Y) = \sum_{1 \le i < j \le n} \left(\|x_i - x_j\| - \|y_i - y_j\| \right)^2 \to \min_X , \qquad (1.29)$$

где $\|x_i - x_j\|$ – евклидово расстояние между высокоразмерными точками x_i и x_j,

$\|y_i - y_j\|$ – евклидово расстояние между низкоразмерными точками y_i и y_j.

В таком варианте задания функции стресса MDS называют метрическим (metric MDS), и он представляет собой линейный вариант снижения размерности.

Также минимизируют функцию стресса, заданную в виде нормированной суммы квадратов отклонений расстояний между новыми точками y_i от расстояний между старыми точками x_i [111]:

$$S(Y) = \frac{1}{\sum_{1 \le i < j \le n} \|x_i - x_j\|} \sum_{1 \le i < j \le n} \frac{\left(\|x_i - x_j\| - \|y_i - y_j\| \right)^2}{\|x_i - x_j\|} \to \to \min_X, \qquad (1.30)$$

которая известна также как функция ошибок Сэммона, а метод снижения размерности – метод отображения Сэммона [112], который относят к неметрическому MDS (nonmetric MDS) [113], где значимыми являются не абсолютные числовые значения оценок попарных расстояний, а только их ранжированный порядок. Функция ошибок Сэммона отличается от предыдущей функции стресса тем, что она делает больший акцент на сохранение попарных расстояний, которые изначально были малы.

Минимизация функции стресса может быть выполнена с помощью различных методов. Так как функция стресса зависит от $d \times D$ переменных, то она имеет большое число локальных минимумов, и ее вычисление является трудоемким. Поэтому во многих алгоритмах многомерного шкалирования в основе лежит итерационное размещение объектов по одному. На каждой итерации оптимизируются координаты только одного из объектов, а координаты

остальных объектов, вычисленные на предыдущих итерациях, остаются фиксированными [110, 111].

Изометрическое отображение

Методы многомерного шкалирования успешно применяют во многих приложениях, но их главный недостаток состоит в том, что в их основе лежит евклидово расстояние, и они не принимают во внимание распределение соседних точек. Если высокоразмерные данные лежат на или в окрестности искривленного многообразия, то по методам многомерного шкалирования две точки могут быть рассмотрены как близкие точки, хотя расстояние между ними по многообразию намного больше, чем евклидово расстояние между ними (рисунок 1.10).

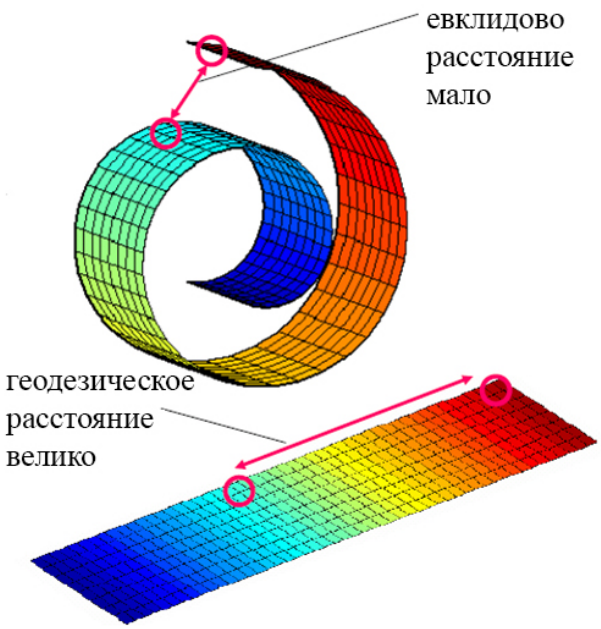


Рисунок 1.10 – Сравнение метрик на искривленном многообразии

Эту проблему решает метод изометрического отображения, Isomap [114], сохраняя попарные геодезические (или криволинейные) расстояния между точками, где геодезическое расстояние представляет собой расстояние между двумя точками, вычисленное по многообразию. В этом методе геодезическое расстояние вычисляется путем построения графа G соседства точек, в котором

каждая точка x_i соединена k ближайшими соседями x_{i_j} из X. Оценка геодезического расстояния между двумя точками осуществляется по величине кратчайшего пути между этими точками, которую можно вычислить с помощью алгоритмов поиска кратчайшего пути Дейкстры [115] или Флойда [116]. Геодезические расстояния между всеми точками из X образуют матрицу попарных геодезических расстояний.

Основной недостаток метода изометрического отображения заключается в его топологической неустойчивости [117]. Isomap может строить неверные связи в графе соседства точек G. Кроме того, качество отображения данных этим методом может страдать при наличии пропусков данных в исходном множестве X, он также может дать неудовлетворительный результат в случае невыпуклого исходного множества [118].

1.3.3 Локально линейные методы

Локально линейные методы снижения размерности базируются исключительно на сохранении свойств в малых окрестностях точек. В этом подразделе представлено описание двух основных глобально нелинейных методов: локально-линейного вложения и карт собственных значений лапласиана. Также к этой группе методов относят локально-линейное вложение с использованием гессиана (Hessian Local Linear Embedding) [119] и метод выравнивания локальных тангенциальных пространств (Local Tangent Space Analysis) [120].

Локально-линейное вложение

Локально-линейное вложение, Local Linear Embedding (LLE) [121], представляет собой локальный метод снижения размерности, который похож на метод изометрического отображения тем, что создает представление графа точек. В отличие от метода Isomap, он нацелен на сохранение только локальных свойств

данных, что делает метод LLE менее чувствительным к замыканиям, чем Isomap. Более того, сохранение локальных свойств позволяет вводить вложение для невыпуклых многообразий. В методе LLE локальные свойства многообразия данных конструируются в виде представления точек как линейной комбинации своих ближайших соседей. В низкоразмерном представлении данных LLE пытается максимально сохранить веса перестраивания в линейных комбинациях.

LLE описывает локальные свойства многообразия вблизи точки x_i путем представления этой точки в виде линейной комбинации W_i (так называемые веса перестраивания) из k ближайших соседей x_{ij}. Таким образом, LLE проводит гиперплоскость через точку x_i и ее ближайших соседей, предполагая, что многообразие является локально линейным. Предположение о локальной линейности следует из того, что веса W_i перестраивания точки x_i инварианты к переносу, повороту и масштабированию. В силу инвариантности к этим преобразованиям любое линейное отображение гиперплоскости в пространство меньшей размерности сохраняет значения весов перестраивания в пространстве меньшей размерности. Другими словами, если низкоразмерное представление данных сохраняет локальную геометрию многообразия, веса W_i перестраивания точки x_i из ее соседей в высокоразмерном представлении также перестраивают точку y_i из ее соседей в низкоразмерном представлении. Как следствие, для нахождения i-размерного представления данных необходимо минимизировать функцию ошибок

$$S(Y) = \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{k} w_{ij} y_{ij} \right)^2 \tag{1.31}$$

Можно показать [121], что координаты низкоразмерного представления y_i, которые минимизируют функцию ошибки, можно найти путем вычисления собственных векторов, соответствующих наименьшим d собственным значениям

$(I - W)^T(I - W)$, где I – единичная матрица размером $n \times n$.

Карты собственных значений лапласиана

Подобно LLE, методом карт собственных значений лапласиана, Laplacian Eigenmaps (LE), находят низкоразмерное представление данных, сохраняя локальные свойства многообразия [122]. В LE локальные свойства базируются на попарных расстояниях между соседями и вычисляют низкоразмерное представление данных, в котором расстояния между точками и их k ближайшими соседями минимальны. Это осуществляется взвешенным образом, т.е. расстояние в низкоразмерном представлении данных между точкой и ее первым ближайшим соседом вносит вклад в функцию ошибок больше, чем расстояние между этой же точкой и вторым ближайшим соседом. Используя спектральную теорию графов минимизация функции ошибок определяется как задача нахождения собственных значений.

В методе карт собственных значений лапласиана сначала конструируют граф G соседства точек, в котором каждая точка x_i соединена с k ближайшими соседями. Для всех точек x_i и x_j в графе G, которые соединены ребром, вычисляют вес w_{ij} ребра, используя гауссову ядерную функцию

$$w_{ij} = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right), \tag{1.32}$$

получая, таким образом, разреженную матрицу смежности W. При вычислении низкоразмерного представления y_i минимизируемая функция ошибок задается следующим образом:

$$S(Y) = \sum_{ij} (y_i - y_j)^2 w_{ij} \tag{1.33}$$

В функции ошибок большие значения весов w_{ij} соответствуют малым расстояниям между точками x_i и x_j. Поэтому разница между y_i и y_j низкоразмерного представления вносит существенный вклад в функцию ошибок. Как следствие, близкие точки в пространстве высокой размерности переносятся близко друг к другу в пространстве низкой размерности. Вычисление матрицы M степеней и дискретного лапласиана L матрицы W весов графа позволяет сформулировать задачу минимизации в виде задачи нахождения собственных значений [123]. Матрица степеней M матрицы W является диагональной, где элементы диагонали равны сумме строки W (т.е. $m_{ii} = \sum_j w_{ij}$). Дискретный лапласиан L, известный также как матрица Лапласа, вычисляется как $L = M - W$. Можно показать [122], что из этого следует

$$S(Y) = \sum_{ij} \left(y_i - y_j\right)^2 w_{ij} = 2Y^T LY \tag{1.34}$$

Поэтому минимизация $S(Y)$ эквивалента минимизации $Y^T LY$.

Таким образом, низкоразмерное представление Y можно найти через решение обобщенной задачи нахождения собственных векторов

$$Lv = \lambda Mv \tag{1.35}$$

для d наименьших ненулевых собственных значений. d собственных векторов v_i, соответствующие наименьшим ненулевым собственным значениями, формируют низкоразмерное представление данных Y.

1.3.4 Расширение вложения для новых точек

Метод расширения вложения для новых точек [2] представляет собой задачу вложения для произвольной точки x_{n+1} и построение вложения h для точек множества

$X \cup x_{new}$, сохраняющего множество $Y = h(X)$.

Для линейных методов снижения размерности расширение вложения для новых точек является не сложным, так как матрица T перехода определяет преобразование данных высокой размерности множества в низкоразмерный эквивалент, которое вычисляется как

$$y_{n+1} = (x_{n+1} - \bar{X})T \tag{1.36}$$

Для рассмотренных нелинейных методов расширение вложения может быть выполнено только с помощью оценочных методов [2]. Рассмотрим общий подход, лежащий в основе этих методов [124]. Первоначально необходимо определить ближайшего соседа x_i к новой точке x_{n+1}. В результате матрица M аффинного преобразования, которая переводит ближайшего соседа x_i в низкоразмерный эквивалент y_i, вычисляется следующим образом

$$M = (y_i - \bar{y}_i)(x_i - \bar{x}_i)^+ \tag{1.37}$$

где $(.)^+$ обозначает псевдообратную матрицу. В силу того, что x_{n+1} близко к x_i, к новой точке можно применить матрицу M перехода. Тогда координаты низкоразмерного эквивалента новой точки y_{n+1}

$$y_{n+1} = \bar{y}_i + M(x_{n+1} - \bar{x}_i) \tag{1.38}$$

Преимуществом такого подхода является его применимость ко всем рассмотренным нелинейным методам.

1.4 Параллельные вычисления с использованием технологии CUDA

Появление многоядерных центральных процессоров (CPU) и многоядерных графических процессоров (GPU) указывает на то, что параллельные системы являются основным направлением для процессорных чипов. Более того, параллелизм этих систем продолжает расти согласно закону Мура [125]. Поэтому на сегодняшний день разработка прикладного программного обеспечения с масштабируемым по количеству процессорных ядер параллелизмом для ускорения общего времени расчетов является актуальной задачей. Одним из вариантов решения этой задачи выступает программно-аппаратный стек параллельной модели программирования CUDA (Compute Unified Device Architecture) [8] с использованием графических карт NVIDIA.

Программно-аппаратный стек CUDA

В основе программно-аппаратного стека CUDA лежат три основные абстракции – иерархия групп потоков, разделяемая (общая) память и барьерная синхронизация. Эти абстракции обеспечивают мелкозернистый параллелизм данных и параллелизм потоков, вложенные в крупнозернистый параллелизм данных и задач. Вследствие этого исходная задача может быть разделена на подзадачи, которые можно выполнять независимо друг от друга на любом количестве процессорных ядер, как показано на рисунке 1.11. Поэтому подзадача будет выполнена быстрее на GPU с большим числом процессорных ядер.

Программная модель CUDA представляет собой расширение языка C, в котором GPU представляет собой вычислительное устройство с отдельной памятью – сопроцессор к CPU. При этом последовательная часть программного кода исполняется на CPU, а параллельная – на GPU. Параллельная часть кода исполняется на большом числе нитей, которые сгруппированы в фиксированные

по размеру блоки. На сетке блоков исполняется вычислительное ядро.

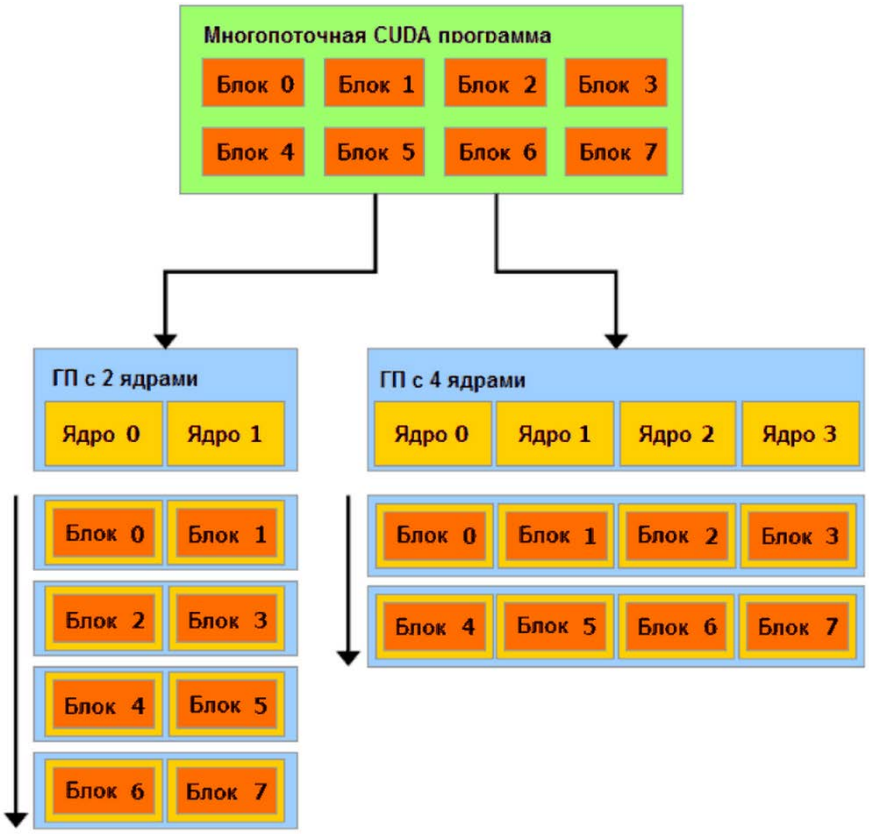


Рисунок 1.11 – Автоматическое масштабирование на GPU [8]

Архитектура NVIDIA Fermi GPU

Графические процессоры NVIDIA Fermi представляют собой масштабируемый массив потоковых мультипроцессоров. Упрощенная архитектура представлена на рисунке 1.12.

Иерархия типов памяти в CUDA

В графической карте память можно разделить на оперативную память DRAM и память, которая физически размещена в потоковых мультипроцессорах GPU. При этом доступные виды памяти обусловлены не только расположением в GPU, но и скоростью работы, а также уровнем доступа на чтение и запись [126]. Иерархия видов памяти представлена на рисунке 1.13.

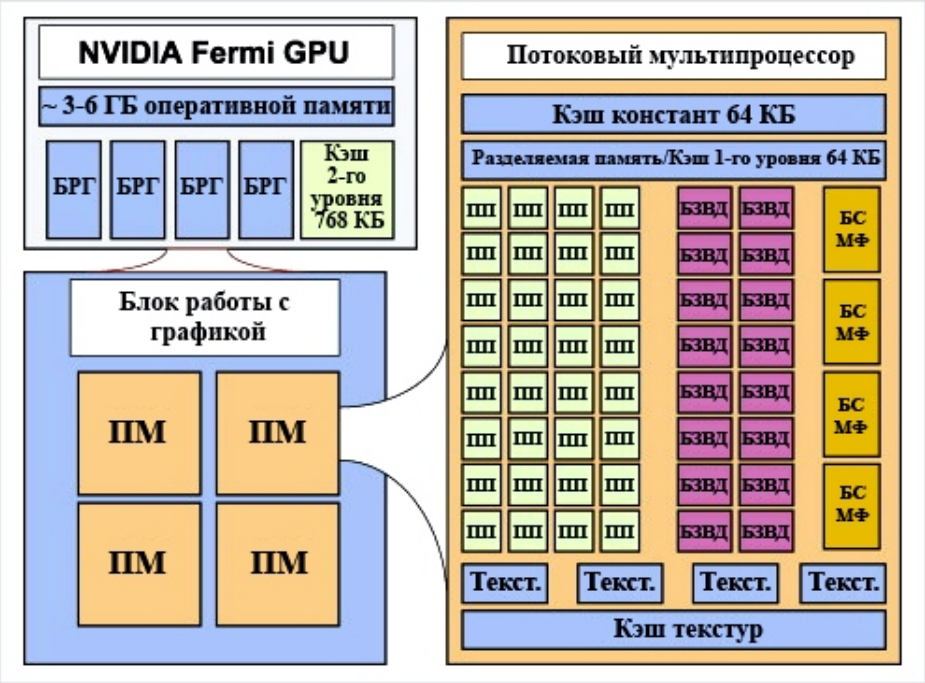


Рисунок 1.12 – Упрощенная архитектура графических процессоров NVIDIA Fermi [127]

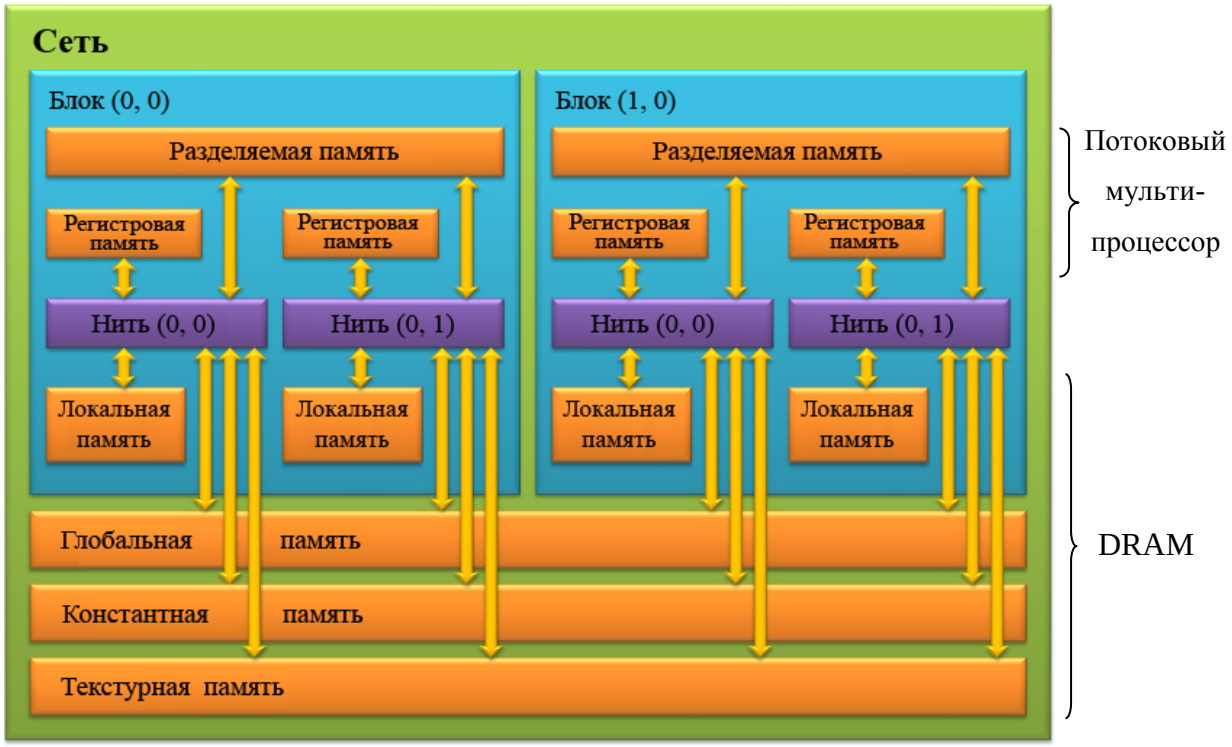


Рисунок 1.13 – Виды памяти графической карты [128]

Глобальная память представляет собой обычную DRAM, которая выделяется с помощью специальных функций на CPU. В нее могут писать все нити сети, но скорость доступа к данным при этом является низкой. Поэтому для повышения производительности число операций с глобальной памятью необходимо сокращать. Этот вид памяти используют для хранения данных большого объема, полученных с CPU по шине PCI-Express.

Разделяемую память относят к быстрому типу памяти. Он расположена в потоковом мультипроцессоре и доступна для всех нитей блока на чтение и запись. Эту память стоит использовать для минимизации обращения к глобальной памяти и хранения локальных переменных.

Константная и текстурная памяти также расположены в DRAM, но они кэшируются, из-за чего обладают высокой скоростью доступа. Запись в эти разделы осуществляется только с CPU, но на чтение они доступны всем нитям сети.

Алгоритм использования модели программирования и эффективная работа с CUDA

Алгоритм использования модели программирования CUDA состоит из следующих этапов:

1. Выделение памяти на графическом процессоре.
2. Копирование данных из памяти хоста CPU в глобальную память устройства GPU.
3. Вызов выполняемых на GPU функций.
4. Копирование полученных данных из глобальной памяти GPU в память хоста CPU.
5. Освобождение глобальной памяти GPU.

Для эффективной работы и получения максимальной производительности выявлен ряд требований и подходов [8, 126, 129, 130]:

1. Минимизация обмена данными между CPU и GPU из-за ограниченной

пропускной способности PCI-Express шины.

2. Минимизация обращения к низкоскоростной глобальной памяти GPU с помощью:
    а. использования высокоскоростной разделяемой и кэшируемой памяти GPU;
    б. объединения запросов к глобальной памяти за счет выравнивания сегментов, используемых для хранения данных.
3. Бесконфликтные обращения нитей в банку распределяемой памяти.

Таким образом, в данной главе представлен обзор существующих вычислительных методов оценки аффинности лиганда к рецептору, отражены главные недостатки этих методов, что указывает на необходимость разработки методов с учетом и минимизацией недостатков известных подходов.

Также рассмотрен широко и успешно применяемый в задачах поиска количественных соотношений «структура-активность» аппарат искусственных нейронных сетей: упрощенная математическая модель биологического нейрона и ее использование в персептроне; представление ИНС как универсальных аппроксиматоров непрерывных функций; алгоритмы обучения с обратным распространением ошибки, а также проблемы переобучения и переподгонки данных и способы их решения. Данный аппарат в дальнейшем был использован для разработки численных методов для оценки аффинности комплексов белок–лиганд.

Одним из подходов предотвращения переобучения ИНС является уменьшение числа ее параметров, которые можно сократить путем снижения размерности входного множества с сохранением максимальной информативности. В данной главе представлены основные линейные и нелинейные методы снижения размерности, а также метод расширения вложения для новых точек, не входящих в исходный набор данных, для которого строилось низкоразмерное отображение.

Далее были рассмотрены основные принципы параллельных вычислений на основе технологии CUDA с использованием графических процессоров NVIDIA,

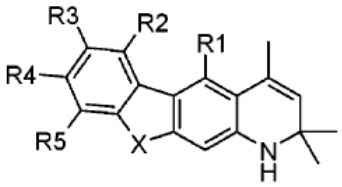
архитектура и особенности использования памяти GPU, алгоритм использования эффективная работа с CUDA. Данная технология была в дальнейшем использована при разработке реализации численного метода оценки аффинности для ускорения вычислений процедур снижения размерности исходного множества и обучения ИНС.

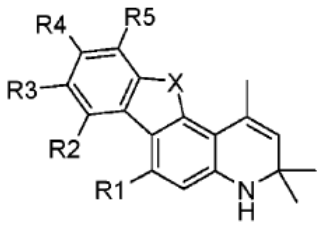
ГЛАВА 2. РАЗРАБОТКА МЕТОДА ОЦЕНКИ АФФИННОСТИ КОМПЛЕКСОВ БЕЛОК-ЛИГАНД

2.1 Объекты исследования

Набор из 63 нестероидных лигандов рецептора прогестерона с известными экспериментальными значениями $pK_i = -lg(K_i)$ был взят из работы [131], набор из 69 нестероидных лигандов рецептора глюкокортикоидов с известными экспериментальными величинами константы ингибирования K_i – из работы [132]. Структурные формулы для этих наборов представлена на рисунке 2.1.

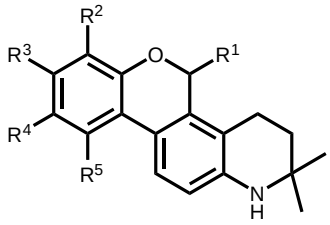
А)





R1, R2, R3	R4	R5	X
H	H	H	CH_2
F	Br	F	O
	Cl	CH_2OH	NH
	F		C≡0
	NO_2		Net
	$COCH_3$		NBu

Б)



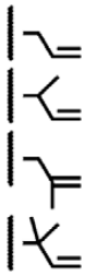
R^1	R^2, R^3, R^4	R^5
	H	

Рисунок 2.1 – Общие базовые структуры рассматриваемых лигандов рецепторов прогестерона (А) и глюкокортикоидов (Б).

Внутриклеточные рецепторы прогестерона и глюкокортикоидов относятся к внутриклеточным рецепторам стероидных гормонов, имеющих общую доменную

структуру (рисунок 2.2) [133]. Рецептор прогестерона в клетке отвечает за рост и развитие женской репродуктивной системы, поддержание беременности, регуляцию центральной нервной системы и иммунной системы, и известные лиганды к этому рецептору используются как лекарства для лечения онкологических заболеваний и для контрацепции. Рецепторы глюкокортикоидов ответственны за регуляцию сердечно-сосудистой системы и всех видов обмена в организме, учувствуют в процессах роста, иммунитета и адаптации к стрессам. А их лиганды используются как лекарства для лечения воспалительных и аутоиммунных заболеваний, например, аллергия, псориаз, системная красная волчанка, ревматоидный артрит, бронхиальная астма.

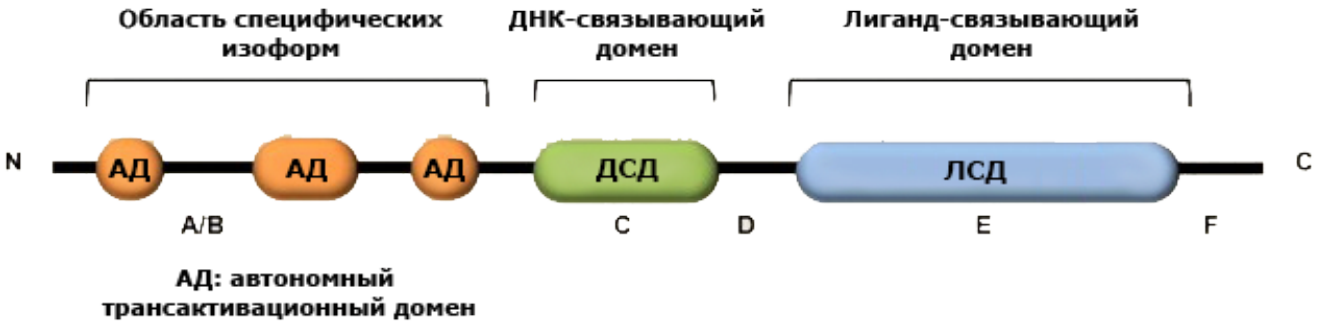


Рисунок 2.2 – Схематическое представление доменной организации ядерных рецепторов стероидных гормонов. Условные обозначения: A/B домен – область автономных трансактивационных доменов; C домен – ДНК-связывающий домен; D домен – шарнирный домен; E/F домен – лиганд-связывающий домен.

Поскольку участком связывания внутриклеточного рецептора с природным лигандом является лиганд-связывающий домен (ЛС-домен), в работе были использованы кристаллические структуры комплексов ЛС-домена рецептора прогестерона с природным лигандом (код в Protein Data Bank – 1A28) [134] и ЛС-домена рецептора глюкокортикоидов с дексаметазоном (код в Protein Data Bank – 1M2Z) [135].

2.2 Молекулярное моделирование

Для оценки аффинности лигандов к рассматриваемым рецепторам всех молекул были построены их комплексы с ЛС-доменом соответствующих рецепторов по следующему алгоритму:

- поиск наиболее успешных «стартовых» комплексов с помощью молекулярного докирования с помощью программный пакета Dock 6.5 [4], в котором оценочная функция Dock энергии взаимодействия комплекса белок-лиганд основана на нековалентных составляющих классических силовых полей молекулярной механики; также данный пакет допускает конформационную подвижность лиганда при сохранении неизменной конформации белка;
- оптимизации структуры комплекса посредством программного пакета молекулярной динамики AMBER 9.0 [6, 46] (поля сил AMBER99 и GAFF):
    - формирование слоя растворителя (вода в явном виде) до границ прямоугольного бокса толщиной не менее 2 Å;
    - минимизация потенциальной энергии системы в периодических граничных условиях – до 500 шагов;
    - нагрев системы от 0 K до 300 K, с этого шага и на всех последующих использовалась процедура моделирования молекулярной динамики. Время симуляции $T_{simulation}$ составляло 10 пс с шагом $\Delta t_{simulation}$ = 2 фс при периодических граничных условиях и заторможенной структуре белка (NTV ансамбль);
    - выравнивание плотности системы при 300 K ($T_{simulation}$ =10 пс, $\Delta t_{simulation}$= 2 фс, NTP ансамбль) при заторможенной структуре белка;
    - уравновешивание системы при 300 K ($T_{simulation}$ =10 пс, $\Delta t_{simulation}$= 2 фс, NTP ансамбль) при периодических граничных условиях. Хотя

$T_{simulation}$ на этом шаге невелико, анализ показывает, что уравновешивание положения лиганда и ближайших боковых радикалов аминокислотных остатков ЛСД рецептора осуществляется за 2-3 пс.

- моделирование молекулярной динамики комплексов методами MM-PBSA/MM-GBSA [47] при 300 K ($T_{simulation}$ =10 пс, $\Delta t_{simulation}$= 2 фс, NTP ансамбль). Последняя молекулярная динамика была взята за основу для вычисления изменения энергии Гиббса комплексов. Значения отдельных составляющих энергии были усреднены по набору из 10 состояний.

2.3 Численный метод оценки аффинности

2.3.1 Входные параметры и выходные значения

Входными параметрами для моделей оценки аффинности «докинг + молекулярная динамика + ИНС» послужили 11 универсальных, не зависящих от химического класса рассматриваемых соединений, молекулярных дескрипторов:

- физико-химические параметры лигандов:
  - молекулярный вес, [г/моль];
  - площадь поверхности, [$Å^2$];
  - площадь полярной поверхности, [$Å^2$];
  - полярный объём, [$Å^3$];
  - общий объём, [$Å^3$];
- усредненные по времени составляющие энергии Гиббса комплексов белок-лиганд [ккал/моль]:
  - изменение энергии электростатического взаимодействия;
  - изменение энергии ван-дер-ваальсовых взаимодействий;
  - вклад гидрофобных взаимодействий в изменение свободной энергии,

рассчитанной по уравнению Пуассона-Больцмана;

- вклад сольватации в изменение свободной энергии, рассчитанной по уравнению Пуассона-Больцмана;
- вклады гидрофобных взаимодействий и сольватации, рассчитанные обобщенным методом Борна.

Для исследуемых рецепторов значения величин pK_i для рассматриваемых комплексов из указанных литературных источников были взяты как выходные целевые значения.

Для решения задачи нелинейной регрессии входных и выходных значений были использованы искусственные нейронные сети.

2.3.2 Предварительная обработка данных

На этапе предварительной обработки данных была проведена стандартизация матрицы данных X размерности $n \times D$, состоящей из n векторов x_i размерности D, во избежание зависимости от выбора единиц измерения используемых дескрипторов:

$$\tilde{x}_i = \frac{x_i - \bar{x}_i}{\sigma_i}, \text{где } \bar{x}_i = \frac{1}{D}\sum_{k=1}^{D} x_{ik},\ {\sigma_i}^2 = \frac{1}{D-1}\sum_{k=1}^{D}(x_{ik} - \bar{x}_i)^2 . \qquad (2.1)$$

В дальнейшем проводилось снижение размерности стандартизованных данных линейным методом – методом главных компонент и четырьмя нелинейными методами – неметрическим многомерным шкалированием, изометрическим отображением, локально-линейным вложением и картами собственных значений лапласиана.

Также для всех пар молекул-лигандов в исходном наборе с помощью пакета SYBYL 8.1 [5] был вычислен коэффициент T_c молекулярного подобия Танимото (Жаккара) [136], который является бинарной мерой сходства, принимает значения

в интервале [0,1] и задается следующей формулой:

$$T_c = \frac{N_{AB}}{N_A + N_B - N_{AB}}, \quad (2.2)$$

где $N_A = n(A)$ – количество структурных фрагментов в молекуле *A*,

$N_B = n(B)$ – количество структурных фрагментов в молекуле *B*,

$N_{AB} = n(A \cap B)$ – количество структурных фрагментов, общих для молекул *A* и *B*.

Количество структурных фрагментов определяется по хэш-функции молекулярных «отпечатков» (fingerprints) химической структуры соединения. Сначала с учетом доноров водородной связи, бензольных колец, наличия конкретного заместителя в определенном положении и других дескрипторов [137, 138] определяются все структурные фрагменты молекулы. Затем вычисляют хеш-функцию (рисунок 2.3), которая представляет собой строку битов (последовательность из «0» и «1»), содержащую информацию о структуре молекулы, где наличие структурного фрагмента кодируется как «1», а его отсутствие – как «0». По набору структурных фрагментов и указанной хеш-функции для каждой молекулы по формуле (2.2) вычисляют коэффициент Танимото T_c для рассматриваемой пары молекул.

Высокая степень молекулярного подобия предполагает близость свойств сравниваемых молекул, и наоборот.

По сумме $\sum_j T_c^j$ было выбрано соединение с наибольшим значением суммы, а значения $\tilde{T}_c^j$ из соответствующей ему строки, в дальнейшем, были использованы при разбиении исходного набора на три выборки: обучающую (70%), контрольную (15%) и тестовую (15%). Разбиение на эти выборки происходило случайным образом, но так, чтобы точки покрывали весь диапазон изменения коэффициента Танимото. Аналогичный подход для разбиения выборки лигандов был применен в [139].

Полученные выборки сжатых данных были использованы для настройки

параметров ИНС с помощью трехвыборочного подхода [92, 94], описанного в первой главе, для предотвращения переобучения нейронной сети.

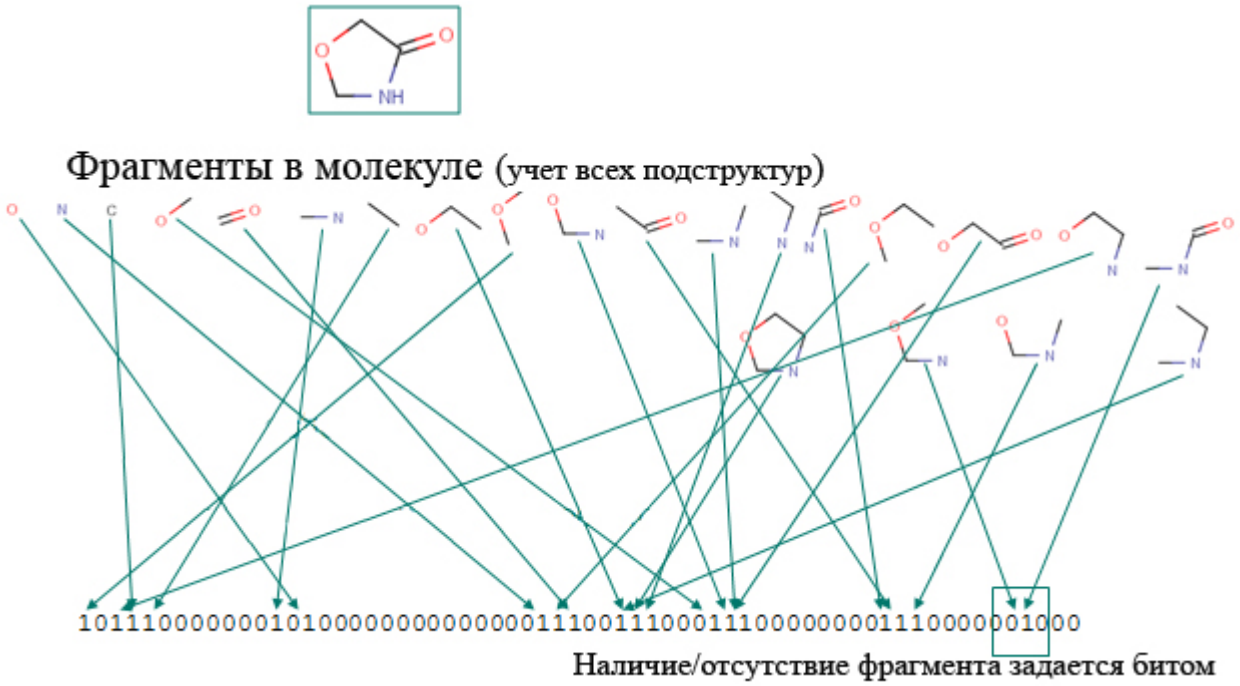


Рисунок 2.3 – Схема вычисления хэш-функции отпечатка химической структуры молекулы.

2.3.3 Структура ИНС и ее оптимизация

В основу структуры нейронной сети лег многослойный персептрон – однонаправленная нейронная сеть с сигмоидальной функцией активации в одном скрытом слое и линейной функцией передачи в выходном слое. Структура сети представлена на рисунке 2.4.

Для настройки весов сети был использован пакетный алгоритм обучения с обратным распространением ошибки и один из наиболее эффективных методов минимизации функционала ошибки (невязки) [76, 82, 90] – метод Левенберга-Марквардта [88]. Входной слой содержал число $N_{(non)linear}$ входных нейронов, соответствующее внутренней размерности d стандартизированных исходных данных, получаемой в результате сжатия данных линейными и нелинейными

методами, и вектор смещений.

Для определения оптимальной архитектуры ИНС варьировались следующие параметры:

- число нейронов в скрытом слое – от $N_{(non)linear} - 3$ до $N_{(non)linear} + 3$ нейронов;
- разбиение выборки на обучающее, контрольное и тестовое множества с учетом подобия молекул – 7 вариантов;
- первоначальные значения весов – 4 варианта.

Скрытый слой также содержал вектор смещений. В ходе настройки сети тестовая выборка была использована для предотвращения переобучения сети. Для этого проводился контроль за изменением среднеквадратичной ошибки на этой выборке. Когда ошибка в течение некоторого числа эпох прекращала уменьшаться, обучение останавливалось, и итоговой ИНС служила та сеть, которая имела минимальную ошибку на тестовой выборке.

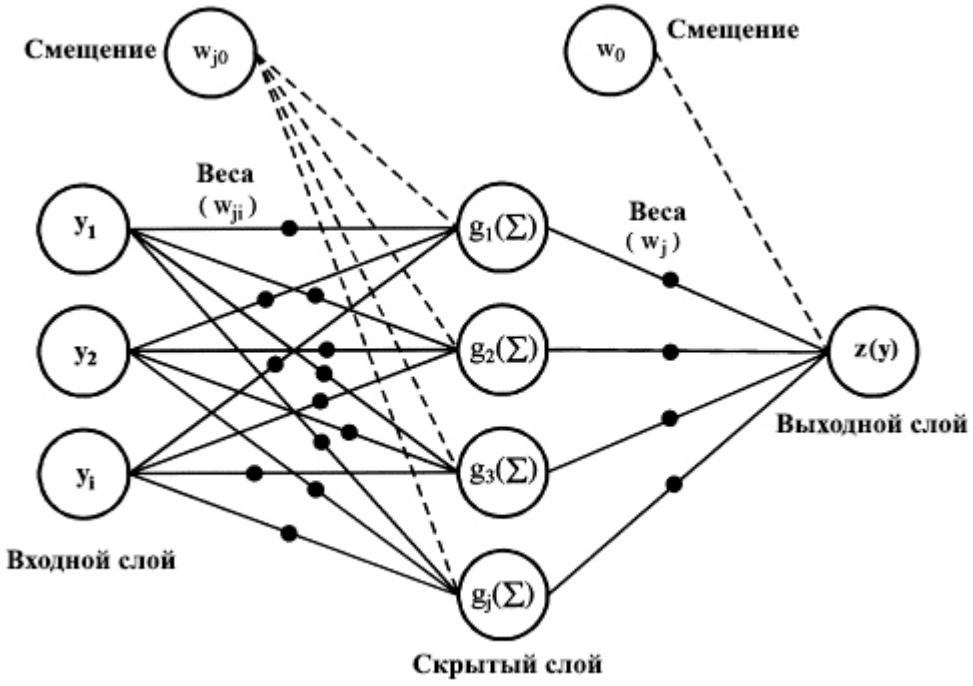


Рисунок 2.4 – Архитектура однонаправленной нейронной сети с сигмоидальной функцией активации $g_j(v) = \tanh(v)$ в скрытом слое и линейной функцией $z(v) = av + b$ передачи в выходном слое.

2.3.4 Параметры оценки моделей

Для оценки модели были выбраны следующие статистические параметры:

1. R^2 – коэффициент детерминации

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}(x_i))^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (2.3)$$

2. Q^2 – коэффициент детерминации предсказания

$$Q^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_{LOO}(x_i))^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (2.4)$$

3. *RMSE* – среднеквадратичная ошибка

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}(x_i))^2}{n}} \quad (2.5)$$

где y_i – экспериментальные значения аффинности,

$\hat{y}(x_i)$ – вычисленные значения по модели оценки аффинности по входным данным x_i,

$\hat{y}_{LOO}(x_i)$ – вычисленные значения по модели оценки аффинности по входным данным x_iпри перекрестном контроле с исключением по одному,

n – количество значений.

2.3.5 Результаты построения моделей с использованием метода главных компонент

По результатам снижения размерности входного множества методом главных компонент удалось размерность $D = 11$ снизить на две единицы, получив внутреннюю размерность $d = 9$. Такой результат объясняется наличием корреляций между вкладами гидрофобных взаимодействий и сольватации в изменение свободной энергии, рассчитанными по уравнению Пуассона-Больцмана (MM-PBSA) и по обобщенной модели Борна (MM-GBSA).

Таким образом, был получен набор линейно-независимых дескрипторов, и число N_{linear} входных нейронов ИНС составило 9. Наилучший вариант по R^2 обучения для обоих рецепторов был достигнут при числе нейронов в скрытом слое $N_{hidden} = 8$. Также, во избежание проблемы переподгонки данных, для каждой модели был проведен перекрестный контроль с исключением по одному (LOO). Результаты представлены в таблице 2.1 и на рисунках 2.5 и 2.6 для рецепторов прогестерона и глюкокортикоидов, соответственно.

Таблица 2.1 – Статистические параметры моделей «докинг + молекулярная динамика + ИНС» при использовании метода главных компонент для снижения размерности входных данных.

Статистические параметры	Рецептор прогестерона	Рецептор глюкокортикоидов
Количество входных параметров	9	9
R^2 для обучающей выборки	0,95	0,96
RMSE для контрольной выборки	0,14	0,09
Q^2 при контроле LOO	0,95	0,93
RMSE при контроле LOO	0,17	0,21

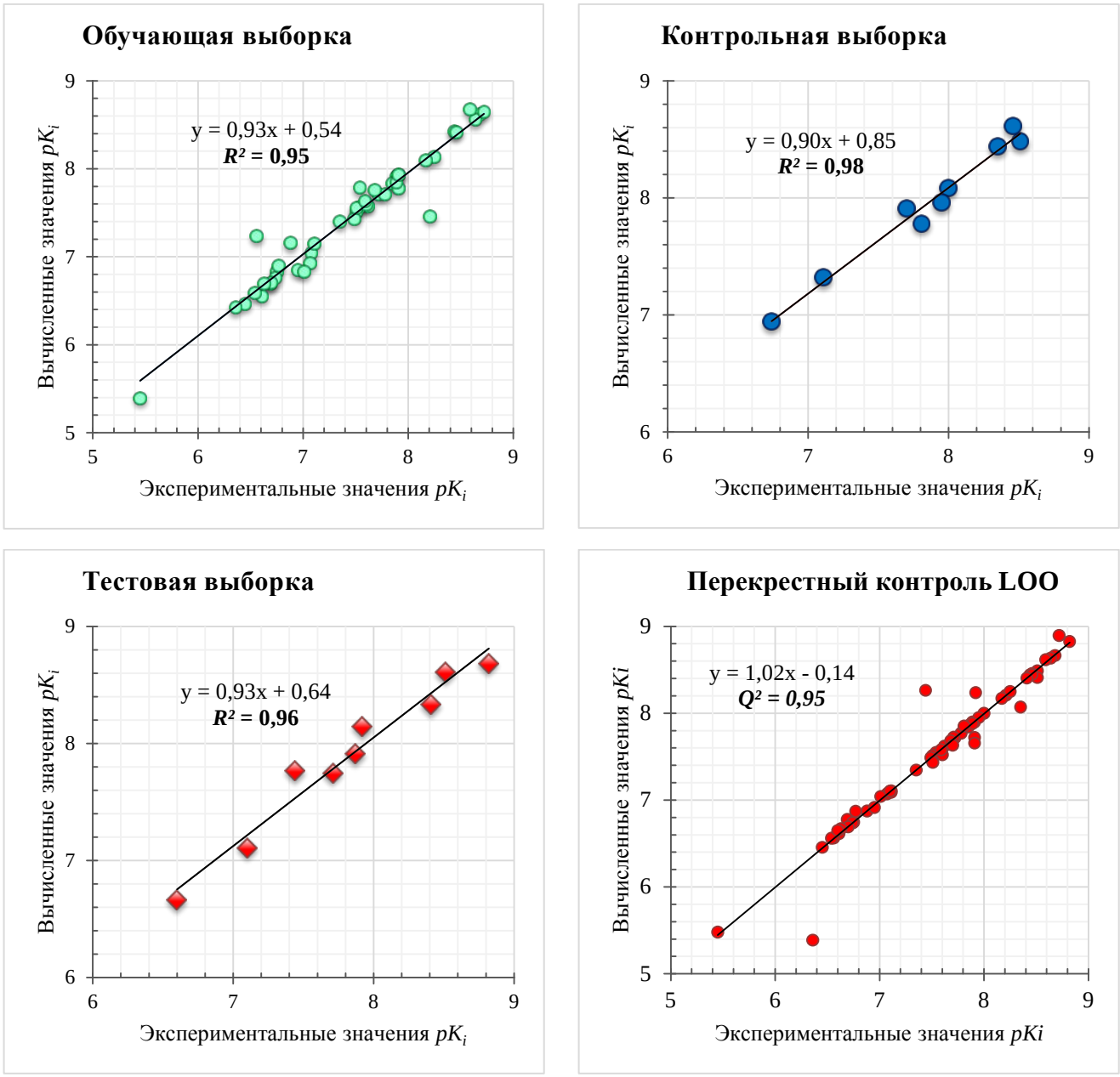


Рисунок 2.5 – Сравнение вычисленных значений pK_i по модели со снижением размерности методом главных компонент и экспериментальных данных pK_i нестероидных лигандов к рецептору прогестерона для обучающей, контрольной и тестовой выборок, а также при перекрестном контроле с исключением по одному.

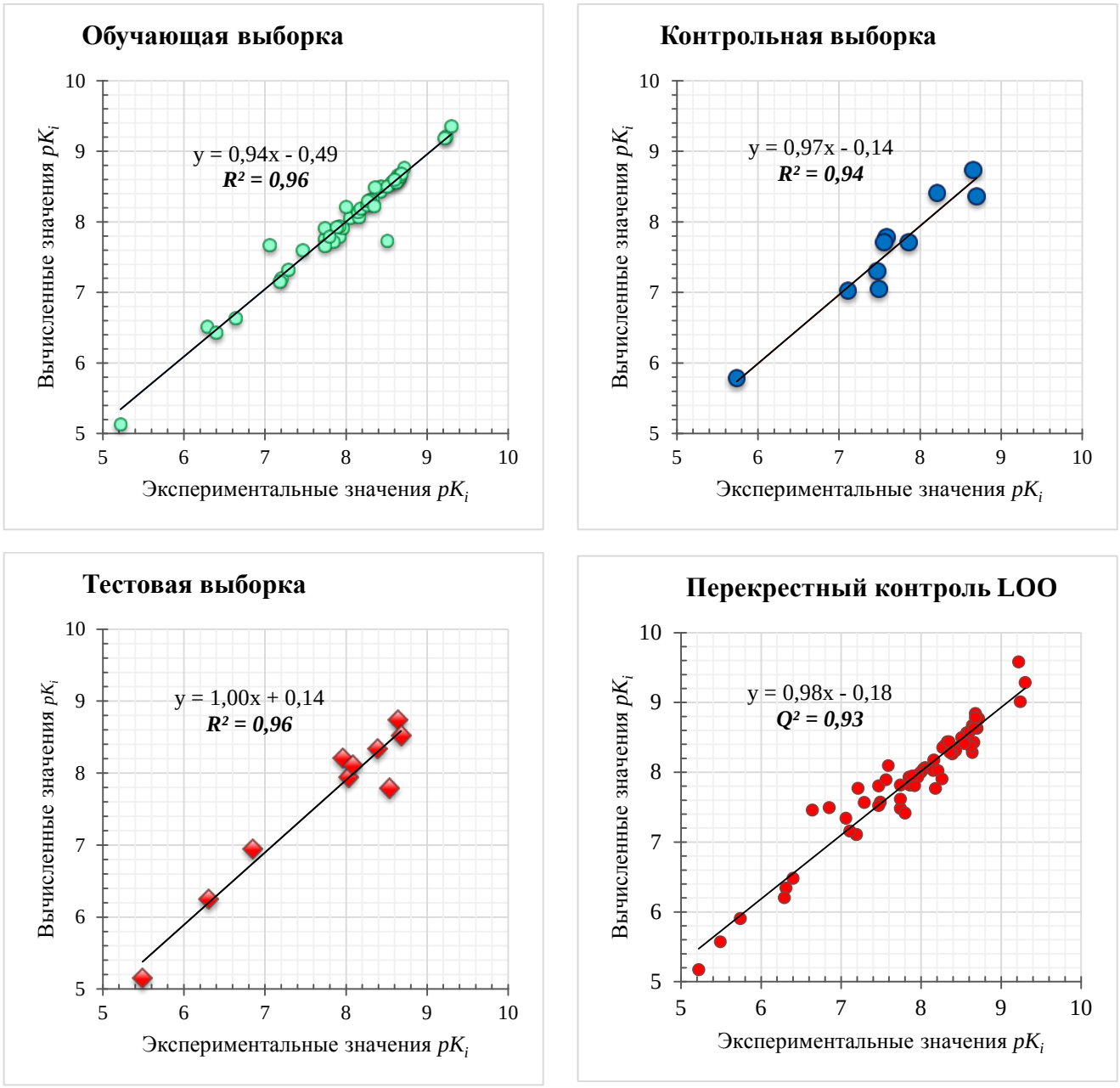


Рисунок 2.6 – Сравнение вычисленных значений pK_i по модели со снижением размерности методом главных компонент и экспериментальных данных pK_i нестероидных лигандов к рецептору глюкокортикоидов для обучающей, контрольной и тестовой выборок, а также при перекрестном контроле с исключением по одному.

Распределения ошибок для обучающей, контрольной и тестовой выборок для рецепторов прогестерона и глюкокортикоидов представлены на рисунке 2.7. Видно, что распределение ошибок удовлетворяет условиям симметричности относительно среднего нулевого значения при минимальном значении среднеквадратичной ошибки.

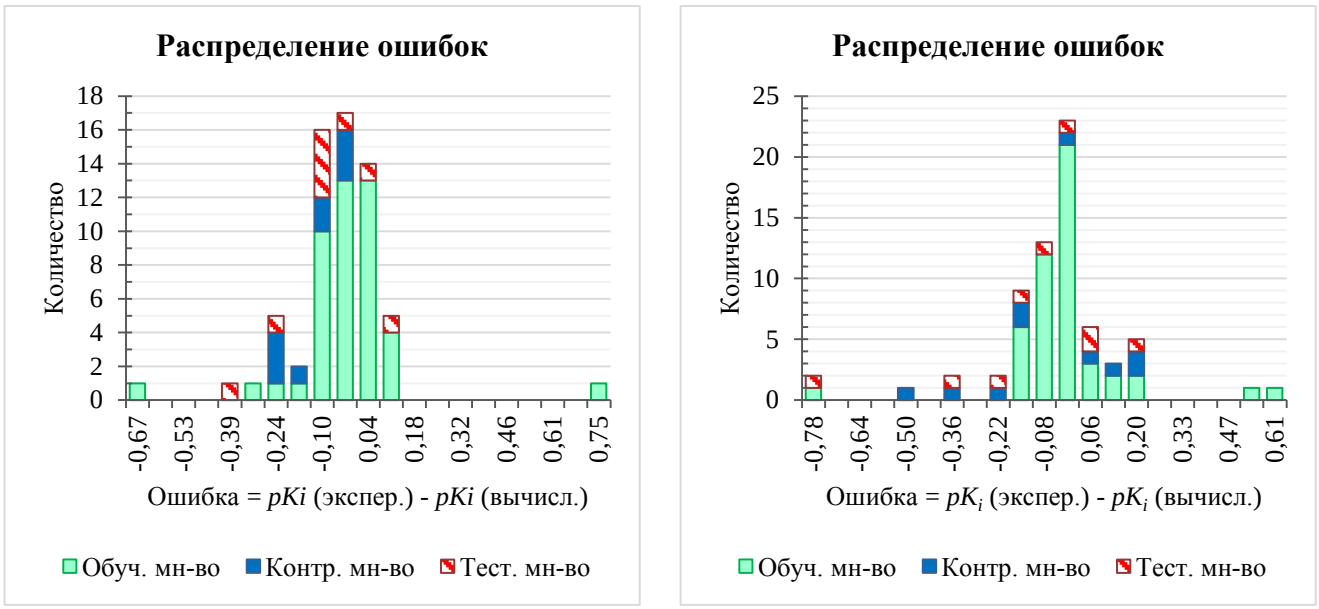


Рисунок 2.7 – Распределение ошибок в процессе обучения ИНС при снижении размерности методом главных компонент входного множества для рецептора прогестерона (слева) и рецептора глюкокортикоидов (справа) на обучающей, контрольной и тестовой выборках.

Дополнительно для обоих рецепторов было выполнено сравнение оценки аффинности по предложенному методу и по оценочным функциям изменения энергии взаимодействия комплексов белок-лиганд в результате молекулярного моделирования (таблица 2.2).

В моделях «Докинг + лин.регр.» была построена линейная регрессия экспериментальных значений и оценочной функции докинга

$$U_{MM} = U_{ele} + U_{vdW}, \qquad (2.6)$$

где U_{ele} – изменение энергии электростатического взаимодействия,

U_{vdW} – изменение энергии ван-дер-ваальсовых взаимодействий.

Таблица 2.2 – R^2 и *RMSE* обучения моделей оценки аффинности по предложенному методу и по оценочным функциям молекулярного моделирования

Модель оценки	Рецептор прогестерона		Рецептор глюкокортикоидов	
	R^2	*RMSE*	R^2	*RMSE*
Докинг + лин.регр.	0,08	0,71	0,03	0,85
Докинг + молек. динамика+ + лин.регр.	0,06	0,69	0,05	0,85
Докинг + молек. динамика + +РСА+ИНС	0,95	0,17	0,93	0,21

В моделях «Докинг + молек. динамика + лин.регр.» была построена линейная регрессия экспериментальных значений и оценочной функции усредненной по времени энергии взаимодействия по методу MM-PBSA/MM-GBSA:

$$\overline{U}_{MM} = \overline{U}_{ele} + \overline{U}_{vdW} + \overline{U}_{SA}, \quad (2.7)$$

где $\overline{U}_{ele}$ – усредненное по времени изменение энергии электростатического взаимодействия,

$\overline{U}_{vdW}$ – усредненное по времени изменение энергии ван-дер-ваальсовых взаимодействий,

$\overline{U}_{SA}$ – усредненный по времени вклад сольватационных взаимодействий.

В моделях «Докинг + молек. динамика+ лин.регр.» была построена регрессионная модель по предлагаемому методу с использованием ИНС на основе усредненных составляющих энергии взаимодействия по методу MM-PBSA/MM-GBSA и физико-химических параметров лиганда.

По результатам сравнения видно, что предлагаемый в работе метод позволяет получить статистически значимые модели для рассматриваемых рецепторов (среднее значение $\overline{R^2} = 0{,}94$) в отличие от моделей по оценочным функциям молекулярного моделирования и линейной регрессии (среднее значение $\overline{R^2} < 0{,}1$).

Также был проведен анализ предсказательной способности предлагаемого метода в зависимости от типа используемых дескрипторов. Для рассматриваемых рецепторов были построены модели только по физико-химическим параметрам лиганда, и только по усредненным составляющим энергии взаимодействия комплекса белок-лиганд. Результаты приведены в таблице 2.3. Полученные модели не достигают точности моделей с учетом обоих типов дескрипторов.

Таблица 2.3 – R^2 и *RMSE* обучения моделей оценки аффинности по предложенному методу на основе групп дескрипторов лигандов и комплекса белок-лиганд

Дескрипторы	Рецептор прогестерона		Рецептор глюкокортикоидов	
	R^2	*RMSE*	R^2	*RMSE*
Физико-химические дескрипторы лиганда	0,59	0,66	0,43	0,91
Составляющие энергии взаимодействия комплекса белок-лиганд	0,69	0,63	0,53	0,69

Таким образом, можно сделать вывод, что предложенный метод позволяет решить основной недостаток оценочных функций методов молекулярного моделирования с использованием линейной регрессии благодаря построению нелинейной зависимости на основе ИНС и неявному учету энтропийной составляющей энергии взаимодействия на основе не только составляющих энергии взаимодействия, но и физико-химических параметров лиганда.

2.3.6 Результаты построения моделей с использованием нелинейных методов снижения размерности

Следующий этап исследования был посвящен применению нелинейных методов снижения размерности для входного множества. Была сформулирована расширенная задача вложения, состоящая из двух пунктов:

- Задача вложения для входного множества: по множеству точек $X_m \subset R^D$ построить отображение во множество точек $Y_m \subset R^d$, лежащее в пространстве меньшей размерности $d < D$ и сохраняющее заданные соотношения между точками множеств.
- Задача вложения для произвольной точки $X_{new} \in X/X_m$.

Для решения задачи вложения для входного множества было проведено сравнение нелинейных методов снижения размерности на основе подходов глобальной (неметрическое многомерное шкалирование – MDS, и изометрическое отображение – Isomap) и локальной (локально-линейное вложение – LLE, и карты собственных значений лапласиана – LE) нелинейности. Для этого на стандартизованном наборе исходных данных поочередно были применены указанные методы с указанием числа нелинейных компонент $N_{nonlinear}$, и для каждого варианта была проведена изложенная выше процедура построения сети с вариацией параметров (число нейронов в скрытом слое – от $N_{nonlinear} - 3$ до $N_{nonlinear} + 3$ нейронов). Результаты представлены в таблицах 2.4 и 2.5.

На основе этого анализа по R^2 обучения было выявлено, что наилучший результат достигается при использовании метода многомерного шкалирования:

- для рецептора прогестерона $N_{nonlinear} = 7,\ N_{hidden} = 6$
- для рецептора глюкокортикоидов $N_{nonlinear} = 6,\ N_{hidden} = 5$

Таким образом, размерность входного множества была снижена до внутренней размерности $d = 7$ и $d = 6$ для рецепторов прогестеронов и глюкокортикоидов, соответственно.

Таблица 2.4 – Коэффициент детерминации R^2 обучения ИНС при использовании нелинейных методов снижения размерности исходных данных для рецептора прогестерона.

$N_{nonlinear}$	5	6	7	8
MDS	0,73	0,91	**0,92**	0,92
Isomap	0,82	0,82	0,82	0,82
LLE	0,82	0,82	0,82	0,82
LE	0,70	0,72	0,72	0,74

Таблица 2.5 – Коэффициент детерминации R^2 обучения ИНС при использовании нелинейных методов снижения размерности исходных данных для рецептора глюкокортикоидов.

$N_{nonlinear}$	5	6	7	8
MDS	0,68	**0,91**	0,79	0,91
Isomap	0,56	0,85	0,67	0,81
LLE	0,67	0,71	0,77	0,82
LE	0,46	0,56	0,43	0,71

Для отобранных моделей была также проведена процедура перекрестного контроля с исключением по одному. Результаты настройки и тестирования сети представлены на рисунках 2.8 и 2.9 для рецепторов прогестерона и глюкокортикоидов, соответственно, и в таблице 2.6. Распределения ошибок для обучающей, контрольной и тестовой выборок для исследуемых рецепторов представлены на рисунке 2.10. Эти распределения также удовлетворяют условиям симметричности относительно среднего нулевого значения при минимальном значении среднеквадратичной ошибки.

Для решения задачи вложения для новых точек был использован метод расширения вложения для новых точек для многомерного шкалирования,

описанный в первой главе.

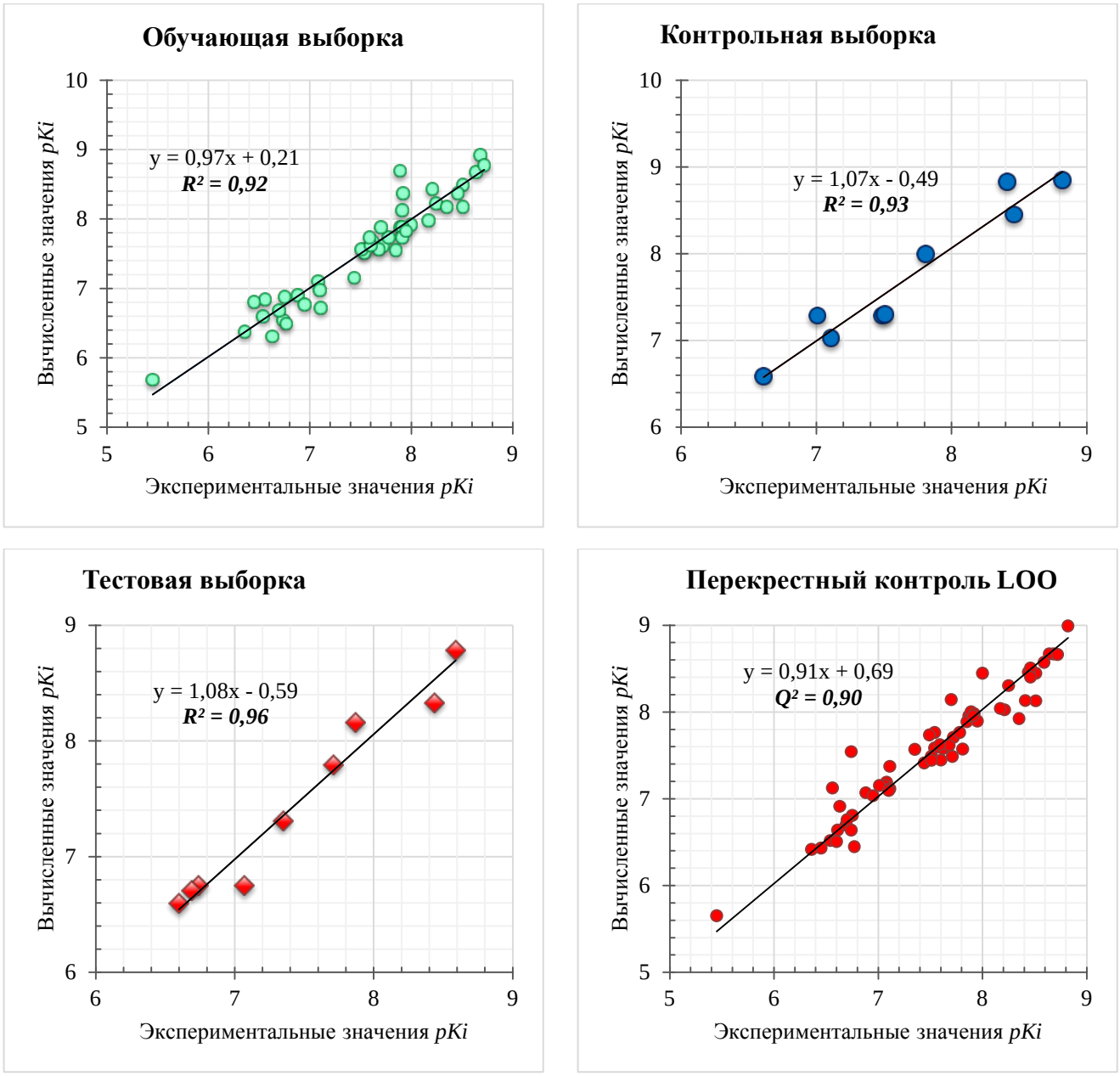


Рисунок 2.8 – Сравнение вычисленных значений pK_i по модели со снижением размерности многомерным шкалированием и экспериментальных данных pK_i нестероидных лигандов к рецептору прогестерона для обучающей, контрольной и тестовой выборок, а также при перекрестном контроле с исключением по одному.

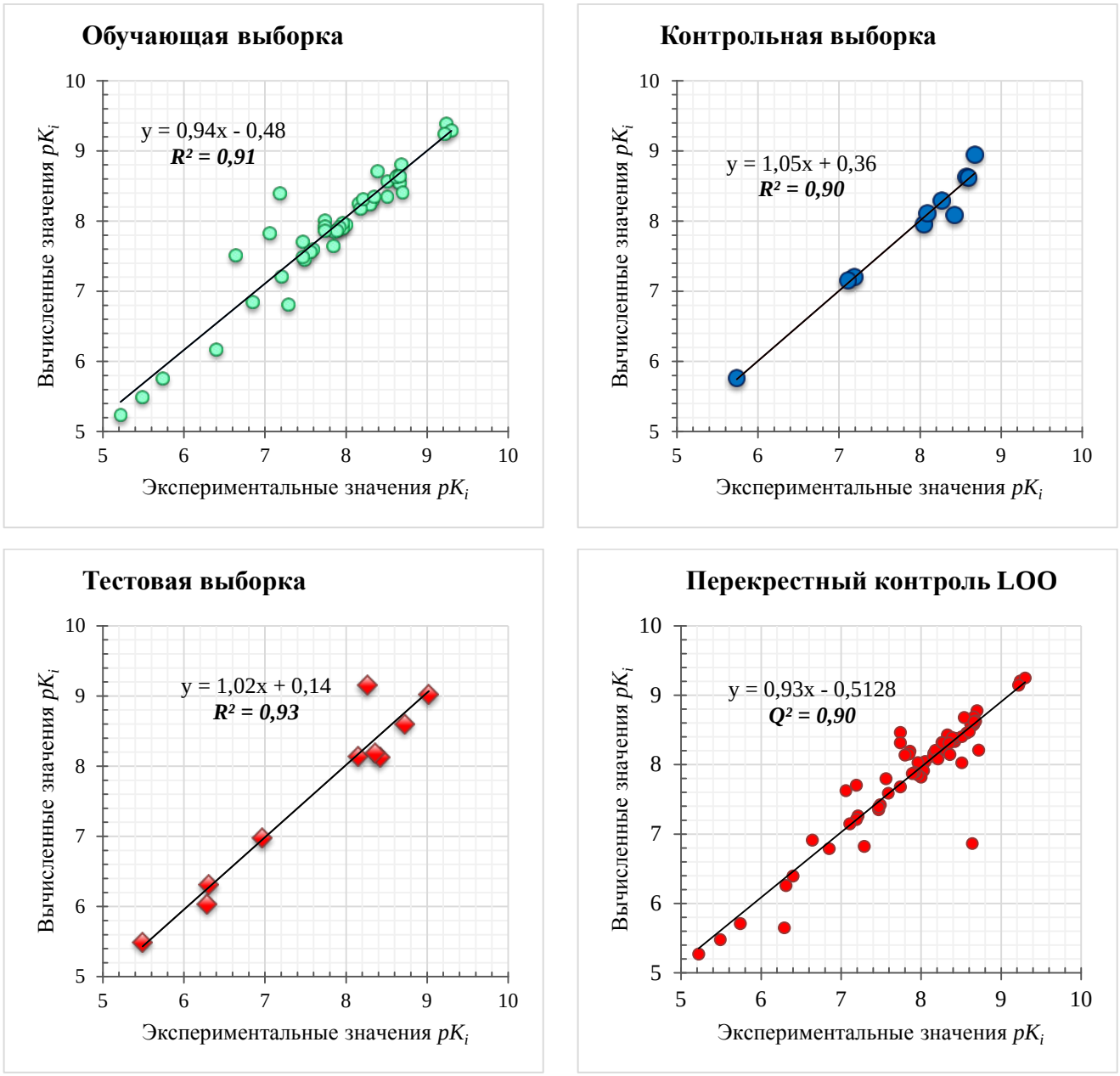


Рисунок 2.9 – Сравнение вычисленных значений pK_i по модели со снижением размерности многомерным шкалированием и экспериментальных данных pK_i нестероидных лигандов к рецептору глюкокортикоидов для обучающей, контрольной и тестовой выборок, а также при перекрестном контроле с исключением по одному.

Таблица 2.6 – Статистические параметры моделей «докинг + молекулярная динамика + ИНС» при использовании метода многомерного шкалирования для снижения размерности входных.

Статистические параметры	Рецептор прогестерона	Рецептор глюкокортикоидов
Количество входных параметров	7	6
R^2 для обучающей выборки	0,92	0,91
RMSE для контрольной выборки	0,19	0,21
Q^2 при контроле LOO	0,90	0,90
RMSE при контроле LOO	0,26	0,25

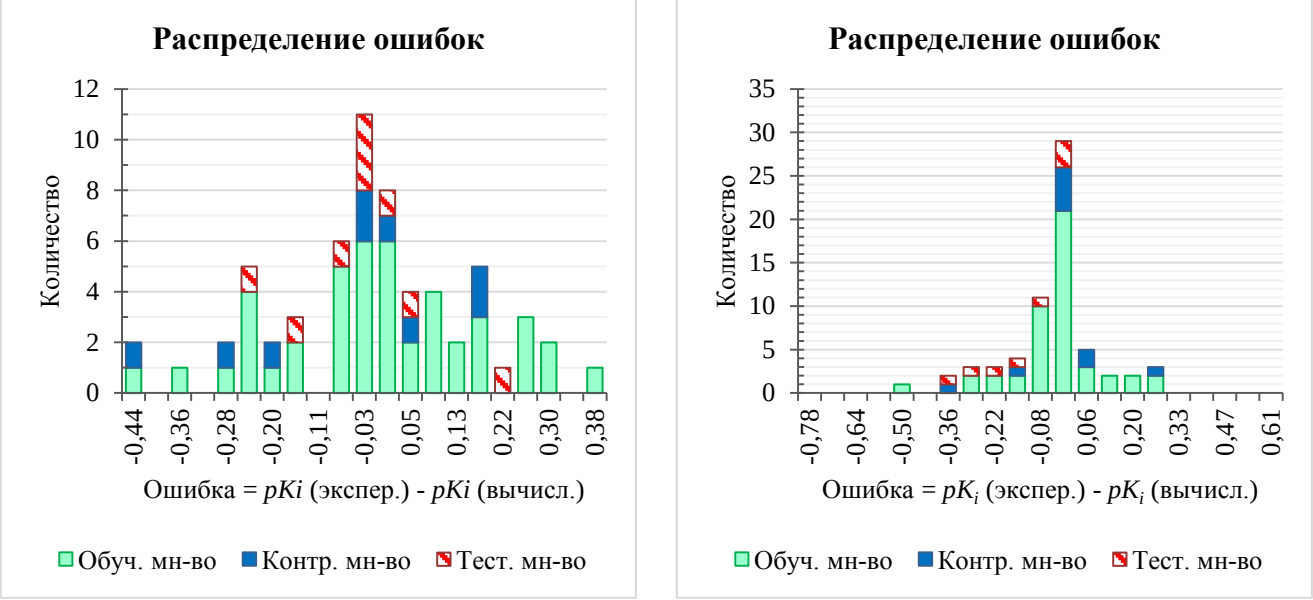


Рисунок 2.10 – Распределение ошибок в процессе обучения ИНС при снижении размерности многомерного шкалирования для рецептора прогестерона (слева) и рецептора глюкокортикоидов (справа) на обучающей, тестовой и контрольной выборках.

Таким образом, разработан комбинированный метод оценки аффинности комплексов белок-лиганд, где в качестве используемых дескрипторов были взяты как физико-химические параметры лигандов, так и составляющие энергии взаимодействия комплексов рецептор-лиганд. На этапе предварительной обработки данных были применены линейный и нелинейные методы снижения размерности. Линейный метод представлял собой метод главных компонент. Среди нелинейных методов были рассмотрены неметрическое многомерное шкалирование, изометрическое отображение, локально-линейное вложение и карты собственных значений лапласиана.

На основе предложенного метода были построены модели для лигандов внутриклеточных рецепторов прогестерона и глюкокортикоидов. По результатам проведения перекрёстного контроля с исключением по одному было установлено, что модели со сжатием данных методом главных компонент имеют высокую точность предсказаний ($\overline{Q^2} = 0{,}94$) при внутренней размерности исходных данных $d = 9$ для обоих рецепторов. При снижении размерности входного множества нелинейными методами, наилучший результат был достигнут при использовании метода неметрического многомерного шкалирования. Его применение приводит к большему сжатию данных ($d = 7$ для рецептора прогестерона и $d = 6$ для рецептора глюкокортикоидов), но построенные модели немного уступает по точности моделям с использованием метода главных компонент – $\overline{Q^2} = 0{,}90$.

Все модели построены с учетом хорошего обобщения ИНС, вследствие чего в моделях не выявлено проблемы переобучения. Отмечено хорошее совпадение рассчитанных по моделям значений pK_i с экспериментальными данными pK_i рассматриваемых комплексов. Дополнительно для обоих рецепторов было проведено сравнение оценки аффинности по предложенному методу (при сжатии данных методом главных компонент) и по оценочным функциям изменения энергии взаимодействия комплексов белок-лиганд в результате молекулярного моделирования. Предложенный в работе метод позволяет получить статистически

значимые модели для рассматриваемых рецепторов (среднее значение $\overline{R^2} = 0{,}94$) в отличие от моделей по оценочным функциям молекулярного моделирования и линейной регрессии (среднее значение $\overline{R^2} < 0{,}1$).

Разработанный метод позволяет решить основной недостаток оценочных функций методов молекулярного моделирования с использованием линейной регрессии благодаря построению нелинейной зависимости на основе ИНС и неявному учету энтропийной составляющей энергии взаимодействия на основе не только составляющих энергии взаимодействия, но и физико-химических параметров лиганда.

Результаты по разработанным моделям оценки аффинности нестероидных соединений к внутриклеточным рецепторам прогестерона и глюкокортикоидов с использованием метода главных компонент опубликованы в [140] и [141], соответственно.

ГЛАВА 3. ТЕСТИРОВАНИЕ РАЗРАБОТАННОГО МЕТОДА ОЦЕНКИ АФФИННОСТИ

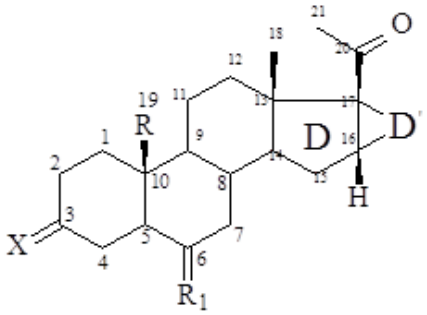
Дополнительно были проведены экспериментальная проверка и сравнительный анализ оценки аффинности по разработанному численному методу и основным методам 3D QSAR.

3.1 Объекты исследования

Для валидации метода оценки аффинности были рассмотрены следующие объекты:

- кристаллическая структура комплекса ЛС-домена рецептора прогестерона с его природным лигандом (код в Protein Data Bank – 1A28) [134], которая была использована при проведении процедуры докинга и продуктивной молекулярной динамики;
- 42 высокоселективных аналога природного лиганда рецептора прогестерона, сочлененные в 16α,17α-положениях с трех-шестичленными карбоциклами (прегна-D´-пентараны), структурная формула которых представлена на рисунке 3.1; синтез и биологическое тестирование этих соединений было выполнено ранее в лаборатории химии стероидных соединений Института органической химии РАН [142-145] и значения относительной конкурентной активности (ОКА) этих соединений к рецепторам прогестерона для крысы и кролика, имеющих высокую гомологию ЛС-домена с рассматриваемым рецептором человека [146] были любезно предоставлены ведущим сотрудником этой лаборатории д.х.н. Левиной И.С.; этот набор был использован для настройки модели, в качестве целевых значений для нейронной сети был взят десятичный логарифм значений ОКА;
- 8 дополнительных прегна-D´-пентаранов, для которых не было данных об

экспериментальной оценке их аффинности к рецептору прогестерона; для этой тестовой выборки были рассчитаны оценки аффинности по разработанной модели и методам 3D QSAR с CoMFA и CoMSIA.



D′ (16α,17α)	**X**	**R**	**R_1**
-CH_2-	O	H	H_2
-$(CH_2)_3$-	H, βOH	CHO	H, αMe
-$(CH_2)_4$-		$CH=CH_2$	Me
-CH_2-CH=CH-		CH_2CH_3	$NO(CH_2)_3CO_2Me$
-CH_2-CH=CH-CH_2-		CH_2OH	(*E*) $NO(CH_2)_3CO_2Me$
-CH_2-CH_2-CH=CH-		CH_2-O-C(6)	(Z) $NO(CH_2)_3CO_2Me$
-CH_2-CH_2-CO-CH_2-		Me	
-CHMe- $(CH_2)_3$-		CH=NOMe	
-CH_2-CMe=CMe-CH_2-		$CH=NO(CH_2)_3CO_2Me$	
-CHMe-CH=CH-CH_2			

Рисунок 3.1 – Структурная формула производных 16α,17α-циклоалканопрогестерона (прегна-D´-пентаранов)

3.2 Модели оценки аффинности

Оценка аффинности рассматриваемых объектов была выполнена с помощью методов 3D QSAR по методам описания взаимодействий CoMFA и CoMSIA, а также на основе физико-химических параметров лигандов и параметров, полученных в результате молекулярного моделирования комплексов рецептор-лиганд, где в качестве оценочной функции была использована линейно-регрессионная модель и модель «докинг + молекулярная динамика + ИНС», изложенная в данной работе.

3.2.1 3D QSAR модели

С помощью модуля SYBYL 8.1 / GALAHAD [147] (с параметрами по умолчанию) было выполнено пространственное выравнивание молекул для 3D QSAR моделей. В итоге было выбрано 20 наилучших вариантов выравниваний, для каждого из которых на основе полей CoMFA и CoMSIA и метода частичных наименьших квадратов были рассчитаны корреляционные уравнения оценки аффинности. В качестве конечного результата использовалась модель с лучшим значением по Q^2.

Для методов CoMFA и CoMSIA указано по два вида моделей, рассчитанных для полной выборки и выборки с учётом наличия «выбросов» – соединений, не попадающих в доверительный интервал. Для этих методов наличие «выбросов» является характерной чертой, поскольку оценка качества построенных моделей осуществляется с помощью перекрестного контроля с исключением по одному. В этом случае наличие в выборке лигандов с боковыми уникальными радикалами или неудачно выравненных молекул может приводить к существенным флуктуациям, поэтому такие соединения часто исключают из вычислений.

3.2.2 Модели на основе молекулярного моделирования

Моделирование комплексов было выполнено путем докирования (программный пакет Dock 6.5 [4]) молекул потенциальных лигандов к участку связывания прогестерона ЛС-домена рецептора прогестерона (код в Protein Data Bank – 1A28) [134, 148]. Далее, до 3-х вариантов виртуальных комплексов, отобранных по оценочной функции программы Dock [4], были оптимизированы посредством программного пакета молекулярной динамики AMBER 9.0 [6, 46] (поля сил AMBER99 и GAFF) по следующему алгоритму:

- формирование слоя растворителя (вода в явном виде) в пределах границ прямоугольного бокса толщиной не менее 4 Å;
- минимизация потенциальной энергии системы в периодических граничных условиях – до 1000 шагов;
- нагрев системы от 0 K до 300 K, с этого шага и на всех последующих использовалась процедура моделирования молекулярной динамики. Время моделирования $T_{simulation}$ составляло 10 пс с шагом $\Delta t_{simulation}$ = 2 фс при периодических граничных условиях и заторможенной структуре белка (NTV ансамбль);
- выравнивание плотности системы при 300 K ($T_{simulation}$ =10 пс, $\Delta t_{simulation}$= 2 фс, NTP ансамбль) при заторможенной структуре белка;
- уравновешивание системы при 300 K ($T_{simulation}$ =10 пс, $\Delta t_{simulation}$= 2 фс, NTP ансамбль) при периодических граничных условиях;
- моделирование молекулярной динамики комплексов методами MM-PBSA/MM-GBSA [47] при 300 K ($T_{simulation}$ =10 пс, $\Delta t_{simulation}$= 2 фс, NTP ансамбль) при периодических граничных условиях.

Усреднение отдельных составляющих энергии взаимодействия комплексов белок-лиганд было проведено по 10 наблюдениям, рассчитанным через одинаковые интервалы времени. В результаты были рассчитаны следующие усредненные составляющие энергии Гиббса [ккал/моль]: изменение энергии электростатического взаимодействия; изменение энергии ван-дер-ваальсовых взаимодействий; гидрофобный вклад в изменение свободной энергии, рассчитанной методом MM-PBSA; вклад сольватации в изменение свободной энергии, рассчитанной методом MM-PBSA; аналогичные двум последним параметрам величины, рассчитанные методом MM-GBSA. Также с помощью модуля AMBER 9 / NMODE [6] были вычислены вклады в энергию Гиббса гармонических колебаний ковалентных связей, углов между смежными ковалентными связями и торсионных взаимодействий. Итоговый вариант комплекса белок-лиганд отбирался по минимальному значению изменения энергии.

Для последующего построения моделей оценки аффинности для каждого лиганда также были вычислены его собственные физико-химические параметры: молекулярный вес, [г/моль]; площадь поверхности, [Å^2]; площадь полярной поверхности, [Å^2]; полярный и общий объёмы [Å^3] молекулы.

Для моделей на основе методов молекулярного моделирования «Докинг + Молекулярная динамика» исходная выборка состояла из 41 соединения, поскольку одно соединение было исключено (в процедуре докинга для него не нашлось решения [149]). На предварительном этапе снижения размерности был использован метод главных компонент, так как было показано, что он дает более высокую точность предсказаний. В целях предотвращения переобучения ИНС было осуществлено увеличение исходной выборки в 4 раза посредством генерации новых точек в пределах погрешности используемых дескрипторов.

3.2.3 Результаты построения моделей

Статистические параметры моделей представлены в таблице 3.1.

Таблица 3.1 – Статистические параметры настройки моделей для прегна-D´-пентаранов.

Статистические параметры	Полная выборка		-10% «выбросов»		Докинг+Молек. динамика+	
	CoMSIA	CoMFA	CoMSIA	CoMFA	+ лин. регр.	+PCA+ ИНС
Число соединений	42	42	37	37	41	41
R^2 обучения	0,82	0,92	0,91	0,93	0,60	0,98
Q^2 при контроле LOO	0,37	0,38	0,59	0,57	0,41	0,91
RMSE при контроле LOO	1,28	1,30	1,08	1,12	1,13	0,29

Из всех построенных моделей вариант «Докинг+Молек. динамика+ РСА+ИНС» показывает наилучший результат ($Q^2 = 0{,}91$), а среднеквадратичная ошибка при LOO в 3,7 раза меньше по сравнению с лучшей моделью 3D QSAR с COMSIA.

3.3 Тестирование моделей оценки аффинности

Для независимой проверки предсказательной способности построенных моделей была использована тестовая выборка из 8 описанных выше пентаранов. Одновременно с вычислительным экспериментом был произведен синтез этих соединений в лаборатории химии стероидных соединений Института органической химии имени Н.Д. Зелинского РАН и их исследование *in vitro* в Московском государственном университете имени М.В. Ломоносова [150-152].

Корреляционные зависимости экспериментальных и вычисленных значений по рассматриваемым методам представлены на рисунке 3.2.

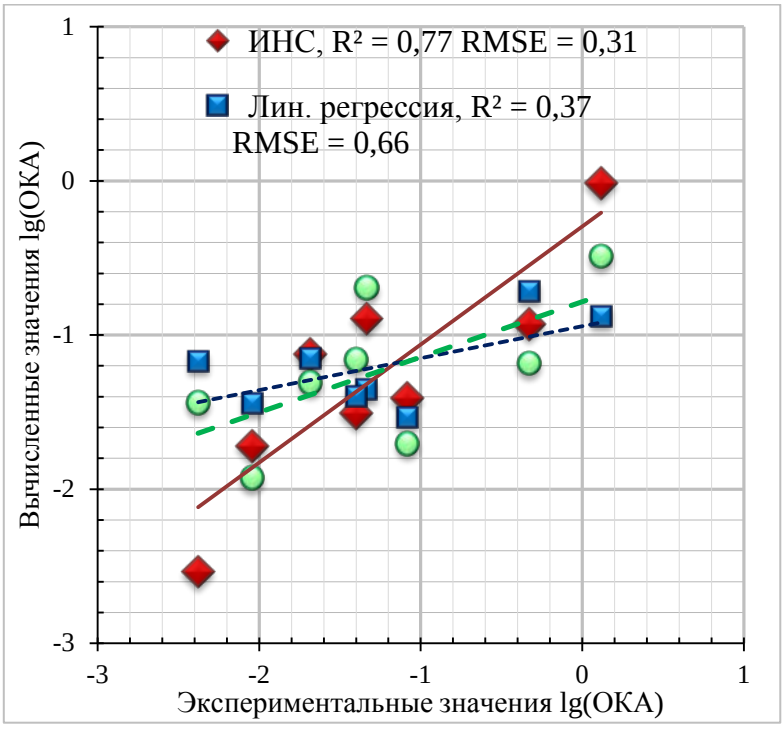


Рисунок 3.2 – Сравнение экспериментально измеренных и вычисленных значений аффинности lg(ОКА) для тестовой выборки молекул прегна-D´-пентаранов.

Метод оценки «Докинг+Молек. динамика+ РСА+ИНС» дает $R^2{}_{test} = 0{,}77$, а для 3D QSAR с COMSIA и «Докинг+Молек. динамика+ лин. регр.» эта величина составляет всего 0,39 и 0,37, соответственно. Но при этом по последним двум моделям можно отличить лиганды с высоким сродством от лигандов с низким сродством. Модель с использование ИНС, в свою очередь, показывает более лучший результат ($R^2{}_{test} = 0{,}77$), что позволяет использовать ее для ранжирования лигандов между собой по сродству к рассматриваемому рецептору.

Таким образом, выполнены сравнительный анализ предсказательной способности предложенного метода и основных методов 3D QSAR на примере лигандов прегна-D´-пентаранов к внутриклеточному рецептору прогестерона и сопоставление полученных оценок с экспериментальными значениями, полученных в результате синтеза и исследования *in vitro* этих соединений [150-152].

Наилучшее соответствие предсказанных значений с экспериментальными данными было выявлено для модели по предложенному в диссертации методу. Эту модель можно применять для ранжирования лигандов по величине связывания в ряду исследуемых соединений. Модели же по основным методам 3D QSAR способны только отличить лиганды с высоким сродством от лигандов с низким сродством.

Полученные результаты опубликованы в [149].

ГЛАВА 4. ПРОГРАММНАЯ РЕАЛИЗАЦИЯ РАЗРАБОТАННОГО МЕТОДА

4.1 Итоговая модель

Первоначально различные модификации были реализованы с помощью математического пакета MATLAB R2012b (встроенное приложение с реализацией ИНС – Neural Network Toolbox [7], внешнее приложение методов снижения размерности - Matlab Toolbox for Dimensionality Reduction [153, 154]). По итогам исследования была установлена типовая модель с конечной структурно-функциональной схемой, для которой в дальнейшем была разработана параллельная реализация с использованием графических процессоров NVIDIA.

Итоговая модель оценки аффинности комплексов белок–лиганд включает в себя следующие компоненты:

1. Стандартизация данных.
2. Два метода снижения размерности:
   2.1. Линейный – метод главных компонент;
   2.2. Нелинейный – многомерное шкалирование.
3. Однонаправленная нейронная сеть с сигмоидальной функцией активации в одном скрытом слое и линейной функцией передачи в выходном слое:
   3.1. Разбиение исходного набора на обучающую (70%), контрольную (15%) и тестовую (15%) выборки с учетом коэффициента молекулярного подобия Танимото.
   3.2. Обучение с минимизацией невязки по методу Левенберга-Марквардта;
   3.3. Вложение произвольного нового лиганда, не входящего в исходный набор данных и оценка его аффинности к рассматриваемому рецептору.

4.2. Параллельная реализация

На этапе разработки параллельной реализации модели была создана последовательная реализация на языке C++, а параллельная форма алгоритма итоговой модели – с использованием CUDA C. Для этого в пунктах 2 и 3.2 функционала модели были выделены независимые операции, по которым вычисления можно проводить по отдельности – в параллельном режиме. Так как параллельный алгоритм был реализован на технологии NVIDIA CUDA, то на этапе построения алгоритма были учтены архитектурные особенности графических карт – SIMT (single-instruction, multiple-thread), пропускная способность передачи данных между центральным процессором и графической картой, объем и скорость чтения/записи различных типов памяти графической карты. В параллельной программной реализации были использованы следующие библиотеки:

- cuBLAS (NVIDIA CUDA Basic Linear Algebra Subroutines) [155], так как большая часть вычислений легко реализуется матричными операциями;
- CULA dense (библиотека линейной алгебры LAPACK, оптимизированная под CUDA) [156] для распараллеливания метода главных компонент;
- cuRAND (NVIDIA CUDA Random Number Generation library) [157] для генерации матриц случайных величин при инициализации весов ИНС на этапе ее обучения;
- HiT-MDS (High-Throughput Multidimensional Scaling) [158] в качестве основы для параллельной реализации многомерного шкалирования;
- LevMar (Levenberg-Marquardt optimization algorithm library) [159] в качестве основы для параллельной реализации обучения с использованием метода Левенберга- Марквардта.

Расчеты проводились на гибридной вычислительной системе на базе серверной платформы HP Proliant G7 (AMD Opteron 6100) и вычислительной

системы Tesla S2050 на базе архитектуры NVIDIA Fermi GPU с использованием технологии NVIDIA CUDA версии 5.0. Характеристики одной графической карты вычислительной системы Tesla S2050 представлены в таблице 4.1.

Таблица 4.1 – Спецификация одной графической карты вычислительной системы Tesla S2050

Частота GPU	1,15 ГГц
Количество потоковых мультипроцессоров	4
Количество CUDA ядер в одном потоковом мультипроцессоре	14 мультипроцессоров × 32 CUDA ядра = 448 CUDA ядер
Глобальная память	3072 МБ
Разделяемая память/L1 кэш	48 КБ/16 КБ
Константная память	64 КБ
Регистровая память	32 КБ
Текстурная память	512 Б

На этапе предварительной обработки данных пункта 2 размерность вектора входных данных составила $63 \times 11 = 693$ для рецептора прогестерона и $69 \times 11 = 759$ для рецептора глюкокортикоидов, что соответствует размеру 5,4 КБ и 5,9 КБ (двойная точность). Поэтому в параллельной реализации PCA (блок-схема на рисунке 4.1) и MDS была задействована разделяемая память графических карт.

На этапе обучения ИНС пункта 3.2 для выбора оптимальной конфигурации варьировались следующие параметры:

1. количество нейронов в скрытом слое – 9 вариантов (для PCA – от 5 до 13 нейронов; для MDS – от 3 до 11 нейронов);
2. разбиение выборки на обучающее, тестовое и контрольное множества с учетом коэффициента Танимото – 7 вариантов;

3. первоначальные значения весов ИНС – 4 варианта.

Таким образом, общее количество сетей, для которых проводилось обучение для каждого рецептора, составило 252. В этом случае в разделяемой памяти хранился набор входных данных, оптимизированный после пункта 2, а также массив индексов дескрипторов для 7 вариантов обучающего, тестового и контрольного множеств. Остальные переменные хранились в глобальной памяти. Вычисления проводились с одинарной точностью.

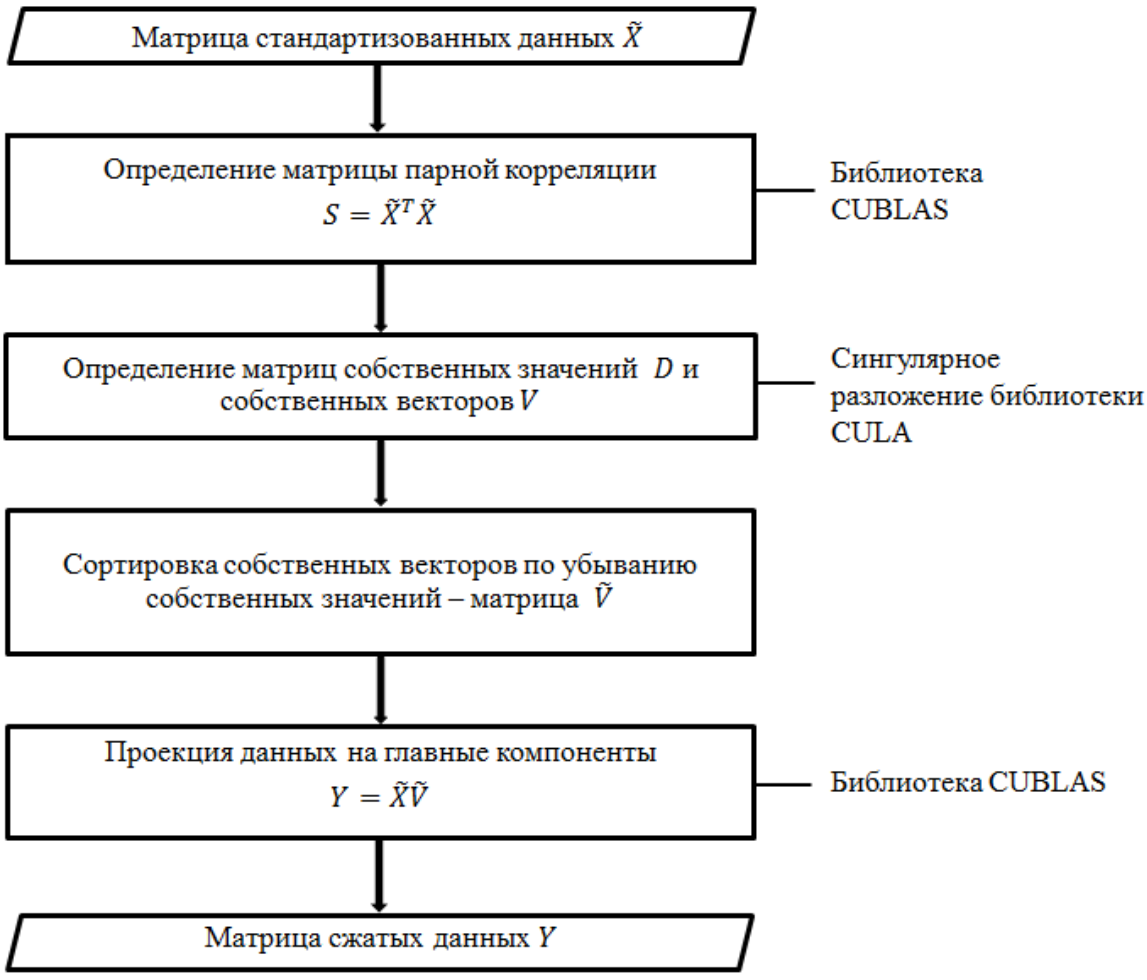


Рисунок 4.1 – Блок-схема программной реализации метода главных компонент с указанием используемых библиотек с поддержкой CUDA в параллельной реализации

Результаты вычислений параллельной версии пунктов 2 и 3.2 практически идентичны результатам, полученным в реализации MATLAB.

Временные характеристики обучения 252 вариантов ИНС для рецептора прогестерона представлены в таблице 4.2.

Таблица 4.2 – Временные характеристики (в сек.) выполнения 252 вариантов ИНС для рецептора прогестерона.

Реализация	MATLAB	C++ CPU	C++ CPU и CUDA GPU
Обучение ИНС при PCA	482,52	334,05	4,87
Обучение ИНС при MDS	398,16	296,02	4,46

Также для обучения сети замерялось время, которое необходимо для обучения 20 эпох. На основании этого времени было вычислено количество эпох обучения одной сети, приходящееся в среднем на единицу времени.

Общая характеристика результатов распараллеливания пунктов 2 и 3.2 представлена в таблице 4.3. По указанным характеристикам видно, что использование графических процессоров позволяет ускорить процедуру сжатия данных в 7 раз для метода главных компонент и в 3,6 раз для многомерного шкалирования по сравнению с реализацией C++ на центральном процессоре CPU. Также видно, что количество эпох обучения для одной сети на CPU+GPU в 3,8 раз ниже, чем на CPU, но из-за распараллеливания на 256 сетей, за единицу времени количество эпох обучения всех 256 сетей в ~69 раз превосходит CPU реализацию.

Данный результат согласуется с архитектурой GPU: большое количество более медленных по сравнению с CPU ядер.

Также для ускорения процедуры обучения при различных методах снижения размерности в реализацию была включена поддержка расчетов на двух графических картах, которая была использована при обучении сети с различными вариантами первоначальных весов для модели с использованием метода PCA и метода MDS для рецептора глюкокортикоидов. Для этого итоговая модель была разделена на 2 блока, каждый из которых вычислялся на отдельной графической

карте. Временные характеристики вычислений представлены в таблице 4.4. Первый блок представлял собой параллельную версию итоговой модели с использованием метода главных компонент, а второй блок – метода многомерного шкалирования. Такой подход позволил получить дополнительное ускорение, соразмерное с количеством используемых графических карт.

Таблица 4.3 – Характеристики выполнения последовательных и параллельных расчетов отдельных функционалов модели для рецептора прогестерона.

Реализация	MATLAB	C++ CPU	C++ CPU и CUDA GPU
Метод главных компонент PCA (сек.)	0,030	0,014	0,002
Многомерное шкалирование MDS (сек.)	0,154	0,103	0,028
Обучение ИНС при PCA (эпох/ед.врем.)	54/1 сеть	78/1 сеть	5396/256 сетей
Обучение ИНС при MDS (эпох/ед.врем.)	61/1 сеть	82/1 сеть	5524/256 сетей

Таблица 4.4 – Временные характеристики выполнения 252 вариантов ИНС для рецептора глюкокортикоидов.

Реализация	MATLAB	C++ CPU	C++ CPU и CUDA GPU
Обучение ИНС при PCA	507,75	347,53	5,04
Обучение ИНС при MDS	415,32	309,15	4,93
Обучение ИНС при PCA + Обучение ИНС при MDS	902,68	645,12	5,18

Таким образом, разработан эффективный параллельный алгоритм метода оценки аффинности комплексов белок-лиганд с применением ИНС и его программная реализация с применением графических карт NVIDIA и технологии CUDA, которая позволяет ускорить процедуру настройки сети в ~69 раз при использовании одного метода снижения размерности. Также использование графических карт позволяет ускорить процедуру сжатия данных в 7 и 3,7 раз для метода главных компонент и многомерного шкалирования, соответственно. Использование двух графических карт с двумя методами снижения размерности дает дополнительный двойной прирост производительности в общем времени выполнения программы C++ и CUDA GPU.

ВЫВОДЫ

1. Предложен численный метод оценки аффинности комплексов лиганд-белок на основе комплексного подхода, совмещающего методы молекулярного моделирования, искусственных нейронных сетей и нелинейного снижения размерности.
2. На основе предложенного метода разработаны модели оценки аффинности лигандов к внутриклеточным рецепторам прогестерона и глюкокортикоидов с высокой предсказательной силой ($\overline{Q^2} = 0{,}94$ при снижении размерности входного множества методом главных компонент, $\overline{Q^2} = 0{,}90$ – методом неметрического многомерного шкалирования).
3. Вычислительный эксперимент по оценке величины связывания прегна-D´-пентаранов с рецептором прогестерона с последующей экспериментальной проверкой показал, что результаты вычислений модели на основе разработанного метода хорошо согласуются с экспериментальными данными и дают существенно лучший результат (${R^2}_{test} = 0{,}77$) по сравнению с предсказаниями по методам 3D QSAR (${R^2}_{test} = 0{,}37$).
4. Разработан эффективный алгоритм и программная реализация численного метода с применением параллельной технологии CUDA. Данная реализация позволяет ускорить процедуру сжатия данных в 7 и 3,7 раз для метода главных компонент и неметрического многомерного шкалирования, соответственно; и ускорить процедуру обучения в ~69 раз при использовании GPU ускорителей.

СПИСОК СОКРАЩЕНИЙ

CoMFA	–	Comparative Molecular Field Analysis (сравнительный анализ молекулярных полей)
CoMSIA	–	Comparative Molecular Similarity Indices Analysis (сравнительный анализ индексов молекулярного подобия)
CPU	–	Central processing unit (центральный процессор)
CUDA	–	Compute Unified Device Architecture
DRAM	–	Dynamic random access memory (динамическая память с произвольным доступом)
GPU	–	Graphics processing unit (графический процессор)
LBDD	–	Ligand-Based Drug Design (компьютерное конструирование лекарств на основе структур лигандов)
LDA	–	Linear Discriminant Analysis (метод дискриминантного анализа)
LE	–	Laplacian Eigenmaps (метод карт собственных значений лапласиана)
LLE	–	Local Linear Embedding (локально-линейное вложение)
LOO	–	перекрестный контроль с исключением по одному
MDS	–	Multidimensional scaling (многомерное шкалирование)
MM-GBSA	–	Molecular Mechanic/ Generalized Born Surface Area (покомпонентный метод расчета изменения свободной энергии на основе симуляции молекулярной динамики по обобщенной модели Борна)
MM-PBSA	–	Molecular Mechanic/ Poisson-Boltzmann Surface Area (покомпонентный метод расчета изменения свободной энергии на основе симуляции молекулярной динамики по уравнению Пуассона-Больцмана)

PCA	–	Principal Components Analysis (метод главных компонент)
QSAR	–	Quantitative structure–activity relationship (поиск количественных соотношений «структура-активность»)
RMSE	–	Root-mean-square error (среднеквадратичная ошибка)
SBDD	–	Structure-Based Drug Design (компьютерное конструирование лекарств на основе структуры белка-мишени)
ИНС	–	Искусственная нейронная сеть
ЛС-домен	–	Лиганд-связывающий домен

СПИСОК ЛИТЕРАТУРЫ

1. Congreve M., Murray C.W., Blundell T.L. Keynote review: Structural biology and drug discovery. // Drug Discovery Today. 2005. V. 10. P. 895-907.
2. Bengio Y., Paiement J.-F., Vincent P., Delalleau O., Le Roux N., Ouimet M. Out-of-sample extensions for lle, isomap, mds, eigenmaps, and spectral clustering. // Advances in Neural Information Processing Systems. 2004. V. 16. P. 177-184.
3. Федюшкина И.В. Предсказание аффинности и спектра действия лигандов ядерных рецепторов стероидных гормонов методами компьютерного моделирования: дис. на соискание ученой степени канд. биол. наук: 03.01.09 / Федюшкина Ирина Викторовна. – М., 2013 – 115 с.
4. Kuntz I.D., Blaney J.M., Oatley S.J., Langridge R., Ferrin T.E. A geometric approach to macromolecule-ligand interactions. // Journal of Molecular Biology. 1982. V. 161. P. 269-288.
5. SYBYL Molecular Modeling System (version 8.0), TRIPOS Associates, 1699 South Hanley Road, Suite 303, St. Louis, MO 63144, U.S.A.
6. Case D.A., Darden T., Cheatham III T.E., Simmerling C., Wang J., Duke R.E., Luo R., Merz K.M., Pearlman D.A., Crowley M. AMBER 9 User's Manual. San Francisco: University of California, 2006. 320 p.
7. Beale M.H., Demuth H.B., Hagan M.T. MATLAB Neural Network Toolbox User's Guide R2012b. Natick, MA: The MathWorks, Inc., 2012. 420 p.
8. NVIDIA CUDA Compute Unified Device Architecture, CUDA C Programming Guide, version 5.0 [Электронный ресурс]. URL: http://docs.nvidia.com/cuda/pdf/CUDA_C_Programming_Guide.pdf (дата обращения: 02.11.2013).
9. Ooms F. Molecular modeling and computer aided drug design. Examples of their applications in medicinal chemistry. // Current Medicinal Chemistry. 2000. V. 7. P. 141-158.

10. Ivanov A.S., Veselovsky A.V., Dubanov A.V., Skvortsov V.S. Bioinformatics Platform Development. Springer, 2006. P. 389-431.
11. Veselovsky A.V., Tikhonova O.V., Skvortsov V.S., Medvedev A.E., Ivanov A.S. An approach for visualization of the active site of enzymes with unknown three-dimensional structures. // SAR and QSAR in Environmental Research. 2001. V. 12. P. 345-358.
12. Flohr S., Kurz M., Kostenis E., Brkovich A., Fournier A., Klabunde T. Identification of nonpeptidic urotensin II receptor antagonists by virtual screening based on a pharmacophore model derived from structure-activity relationships and nuclear magnetic resonance studies on urotensin II. // Journal of Medicinal Chemistry. 2002. V. 45. P. 1799-1805.
13. Schleifer K.-J. Pseudoreceptor model for ryanodine derivatives at calcium release channels. // Journal of Computer-Aided Molecular Design. 2000. V. 14. P. 467-475.
14. Kubiny H. Variable selection in QSAR studies. I. An evolutionary algorithm. // Quantitative Structure-Activity Relationships. 1994. V. 13. P. 285-294.
15. Verma J., Khedkar V.M., Coutinho E.C. 3D-QSAR in drug design-a review. // Current Topics in Medicinal Chemistry. 2010. V. 10. P. 95-115.
16. Lill M.A. Multi-dimensional QSAR in drug discovery. // Drug Discovery Today. 2007. V. 12. P. 1013-1017.
17. Richardson B.J. Physiological research on alcohols. // Medical Times and Gazzette. 1869. V. 2. P. 703-706.
18. Overton E. Ueber die allgemeinen osmotischen Eigenschaften der Zelle, ihre vermutlichen Ursachen u. ihre Bedeutg fd Physiologie. Berlin: Fäsi & Beer, 1899. 48 p.
19. Kuz’min V.E., Muratov E.N., Artemenko A.G., Gorb L., Qasim M., Leszczynski J. The effect of nitroaromatics’ composition on their toxicity in vivo: novel, efficient non-additive 1D QSAR analysis. // Chemosphere. 2008. V. 72. P. 1373-1380.

20. Hansch C., Muir R.M., Fujita T., Maloney P.P., Geiger F., Streich M. The correlation of biological activity of plant growth regulators and chloromycetin derivatives with Hammett constants and partition coefficients. // Journal of the American Chemical Society. 1963. V. 85. P. 2817-2824.
21. Fujita T., Ban T. Structure-activity relation. 3. Structure-activity study of phenethylamines as substrates of biosynthetic enzymes of sympathetic transmitters. // Journal of Medicinal Chemistry. 1971. V. 14. P. 148-152.
22. Сухачев Д.В., Пивина Т.С., Шляпочников В.А., Петров Э.А., Палюлин В.А., Зефиров Н.С. Исследование количественных соотношений "структура-чувствительность к удару" органических полиазотистых веществ. // Доклады РАН. 1993. № 328. С. 50-57.
23. Kubinyi H. 3D QSAR in Drug Design: Volume 1: Theory Methods and Applications. Leiden: ESCOM, 1993. 759 p.
24. Kubinyi H., Folkers G., Martin Y.C. 3D QSAR in Drug Design: Volume 2: Ligand-Protein Interactions and Molecular Similarity. Dordrecht: Kluwer/ESCOM, 1998. 417 p.
25. Kubinyi H., Folkers G., Martin Y.C. 3D QSAR in Drug Design: Volume 3: Recent Advances. Dordrecht: Kluwer/ESCOM, 1998. 368 p.
26. Bahl A., Joshi P., B Bharate S., Chopra H. Pharmacophore Based 3D-QSAR Modeling and Molecular Docking of Leucettines as Potent Dyrk1A Inhibitors. // Letters in Drug Design & Discovery. 2013. V. 10. P. 719-726.
27. Hopfinger A., Wang S., Tokarski J.S., Jin B., Albuquerque M., Madhav P.J., Duraiswami C. Construction of 3D-QSAR models using the 4D-QSAR analysis formalism. // Journal of the American Chemical Society. 1997. V. 119. P. 10509-10524.
28. Andrade C.H., Pasqualoto K.F., Ferreira E.I., Hopfinger A.J. 4D-QSAR: perspectives in drug design. // Molecules. 2010. V. 15. P. 3281-3294.
29. Vedani A., Dobler M. 5D-QSAR: the key for simulating induced fit? // Journal of Medicinal Chemistry. 2002. V. 45. P. 2139-2149.

30. Vedani A., Dobler M., Dollinger H., Hasselbach K.-M., Birke F., Lill M.A. Novel ligands for the chemokine receptor-3 (CCR3): a receptor-modeling study based on 5D-QSAR. // Journal of Medicinal Chemistry. 2005. V. 48. P. 1515-1527.
31. Vedani A., Dobler M., Lill M.A. Combining protein modeling and 6D-QSAR. Simulating the binding of structurally diverse ligands to the estrogen receptor. // Journal of Medicinal Chemistry. 2005. V. 48. P. 3700-3703.
32. Gasteiger J. Handbook of Chemoinformatics: From Data to Knowledge in 4 Volumes. Weinheim: Wiley-VCH, 2003. 1930 p.
33. Cramer R.D., Patterson D.E., Bunce J.D. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. // Journal of the American Chemical Society. 1988. V. 110. P. 5959-5967.
34. Wold S., Sjöström M., Eriksson L. PLS-regression: a basic tool of chemometrics. // Chemometrics and Intelligent Laboratory Systems. 2001. V. 58. P. 109-130.
35. Klebe G., Abraham U. On the prediction of binding properties of drug molecules by comparative molecular field analysis. // Journal of Medicinal Chemistry. 1993. V. 36. P. 70-80.
36. Kearns M., Ron D. Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. // Neural Computation. 1999. V. 11. P. 1427-1453.
37. Sippl W. Receptor-based 3D QSAR analysis of estrogen receptor ligands–merging the accuracy of receptor-based alignments with the computational efficiency of ligand-based methods. // Journal of Computer-Aided Molecular Design. 2000. V. 14. P. 559-572.
38. Sippl W., Contreras J.-M., Parrot I., Rival Y.M., Wermuth C.G. Structure-based 3D QSAR and design of novel acetylcholinesterase inhibitors. // Journal of Computer-Aided Molecular Design. 2001. V. 15. P. 395-410.
39. Kubinyi H. Comparative molecular field analysis (CoMFA). // Handbook of Chemoinformatics: From Data to Knowledge in 4 Volumes. 2008. P. 1555-1574.
40. Klebe G., Abraham U., Mietzner T. Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. // Journal of Medicinal Chemistry. 1994. V. 37. P. 4130-4146.

41. Joseph-McCarthy D. Structure-based lead optimization. // Annual Reports in Computational Chemistry. 2005. V. 1. P. 169-183.
42. Seifert M.H., Wolf K., Vitt D. Virtual high-throughput *in silico* screening. // Biosilico. 2003. V. 1. P. 143-149.
43. Zsoldos Z., Reid D., Simon A., Sadjad S.B., Johnson A.P. eHiTS: a new fast, exhaustive flexible ligand docking system. // Journal of Molecular Graphics and Modelling. 2007. V. 26. P. 198-212.
44. Hubbard R.E. Can drugs be designed? // Current Opinion in Biotechnology. 1997. V. 8. P. 696-700.
45. Burkert U., Allinger N.L. Molecular mechanics. Washington: American Chemical Society 1982. 339 p.
46. Pearlman D.A., Case D.A., Caldwell J.W., Ross W.S., Cheatham III T.E., DeBolt S., Ferguson D., Seibel G., Kollman P. AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. // Computer Physics Communications. 1995. V. 91. P. 1-41.
47. Kollman P.A., Massova I., Reyes C., Kuhn B., Huo S., Chong L., Lee M., Lee T., Duan Y., Wang W. Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. // Accounts of Chemical Research. 2000. V. 33. P. 889-897.
48. Luo R., David L., Gilson M.K. Accelerated Poisson-Boltzmann calculations for static and dynamic systems. // Journal of Computational Chemistry. 2002. V. 23. P. 1244-1253.
49. Tsui V., Case D.A. Theory and applications of the generalized Born solvation model in macromolecular simulations. // Biopolymers. 2000. V. 56. P. 275-291.
50. Chong L.T., Pitera J.W., Swope W.C., Pande V.S. Comparison of computational approaches for predicting the effects of missense mutations on p53 function. // Journal of Molecular Graphics and Modelling. 2009. V. 27. P. 978-982.

51. Stahl M., Todorov N.P., James T., Mauser H., Boehm H.J., Dean P.M. A validation study on the practical use of automated de novo design. // Journal of Computer-Aided Molecular Design. 2002. V. 16. P. 459-478.
52. Баскин И.И., Палюлин В.А., Зефиров Н.С. Применение искусственных нейронных сетей в химических и биохимических исследованиях. // Вестник Московского университета. Серия 2. Химия. 1999. № 40. С. 323-326.
53. Baskin I.I., Palyulin V.A., Zefirov N.S. Neural networks in building QSAR models. // Artificial Neural Networks: Methods and Applications. 2009. P. 133-154.
54. Гиллер С.А., Глаз А.Б., Растригин Л.А. Распознавание физиологической активности химических соединений на перцептроне со случайной адаптацией структуры. // Доклады Академии наук СССР. 1971. № 199. С. 851-853.
55. Rosenblatt F. The perceptron: a probabilistic model for information storage and organization in the brain. // Psychological Review. 1958. V. 65. P. 386.
56. Aoyama T., Suzuki Y., Ichikawa H. Neural networks applied to structure-activity relationships. // Journal of Medicinal Chemistry. 1990. V. 33. P. 905-908.
57. Aoyama T., Suzuki Y., Ichikawa H. Neural networks applied to pharmaceutical problems. III. Neural networks applied to quantitative structure-activity relationship (QSAR) analysis. // Journal of Medicinal Chemistry. 1990. V. 33. P. 2583-2590.
58. Галушкин А.И., Судариков В.А., Шабанов Е.В. Нейроматематика: методы решения задач на нейрокомпьютерах. // Математическое моделирование. 1991. № 3. С. 93-111.
59. Devillers J. Neural Networks in QSAR and Drug Design. London: Academic Press, Inc., 1996. 304 p.
60. Niculescu S.P. Artificial neural networks and genetic algorithms in QSAR. // Journal of Molecular Structure: THEOCHEM. 2003. V. 622. P. 71-83.
61. Cheng F., Sutariya V. Applications of Artificial Neural Network Modeling in Drug Discovery. // Clinical & Experimental Pharmacology. 2012.

62. Golbraikh A., Wang X.S., Zhu H., Tropsha A. Predictive QSAR Modeling: Methods and Applications in Drug Discovery andChemical Risk Assessment. Springer, 2012. P. 1309-1342.
63. Портал искусственного интеллекта: Нейронные сети [Электронный ресурс]. URL: http://www.aiportal.ru/articles/neural-networks/neural-networks.html (дата обращения: 12.11.2013).
64. Портал искусственного интеллекта: Модель нейрона [Электронный ресурс]. URL: http://www.aiportal.ru/articles/neural-networks/model-neuron.html (дата обращения: 12.11.2013).
65. McCulloch W., Pitts W. A logical calculus of the ideas immanent in nervous activity. // The bulletin of mathematical biophysics. 1943. V. 5. P. 115-133.
66. Портал искусственного интеллекта: Персептрон [Электронный ресурс]. URL: http://www.aiportal.ru/articles/neural-networks/perceptron.html (дата обращения: 12.11.2013).
67. Minsky M., Papert S. Perceptron: an introduction to computational geometry. Cambridge: MIT Press, 1969. 88 p.
68. Портал искусственного интеллекта: Многослойный персептрон [Электронный ресурс]. URL: http://www.aiportal.ru/articles/neural-networks/multi-perceptron.html (дата обращения: 12.11.2013).
69. Колмогоров А.Н. О представлении непрерывных функций нескольких переменных в виде суперпозиции непрерывных функций одного переменного. // Доклады Академии наук СССР. 1957. № 114. С. 953-956.
70. Hecht-Nielsen R. Kolmogorov’s mapping neural network existence theorem. // Proceedings of the international conference on Neural Networks. 1987. P.
71. Горбань А.Н., Дунин-Барковский В.Л., Кирдин А.Н., Миркес Е.М., Новоходько А.Ю., Россиев Д.А., Терехов С.А., Сенашова М.Ю., Царегородцев В.Г. Нейроинформатика. Новосибирск: Наука. Сибирское предприятие РАН, 1998. 296 с.
72. Girossi F., Poggio T. Representation qualities of neural networks: Kolmogorov's theorem is irrelevant. // Neural Computation. 1989. V. 1. P. 465-469.

73. Kůrková V. Kolmogorov's theorem and multilayer neural networks. // Neural Networks. 1992. V. 5. P. 501-506.
74. Kohonen T. Self-Organizing Maps. Berlin Heidelberg: Springer-Verlag, 2001. 501 p.
75. Rumelhart D.E., Hintont G.E., Williams R.J. Learning representations by back-propagating errors. // Nature. 1986. V. 323. P. 533-536.
76. Иванов В.В., Пурэвдорж Б., Пузырин И.В. Методы второго порядка для обучения многослойного перцептрона. // Математическое моделирование. 1998. № 10. С. 117-124.
77. Ясницкий Л.Н. Введение в искусственный интеллект. М.: Академия, 2005. 176 с.
78. Rumelhart D.E., McClelland J.L. Parallel distributed processing: explorations in the microstructure of cognition. Volume 1. Foundations. // 1986.
79. Hestenes M.R., Stiefel E. Methods of conjugate gradients for solving linear systems. // Journal of Research of the National Bureau of Standards. 1952. V. 49. P. 409-432.
80. Fletcher R., Reeves C.M. Function minimization by conjugate gradients. // The computer journal. 1964. V. 7. P. 149-154.
81. Polak E., Ribiere G. Note on convergence of conjugate direction methods. // Revue Francaise d'Informatique de Recherche Operationnelle. 1969. V. 3. P. 35-43.
82. Haykin S. Neural Networks: A Comprehensive Foundation. New Jersey: Prentice Hall International, 1999. 842 p.
83. Ермаков В.В., Калиткин Н.Н. Оптимальный шаг и регуляризация метода Ньютона. // Журнал вычислительной математики и математической физики. 1981. № 21. С. 491-497.
84. Saarinen S., Bramley R., Cybenko G. Ill-conditioning in neural network training problems. // SIAM Journal on Scientific Computing. 1993. V. 14. P. 693-714.
85. Тихонов А.Н., Арсенин В.Я. Методы решения некорректных задач. М.: Наука, 1979. 284 с.

86. Жанлав Т., Пузынин И.В. О сходимости итераций на основе непрерывного аналога метода Ньютона. // Журнал вычислительной математики и математической физики. 1992. № 32. С. 846-856.
87. Dennis J.J.E., Schnabel R.B. Numerical methods for unconstrained optimization and nonlinear equations. New Jercy: Prentice-Hall: Englewood Cliffs, 1983. 378 p.
88. Hagan M.T., Menhaj M.B. Training feedforward networks with the Marquardt algorithm. // Neural Networks, IEEE Transactions on. 1994. V. 5. P. 989-993.
89. Battiti R. First-and second-order methods for learning: between steepest descent and Newton's method. // Neural Computation. 1992. V. 4. P. 141-166.
90. Becker S., Le Cun Y. Improving the convergence of back-propagation learning with second order methods. // Proceedings of the 1988 connectionist models summer school. 1988. P. 29-37.
91. Andrea T.A., Kalayeh H. Applications of neural networks in quantitative structure-activity relationships of dihydrofolate reductase inhibitors. // Journal of Medicinal Chemistry. 1991. V. 34. P. 2824-2836.
92. Tetko I.V., Livingstone D.J., Luik A.I. Neural network studies. 1. Comparison of overfitting and overtraining. // Journal of Chemical Information and Computer Sciences. 1995. V. 35. P. 826-833.
93. Livingstone D.J., Manallack D.T., Tetko I.V. Data modelling with neural networks: advantages and limitations. // Journal of Computer-Aided Molecular Design. 1997. V. 11. P. 135-142.
94. Баскин И.И. Моделирование свойств химических соединений с использованием искусственных нейронных сетей и фрагментативных дескрипторов: дис. на соискание ученой степени докт. физ.-мат. наук: 02.00.17 / Баскин Игорь Иосифович. – М., 2009 – 365 с.
95. Bishop C.M., Nasrabadi N.M. Pattern recognition and machine learning. New York: Springer, 2006. 738 p.
96. Van der Maaten L.J.P. An introduction to dimensionality reduction using matlab. Maastricht: Universiteit Maastricht, 2007. 44 p.

97. Pearson K. LIII. On lines and planes of closest fit to systems of points in space. // The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science. 1901. V. 2. P. 559-572.
98. Hotelling H. Analysis of a complex of statistical variables into principal components. // Journal of Educational Psychology. 1933. V. 24. P. 417.
99. Abdi H., Williams L.J. Principal component analysis. // Wiley Interdisciplinary Reviews: Computational Statistics. 2010. V. 2. P. 433-459.
100. Fisher R.A. The use of multiple measurements in taxonomic problems. // Annals of eugenics. 1936. V. 7. P. 179-188.
101. Agrafiotis D.K. Stochastic proximity embedding. // Journal of Computational Chemistry. 2003. V. 24. P. 1215-1221.
102. Sun J., Boyd S., Xiao L., Diaconis P. The fastest mixing Markov process on a graph and a connection to a maximum variance unfolding problem. // SIAM review. 2006. V. 48. P. 681-699.
103. Schölkopf B., Smola A., Müller K.-R. Nonlinear component analysis as a kernel eigenvalue problem. // Neural Computation. 1998. V. 10. P. 1299-1319.
104. Baudat G., Anouar F. Generalized discriminant analysis using a kernel approach. // Neural Computation. 2000. V. 12. P. 2385-2404.
105. Lafon S., Lee A.B. Diffusion maps and coarse-graining: A unified framework for dimensionality reduction, graph partitioning, and data set parameterization. // Pattern Analysis and Machine Intelligence, IEEE Transactions on. 2006. V. 28. P. 1393-1403.
106. Nadler B., Lafon S., Coifman R.R., Kevrekidis I.G. Diffusion maps, spectral clustering and reaction coordinates of dynamical systems. // Applied and Computational Harmonic Analysis. 2006. V. 21. P. 113-127.
107. Hinton G.E., Roweis S.T. Stochastic neighbor embedding. // Advances in Neural Information Processing Systems. 2002. P. 833-840.
108. Hinton G.E., Salakhutdinov R.R. Reducing the dimensionality of data with neural networks. // Science. 2006. V. 313. P. 504-507.

109. Kruskal J.B. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. // Psychometrika. 1964. V. 29. P. 1-27.
110. Cox T., Cox M. Multidimensional Scaling. London: Chapman & Hall, 1994. 328 p.
111. Александров В.В., Горский Н.Д. Алгоритмы и программы структурного метода обработки данных. Ленинград: Наука, 1983. 208 с.
112. Sammon Jr J.W. A nonlinear mapping for data structure analysis. // Computers, IEEE Transactions on. 1969. V. 100. P. 401-409.
113. Borg I., Groenen P. Modern multidimensional scaling: Theory and applications. New York: Springer, 2005. 614 p.
114. Tenenbaum J., de Silva V., Langford J. A global geometric framework for nonlinear dimensionality reduction. // Science. 2000. V. 290. P. 2319-2323.
115. Dijkstra E.W. A note on two problems in connexion with graphs. // Numerische mathematik. 1959. V. 1. P. 269-271.
116. Floyd R.W. Algorithm 97: shortest path. // Communications of the ACM. 1962. V. 5. P. 344-348.
117. Balasubramanian M., Schwartz E.L. The isomap algorithm and topological stability. // Science. 2002. V. 295. P. 7-7.
118. Tenenbaum J.B., De Silva V., Langford J.C. A global geometric framework for nonlinear dimensionality reduction. // Science. 2000. V. 290. P. 2319-2323.
119. Donoho D.L., Grimes C. Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. // Proceedings of the National Academy of Sciences. 2003. V. 100. P. 5591-5596.
120. Zhang Z., Zha H. Principal manifolds and nonlinear dimension reduction via local tangent space alignment. // SIAM Journal of Scientific Computing. 2004. V. 26. P. 313–338.
121. Roweis S.T., Saul L.K. Nonlinear dimensionality reduction by locally linear embedding. // Science. 2000. V. 290. P. 2323-2326.
122. Belkin M., Niyogi P. Laplacian eigenmaps for dimensionality reduction and data representation. // Neural Computation. 2003. V. 15. P. 1373-1396.

123. Anderson Jr W.N., Morley T.D. Eigenvalues of the Laplacian of a graph∗. // Linear and Multilinear Algebra. 1985. V. 18. P. 141-145.
124. Teng L., Li H., Fu X., Chen W., Shen I.-F. Dimension reduction of microarray data based on local tangent space alignment. // Fourth IEEE Conference on Cognitive Informatics. 2005. С. 154-159.
125. Воеводин В.В., Воеводин В.В. Параллельные вычисления. СПб.: БХВ-Петербург, 2002. 600 с.
126. Боресков А.В., Харламов А.А. Основы работы с технологией CUDA. М.: ДМК Пресс, 2010. 232 с.
127. Stone J.E., Hardy D.J., Ufimtsev I.S., Schulten K. GPU-accelerated molecular modeling coming of age. // Journal of Molecular Graphics and Modelling. 2010. V. 29. P. 116-125.
128. Хлопов Д.И., Авксентьева О.А. CUDA реализация алгоритмов воксельного представления отрезков прямых для объемных 3D дисплеев. // Материалы VI научно-технической конференции молодых ученых и студентов. Информатика и компьютерные технологии-2010. Донецк. 2010. С. 132-138.
129. CUDA C Best Practices Guide, version 5.0 [Электронный ресурс]. URL: http://docs.nvidia.com/cuda/pdf/CUDA_C_Best_Practices_Guide.pdf (дата обращения: 02.11.2013).
130. Sanders J., Kandrot E. CUDA by example: an introduction to general-purpose GPU programming. Upper Saddle River, New Jersey: Addison-Wesley Professional, 2010. 312 p.
131. Söderholm A.A., Lehtovuori P.T., Nyrönen T.H. Docking and three-dimensional quantitative structure-activity relationship (3D QSAR) analyses of nonsteroidal progesterone receptor ligands. // Journal of Medicinal Chemistry. 2006. V. 49. P. 4261-4268.
132. Xu Y., Zhang T., Chen M. Combining 3D-QSAR, docking, molecular dynamics and MM/PBSA methods to predict binding modes for nonsteroidal selective modulator to glucocorticoid receptor. // Bioorganic and Medicinal Chemistry Letters. 2009. V. 19. P. 393-396.

133. Bourguet W., Germain P., Gronemeyer H. Nuclear receptor ligand-binding domains: three-dimensional structures, molecular interactions and pharmacological implications. // Trends in Pharmacological Sciences. 2000. V. 21. P. 381-388.
134. Williams S.P., Sigler P.B. Atomic structure of progesterone complexed with its receptor. // Nature. 1998. V. 393. P. 392-396.
135. Bledsoe R.K., Montana V.G., Stanley T.B., Delves C.J., Apolito C.J., McKee D.D., Consler T.G., Parks D.J., Stewart E.L., Willson T.M. Crystal structure of the glucocorticoid receptor ligand binding domain reveals a novel mode of receptor dimerization and coactivator recognition. // Cell. 2002. V. 110. P. 93-105.
136. Willett P., Barnard J.M., Downs G.M. Chemical similarity searching. // Journal of Chemical Information and Computer Sciences. 1998. V. 38. P. 983-996.
137. Раевский О.А. Дескрипторы молекулярной структуры в компьютерном дизайне биологически активных веществ. // Успехи химии. 1999. № 68. С. 555-575.
138. Filimonov D., Poroikov V., Borodina Y., Gloriozova T. Chemical similarity assessment through multilevel neighborhoods of atoms: definition and comparison with the other descriptors. // Journal of Chemical Information and Computer Sciences. 1999. V. 39. P. 666-670.
139. Федюшкина И.В., Ромеро Рейес И.В., Лагунин А.А., Скворцов В.С. Предсказание спектра действия лигандов рецепторов стероидных гормонов. // Биомедицинская химия. 2013. № 59. С. 591-599.
140. Romero Reyes I., Fedyushkina I., Skvortsov V., Filimonov D. Prediction of progesterone receptor inhibition by high-performance neural network algorithm // International journal of mathematical models and methods in applied sciences. 2013. V. 7. P. 303-310.
141. Fedyushkina I.V., Romero Reyes I.V. Prediction of glucocorticoid receptor inhibition by high-performance neural network algorithm. // Advances in Mathematical and Computational Methods. 2012. V. 4. P. 203-207.

142. Левина И.С., Куликова Л.Е., Камерницкий А.В., Покровская Е.В., Смирнов А.Н. Синтез 19-замещенных стероидов ряда 16α,17α-циклогексанопрегнанов и изучение их взаимодействия с белками цитозоля матки и сыворотки крови крысы. // Известия Академии Наук, Серия Химическая. 2005. № 11. С. 2579-2584.

143. Levina I.S., Kulikova L.E., Kamernitskii A.V., Shashkov A.S., Smirnov A.N., Pokrovskaya E.V. Synthesis of 6 (E)-and 6 (Z)-(3-ethoxycarbonylpropyl) oximes of 16α, 17α-cyclohexanopregn-4-ene-3, 6, 20-trione and study of their interaction with proteins of the rat uterine cytosol and blood serum. // Russian chemical bulletin. 2002. V. 51. P. 703-708.

144. Kamernitskii A.V., Levina I.S. [Pregna-D'-pentaranes--progestins and antiprogestins: I. Differentiation of biological functions of steroid hormones]. // Bioorganicheskaia Khimiia. 2005. V. 31. P. 115-129.

145. Smirnov A.N., Pokrovskaya E.V., Kogteva G.S., Shevchenko V.P., Levina I.S., Kulikova L.E., Kamernitzky A.V. The size and/or configuration of the cycloalkane D′ ring in pentacyclic progesterone derivatives are crucial for their high-affinity binding to a protein in addition to progesterone receptor in rat uterine cytosol☆. // Steroids. 2000. V. 65. P. 163-170.

146. Смирнов А.Н., Покровская Е.В., Шевченко В.П., Нагаев И.Ю., Мясоедов Н.Ф., Левина И.С., Куликова Л.Е., Камерницкий А.В. Видовые и тканевые особенности распределения белков, связывающих 16а, 17а-циклоалкановые производные прогестерона. // Биоорганическая химия. 2002. № 28. С. 251-257.

147. Richmond N.J., Abrams C.A., Wolohan P.R., Abrahamian E., Willett P., Clark R.D. GALAHAD: 1. Pharmacophore identification by hypermolecular alignment of ligands in 3D. // Journal of Computer-Aided Molecular Design. 2006. V. 20. P. 567-587.

148. Berman H., Henrick K., Nakamura H. Announcing the worldwide Protein Data Bank. // Nature Structural Biology. 2003. V. 10. P. 980.

149. Федюшкина И.В., Скворцов В.С., Ромеро Рейес И.В., Левина И.С. Молекулярный докинг и 3D-QSAR производных 16α,17α-циклоалканопрогестерона как лигандов рецептора прогестерона // Биомедицинская химия. 2013. № 2013. С. 622-635.

150. Левина И.С., Куликова Л.Е., Шулишов Е.В., Томилов Ю.В., Смирнов А.Н. Синтез, структура и биологические свойства замещенных [16α,17α]-циклопропапрегн-4-ен-3,20-дионов. // Известия Академии Наук, Серия Химическая. 2013. № 6. С. 1449-1453.

151. Levina I.S., Pokrovskaya E.V., Kulikova L.E., Kamernitzky A.V., Kachala V.V., Smirnov A.N. 3- and 19-oximes of 16alpha,17alpha-cyclohexanoprogesterone derivatives: synthesis and interactions with progesterone receptor and other proteins. // Steroids. 2008. V. 73. P. 815-827.

152. 4-Гетеро-16α, 17α-Циклогексанопрегнаны: пат. 2426737 Рос. Федерация: МПК C07J 53/00 C07J 73/00 A61P 35/00 / Левина И.С., Куликова Л.Е., Смирнов А.Н., Шимановский Н.Л., Семейкин А.В., Карева Е.Н., Федотчева Т.А., Болотова Е.Н.; заявитель и патентообладатель ИОХ РАН, ГОУ ВПО РГМУ Росздрава. – № 2009146877/04; заявл. 17.12.2009; опубл. 20.08.2011, Бюл. № 23. – 7 с.

153. Matlab Toolbox for Dimensionality Reduction (v0.8.1 - March 2013) [Электронный ресурс]. URL: http://homepage.tudelft.nl/19j49/Matlab_Toolbox_for_Dimensionality_Reduction.html (дата обращения: 08.07.2013).

154. Van der Maaten L.J.P., Postma E.O., Van Den Herik H.J. Dimensionality reduction: A comparative review. // Journal of Machine Learning Research. 2009. V. 10. P. 1-41.

155. CUBLAS Library, version 5.0 [Электронный ресурс]. URL: http://docs.nvidia.com/cuda/pdf/CUBLAS_Library.pdf (дата обращения: 02.11.2013).

156. Humphrey J.R., Price D.K., Spagnoli K.E., Paolini A.L., Kelmelis E.J. CULA: hybrid GPU accelerated linear algebra routines. // SPIE Defense and Security Symposium (DSS). 2010. P. 770502-770507.
157. CURAND Library, version 5.0 [Электронный ресурс]. URL: http://docs.nvidia.com/cuda/pdf/CURAND_Library.pdf (дата обращения: 02.11.2013).
158. Fester T., Schreiber F., Strickert M., Gatersleben I. CUDA-based Multi-core Implementation of MDS-based Bioinformatics Algorithms. // Proceedings of German Conference on Bioinformatics. 2009. P. 67-79.
159. levmar: Levenberg-Marquardt nonlinear least squares algorithms in C/C++ [Электронный ресурс]. URL: http://users.ics.forth.gr/~lourakis/levmar/ (дата обращения: 15.02.2013).

БЛАГОДАРНОСТИ

Считаю своим приятным долгом выразить глубокую благодарность моему научному руководителю кандидату физико-математических наук Дмитрию Алексеевичу Филимонову, а также заведующему лабораторией параллельных вычислений и информационных вычислений кандидату биологических наук Владлену Станиславовичу Скворцову за постоянное внимание, творческие советы, конструктивные замечания и помощь при выполнении работы.

Искренне благодарю коллегу и соавтора ряда публикаций кандидата биологических наук Ирину Викторовну Федюшкину за плодотворные совместные обсуждения идей диссертации.

Отдельная благодарность доктору биологических наук А.В. Веселовскому, заведующему лабораторией структурной биоинформатики ФГБУ «ИБМХ» РАМН, и сотрудникам сектора «Интеллектуальный анализ данных и моделирование» ИППИ РАН за большой интерес к теме диссертационной работы и консультации; доктору физико-математических наук П.Н. Вабищевичу, заведующему лабораторией разработки интегральных расчётных кодов ИБРАЭ РАН, и всему сплоченному коллективу лаборатории вычислительной теплогидродинамики ИБРАЭ РАН за всестороннюю помощь в организационных вопросах, а также сотрудникам Лаборатории информационных технологий ОИЯИ докторам физико-математических наук И.М. Иванченко, В.В. Иванову и Е.В. Земляной за всестороннюю помощь при подготовке материалов диссертации.