МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ
БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ФАКУЛЬТЕТ ПРИКЛАДНОЙ МАТЕМАТИКИ И ИНФОРМАТИКИ
Кафедра биомедицинской информатики

АНИЩЕНКО
Арсений Игоревич

РАЗРАБОТКА АЛГОРИТМА ОЦЕНКИ ИНГИБИТОРНОЙ АКТИВНОСТИ ХИМИЧЕСКОГО СОЕДИНЕНИЯ

Дипломная работа

Научный руководитель:
старший преподаватель кафедры БМИ
Г. И. Николаев

Допущен к защите
« » 2020 г
Зав. кафедрой биомедицинской информатики
кандидат физико-математических наук, доцент Ю. Л. Орлович

Минск, 2020

РЕФЕРАТ

Дипломная работа, 52 страницы, 17 рисунков, 1 таблица, 9 формул, 35 источников.

ДРАГ-ДИЗАЙН, МАШИННОЕ ОБУЧЕНИЕ, ИСКУССТВЕННЫЕ НЕЙРОННЫЕ СЕТИ, ГЛУБОКОЕ ОБУЧЕНИЕ, РЕГРЕССИОННЫЕ МОДЕЛИ, МОЛЕКУЛЯРНЫЙ ФИНГЕРПРИНТ, ИНГИБИТОРЫ ПРОНИКНОВЕНИЯ ВИЧ-1, БЕЛОК GP120, МЕТОДЫ МОЛЕКУЛЯРНОГО МОДЕЛИРОВАНИЯ.

Объект исследования — методы машинного обучения, которые можно применить для разработки новых лекарственных соединений, и способы оценки ингибиторной активности химического соединения.

Цель работы — создание нового алгоритма оценки ингибиторной активности химических соединений на основе технологии машинного обучения и молекулярных фингерпринтов.

Методы исследования: изучение тематической литературы и научных исследований, анализ данных о химических соединениях и энергии связи их комплекса с белком оболочки ВИЧ-1 gp120, создание базы данных молекулярных фингерпринтов, обучение нейронных сетей, статистическая оценка результатов.

Результатом работы являются новые методы поиска потенциальных ингибиторов белка оболочки ВИЧ-1 gp120. Была создана специальная база данных молекулярных фингерпринтов, которая насчитывает 166.416.851 соединений с соответствующими им молекулярными фингерпринтами MACCS, служащая для поиска химических соединений. Было предложено несколько решений регрессионной задачи прогнозирования потенциальной энергии связи комплекса лиганд с белком gp120 при помощи нейронных сетей.

Области применения: поиск новых лекарственных препаратов, виртуальный скрининг, оценка ингибиторной активности химических соединений.

РЭФЕРАТ

Дыпломная праца, 52 старонкі, 17 відарысаў, 1 табліца, 9 формул, 35 крыніц.

ДРАГ-ДЫЗАЙН, МАШЫННАЕ НАВУЧАННЕ, ШТУЧНЫЯ НЕЙРОННЫЯ СЕТКІ, ГЛЫБОКАЕ НАВУЧАННЕ, РЭГРЕСІЙНЫЯ МАДЭЛІ, МАЛЕКУЛЯРНЫЯ ФІНГЕРПРЫНТЫ, ІНГІБІТАРЫ ПРАНІКНЕННЯ ВІЧ-1, БЯЛОК GP120, МЕТАДЫ МАЛЕКУЛЯРНАГА МАДЭЛЯВАННЯ.

Аб'ект даследавання — метады машыннага навучання, якія можна выкарыстоўваць для распрацоўкі новых лекавых злучэнняў, і спосабы ацэнкі інгібітарнай актыўнасці хімічнага злучэння.

Мэта работы — стварэнне новага алгарытму ацэнкі інгібітарнай актыўнасці хімічных злучэнняў на аснове тэхналогіі машыннага навучання і малекулярных фінгерпынтаў.

Метады даследавання: вывучэнне тэматычнай літаратуры і навуковых даследаванняў, аналіз дадзеных аб хімічных злучэннях і энергіі сувязі іх комплексу з бялком абалонкі ВІЧ-1 gp120, стварэнне базы дадзеных малекулярных фінгерпрынтаў, навучанне нейронавых сетак, статыстычная ацэнка вынікаў.

Вынікам працы з'яўляюцца новыя метады пошуку патэнцыйных інгібітараў бялку абалонкі ВІЧ-1 gp120. Была створана спецыяльная база дадзеных малекулярных фінгерпрынтаў, якая налічвае 166.416.851 злучэнняў з адпаведнымі ім малекулярнымі фінгерпрынтамі MACCS, якая патрэбна для пошуку хімічных злучэнняў. Было прапанавана некалькі рашэнняў рэгрэсійнай задачы прагназавання патэнцыйнай энергіі сувязі комплексу ліганд з бялком gp120 пры дапамозе нейронных сетак.

Вобласці прымянення: пошук новых лекавых прэпаратаў, віртуальны скрынінг, ацэнка інгібітарнай актыўнасці хімічных злучэнняў.

ANNOTATION

Graduate work, 52 pages, 17 pictures, 1 table, 9 formulas, 35 sources.

DRUG DESIGN, MACHINE LEARNING, ARTIFICIAL NEURAL NETWORKS, DEEP LEARNING, REGRESSION MODELS, MOLECULAR FINGERPRINT, HIV-1 ENTRY INHIBITORS, GP120 PROTEIN, MOLECULAR MODELING.

The object of the research is machine learning methods, which can be used for new drug compounds development, and methods for inhibitory activity of a chemical compound evaluation.

Purpose of the work is to create a new algorithm for assessing the inhibitory activity of chemical compounds, based on machine learning technology and molecular fingerprints.

Research methods: subject literature and scientific research study, chemical compounds and the binding energy of their complex with HIV-1 gp120 envelope protein data analysis, molecular fingerprint database creation, neural networks training, results statistical evaluation.

The result of the work is new methods for searching for potential HIV-1 gp120 coat protein inhibitors. A special database of molecular fingerprints was created, which has 166,416,851 compounds with their corresponding molecular fingerprints MACCS, which is used for chemical compounds search. Several solutions to the regression problem of potential binding energy of the ligand complex with the gp120 protein prediction using neural networks was proposed.

The scopes are: drug search, virtual screening, inhibitory activity of chemical compounds evaluation.

ОГЛАВЛЕНИЕ

ОГЛАВЛЕНИЕ 5
ВВЕДЕНИЕ 7
ГЛАВА 1. ВИЧ-1 9
1.1 Общие сведения 9
1.2 Строение вируса 9
1.3 Лечение 10
ГЛАВА 2. КОМПЬЮТЕРНОЕ МОДЕЛИРОВАНИЕ ЛЕКАРСТВ 12
1.1 Скрининг 12
1.2 Докинг 13
1.3 Квантово-химическое моделирование 14
1.4 Молекулярная динамика 14
ГЛАВА 3. НЕЙРОННЫЕ СЕТИ 16
3.1 Общая информация о нейронных сетях 16
3.2 Сверточные нейронные сети 17
3.3 Генеративно-состязательные сети 18
3.4 Автоэнкодер 18
3.5 Градиентный спуск 19
3.6 Оптимизаторы градиентного спуска 19
3.5.1 RMSProp 19
3.5.2 SGD 20
3.5.3 Adam 20
3.7 Метрики 21
ГЛАВА 4. ГЕНЕРАЦИЯ ЛЕКАРСТВЕННЫХ СОЕДИНЕНИЙ ПРИ ПОМОЩИ НЕЙРОННЫХ СЕТЕЙ 24
4.1 Применение состязательного автоэнкодера для создания новых лекарств в области онкологии 24
4.2 Подготовка данных 26
4.3 Автоэнкодер для генерации лекарственных соединений ВИЧ-1 27
4.4 Интерпретация результатов 28
ГЛАВА 5. РАЗРАБОТКА АЛГОРИТМА ПРЕДСКАЗАНИЯ ИНГИБИТОРНОЙ АКТИВНОСТИ ХИМИЧЕСКИХ СОЕДИНЕНИЙ 31
5.1 PoseScore и RankScore статистические функции оценки 31
5.2 NNscore: оценочная функция для характеристики белок-лигандных комплексов 32
5.3 Регрессионная задача — оценка ингибиторной активности химического соединения 33
5.4 Регрессионные модели машинного обучения 33
5.4.1 Линейная регрессия 33
5.4.2 Регрессионное дерево решений 34
5.4.3 Случайный регрессионный лес 35
5.5 Разработка нейронной сети для предсказания энергии связи 36

5.5.1 Анализ данных 36
5.5.2 Полносвязные нейронные сети 38
5.5.3 Сверточные нейронные сети 41
5.5.4 Оценка моделей 45
ЗАКЛЮЧЕНИЕ 46
СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ 48
СПИСОК ПУБЛИКАЦИЙ 51
ПРИЛОЖЕНИЯ 52
Приложение 1. Сравнительная таблица результатов обучения нейронных сетей 52

ВВЕДЕНИЕ

На сегодняшний день проблема разработки и поиска новых лекарственных препаратов остается одной из наиболее важных проблем человечества. Появление больших вычислительных мощностей позволило куда более эффективно разрабатывать новые лекарственные препараты, отсеивая неподходящие химические соединения еще на этапах компьютерного моделирования — in silico. Такой подход называют драг-дизайном.

Ключевыми понятиями драг-дизайна являются мишень и лекарство. Мишень представляет собой биологическую макромолекулу, которая связана с определенной функцией, которая прямым или косвенным путем вызывает заболевание. Мишень чаще всего оказывается некоторым белком - рецептором либо ферментом. Лекарством же называют химическое соединение (как правило, низкомолекулярное), которое специфически взаимодействует со своей мишенью, тем самым влияя на процессы внутри клетки, что блокирует заболевание.

Вещество, которое замедляет протекание реакции, имеет общепринятое название - ингибитор. В свою очередь ингибирование делится на три типа: конкурентное, неконкурентное и бесконкурентное. В конкурентном ингибировании ингибитор связывается в активном центре ингибируемой молекулы, при этом составляя конкуренцию субстрату либо другому взаимодействующему веществу. Ингибиторы такого типа часто имеет схожую структурную форму с субстратом, однако не имеют в своем составе необходимые функциональные группы для катализации реакции. Неконкурентный ингибитор не мешает связыванию субстрата с ферментом, но вызывает такие конформационные изменения, которые делают фермент не пригодным для превращения субстрата в продукт. Неконкурентный ингибитор способен присоединяться и к свободному ферменту, и к фермент-субстратному комплексу. Бесконкурентный ингибитор связывается только с фермент-субстратным комплексом, но не со свободным ферментом. При связывании субстрата с ферментом конформация фермента изменяется, делая возможным связывание комплекса с ингибитором. Затем ингибитор, связываясь с ферментом, меняет его конформацию, что приводит к блокированию катализа. На практике поиск конкурентных ингибиторов становится возможным за счёт того, что структура потенциального ингибитора может быть схожа со структурой субстрата.

Целью драг-дизайна является создание новых препаратов для лечения заболеваний. Как правило, лекарством является некоторый ингибитор, способный остановить процессы развития и распространения заболевания.

Этиологический агент СПИДа — вирус иммунодефицита человека типа 1 (ВИЧ-1) — является одним из наиболее хорошо изученных вирусов, однако эффективные лекарства для профилактики и лечения этого заболевания до сих пор не созданы [1].

На сегодняшний день вирус иммунодефицита уже унес более 35 миллионов человеческих жизней, а число инфицированных уже превышает 37 миллионов человек [1].

Поэтому поиск новых методов разработки лекарственных препаратов является важной задачей для борьбы с такими серьезными заболеваниями.

ГЛАВА 1. ВИЧ-1

В данной главе рассмотрим основные особенности вируса иммунодефицита человека. Определим основные мишени, которые используются для разработки лекарственных препаратов против ВИЧ. А также рассмотрим существующие методы лечения.

1.1 Общие сведения

Вирус иммунодефицита человека (ВИЧ) является РНК вирусом, относится к семейству ретровирусов (лат. Retroviridae) и подсемейству лентивирусов (лат. Lentivirinae). Существует два типа вируса, которые могут вызывать СПИД, — ВИЧ-1 и ВИЧ-2. Во всем мире чаще всего причиной возникновения СПИД является вирус иммунодефицита человека первого типа [2]. Поэтому в данной работе будем рассматривать именное его с целью поиска эффективных лекарственных соединений для борьбы с ним.

Характерными особенностями ВИЧ являются длительный инкубационный период и медленная, неоднородная скорость инфицирования. Сам вирус не приводит к летальным случаям, он воздействует на иммунную систему носителя, поражая иммунные клетки имеющие сайд связывания CD4, что впоследствии увеличивает чувствительность организма к другим инфекциям и даже онкологическим заболеваниям, которым организм со здоровой иммунной системой мог бы противостоять. Такую невозможность противостоять инфекционным заболеваниям, вызванной вирусом иммунодефицита человека называют синдром приобретенного иммунодефицита (СПИД). СПИД — самая поздняя стадия ВИЧ-инфекции [1].

Также существует проблема диагностики ВИЧ, так как на начальных стадиях развития вируса никакие симптомы не проявляются, и узнать о заражении можно только при помощи специальных тестов. Не смотря на отсутствие явных симптомов, носители ВИЧ наиболее заразны первые несколько месяцев после заражения [1].

Передача вируса от человека к человеку возможно только через жидкости организма, которые обладают высокой концентрацией вируса — кровь, грудное молоко, семенная жидкость и вагинальные выделения [1]. Поэтому вирус обладает не столь высокой степенью заразности.

1.2 Строение вируса

Вирионы ВИЧ имеют форму икосаэдра, который покрыт шиповидными выростами. Размер вириона вирусной частицы варьируется от 100 до 120 нанометров. Оболочка вируса состоит из 2000 матриксных белков p17. Внутри оболочки находятся две нити РНК, которые связаны с белком p7. Шиповидные выросты состоят из двух белков внешней оболочки — gp120 и gp41. Белок gp120 и отвечает за связывание с рецептором CD4 иммунных клеток организма [2].

Данные белки вириона и являются мишенями драг-дизайна. И они активно изучаются для поиска лекарств и вакцин, направленных на борьбу с ними. В данной работе за мишень был выбран белок оболочки ВИЧ-1 gp120.

1.3 Лечение

Вирус иммунодефицита человека, к сожалению, не поддается лечению. Основными принципами лечения людей с ВИЧ-инфекцией являются стимулирование и укрепление иммунной системы и замедление развития и распространения вируса по организму. С этой целью применяют высокоактивную антиретровирусную терапию (ВААРТ), задача которой блокировать различные этапы репликационного цикла вируса. Такая терапия включает в себя прием сразу нескольких антиретровирусных препаратов [1]. Главными проблемами применения ВААРТ являются выработка устойчивости к используемым препаратам со стороны вируса путем мутаций (ВИЧ обладает высокой мутационной способностью) и серьезные побочные эффекты. Антиретровирусные препараты обладают высокой токсичностью, некоторые из них имеют в составе достаточно крупные молекулы, что требует инъекцию препарата в кровь пациента.

В настоящее время более 40 моноклональных антител (МКА) с широкой вирусной нейтрализацией являются потенциальными кандидатами для разработки безопасной и эффективной вакцины против ВИЧ-1. Эти антитела блокируют четыре функционально консервативных эпитопа оболочки вируса, которые включают 1) CD4-связывающий сайт белка gp120, 2) сегменты V1/V2 и V3 этого гликопротеина, 3) проксимальную внешнюю область MPER белка gp41 и 4) четвертичный интерфейс gp120-gp41. Среди кросс-реактивных анти-ВИЧ антител следует особо отметить МКА N6, которое нейтрализует до 98% протестированных штаммов ВИЧ-1, блокирующим CD4-связывающий участок белка gp120 [3].

В настоящий момент значительный прогресс был достигнут в идентификации анти-ВИЧ антител, которые обладают широкой нейтрализующей активностью. Однако многочисленные попытки разработать иммуноген, индуцирующий кросс-реактивные антитела к ВИЧ-1 не были успешны, разработанные вакцины-кандидаты не могут стимулировать индукцию нейтрализующих антител против большинства различных штаммов ВИЧ-1, которые распространены в мире [3].

В последние годы было разработано и протестировано большое число ингибиторов проникновения ВИЧ-1 с различными механизмами действия, но только два из них — маравирок и энфувиртид — были одобрены для клинического использования. Однако клиническое применение энфувертида ограничено его относительно низкой активностью, низким генетическим барьером лекарственной устойчивости и коротким периодом полувыведения, а маравирок вовсе взаимодействует с корецептором CCR5 клетки-мишени, а не с молекулярной мишенью [3].

Вышеописанные проблемы подтверждают тот факт, что поиск эффективных лекарственных препаратов с высокой устойчивостью против ВИЧ-1 остается актуальной задачей.

ГЛАВА 2. КОМПЬЮТЕРНОЕ МОДЕЛИРОВАНИЕ ЛЕКАРСТВ

Часто используемом подходом в разработке лекарственных препаратов является компьютерное моделирование, позволяющее отбирать наиболее подходящие соединения для дальнейших испытаний и исследований. Как было сказано ранее, драг-дизайн подразумевает поиск лекарства для конкретной мишени.

Поиск потенциальных лекарств начинается с выбора баз данных лиганд. Затем проводится виртуальный скрининг, результатами которого является некоторое число соединений. Затем это число уменьшается посредствам уточнения расчетов. Таким образом за несколько этапов моделирования будет получено несколько наиболее выгодных в плане энергии связи и других параметров соединений.

Как правило, сначала проводят виртуальный скрининг, затем молекулярный докинг, расчет квантово-химических характеристик и молекулярную динамику.

В данной главе опиши основные этапы компьютерного моделирования и основанного на нем отбора потенциальных лекарств.

1.1 Скрининг

Первым этапом уже непосредственно поиска потенциальных лекарств является виртуальный скрининг. Виртуальный скрининг — это автоматизированная процедура поиска химических соединений, которые обладают заданными физико-химическими свойствами, в различных базах данных.

Для формирование параметров поиска, как правило, используют фармакофорную модель. Фармакофором называют пространственные и сруктурные особенности потенциального лекарсва, которые необходимы для обеспечения оптимальных супрамолекулярных взаимодействий со структурой мишени. Выбор фармакофорной модели для поиска потенциальных лекарств может быть основан на знании о пространственном строении мишени и на данных о взаимодействии уже существующих ингибиторов с рассматриваемым белком.

Виртуальный скрининг использует такую процедуру, как молекулярный докинг, ключевой целью которого является предсказание пространственного строения лиганд-белкового комплекса. Таким образом для большого числа лигандов производиться “грубый” докинг, результатом которого является структура лиганд-ферментного комплекса. Затем при помощи специальных оценочных функций для каждого такого комплекса рассчитывается константа связывания, которая характеризует склонность комплекса к диссоциации. На основе предсказанных констант связывания формируется группа лигандов для дальнейшего исследования.

Также для виртуального скрининга можно использовать правило Липинского [4], так называемого “правило пяти”, которое включает в себя следующие требования:

1. молекулярный вес не должен превышать 500;
2. коэффициент липофильности менее 5;
3. менее 5 потенциальных атомов-доноров водородной связи в структуре;
4. должно быть не более 10 атомов азота и кислорода в структуре лиганда.

Последнее требование является грубой оценкой количества акцепторов водородной связи.

Для проведения виртуального скрининга можно использовать некоммерческих веб сервис Pharmit [5]. Он позволяет проводить поиск потенциальных лигандов посредствам высокопроизводительного скрининга. Pharmit имеет доступ к большому числу баз данных химических соединений, в том числе PubChem [6], которая содержит более 440 миллионов конформеров более чем 90 миллионов молекул.

1.2 Докинг

После проведения виртуального скрининга мы получим некоторый набор лигандов, из которого был выбран набор наиболее релевантных лигандов для дальнейшего изучения. Молекулярный докинг будет следующим этапом отбора химических соединений.

Молекулярный докинг представляет собой процедуру, которая позволяет предсказать наиболее выгодную для образования устойчивого комплекса пространственную ориентацию и положение одной молекулы по отношению к другой. Данный метод молекулярного моделирования еще называют молекулярной стыковкой.

Для проведения докинга существует два концептуально разных подхода. Первый подход называют жестким докингом [7]. В нем для белка и лиганда проверяется геометрическое соответствие (метод взаимозависимости формы). Для этого описывается ряд особенностей белка и лиганда, которые позволяют их стыковать. Второй подход — симуляция. В нем полагают, что лиганд находится на некотором расстоянии от белка, а затем последовательно его приближают. На каждом шаге приближения лиганд перемещается и вращается в пространстве, учитываются внутренние структурные изменения лиганда. После каждого движения энергетическая оценка системы вычисляется заново. Таким образом симулируется реальное поведение молекул, что дает более точные результаты. Такой вариант докинга еще называют гибким докингом. Существенным недостатком гибкого докинга является время, которое он занимает.

Для проведения молекулярного докинга существует достаточно много программного обеспечения. Одним из самых популярных программных пакетов для проведения молекулярного докинга является AutoDock [8], который в основном применяется для проведения белок-лигандного докинга.

1.3 Квантово-химическое моделирование

Квантово-химическое моделирование является следующим этапом в оценке энергетических характеристик соединения. И по результатам квантово-химического моделирования будут отобраны потенциальные лекарства для дальнейшего исследования. Расчеты и моделирование квантово-химических характеристик по сравнению с докингом более точная и высокозатратная в плане вычислительных мощностей процедура.

Для осуществления данных расчетов используют такое понятие как квантово-химическая модель. В химических моделях необходим учет большого числа физических явлений: движения электронов в поле ядер, взаимодействия электронов между собой, релятивистское увеличения массы электронов за счет ускорения вблизи ядра, квантовые эффекты неопределенности положения и импульса, спиновые и другие эффекты [9]. Одновременно учитывать все явление не представляется возможным, поэтому используют математические модели, в которых учитываются наиболее простые и важные эффекты.

Существует множество методов для осуществления данных расчетов. Они деляться на неэмпирические и полуэмпирические. В неэмпирических методах различными приближенными методами решается уравнение Шредингера. В зависимости от приближения и выбранной квантово-химической модели возможно получить результаты разной точности. Но данные методы не всегда оказываются эффективными ввиду громоздкости и сложности.

В полуэмпирических методах за основу берется модель неэмпирических методов и затем, путем замены некоторого числа взаимодействий на подгоночные параметры, модель упрощается, позволяя значительно снизить сложность вычислений. Параметры подбираются на основе сравнения экспериментальных данных с расчетными. Также выбор эмпирических параметров обуславливается опытом неэмпирических методов. Введение параметров способно аппроксимировать влияние внутренних электронов атомов на валентные электроны и другие неточности, возникающие с введением допущений квантово-химической модели. Увеличение производительности расчетов полуэмпирическими методами неоспоримо, однако оно достигается за счет снижения строгости, что может привести к непредсказуемым результатам.

Для проведение расчетов многоатомных химических систем полуэмпирические методы являются самым приемлемым, в плане сложности и точности, способом.

На сегодняшний момент одним из наиболее популярных программных пакетов для квантово-химического моделирования является MOPAC (Molecular Orbital PACkage) [10].

1.4 Молекулярная динамика

Компьютерная молекулярная динамика один из наиболее мощных вычислительных методов моделирования физических систем. Методы молекулярной динамики делают возможным вычислять траектории отдельных атомов и полимерных цепей, а также исследовать динамику взаимодействия частиц в конденсированных системах на молекулярном уровне.

Молекулярная динамика обладает высоким пространственно-временным разрешением и позволяет получить информацию о процессах, происходящих в атомно-молекулярных масштабах и на временах порядка нескольких наносекунд [11]. Молекулярная динамика позволяет теоретически изучить структуру и траекторию различных биологических объектов в пространстве: от функциональных групп низкомолекулярных соединений до макромолекулярных систем. В основе метода лежит расчет ньютоновских траекторий движения макромолекулы и импульсов ее атомов в пространстве координат, а молекула рассматривается как система классических частиц, которые взаимодействуют.

Методы молекулярной динамики требуют огромных вычислительных мощностей, и проведение их без специальных электронных вычислительных машин не является рациональным.

ГЛАВА 3. НЕЙРОННЫЕ СЕТИ

В этой главе будут рассмотрены основные типы нейронных сетей, которые можно применить на практике для решения различных прикладных задач. Технология нейронных сетей используются для решения задач связанных с прогнозированием. Решения основанные на использовании данной технологии показывают хорошие результаты в самых различных областях человеческого знания, таких как распознавание речи, анализ текста, анализ изображений.

3.1 Общая информация о нейронных сетях

Сама по себе искусственная нейронная сеть является математической моделью, построенной с учетом принципов, по которым функционируют биологические сети. Вдохновением для построения моделей нейронных сетей является живой организм, а именно - нервная система. Как по нервной системе посредством электрических возмущений по синапсам нейронов передаются сигналы, которые в совокупности формируют более сложную систему.

Первая простейшая модель искусственной нейронной сети, представленная одним нейроном, была разработана Маккаллоком и Питтсом уже в 1943 году. Данная модель получила название пороговый нейрон. Он вычисляет произведение входного вектора и вектора весов нейрона и, в случае если полученное произведение превосходит некоторое пороговое значение, генерирует на выходе 1, а в противном случае — 0 [12]. В настоящее же время сети умеют решать куда более широкий ряд задач, облегчая жизнь современных людей.

Таким образом, искусственная нейронная сеть функционируют следующим образом. На первый входной слой подается вектор входных данных. Каждый нейрон принимает сигнал и передает импульс следующему слою. Он передает каждому из нейронов на следующем слое сигнал с некоторым весом. Последний слой называется выходным и несет основную информацию, например, вероятность принадлежности определенному классу.

Обучаются такие сети, как правило, методом обратного распространения ошибки (backpropagation), который заключается оптимизации всех весов нейронной сети начиная с последнего слоя до первого, используя градиентный спуск.

Градиентным спуском называется такой метод поиска минимума функции, при котором на каждом шаге выясняется, в каком направлении относительно каждого неизвестного параметра следует двигаться по функции, чтобы достичь её минимума. Скорость такого движения (learning rate) является гиперпараметром, значение которого можно изменять для корректирования результатов обучения. Чем больше значение этого параметра, тем большими шагами будет осуществляться движение по функции ошибки, однако значение параметра не должно быть слишком большим или слишком маленьким. Слишком большое значение может привести к тому, что минимума будет

невозможно достичь, в силу того, что шаг будет слишком велик, а слишком маленькое будет тренировать сеть чрезвычайно долго.

Суть метода градиентного спуска состоит в следующем. На каждом шаге обучения для уточнения значения каждого неизвестного параметра вычисляется частная производная целевой функции ошибки по каждому из всех неизвестных параметров. С её учётом и с учётом скорости обучения происходит пересчёт каждого параметра.

Пересчёт параметров выходного слоя не представляет особой сложности, но чтобы пересчитать параметры слоёв за ним, приходится проходить через нелинейности, производные от которых вносят свой вклад. Это принцип обратного распространения ошибки.

Однако задача поиска глобального минимума функции не так легко разрешима. Скорость обучения может быть слишком маленькой или слишком большой, у целевой функции (в данном случае функции ошибки сети) может быть сложный ландшафт, есть вероятность застревать в локальных минимумах и пр.

Существует много подходов к оптимизации градиентного спуска, таких как накопление импульса, заглядывание вперёд или же, например, введение запрета на значимое изменение значений тех весов, которые слишком часто обновляются и разрешение на изменение в полную силу тех параметров, которые почти совсем не участвовали в обновлении.

3.2 Сверточные нейронные сети

Структуру сверточной нейронной сети можно разделить на два блока: блок свертки (convolution layers) и полносвязный блок (fully-connected layers).

В первом блоке каждый сверточный слой представляет собой набор матриц-фильтров, ядер для свертки. Число таких слоев, фильтров и их размеры выбираются в зависимости от задачи методом математического подбора. В некоторых сетях число слоев может доходить до десятков и даже сотен, но не имея соответствующей вычислительной мощности их использование может быть не оправдано ввиду трудности обучения.

Впервые архитектура сверточной нейронной сети была предложена еще в 1988 году Яном Лекуном [12]. Данная архитектура в первую очередь нацелена на эффективное распознавание образов и является основной частью технологии глубокого обучения (deep learning).

Смысл сверточных сетей состоит в том, что последовательное выполнение операций свертки и пулинга (pooling) уменьшает размерность входных данных, извлекая из них необходимые признаки, которые далее могут быть использованны для решение, например, задач классификации.

Блок состоящий из полносвязной сети по сути может и не использоваться. Чаще всего он является классификатором, который относит объект к тем либо иным классам исходя из вектора признаков, полученным сверточной частью.

3.3 Генеративно-состязательные сети

Генеративно-состязательные сети широко используются в биомедицинской информатике. Они представляют собой две сети, которые обучаются вместе.

Первая сеть — это генератор, задача которой генерировать из случайного шума данные. Вторая сеть — дискриминатор, на вход которому подаются настоящие данные и данные, созданные сетью генератором. Задача дискриминатора отличить реальные данные от искусственных [13].

Данная модель может применяться в конструировании новых химических соединений, когда сеть генератор будет пытаться создавать соединения похожие на реальный, а сеть дискриминатор будет пытаться их различить.

Таким образом две сети будто соперничают, состязаются, генератор пытается создавать все более похожие на реальные данные искусственные данные, а дискриминатор пытается все лучше различать реальные данные от сгенерированных.

Обучив такую модель можно получить сразу два инструмента: генератор и дискриминатор. Генератор из случайного шума создает искусственные данные, что в области медицины является очень полезным ввиду конфиденциальности генетических и медицинских данных, которые необходимы в больших количествах для обучения сетей, решающих задачи биологической и медицинской направленности.

3.4 Автоэнкодер

Автоэнкодеры — это нейронные сети прямого распространения, которые восстанавливают входной сигнал на выходе. Внутри у них имеется скрытый слой, иногда называемый латентный слой, который представляет собой вектор признаков, описывающих входные данные.

Автоэнкодеры конструируются таким образом, чтобы не иметь возможность точно скопировать вход на выходе. Обычно их ограничивают в размерности вектора признаков, он имеет меньшую размерность чем входной вектор данных, или штрафуют за активации. Входной сигнал восстанавливается с ошибками из-за потерь при кодировании, но, чтобы их минимизировать, сеть вынуждена учиться отбирать наиболее важные признаки.

Таким образом автоэнкодеры состоят из двух частей: энкодера и декодера. Энкодер сжимает входные данные в вектор признаков на латентном слое, а декодер восстанавливает из вектора признаков сами данные. Следовательно сеть сама кодирует и декодирует данные [13].

Автоэнкодеры могут применяться с различными целями. Например, выявив основные признаки и восстановив по ним данные, можно избавиться от шума, или тренирую определенным образом, можно заставить автоэнкодер преобразовывать входные данные необходимым образом.

3.5 Градиентный спуск

Градиентный спуск является методом поиска минимума функции. Основной принцип которого заключается в том, чтобы на каждом очередном шаге продвигаться в сторону глобального минимума функции в правильном направлении по каждому ее параметру. Размер шага, или скорость движения, — это изменяемый параметр. Чем больше значение этого параметра, тем большими шагами мы будем продвигаться по функции ошибки, однако значение параметра не должно быть слишком большим или слишком маленьким. Слишком большое значение не сможет дать возможность правильно подбирать весовые коэффициенты модели, а слишком маленькое может привести к застреванию в локальных минимумах и требует больше вычислительных ресурсов для обучения модели.

Далее рассмотрим алгоритм градиентного спуска. На каждом шаге обучения вычислим частную производную от целевой функции ошибки по каждому неизвестному параметру. Далее, учитывая значение производной и скорость обучения, произведем пересчет каждого весового коэффициента. Ниже приведены формулы классического градиентного спуска:

$$\Delta\theta = -\eta * \nabla_{\theta} * J(\theta) \quad (1)$$

$$\theta_{next} = \theta_{current} + \Delta\theta \quad (2)$$

где θ — параметры сети, $J(\theta)$ — функция потерь в случае машинного обучения, а η — скорость обучения [28].

Если применять градиентный спуск для оптимизации весовых коэффициентов моделей нейронных сетей, то для их пересчета необходимо проходить через нелинейности, производные от которых вносят свой вклад. Это принцип обратного распространения ошибки.

Однако, задача поиска глобального минимума функции нетривиальна. Необходимо эмпирически подбирать правильную скорость обучения, так как у целевой функции ошибки может быть сложный ландшафт, и есть вероятность, например, застрять в локальном минимуме.

3.6 Оптимизаторы градиентного спуска

Существует много подходов к оптимизации градиентного спуска. Каждый подход имеет свои особенности и может быть применен с разного рода функциями потерь. Рассмотрим три из таких подходов к оптимизации градиентного спуска.

3.5.1 RMSProp

RMSProp (Root Mean Square Propagation) — среднеквадратичное распространение, является некоторой модификацией еще одного известного алгоритма AdaGrad (Adaptive Gradient) [23]. Суть метода заключается в том, что с каждой итерацией оптимизирующего алгоритма имеется возможность накапливать по каждому из обновляемых параметров своеобразный импульс. То есть можно отслеживать насколько часто и насколько сильно каждый

обновляемый параметр подвергается изменению. У часто изменяющегося параметра, как следствие, будет накапливаться большой импульс, что при последующих шагах алгоритма будет учтено, и далее этот параметр не будет подвергаться столь частым изменениям. В данном методе была заимствована основная идея метода адаптивного градиента, однако в нем происходит накопление полного импульса всех параметров, что очень быстро может привести к тому, что все параметры почти перестают обновляться. В алгоритме RMSProp используется среднеквадратичное накопление изменений. Отсюда и его название.

3.5.2 SGD

SGD (Stochastic Gradient Descent) — один из типичных видов градиентного спуска — стохастический. Само прилагательное “стохастический” предполагает наличие какой-либо произвольной вероятности для системы или модели. В данном случае при наличии обучающей выборки шаг работы оптимизационного алгоритма зависит от произвольным образом выбранных элементов тренировочных данных. В данном подходе есть интуитивное преимущество в количестве вычислений в отличие от, например, пакетного (batch) градиентного спуска. Использование последовательно с каждой итерацией всего набора данных действительно полезно в задаче достижения глобального минимума. Однако, проблема возникает, когда наборы данных становятся действительно огромными. При оптимизации целевой функции с набором данных размеров, превышающих несколько миллионов, нельзя не использовать весь имеющийся набор в самой классической технике градиентного спуска. За этим следует проблема использования большого количества вычислительных ресурсов, и как следствие, большое время на оптимизацию [24].

Вычислительно очень дорогой типичный подход было решено модифицировать путем внедрения произвольного рандомизированного выбора очередного фиксированного пакета из выборки. Тем не менее, стохастический градиентный спуск не всегда может оказаться хорош ввиду того, что для оптимизации зачастую требуется не намного меньше вычислительных ресурсов, и число итераций для оптимизации целевой функции все равно, в общем случае, остается довольно большим.

3.5.3 Adam

Алгоритм Adam (Adaptive moment estimation) — является широко используемым алгоритмом оптимизации градиентного спуска, применяемый в самых разных моделях, будь то, например, полносвязная модель нейронной сети, или же сверточная. Этот алгоритм базируется и является комбинацией двух выше рассмотренных алгоритмов: RMSProp и SGD.

Данный метод адаптивной скорости обучения. Это значит то, что скорость изменения параметров адаптивно подбирается для каждого из них. Само

название говорит об использовании адаптивной оценки момента. Используются моменты первого и второго порядков вектора градиента.

Данный подход сочетает в себе идею накопления движения и идею более слабого обновления весов для типичных признаков. Его особенность состоит в том, что накапливаются не приращения градиента в качестве импульса, а его значения [25].

Отметим некоторые свойства алгоритма Adam. В первую очередь, фактический размера шага изменения каждого весового коэффициента сверху ограничен заданной скоростью обучения. Во-вторых, в правиле обновления Adam размер шага не зависит от величины градиента. Это значительное преимущество, которое помогает при прохождении областей с очень маленькими градиентами, будь то седловые точки или овраги в ландшафте целевой функции потерь. Заимствование подходов SGD помогает быстро перемещаться по таким областям. И в третьих, Adam является комплексным алгоритмом с идеями AdaGrad, который хорошо работает с разреженными градиентами, и RMSProp, который эффективно работает по факту с ходом каждого шага алгоритма для каждого подвергшегося изменению весового коэффициента. Наличие всех описанных преимуществ позволяет использовать Adam для широкого круга оптимизационных задач в различных моделях.

Тем не менее, данный алгоритм, как и все существующие, не может быть применим как эталонный подход ко всем задачам. Существует несколько модификаций данного алгоритма, как, например, NAdam (Nesterov Adam), сочетающий в себе целиком идею вышеизложенного подхода, а также накопление нестеровского импульса, или же ND-Adam (Normalized Direction-Preserving Adam) [26] — нормализованный Adam, сохраняющий направление, борющийся с проблемами стохастического градиентного спуска.

3.7 Метрики

Решения задач машинного обучения не имеет фактического смысла, если не представляется возможным оценить насколько хорошо была обучена модель. Для разных задач используются разные метрики.

Простейшими метриками можно считать те, что используются в задачах классификации. Целью задач классификации является ответить на вопрос принадлежит ли объект к определенному классу или нет. Очевидно, что классификатор может давать истинные и ложные ответы, которые также делятся на две группы. Таким образом получаем четыре группы ответов: истинные положительные (TP — true positive), истинные отрицательные (TN — true negative), ложные положительные (FP — false positive) и ложные отрицательные (FN — false negative). Данные значения для удобства и понимания целостной картины обученной модели записывают в специальную матрицу (confusion matrix) [32].

С использованием данных значений для классификатора можно ввести следующие метрики: accuracy, recall, precision. Существуют и другие, рассмотрим самые часто встречаемые.

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

При решении задач классификации метрику accuracy зачастую используют по умолчанию. Это обусловлено тем, что она интуитивно понятна всем. Значение данной метрики означает какая часть объектов была классифицирована правильно [32]. Однако в задачах, где количество объектов принадлежащих разным классам существенно отличается, данная метрика будет не совсем корректно отражать качество обученной модели.

$$recall = \frac{TP}{TP+FN} \quad (4)$$

Метрика recall показывает какую часть объектов с позитивной меткой модель смогла правильно классифицировать [32]. Данная метрика может быть применена, например, в задаче определения является ли человек больным или нет. Вероятность диагностировать заболевание у здорового человека не так плоха, как вероятность не диагностировать заболевание у больного.

$$precision = \frac{TP}{TP+FP} \quad (5)$$

Precision показывает сколько из всех объектов, которые были классифицированы как положительные, являются положительными на самом деле [32]. Примером использования может послужить задача фильтрации спам сообщений, так как пользователю действительно не хотелось бы, чтобы какое-то из его важных сообщений определилось как спам.

Далее рассмотрим метрики, которые можно использовать для оценки результатов регрессионных моделей. В задачах регрессионного типа выходным значением является некоторое вещественное число. И для того, чтобы оценить качество предсказаний нужно ответить на сколько это предсказанное число близко к реальному значению.

Самой простой метрикой в случае регрессионной задачи является средняя абсолютная ошибка (MAE — mean absolute error), которая вычисляется как средняя сумма абсолютных разниц между фактическими значениями и прогнозируемыми значениями [31].

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |original_i - predict_i| \quad (6)$$

Чаще используют среднюю квадратическую ошибку (MSE — mean squared error).

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (original_i - predict_i)^2 \quad (7)$$

Данная метрика считается как средняя сумма квадрата разницы фактического значения с прогнозируемым [31], что по сути выделяет большие ошибки над меньшими, более точно характеризуя качество обученной модели. Более того, данная функция является дифференцируемой, что позволяет более

эффективно использовать математические инструменты для минимизации. Чем меньше mse, тем качество предсказаний лучше. Если значение mse равно нулю, то модель работает идеально.

Среднеквадратическая ошибка (RMSE — root mean squared error) является модернизацией mse и представляет собой корень квадрата ошибки [31].

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(original_i - predict_i)^2} \quad (8)$$

Метрика rmse удобнее интерпретируется, чем mse, так как она имеет те же единицы измерения, что и исходные данные.

У данных метрик есть определенная проблема связанная с не нормированностью их. То есть глядя на них нельзя сказать насколько точны предсказанные значения относительно исходных значений. Так для одной задачи, которая оперирует большими числами, значение mae может совпадать со значением mae для другой задачи, которая оперирует маленькими числами. Но это не значит, что качество двух моделей будет одинаково. Так для задачи с большими числами mae будет невелика относительно ее значений, а для задачи с малыми числами наоборот.

Для получение нормированой оценки точности модели можно использовать, например, коэффициент детерминации или, как его еще называют, R-квадрат (R^2 score). Значение коэффициента детерминации находятся в диапазоне от 0 до 1; 0 указывает на то, что связь между исходными и прогнозируемыми значениями отсутствует, а в случае с 1, наоборот, модель работает очень хорошо [31].

ГЛАВА 4. ГЕНЕРАЦИЯ ЛЕКАРСТВЕННЫХ СОЕДИНЕНИЙ ПРИ ПОМОЩИ НЕЙРОННЫХ СЕТЕЙ

В данной главе речь пойдет про использования искусственных нейронных сетей для генерации новых потенциальных лекарственных соединений. Рассмотрим уже существующую технологию использования состязательного автоэнкодера для поиска лекарственных соединений в области онкологии и предложим свою модернизацию данного способа для генерации лекарственных соединений против ВИЧ-1.

4.1 Применение состязательного автоэнкодера для создания новых лекарств в области онкологии

Последние достижения в области глубокого обучения и, в частности, в генеративно-состязательных нейронных сетях продемонстрировали удивительные результаты в воспроизведении изображений, видео, аудио достаточно близким к реальным, не искусственным. Данная технология нашла свое применения даже в области обработки естественных языков.

Несмотря на многочисленные достижения в области биомедицинских наук, производительность и эффективность прикладного программного обеспечения для исследований и разработок в фармацевтической промышленности снижается. Менее 10% полученных данным способом лекарственных препаратов допускаются далее для клинических испытаний для всех категорий заболеваний. Для онкологических заболеваний это число составляет вовсе 5,1% вероятность одобрения. Одной из основных причин такой высокой частоты отказов является неэффективный процесс первичных доклинических испытаний. Для оптимизации поиска лекарственных препаратов осуществлялись попытки модификации и комбинирование скрининга баз данных химических соединений с целью кластеризации их на соответствующее лекарственные группы. Однако такой скрининг остается слепым поиском. Несмотря на многие предыдущие неудачи, подходы на основе in silico обещают привлекательную альтернативу для расширения возможностей промышленности с помощью более эффективных методов скрининга, способных обеспечить более надежные результаты при меньших затратах и временных рамках.

Технологии в основе с генеративно-состязательными нейронными сетями также нашли свое применения в области разработки лекарственных препаратов. Далее рассмотрим один из существующих подходов применения глубокого обучения в разработке лекарственных препаратов.

Исследователи поставили перед собой задачу при помощи глубокого обучения создать модель, позволяющую генерировать новые соединения, которые будут обладать противоонкологическими свойствами.

Была предложена архитектура состязательного автоэнкодера для решения данной задачи (см. рис. 4.1).

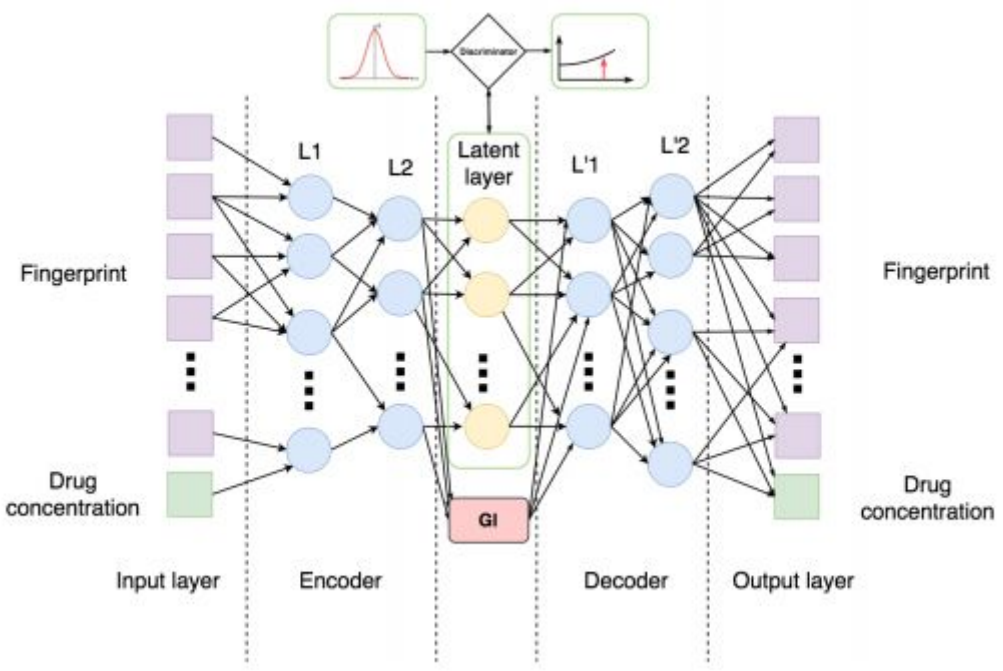

Рис. 4.1. Модель состязательного автоэнкодер для создания новых лекарств в области онкологии [14].

В качестве входа состязательного автоэнкодер используется вектор бинарных фингерпринтов, специальный формат представления пространственной структуры химического соединения, и логарифмическая концентрация вещества. На выводе декодера также получаем концентрацию и вектор, состоящий из вероятностей того, что на данном месте в новом фингерпринте будет находиться еденица. В латентный слой, промежуточный слой с вектором признаков, был также добавлен нейрон, который ответственен за то, как хорошо соединение будет ингибировать мишень (GI). Отрицательные значения GI указывают на уменьшение количества опухолевых клеток после медикаментозного лечения. Данный состязательный автоэнкодер обучался на наборе из 6252 фингерпринтов, данных о концентрации и качестве ингибирования. Далее было отобрано 640 векторов латентного слоя для дальнейшей генерации при помощи декодера фингерпринтов. Было получено 32 вектора с вероятностями, по которым были созданы фингерпринты. Проводился скрининг по базе данных из 72 миллионов соединений, полученных из Pubchem. Использовалась функция максимального правдоподобия, чтобы выбрать 10 лучших совпадений для каждого из 32 фингерпринтов. Было получено 69 уникальных соединений. Для оценки биологическую значимость результатов, соединения идентифицировались при помощи базы данных Pubchem BioAssay, и анализировалась на основе уже полученных экспериментальных или моделированных данных их противораковая активность и другие соответствующие биомедицинские свойства, представляющие интерес. Полученный набор данных включал в себя несколько запатентованных соединений. Но информация о потенциальном лекарственном действии и противораковой активности доступна не для всех

соединений. Однако, некоторые из них уже известны как противораковые агенты различных видов. Большинство из этих соединений связаны с антрациклинами (или антрациклиновыми антибиотиками). Антрациклины используются в химиотерапии рака для лечения многих видов рака, включая лейкемии, лимфомы, рак желудка, матки, яичников, рак молочной железы и рак легких [14].

Этот подход является доказательством концепции искусственно-интеллектуального механизма обнаружения лекарств, в котором состязательный автоэнкодер используется для создания новых молекулярных фингерпринтов, соответствующих желаемыми химическим и биологическим свойствам.

4.2 Подготовка данных

В качестве данных для обучения мы имеем набор файлов в формате pdbqt, которые представляют собой пространственную структуру соединения и рассчитанное значение энергии связи комплекса лиганда с белком оболочки ВИЧ-1 gp120.

На вход нейронной сети мы не можем подавать пространственные структуры молекул в форматах pdb, pdbqt, как это происходит для докинга. Модели машинного обучения требуют более нормализованные входные данные, имеющие однородную структуру и постоянную длину. Подходящим для таких целей форматом для записи информации об молекуле являются молекулярный фингерпринт. В данной работе будем использовать фингерпринты формата MACCS. Фингерпринты кодируют свойства пространственной структуры молекул. Они могут использоваться в таких задачах, как поиск сходства молекул, молекулярная характеристика, молекулярное разнообразие и химическая кластеризация базы данных. MACCS фингерпринты — это 166-битные структурные дескрипторы ключей, в которых каждый бит связан с некоторым шаблоном [15].

Датасет был сформирован сначала преобразовывая молекулы формата pdb в формат SMILES (Simplified Molecular Input Line Entry System). Данное представление является строковым и включает в себя специальные символы двух типов: атомы и связи между ними, записывая таким образом структурное представление об молекуле [16]. И затем при помощи программного пакета RDKit [15] преобразован в MACCS фингерпринты.

Удобным инструментом для перевода из одного химического формата в другом является OpenBabel [17]. Он представляет собой свободную химическую экспертную систему, в основном используемая для преобразования форматов файлов. Имеет как графический интерфейс так и возможность использовать через командную консоль.

Для реализации перевода из pdbqt формата в SMILES был написан скрипт на python 3, с использованием функционала OpenBabel. Каждое соединение хранится в отдельной папке, которая первоначально содержала pdbqt файл и

результаты докинга. Скрипт создает для каждого соединение файл формата smi, где записан SMILES для каждого соединения.

Следующим этапом необходимо для каждого соединения по его SMILES записи генерировать фингерпринт, которые в дальнейшем будут подаваться на вход нейронной сети. OpenBabel также позволяет переводить SMILES в фингерпринты, но входе использования возникли сложности с интерпретацией результатов, и было принято решение использовать библиотеку RDKit [15].

Был написан скрипт, позволяющий автоматически распарсить результаты докинга, выделив энергию связи из выходного файла AutoDock, и перевести SMILES в массив из нулей и единиц, представляющий собой MACCS фингерпринт. Особенностью RDKit является то, что при переводе фингерпринт получается из 167 битов, это обусловлено тем, что нулевой бит всегда равен нулю, так как нумерация битов фингерпринта начинается с единицы. Поэтому было необходимо отбросить нулевой бит.

Таким образом входные данные представляют собой бинарный массив длинны 166. Датасет включает в себя 82710 соединений преобразованных в формат MACCS с соответствующим им энергиями связи для комплекса с молекулой мишенью.

4.3 Автоэнкодер для генерации лекарственных соединений ВИЧ-1

За основу была взята модель состязательного автоэнкодера, который был рассмотрен в предыдущем пункте. Однако, в нашем случае существует несколько концептуальных отличий, которые необходимо учесть в проектировании сети.

В качестве входных данных мы не имеем концентрацию вещества. Так что входной слой, на данном этапе, будет представлять собой вектор из 166 значений, представляющий собой бинарный фингерпринт, сгенерированный для каждого химического соединения.

Латентный слой в рассмотренном примере содержал специальный GI нейрон, отвечающий за качество соединения. Качество соединения оценивалось тем, как хорошо оно будет ингибировать выбранную мишень. В случае с нашим автоэнкодером таким показателем будет являться энергия связи. Таким образом на вход декодеру в нашей модели будет подаваться также энергия связи.

Важно отметить, что нейронная сеть из рассмотренной статьи использовалась только для генерации новых фингерпринтов при помощи декодера. Мы также можем использовать нашу модель для генерации фингерпринтов путем подачи случайных чисел на вход декодеру. Однако, мы также предполагаем следующее использование: подать энкодеру некоторую химическую структуру, а декодеру дать энергию заведомо лучшую, чем у исходного соединения. Таким образом, декодер попробует улучшить структуру соединения для получения лучшего результата.

Таким образом получаем следующую структуру сети.

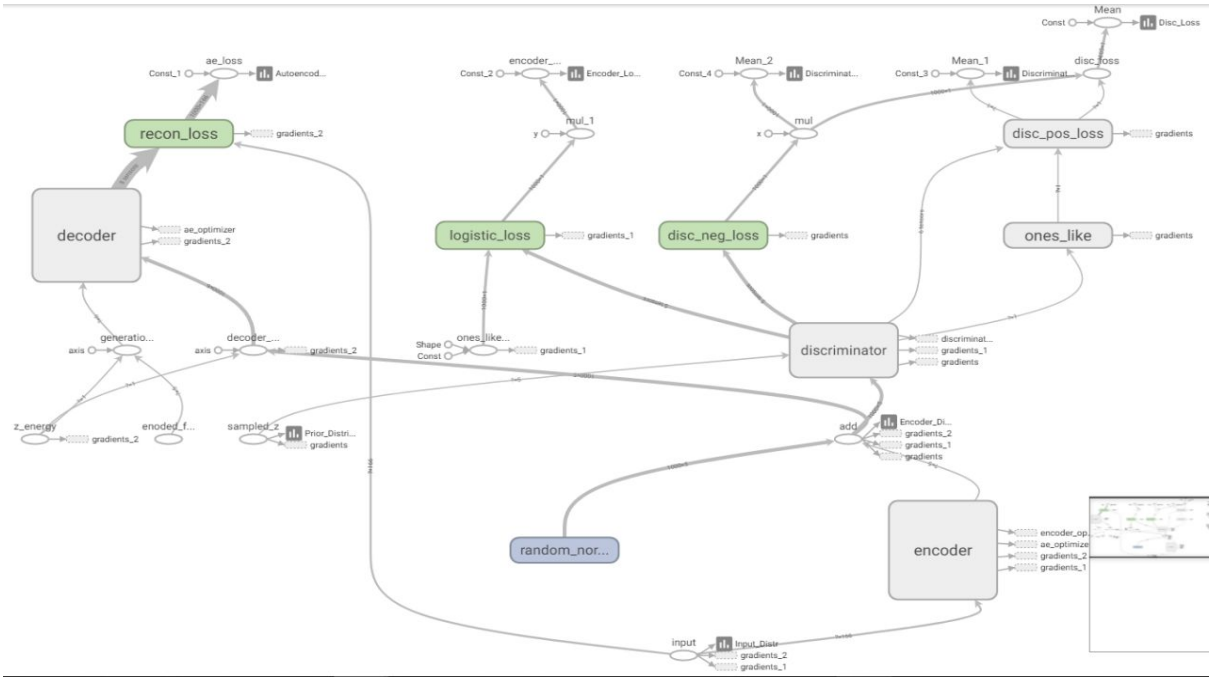

Рис. 4.3. Модель состязательного автоэнкодер для создания новых лекарств ВИЧ-1.

Описанные выше изменения классической модели автоэнкодера привели к изменению способа обучения состязательного автоэнкодера. Классическое трехступенчатое обучение автоэнкодера для данной модели превратилось в пятиступенчатое.

Первым этапом дискриминатор обучается распознавать заданное скрытое распределение и закодированные представления. Затем энкодер учится путать дискриминатор. Данный способ обучения представляет обучение генеративно-состязательной сети. Однако, на следующем этапе энкодер и декодер обучаются совместно как автоэнкодер.

4.4 Интерпретация результатов

Для рассмотренного выше состязательного автоэнкодера существует проблема интерпретации результатов. Когда как на вход подаются конкретные соединения переведенные в формат фингерпринтов, то на выходе получается вектор из 166 бит, который является фингерпринтом неизвестного соединения. Зная пространственную структуру либо SMILES можно вычислить фингерпринт для данной молекулы, но, так как один фингерпринт может соответствовать разным соединениям, определить по нему молекулу не представляется возможным. Для решения этой задачи можно преподсчитать фингерпринты для интересующего нас класса соединений и осуществлять поиск по данной базе.

Имеет смысл наполнять данную базу только теми соединениями, которые можно будет использовать в качестве лекарства. Было принято решение использовать базу данных белков ZINC [18], которая насчитывает более миллиарда соединений. Данный ресурс позволяет выбрать определенную категорию соединений, для наших целей подходит категория потенциально лекарственных (drug like).

ZINC позволяет скачать соединения в разных форматах. Наиболее подходящим для наших целей является SMILES, так как его просто перевести в формат фингерпринтов и он занимает меньше места по сравнению с форматами содержащие пространственную структуру молекулы.

В качестве базы данных была выбрана реляционная база данных Ms SQL Server [19]. Таблица с соединениями имеет 3 колонки: zind_id (уникальный номер присвоенный в базе данных zinc), smiles (соединение в формате SMILES) и fp (фингерпринт представляющий собой массив бит).

Был написан скрипт на python, который скачивает соединения по специальным ссылкам ZINC, затем конвертирует их в фингерпринты и записывает в базу данных.

Таким образом в базу было добавлено 166 416 851 соединений и для каждого из них был рассчитан фингерпринт.

Так как база данных содержит достаточно большое число соединений, то для ускорения поиска был создан некластеризованный индекс на колонку с фингерпринтом, что позволяет искать соеденения соответвующие данному фингерпринту не за O(n), а за O(log(n)).

Однако может быть такая ситуация, что соединения с точно таким же фингерпринтом в базе не обнаружено. Тогда возникает необходимость искать максимально похожие. Для массива байт нет смысла придумывать какую-то особенную меру “похожести”, поэтому можно использовать просто расстояние Хэмминга (количество несовпадающих символов) [20]. Данный поиск занимает гораздо большее время, так как необходимо перебрать все соединения в базе.

Схему базы можно оптимизировать для такого поиска. Исходная база данных ZINC содержит соединения вместе со своими конформерами, которым соответствуют одни и те же молекулярные фингерпринты. Для того, чтобы не искать расстояния для одних и тех же фингерпринтов, можно добавить таблицу, которая будет содержать одну колонку — фингерпринт. Поиск похожих будет осуществляться по данной таблицы. Затем, когда наиболее подходящие под запрос фингерпринты из базы будут найдены, для сопоставления их реальным соединениям можно сделать объединение таблиц. Для оптимизации операции объединения, в таблице, которая содержит фингерпринт и идентификационный номер из ZINC, можно создать некластеризованный индекс по колонке с молекулярными фингерпринтами.

Для решения задачи интерпретации результатов сгенерированных при помощи состязательного автоэнкодера можно использовать следующий подход. Вектор полученный на выходе сети имеет не дискретное распределение из нулей и единиц, на некоторых позициях могут стоять значение одинаково близкие и к нулю и к единице. При поиске по базе данных можно перебрать разные варианты интерпретации данного вектора, что по времени может оказаться выгоднее, чем поиск всех похожих на один конкретный.

Подробнее о результатах проведенной работы можно узнать в статье [1], которая также содержит перечень потенциальных лекарственных препаратов, которые были найдены при помощи базы данных молекулярных фингерпринтов на основе результатов работы состязательного автоэнкодера.

ГЛАВА 5. РАЗРАБОТКА АЛГОРИТМА ПРЕДСКАЗАНИЯ ИНГИБИТОРНОЙ АКТИВНОСТИ ХИМИЧЕСКИХ СОЕДИНЕНИЙ

В данное главе рассмотрим другое применение технологии машинного обучения в разработке новых лекарственных соединений. Рассмотрим несколько существующих оценочных функций базирующихся на методах машинного обучения. И разработаем свой способ оценки ингибиторной активности химического соединения.

5.1 PoseScore и RankScore статистические функции оценки

Функции оценки белка-лигандного комплекса требуются для различных приложений структурной биологии и медицинской химии. Данная функция используется в двух различных задачах: ранжирование различных положений маленькой молекулы в сайте связывания белка и ранжирование различных малых молекул по их комплементарности с сайтом белка.

Используя теорию вероятностей, были разработаны две функции статистического скоринга, зависящие от атомного расстояния: PoseScore был оптимизирован для распознавания нативной геометрии связывания лигандов, а RankScore был оптимизирован для различения лигандов от несвязывающих молекул.

Обе оценки основаны на датасете из 8 885 кристаллографических структур белково-лигандных комплексов, но отличаются значениями трех ключевых параметров. Были исследованы факторы, влияющие на точность оценки, включая максимальное атомное расстояние и геометрию ненативных лигандов, используемых для оценки, а также использование белковых моделей вместо кристаллографических структур для обучения и тестирования функции оценки. Для набора тестов из 19 мишеней RankScore улучшил баллы по обогащению лигандов (logAUC) и по раннему обогащению (EF1), рассчитанные с помощью DOCK 3.6 для 13 и 14 мишеней, соответственно. Кроме того, RankScore показала лучшие результаты при восстановлении, чем каждая из семи других протестированных функций оценки. Принимая как кристаллическую структуру, так и геометрию активных сайтов со среднеквадратичными отклонениями всех атомов до 2 Å от кристаллической структуры в качестве правильных положений связывания, PoseScore дал лучший результат [21].

Точность PoseScore сопоставима с точностью DrugScoreCSD и ITScore / SE и превосходит 12 других протестированных функций оценки. Таким образом, RankScore может облегчить обнаружение лигандов путем ранжирования комплексов мишени с различными маленькими молекулами. PoseScore можно использовать для прогнозирования сложной структуры белок-лиганд путем ранжирования различных конформаций данной пары белок-лиганд.

5.2 NNscore: оценочная функция для характеристики белок-лигандных комплексов

Важным после проведения виртуального скрининга и высокопроизводительного докинга является правильная оценка результатов. Для оценки результатов используют так называемую оценочную (scoring) функцию. NNscore представляет собой оценочную функцию основанную на нейронной сети, которая позволяет как самостоятельно, так и в сочетании с другими более классическим скоринг функциями производить качественную оценку моделируемых лиганд-белковых комплексов.

Текущие скоринговые функции делятся на три основных класса. Первый класс, основанный на молекулярных силовых полях, предсказывает энергию связи, явно оценивая электростатические и Ван-Дер-Ваальсовые силы. Программами для проведения докинга, которые используют функцию оценки основанную на силовых полях, являются AutoDock, Dock, Glide. Вторым типом оценочных функций являются эмпирические функции оценки. Они оценивают энергию связи, вычисляя взвешенную сумму всех водородных связей и гидрофобных контактов. И третьим типом можно назвать оценочную функцию основанную на статистических данных. Данные оценочные функции опираются на статистический анализ баз данных кристаллической структуры. Пары типов атомов, которые часто встречаются в непосредственной близости, считаются энергетически выгодными. Примеры включают в себя статистический потенциал Astex и функцию статистической оценки SF.

Для создание оценочной функции нового типа было предложено использовать нейронные сети. Все нейронные сети имеют как минимум два слоя. Первый, называемый входным слоем, получает информацию о системе, которую должна проанализировать сеть. Второй, называемый выходным слоем, кодирует результаты этого анализа. Кроме того, дополнительные скрытые слои получают входные данные от входного слоя и передают его на выходной слой, что обеспечивает еще более сложное поведение.

Для решение поставленной задачи NNScore использует в качестве входного слоя вектор из 194 параметров. Данными 194 признаками разработчики NNscore охарактеризовали лигад-белковый комплекс. На выходе есть всего два нейрона для того, чтобы охарактеризовать качество соединения. Выход (1, 0) указывает, что данный комплекс белок-лиганд имеет константу диссоциации Kd < 25 мкМ, и (0, 1) указывает, что данный комплекс имеет Kd > 25 мкМ [22]. Данное пороговое значение было получено из экспериментальных данных и является разумным для разграничения ингибиторов, которые могут быть потенциально эффективными.

Для обучения данной сети использовалось 4141 лиганд-белковых комплексов, которые были загружены из базы данных белков PDB. После тренировки одной сети и оценив ее результаты, разработчики NNScore обучили

дополнительные сети для увеличения точности. 10 сетей с обучающими наборами из 4000 случайно выбранных комплексов были созданы в предварительных исследованиях, описанных выше. Чтобы определить, можно ли обучить еще более точные сети, были созданы дополнительно 1000 независимых нейронных сетей с аналогичными обучающими наборами из 4000 случайно выбранных белково-лигандных комплексов. В каждом случае оставшиеся 141 комплекс снова использовались в качестве проверочного набора. Три из этих 1000 сетей оказались наиболее точными (точность набора достоверности 89,4%); 24 имели достоверность, установленную для валидации, более 87,5%. Каждая сеть уникальна и согласованные результаты можно получить при рассмотрении среднего прогноза для нескольких сетей. Таким образом проведя усреднение результатов для вышеописанных 24 сетей определили единую оценку, которую назвали NNScore.

NNScore используется для переоценки результатов докинга, то есть уже после проведения расчетов высокопроизводительного докинга.

5.3 Регрессионная задача — оценка ингибиторной активности химического соединения

Теперь сформулируем задачу. Имеется набор химических соединений, требуется предсказать их ингибиторную активность.

В качестве входных данных будем использовать молекулярные фингерпринты. Ингибиторную активность будем измерять энергией связи. Энергия связи представлена некоторым рациональным числом.

Мы уже имеем подходящий набор данных, который был создан для обучения состязательного автоэнкодера (см. 4.2). Данный датасет состоит из 82710 соединений в формате MACCS и соответствующей их комплексу с белком оболочки ВИЧ-1 gp120 энергией связи.

Данная задача приобретает вид регрессионной, где для набора входных данных требуется предсказать некоторое рациональное число, выходные данные. В нашем случае по 166 бинарным признакам требуется предсказать одно выходное рациональное значения — энергию связи.

5.4 Регрессионные модели машинного обучения

Далее рассмотрим основные регрессионные модели машинного обучения, которые можно применить для решения задачи прогнозирования ингибиторной активности химических соединений, описанной в предыдущей главе.

5.4.1 Линейная регрессия

Линейная регрессия — одна из простейших моделей машинного обучения, предназначенная для решения регрессионных задач. Целевой величиной является действительное число, в нашем конкретном случае потенциальной энергией связи. Данная линейная модель подразумевает линейные отношения между входными параметрами и выходным значением. В общем случае

выходное значение является линейной комбинацией входного вектора и параметров модели, которые еще называют весами (регрессорами).

Линейная модель множественной регрессии описывается следующим уравнением:

$$y(x) = \omega_0 + \omega_1 x_1 + \omega_2 x_2 + ... + \omega_{42} x_{42} + ... + \omega_n x_n \quad (9)$$

Данное уравнение представляет собой уравнение гиперплоскости. В простейшем случае при наличии единственного входного параметра и выходного параметра гиперплоскостью будет являться прямая, а при наличии двух входных параметров — плоскостью. Задача состоит в том, чтобы подобрать коэффициенты (веса в нашем случае) гиперплоскости так, чтобы она наилучшим образом описывала данную входную выборку.

Для построения такой гиперплоскости используют построение функции стоимости. Разные значения весов регрессионной модели дают нам различные гиперплоскости, а задача состоит в построении наиболее подходящей. Для сравнения различных построенных плоскостей и используется функция стоимости. То есть для того, чтобы понять, какая из имеющихся лучше. Или же для выяснения того, насколько близкое к правильному значению выходного параметра дают подобранные веса.

Для линейной регрессионной модели зачастую используется среднеквадратичная ошибка (MSE - mean squared error). Данная функция стоимости — это среднее всех квадратичных ошибок между всеми наборами (x, y) входных данных. Это является своеобразным расстоянием или стоимостью между требуемым ответом и предсказанной моделью величиной.

Для каждой пары (x, y) и имеющимися весами w можно посчитать невязку (w(x) - y). В идеальном случае, требуется получения равенства w(x) = y для каждого набора из выборки. Это значит, что требуется минимизировать функцию стоимости, ее еще называют функцией ошибки или функцией потерь. Тогда регрессионная модель будет абсолютно точной. Зачастую на реальных данных такого добиться практически невозможно. Тогда речь идет о минимизации невязки, то есть минимизации несоответствия между предсказанной величиной и имеющейся требуемой [27].

Данная минимизация достигается за счет, например, применения метода градиентного спуска (см. 3.5) в совокупности с оптимизаторами градиентного спуска (см. 3.6).

5.4.2 Регрессионное дерево решений

Все регрессионный техники имеют единственный ожидаемый выходной параметр от одного или многих входных параметров. Выходным параметром является число. В общем случае, методология построения регрессионного дерева позволяет входному набору параметров быть смесью некоторого непрерывного набора, как например в задаче предсказания ингибиторной активности химических соединений. Рассматривая же дерево решений, отметим, что оно строится таким образом, что каждая вершина (узел) в дереве

содержит тест или вопрос для значения какой-либо конкретной переменной из входного элемента. Конечные узлы дерева, то есть его листья, содержат предсказанные значения выходных переменных.

Регрессионное дерево также может рассматриваться как вариант дерева решений, но предназначенный для аппроксимации вещественных функций вместо использования его как метода классификации.

Способ его построения называется двоичным рекурсивным разбиением. Это итеративный процесс, который работает разбивая данные на ветви, а затем продолжает разделять каждый вновь образовавшийся раздел на еще более мелкие, по мере рекурсивного возвращения к корню. На первом этапе алгоритма построения регрессионного дерева все данные заносятся в один исходный раздел. Далее алгоритм разделяет выборку на две части, используя все возможные двоичные разделения по каждому параметру, выбирая разделение, минимизирующее сумму квадратов отклонений от среднего в двух образовавшихся частях. Это правило разделения применяется к вновь образовавшимся ветвям. Процесс продолжается до тех пор, пока каждая вершина не достигнет заранее определенного минимального размера, или же когда сумма квадратов отклонений от среднего не будет равной нулю. В таком случае вершина назначается листом по умолчанию.

Существует также после построения дерева необходимость в его обрезании. Данная необходимость возникает ввиду того, что дерево было выращено целиком на тренировочном наборе данных и велика вероятность переобучения на входном наборе. Это переобучения сказывается на последующем применении дерева на тестовых выборках и в эффективности использования в целом. Дерево подвергается обрезанию с помощью валидационной выборки и основывается на принципе, в котором первым удалению подвергается тот узел дерева, которым был образован последним [29].

5.4.3 Случайный регрессионный лес

Random forest (случайный лес) включает в себя использование деревьев решений, описанных в предыдущей главе. Случайный лес состоит из ансамбля или комитета деревьев решений. По отдельности деревья могут давать не очень точные результаты, однако, используя сразу несколько деревьев, предсказательную мощность такой регрессионной модели можно значительно увеличить.

Для обучения такой модели тренировочная выборка делится на случайные наборы, и каждое дерево решений из ансамбля тренируется независимо от остальных на своем наборе данных. Такой подход в сравнении с использованием одиночного дерева решений позволяет избежать проблему переобучение модели и повысить точность путем агрегирования нескольких ответов, полученными разными деревьями из ансамбля [30].

5.5 Разработка нейронной сети для предсказания энергии связи

На практике модели машинного обучения базирующиеся на технологии искусственных нейронных сетей показывают хорошие результаты в решении множества прикладных задач. Далее приведем результаты практического исследования возможностей искусственных нейронных сетей для предсказания ингибиторной активности химических соединений.

В качестве датасета имеем 82710 соединений в формате фингерпринтов с соответствующей их комплексам с белком оболочки ВИЧ-1 gp120 энергиями связей. Датасет изначально разбили в пропорции 4 к 1 на тренировочный и тестовый набор, соответственно 66 168 на тренировочный и валидационный и 16 542 тестовый набор. Затем тренировочную выборку разбивали еще на две части: непосредственно саму тренировочную — 59551 и валидационную — 6617.

Удобным инструментом для создания моделей нейронных сетей является библиотека Keras [33], которая и была использована в данной работе.

Для обучения нейронных сетей требуется определить параметр, который мы будем стремиться минимизировать. Данный параметр называют функцией ошибки. Подробнее про функции ошибки было сказано в пункте 3.7. В качестве функции ошибки будем использовать среднюю квадратическую ошибку (MSE). Для более точной оценки моделей, после обучения будем также использовать среднюю абсолютную ошибку (MAE) и R-квадрат (R2 score). Данные метрики доступны в библиотеке машинного обучения Scikit-learn [34].

В качестве оптимизатора градиентного спуска использовался Adam (см. 3.5.3). Также для исключений бесполезного обучения была использована функция ранней остановки обучений (EarlyStopping) из библиотеки Keras [33].

5.5.1 Анализ данных

Первым этапом было принято решение редуцировать набор входных признаков с 166 мерного пространства на двумерное с целью визуализации данных.

В первом варианте был использован метод главных компонент (principal component analysis - PCA) для редукции признаков. На рисунке 5.1 показаны данные редуцированные на плоскость, цветом показана энергия связи (синий — наибольшая энергия связи, красный — наименьшая). Можно заметить, что никакой четкой корреляции между признаками и значением не наблюдается.

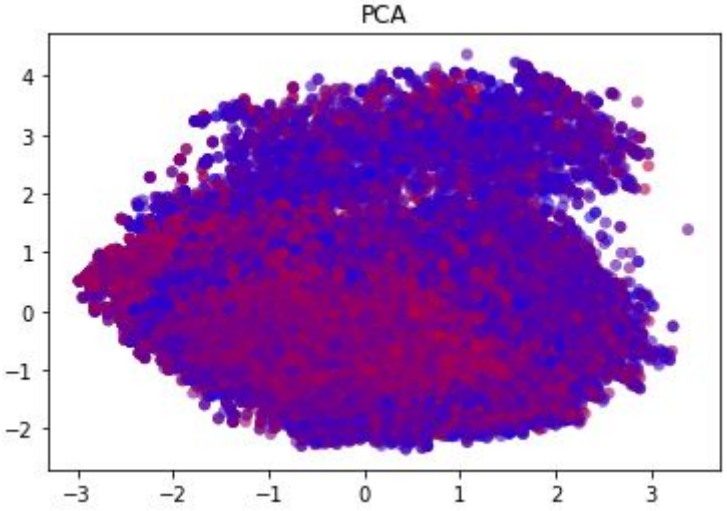

Рис. 5.1. Редуцированные входные данные при помощи PCA.

Далее был применен метод TSNE (t-distributed Stochastic Neighbor Embedding) — стохастическое вложение соседей с t-распределением. Данный метод хорошо подходит для редукции признаков с большой размерностью, в нашем случае — 166, на маленьке размерности (2, 3 мерные). Метод является не линейным. На рисунке 5.2. показаны результаты применения данного метода на наших входных данных.

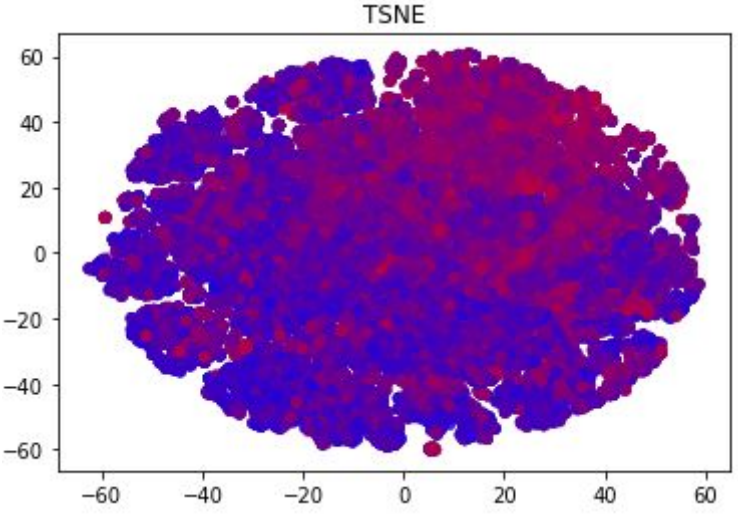

Рис 5.2. Редуцированные входные данные при помощи TSNE.

Визуально наилучшие результаты показал линейный дискриминантный анализ — LDA (Linear Discriminant Analysis). Результаты полученные данным методом изображены на рисунке 5.3.

По результатам, полученным путем редукции признаков, можно заметить, что не существует четко просматриваемой корреляции между входными и выходными данными.

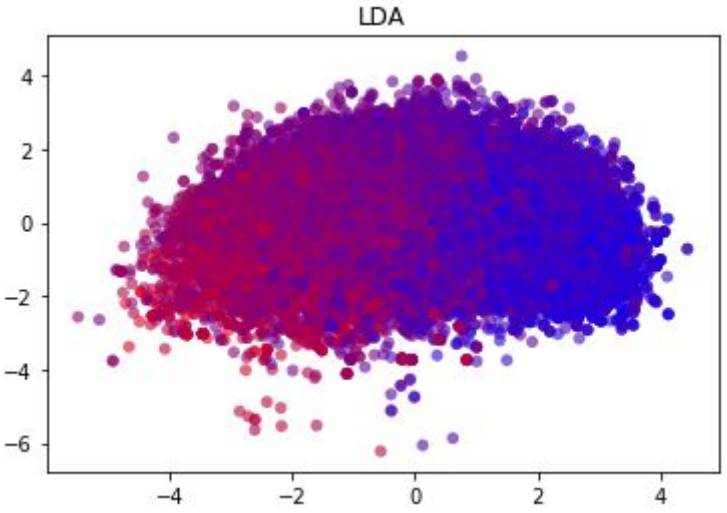

Рис. 5.3. Редуцированные входные данные при помощи LDA.

5.5.2 Полносвязные нейронные сети

Первая обученная модель имела самую простую архитектуру, представляющую собой один полносвязный слой. На вход подавалось 166 параметров (биты фингерпринта), и на выход соответственно одно выходное число — потенциальная энергия связи.

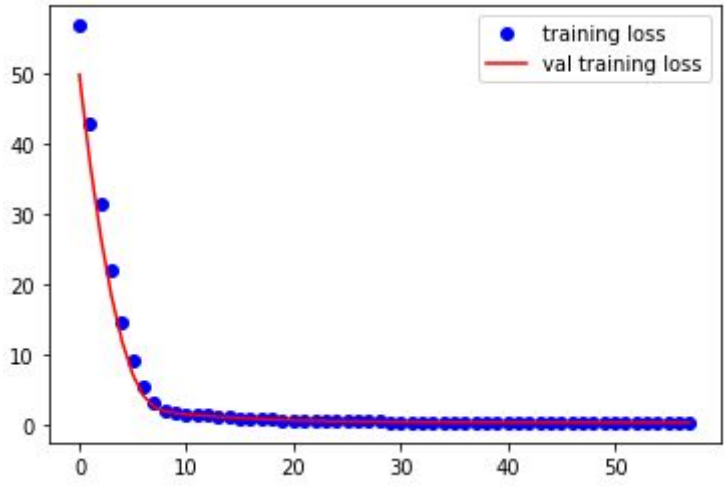

Рис. 5.4. График функции потерь для однослойной полносвязной модели.

На рисунке 5.4 изображен график зависимости значения функции потерь на тренировочном и валидационном наборе данных соосветственно. Для тренировки сети использовалась скорость обучения — 0.00001.

Далее была попытка применить функцию активации для этого единственного слоя. Однако результаты оказались хуже, что может быть обусловлено спецификой входных данных, которые представлены набором 0 и 1. В качестве примера приведем результаты полученные для однослойной полносвязной модели с функцией активации — сигмоид.

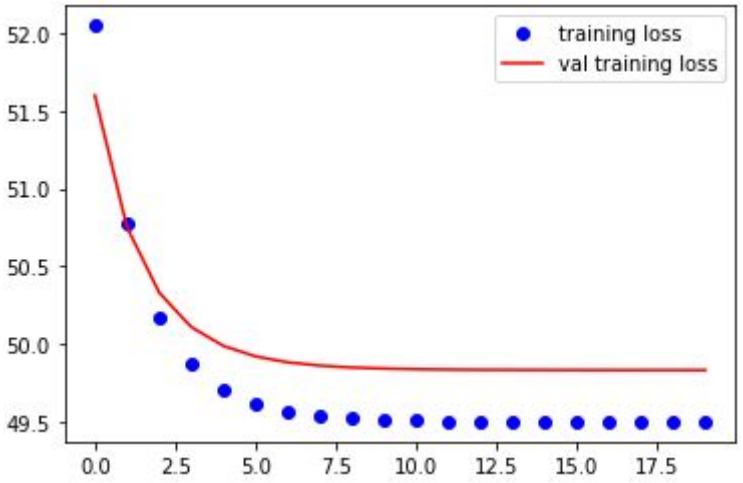

Рис. 5.5. График функции потерь для однослойной полносвязной модели с функцией активации — сигмоид.

Далее были использованы более глубокие сети. Первая пробная включала в себя 4 слоя по 142, 120, 80, 1 нейрон соответственно. И имела 50,635 тренируемых параметра. Для ее обучения был использован шаг 0.0000005. Размер шага подбирается эмпирически.

Для второй глубокой нейронной сети было изменено количество нейронов в слоях: 166, 166, 166, 1. Что в свою очередь увеличило число тренируемых параметров до 83,333. Точность такой модели не значительным образом отличается от предыдущей.

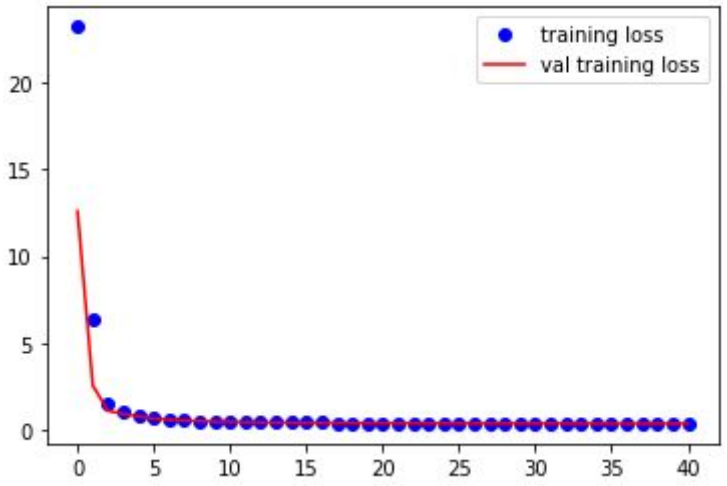

Рис. 5.6. График функции потерь для глубокой полносвязной сети № 1.

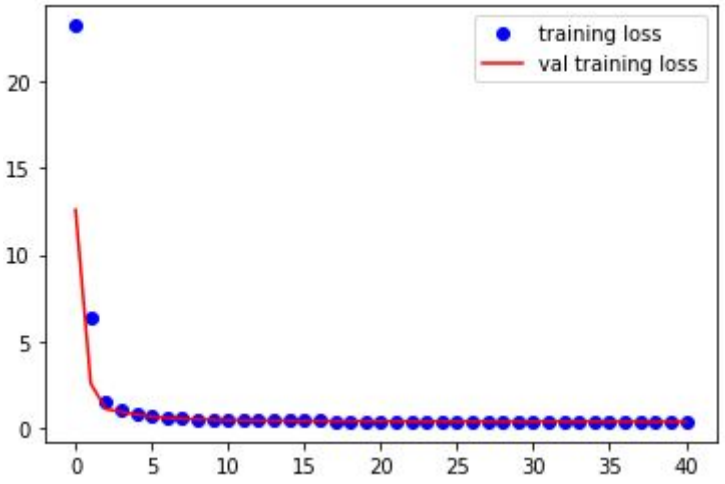

Рис. 5.7. График функции потерь для глубокой полносвязной сети № 2.

В качестве эксперимента также была обучена сеть с большим числом слоев, например — 5. В такой модели присутствовало уже 111,055 тренируемых параметра.

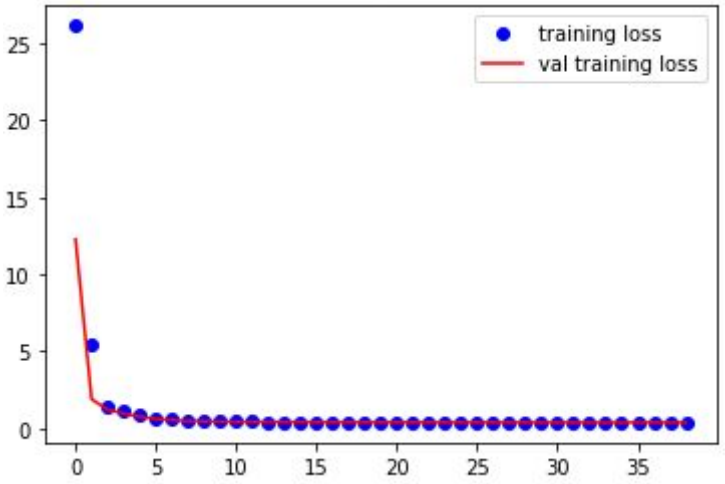

Рис. 5.8. График функции потерь для глубокой полносвязной сети № 3.

Увеличение числа слоев не дало прироста точности. А если сравнивать две сети с одинаковой точность, но разного размера, то по производительности лучше будет сеть меньше, за счет меньшего числа требуемых опираций для получений результата.

Также была попытка использовать функции активации, но, как и в случае с однослойной моделью, многослойная модель с функцией активации показала результат хуже.

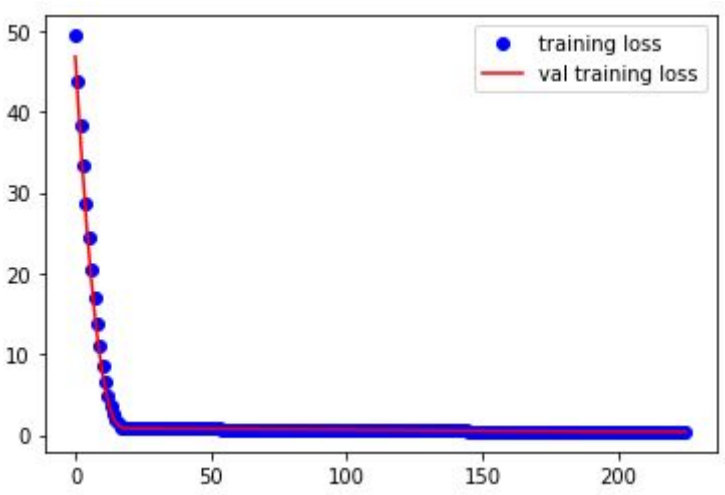

Рис. 5.9. График функции потерь для глубокой полносвязной сети № 2 с функцией активации — сигмоид.

5.5.3 Сверточные нейронные сети

Сверточные нейронные сети являются мощным инструментом позволяющим решать различные прикладные задачи. Структуру сверточной нейронной сети можно разделить на два блока: блок свертки (convolution layers) и полносвязный блок (fully-connected layers).

В первом блоке каждый сверточный слой представляет собой набор матриц-фильтров, ядер для свертки. Число таких слоев, фильтров и их размеры выбираются в зависимости от задачи методом математического подбора. В некоторых сетях число слоев может доходить до десятков и даже сотен, но не имея соответствующей вычислительной мощности их использование может быть не оправдано ввиду трудности обучения.

Смысл сверточных сетей состоит в том, что последовательное выполнение операций свертки и пулинга (pooling) уменьшает размерность входных данных, извлекая из них необходимые признаки, которые далее могут быть использованы для решения поставленной задачи.

За основу первой использованной сверточной сети была взята архитектура AlexNet [35]. Однако AlexNet была применена на изображениях, где входными параметрами являются матрицы чисел (интенсивности света). Поэтому требовалось изменить архитектуру таким разом, чтобы ее можно было использовать на векторах фингерпринтов — матрицы размерностью 166 на 1. Все 2d свертки и 2d пулинги были заменены на 1d свертки и 1d пулинги. Размеры фильтров на сверточных слоях также были изменены на одномерные, но остались прежних значений. На рисунке 5.10 изображена полная архитектура полученной сети. Данная модель имеет 1,796,873 тренируемых параметра.

Layer (type)	Output Shape	Param #
conv1d_1 (Conv1D)	(None, 156, 3)	36
conv1d_2 (Conv1D)	(None, 152, 96)	1536
max_pooling1d_1 (MaxPooling1	(None, 76, 96)	0
conv1d_3 (Conv1D)	(None, 74, 256)	73984
max_pooling1d_2 (MaxPooling1	(None, 37, 256)	0
conv1d_4 (Conv1D)	(None, 35, 384)	295296
conv1d_5 (Conv1D)	(None, 33, 384)	442752
max_pooling1d_3 (MaxPooling1	(None, 16, 384)	0
flatten_1 (Flatten)	(None, 6144)	0
dense_1 (Dense)	(None, 156)	958620
dense_2 (Dense)	(None, 156)	24492
dense_3 (Dense)	(None, 1)	157

Рис. 5.10. Архитектура сверточной сети для оценки ингибиторной активности химических соединений, основанная на архитектуре AlexNet.

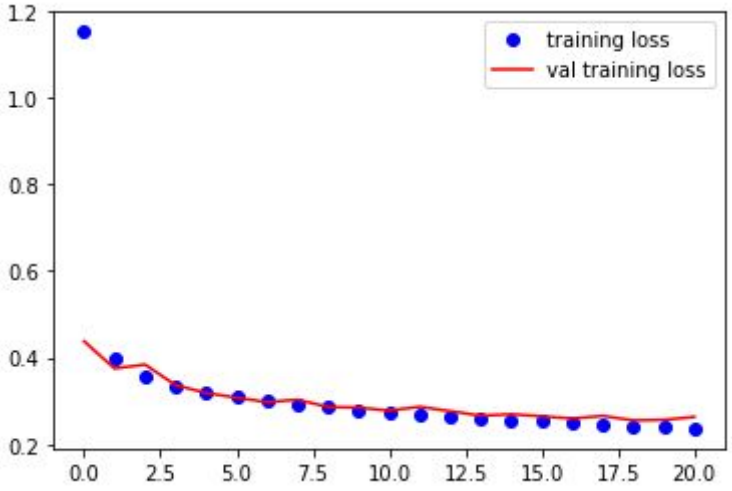

Рис. 5.11. График функции потерь для сверточной сети основанной на архитектуре AlexNet.

На рисунке 5.11 изображен график функции потерь при обучении данной сети. Результаты превзошли все те, которые были получены при помощи полносвязных сетей.

Также была в качестве эксперимента спроектирована сеть более простой архитектуры (см. рис. 5.12). Данная сеть имеет 707,881 тренируемых параметра, что значительно меньше, чем в предыдущей модели.

Layer (type)	Output Shape	Param #
conv1d_1 (Conv1D)	(None, 160, 64)	512
max_pooling1d_1 (MaxPooling1	(None, 80, 64)	0
conv1d_2 (Conv1D)	(None, 76, 128)	41088
max_pooling1d_2 (MaxPooling1	(None, 38, 128)	0
conv1d_3 (Conv1D)	(None, 36, 256)	98560
max_pooling1d_3 (MaxPooling1	(None, 18, 256)	0
flatten_1 (Flatten)	(None, 4608)	0
dense_1 (Dense)	(None, 120)	553080
dense_2 (Dense)	(None, 120)	14520
dense_3 (Dense)	(None, 1)	121

Рис. 5.12. Архитектура сверточной сети для оценки ингибиторной активности химических соединений.

Результаты для такой архитектуры оказались хуже, чем для сети основанной на архитектуре AlexNet.

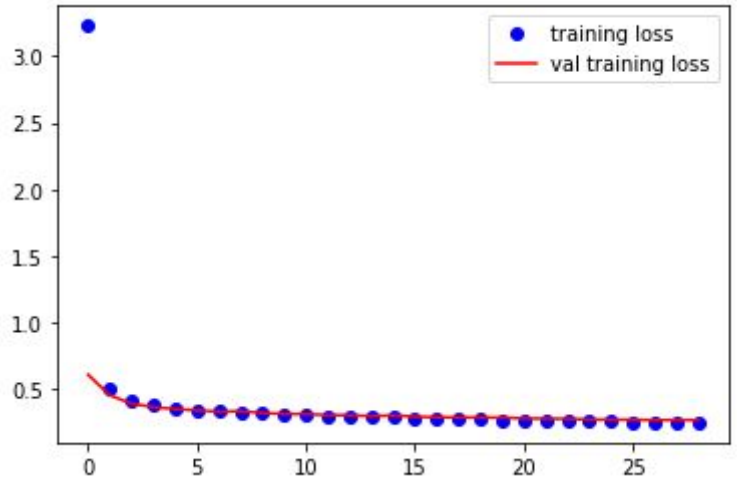

Рис. 5.13. График функции потерь для сверточной сети.

Достаточно мощными сетями является группа сетей для анализа изображений VGG [35]. На основе архитектуры таких сетей была спроектирована и сеть для предсказания энергии связи химических соединений.

Layer (type)	Output Shape	Param #
conv1d_1 (Conv1D)	(None, 164, 64)	256
conv1d_2 (Conv1D)	(None, 162, 64)	12352
max_pooling1d_1 (MaxPooling1	(None, 81, 64)	0
conv1d_3 (Conv1D)	(None, 79, 128)	24704
conv1d_4 (Conv1D)	(None, 77, 128)	49280
max_pooling1d_2 (MaxPooling1	(None, 38, 128)	0
conv1d_5 (Conv1D)	(None, 36, 256)	98560
conv1d_6 (Conv1D)	(None, 34, 256)	196864
conv1d_7 (Conv1D)	(None, 32, 256)	196864
max_pooling1d_3 (MaxPooling1	(None, 16, 256)	0
conv1d_8 (Conv1D)	(None, 14, 512)	393728
conv1d_9 (Conv1D)	(None, 12, 512)	786944
conv1d_10 (Conv1D)	(None, 10, 512)	786944
max_pooling1d_4 (MaxPooling1	(None, 5, 512)	0
flatten_1 (Flatten)	(None, 2560)	0
dense_1 (Dense)	(None, 120)	307320
dense_2 (Dense)	(None, 120)	14520
dense_3 (Dense)	(None, 1)	121

Рис. 5.14. Архитектура сверточной сети для оценки ингибиторной активности химических соединений, основанная на архитектуре VGG.

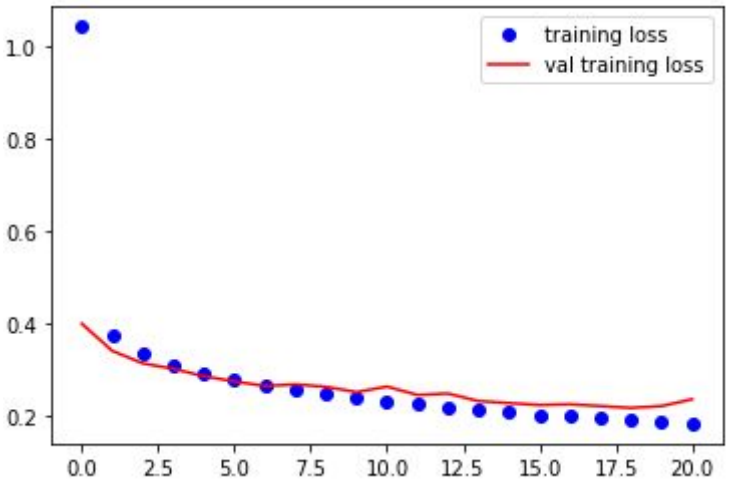

Рис. 5.15. График функции потерь для сверточной сети основанной на архитектуре VGG.

Как и в случае с AlexNet, сети группы VGG используются для работы с изображениями, поэтому архитектуру их необходимо переделывать таким образом, чтобы она могла работать с массивами, а не матрицами. Также было сокращено число сверточных слоев с целью уменьшения веса сети. Данная сеть показала наилучшие результаты, но также она является самой тяжеловесной, и имеет 2,868,457 тренируемых параметра. На рисунке 5.15. изображен график зависимости значения функции потерь от номера эпохи.

5.5.4 Оценка моделей

Для описанных выше сетей были ещё посчитаны средняя абсолютная ошибка (MAE) и R-квадрат (R2 score). Значения данных метрик для всех рассмотренных моделей были занесены в таблицу (см. приложение 1).

Как сказано выше, лучшие значения метрик получились у модели на основе VGG. Для нее значение MSE, MAE и R-квадрат на тестовой выборке соответственно равны 0.217, 0.358 и 0.727. На тестовой выборке прогнозируемые значения энергии связи отклоняются более чем на 0.5 ккал от реальных значений менее чем в 30% случаях.

ЗАКЛЮЧЕНИЕ

Данная работа представляет собой совокупность различных подходов, базирующихся на популярной в данный момент времени технологии машинного обучений, которые можно применить в области разработки новых лекарственных препаратов.

Первый подход заключается в использовании состязательного автоэнкодера для генерации новых потенциально лекарственных химических соединений для борьбы с ВИЧ-1. Для обучения автоэнкодера был создан специальный датасет, который включает в себя 82710 химических соединений с просчитанной при помощи высокопроизводительного докинга энергии связи комплекса лиганда с белком оболочки ВИЧ-1 gp120. Молекулы в данном датасете представлены в формате молекулярных фингерпринтов MACCS.

Использование в качестве формата представления химических соединений молекулярных фингерпринтов повлекло за собой проблему интерпретации результатов работы автоэнкодера. С целью решения данной проблемы была создана база данных для поиска химических соединений по их молекулярному фингерпринту, которая включает в себя 166.416.851 соединения с посчитаными фингерпринтами. Данная база была создана на основе базы потенциальных лекарственных соединений ZINC.

Вторым подходом, рассмотренном в данной работе, является использование машинного обучения для предсказания ингибиторной активности химического соединения. Эта задача представляет собой регрессионную задачу, которая была решена с достаточно высокой степенью точности при помощи глубокого обучения. Было разработано несколько моделей нейронных сетей как простых полносвязных, так и более сложных сверточных на основе архитектуры хорошо зарекомендовавших себя моделей.

Таким образом был разработан алгоритм предсказания энергии связи комплекса лиганда с белком gp120 на основе машинного обучения. Данный подход является менее точным, чем использование молекулярного докинга, однако он может применяться для скрининга баз данных, которые содержат большое число химических соединений, на предмет поиска потенциальных ингибиторов белка оболочки ВИЧ-1 gp120, так как данный метод значительно менее ресурсоемкий и обладает большей скоростью в сравнении с молекулярным докингом.

Имеет место также проведение скрининга базы данных фингерпринтов, созданной с целью интерпретации работы состязательного автоэнкодера. Поиск может осуществляться при помощи нейронной сети, обученной для прогнозирования потенциальной энергии связи.

В данной работе были рассмотрены далеко не все способы применения машинного обучения в разработке новых лекарственных соединений, однако полученные результаты подтверждают сравнительно высокую эффективность

машинного обучения для решения ряда прикладных задач биомедицинской информатики.

На сегодняшний день актуальность данного направления исследований не может ставиться под сомнения. Разработка новых способов поиска лекарственных соединений является одной из ключевых задач для всего человечества.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. World Health Organization. HIV/AIDS — Режим Доступа: https://www.who.int/news-room/fact-sheets/detail/hiv-aids — Дата доступа: 29.04.2020
2. Э. Фаучи, К. Лэйн. ВИЧ-ИНФЕКЦИЯ.
3. A.M. ANDRIANOV, G.I. NIKOLAEV, YU.V. KORNOUSHENKO, J. HUANG, S.JIANG. VIRTUAL SCREENING AND IDENTIFICATION OF POTENTIAL HIV-1 INHIBITORS BASED ON THE CROSS-REACTIVE NEUTRALIZING ANTIBODY N6.
4. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings // 2001.
5. Pharmit: interactive exploration of chemical space — Режим Доступа: http://pharmit.csb.pitt.edu — Дата доступа: 30.04.2020
6. PubChem — Режим Доступа: https://pubchem.ncbi.nlm.nih.gov — Дата доступа: 30.04.2020
7. Гуреев М.А., Кадочников В.В., Порозов Ю.Б. Молекулярный докинг и его верификация в контексте виртуального скрининга // Университет ИТМО 2018.
8. AutoDock — Режим Доступа: http://autodock.scripps.edu — Дата доступа: 30.04.2020
9. С.К. Игнатов. Квантово-химическое моделирование молекулярной структуры, физико-химических свойств и реакционной способности // 2006.
10. James J. P. Stewart. General Description of MOPAC — Режим Доступа: http://openmopac.net/manual — Дата доступа: 30.04.2020
11. Х.Т. Холмуродов, М.В. Алтайский, И.В. Пузынин; Т. Дардин; Ф.П. Филатов. МЕТОДЫ МОЛЕКУЛЯРНОЙ ДИНАМИКИ ДЛЯ МОДЕЛИРОВАНИЯ ФИЗИЧЕСКИХ И БИОЛОГИЧЕСКИХ ПРОЦЕССОВ.
12. Krzysztof J. Cios. Deep Neural Networks — A Brief History — Режим Доступа: https://arxiv.org/abs/1701.05549 — Дата доступа: 01.05.2020
13. Yongjun Hong, Uiwon Hwang, Jaeyoon Yoo, Sungroh Yoon. How Generative Adversarial Networks and Their Variants Work: An Overview — Режим Доступа: https://arxiv.org/abs/1711.05914 — Дата доступа: 01.05.2020
14. Artur Kadurin, Alexander Aliper, Andrey Kazennov, Polina Mamoshina, Quentin Vanhaelen, Kuzma Khrabrov, Alex Zhavoronkov. The cornucopia of meaningful leads: Applying deep adversarial autoencoders for new molecule development in oncology.
15. The RDKit Documentation — Режим Доступа: https://www.rdkit.org/docs/index.html — Дата доступа: 03.05.2020

16. SMARTS — A Language for Describing Molecular Patterns. Daylight — Режим Доступа: https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html — Дата доступа: 03.05.2020
17. Open Babel: The Open Source Chemistry Toolbox — Режим Доступа: http://openbabel.org/wiki/Main_Page Дата доступа: 03.05.2020
18. ZINC — Режим Доступа: https://zinc.docking.org Дата доступа: 03.05.2020
19. SQL Server technical documentation — Режим Доступа: https://docs.microsoft.com/en-us/sql/sql-server/?view=sql-server-ver15 Дата доступа: 03.05.2020
20. Jeffrey Shallit. Hamming Distance for Conjugates — Режим Доступа: https://arxiv.org/abs/0710.1234 Дата доступа: 03.05.2020
21. Hao Fan, Dina Schneidman-Duhovny, John J. Irwin, Guangqiang Dong, Brian K. Shoichet and Andrej Sali1. Statistical Potential for Modeling and Ranking of Protein-Ligand Interactions.
22. Jacob D. Durrant and J. Andrew McCammon. NNScore: A Neural-Network-Based Scoring Function for the Characterization of Protein−Ligand Complexes.
23. Sebastian Ruder. An overview of gradient descent optimization algorithms — Режим Доступа: https://arxiv.org/abs/1609.04747 — Дата доступа: 03.05.2020
24. Chen Xing, Devansh Arpit, Christos Tsirigotis, Yoshua Bengio. A walk with SGD — Режим Доступа: https://arxiv.org/abs/1802.08770 — Дата доступа: 03.05.2020
25. Diederik P. Kingma, Jimmy Ba. Adam: A Method for Stochastic Optimization — Режим Доступа: https://arxiv.org/abs/1412.6980 — Дата доступа: 03.05.2020
26. Zijun Zhang, Lin Ma, Zongpeng Li, Chuan Wu. Normalized Direction-preserving Adam — Режим Доступа: https://arxiv.org/abs/1709.04546 — Дата доступа: 03.05.2020
27. Arun K. Kuchibhotla, Lawrence D. Brown, Andreas Buja, Junhui Cai. All of Linear Regression — Режим Доступа: https://arxiv.org/abs/1910.06386 — Дата доступа: 04.05.2020
28. Kartik Chandra, Erik Meijer, Samantha Andow, Emilio Arroyo-Fang, Irene Dea, Johann George, Melissa Grueter, Basil Hosmer, Steffi Stumpos, Alanna Tempest, Shannon Yang. Gradient Descent: The Ultimate Optimizer — Режим Доступа: https://arxiv.org/abs/1909.13371 — Дата доступа: 04.05.2020
29. Shahan Ali Memon, Wenbo Zhao, Bhiksha Raj, Rita Singh. Neural Regression Trees — Режим Доступа: https://arxiv.org/abs/1810.00974 — Дата доступа: 04.05.2020

30. Haozhe Zhang, Dan Nettleton, Zhengyuan Zhu. Regression-Enhanced Random Forests — Режим Доступа: https://arxiv.org/abs/1810.00974 — Дата доступа: 04.05.2020
31. Alexei Botchkarev. Performance Metrics (Error Measures) in Machine Learning Regression, Forecasting and Prognostics: Properties and Typology — Режим Доступа: https://arxiv.org/abs/1809.03006 — Дата доступа: 06.05.2020
32. Federico Cabitza, Andrea Campagner. Who wants accurate models? Arguing for a different metrics to take classification models seriously — Режим Доступа: https://arxiv.org/abs/1910.09246 — Дата доступа: 07.05.2020
33. Keras: The Python Deep Learning library — Режим Доступа: https://keras.io/ — Дата доступа: 12.05.2020
34. Scikit-learn Machine Learning in Python — Режим Доступа: https://scikit-learn.org/stable/ — Дата доступа: 12.05.2020
35. Md Zahangir Alom, Tarek M. Taha, Christopher Yakopcic, Stefan Westberg, Paheding Sidike, Mst Shamima Nasrin, Brian C Van Esesn, Abdul A S. Awwal, Vijayan K. Asari. The History Began from AlexNet: A Comprehensive Survey on Deep Learning Approaches — Режим Доступа: https://arxiv.org/abs/1803.01164 — Дата доступа: 12.05.2020

СПИСОК ПУБЛИКАЦИЙ

1. Николаев Г.И., Шульдов Н.А., Анищенко А.И., Тузиков А.В., Андрианов А.М. Разработка генеративной состязательной нейронной сети для идентификации потенциальных ингибиторов ВИЧ-1 методами глубокого обучения. Информатика. 2020;17(1):7-17. https://doi.org/10.37661/1816-0301-2020-17-1-7-17

ПРИЛОЖЕНИЯ

Приложение 1. Сравнительная таблица результатов обучения нейронных сетей

Модель	MSE		MAE		R2 score	
	Трениро вочная	Тестовая	Трениров очная	Тестовая	Трениро вочная	Тестовая
Однослойная полносвязная.	0.395	0.389	0.499	0.493	0.504	0.511
Однослойная полносвязная с сигмоидом.	49.533	49.597	6.981	6.986	-61.202	-61.423
Глубокая полносвязная сеть № 1.	0.376	0.370	0.486	0.480	0.528	0.535
Глубокая полносвязная сеть № 2.	0.378	0.371	0.487	0.481	0.525	0.533
Глубокая полносвязная сеть № 3.	0.378	0.372	0.487	0.482	0.525	0.532
Глубокая полносвязная сеть № 2 с сигмоидом.	0.415	0.406	0.512	0.506	0.479	0.488
На основе AlexNet.	0.234	0.252	0.377	0.391	0.707	0.682
Сверточная сеть рис. 3.12.	0.252	0.265	0.391	0.401	0.684	0.666
На основе VGG.	0.180	0.217	0.325	0.358	0.774	0.727