МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
имени М.В. ЛОМОНОСОВА

На правах рукописи

Богданова Елизавета Александровна

Предсказание аффинности в белок-белковых комплексах на основе межатомных расстояний с использованием трёхмерной свёрточной нейронной сети

1.5.8. – Математическая биология, биоинформатика

АВТОРЕФЕРАТ
Диссертации на соискание ученой степени
кандидата биологических наук

Москва – 2025

Работа выполнена на кафедре биоинженерии Биологического факультета Федерального государственного бюджетного образовательного учреждения высшего образования «Московский государственный университет имени М.В. Ломоносова».

Научный руководитель:	**Новоселецкий Валерий Николаевич** *кандидат физико-математических наук*
Официальные оппоненты:	**Коваленко Илья Борисович** *доктор физико-математических наук, ведущий научный сотрудник кафедры биофизики Биологического факультета Федерального государственного бюджетного образовательного учреждения высшего образования «Московский государственный университет имени М.В. Ломоносова»* **Хренова Мария Григорьевна** *доктор физико-математических наук, профессор, профессор кафедры физической химии Химического факультета Федерального государственного бюджетного образовательного учреждения высшего образования «Московский государственный университет имени М.В. Ломоносова»* **Попцова Мария Сергеевна** *кандидат физико-математических наук, заведующий международной лабораторией биоинформатики, доцент Факультета компьютерных наук Федерального государственного автономного образовательное учреждения высшего образования «Национальный исследовательский университет «Высшая школа экономики»*

Защита диссертации состоится 5 марта 2025 года в 15:30 на заседании диссертационного совета МГУ.015.10 Московского государственного университета имени М.В. Ломоносова по адресу: 119234, Москва, Ленинские горы, д. 1, стр. 73, Факультет биоинженерии и биоинформатики, ауд. 221.

E-mail: dissovet@belozersky.msu.ru

С диссертацией можно ознакомиться в отделе диссертаций Научной библиотеки МГУ имени М.В. Ломоносова (Москва, Ломоносовский просп., д. 27) и на сайте портале: https://dissovet.msu.ru/dissertation/3297.

Автореферат разослан «___» февраля 2025 года.

Ученый секретарь диссертационного совета,
кандидат химический наук И.В. Шаповалова

СПИСОК СОКРАЩЕНИЙ

ACE2 – Angiotensin-Converting Enzyme 2/ Ангиотензинпревращающий фермент 2, SARS-CoV – Severe Acute Respiratory Syndrome-related Coronavirus/ Коронавирус тяжёлого острого респираторного синдрома, FRET – Förster resonance energy transfer/ резонансный перенос энергии флуоресценции, IC_{50} – half maximal Inhibitory Concentration/ концентрация полумаксимального ингибирования, MAE – Mean Absolute Error/ Средняя абсолютная ошибка, MSE – Mean Squared Error/ Средняя квадратичная ошибка, RMSE – Root Mean Squared Error/ корень средней квадратичной ошибки, RBD – Receptor Binding Domain/ рецептор-связывающий домен, ReLU – Rectified Linear Unit, PDB – Protein Data Bank, МД – молекулярная динамика, ЯМР – Ядерный Магнитный Резонанс.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы исследования

Белок-белковые взаимодействия образуются в результате возникновения стереохимических контактов между поверхностями белковых молекул в области, называемой интерфейсом связывания. Данные взаимодействия возникают в процессе сборки четвертичных структур и функциональных макромолекулярных комплексов (Bryant et al., 2022).

Многие физиологические клеточные процессы зависят от скоординированного формирования таких взаимодействий (Lucero et al., 2023). К примерам таких динамических процессов можно отнести репликацию ДНК и другие реакции матричного синтеза, регуляцию экспрессии генов, сплайсинг мРНК в эукариотических клетках, формирование внутриклеточных белковых структур, а также многие процессы, связанные с внутри- и межклеточной сигнализацией (Boike et al., 2022; Lucero et al., 2023).

Также взаимодействия между определенными белковыми молекулами могут быть ответственны за развитие патологических процессов, таких как болезнь Альцгеймера, прионные, аутоиммунные заболевания (Goncearenco et al.,2017), некоторые формы рака и другие (Lu et al., 2020). Кроме того, взаимодействия между вирусными белками и клеточными факторами ответственны за заражение клетки и происходят в процессе реализации вирусной генетической информации в клетках–хозяевах (Loregian et al., 2002).

Следовательно, использование белок-белковых взаимодействий в качестве мишени для терапевтического вмешательства является крайне актуальным и важным направлением в фармакологии. Однако данная задача представляет высокую сложность в связи с рядом факторов, к которым можно отнести пространственные особенности интерфейсов связывания, такие как их размер, форма и др. Так, для плоских интерфейсов, лишенных карманов связывания, возникают сложности в функциональном анализе взаимодействующих молекул. Кроме того, многие существующие лекарственные средства могут оказывать разноплановое воздействие на данные мишени, оказывая положительное влияние на связывание молекул, или, наоборот, ингибируя возможные взаимодействия. В частности, могут разрабатываться лекарственные средства, терапевтический эффект

которых основан на их высокоспецифичном связывании с целевым белковым комплексом (Goncearenco et al.,2017).

Для успешной разработки терапевтических и диагностических средств, основанных на работе белок-белковых комплексов, решающее значение имеет достоверная информация об энергии белковых взаимодействий и их наличии в физиологических и патофизиологических процессах.

Одной из основных характеристик белок-белковых взаимодействий является аффинность связывания. Данный параметр представляет собой количественную меру энергии взаимодействия между двумя или более молекулами, при условии обратимости их связывания. Наиболее точными методами определения аффинности являются экспериментальные методы, такие как изотермическая титрационная калориметрия (Ladbury et al., 1996) поверхностный плазмонный резонанс (Willander et al., 2009) и резонансный перенос энергии флуоресценции (Phillip et al., 2012). Однако, данные методы требуют дорогостоящих экспериментальных установок и являются затратными в плане временных ресурсов (Zheng et al., 2023).

Таким образом, предсказание аффинности связывания в белковых комплексах является одной из фундаментальных задач биоинформатики и вычислительной биологии в целом (Soleymani et al., 2022). Создание высокоточных алгоритмов оценки энергии взаимодействия позволило бы, в частности, более эффективно проводить направленный мутагенез взаимодействующих белков (Zhang et al., 2020), что имеет существенное значение для создания медицинских препаратов белковой природы, включая антитела (Zhang et al., 2018).

В настоящее время в биоинформатике всё больше находят широкое распространение такие методы машинного обучения, как нейронные сети, относящиеся к подходам глубинного обучения. За последнее десятилетие было предложено большое число предсказательных алгоритмов, решающих задачу оценки связывания в белковых комплексах. Однако, в связи с рядом ограничений, таких как недостаточный объем данных для многих комплексов, влияние внешних факторов на связывание и др., использование предсказательных алгоритмов на практике не имеет широкого применения. При преодолении вышеупомянутых ограничений станет возможным конструировать более универсальные алгоритмы предсказания, что позволило бы значительно продвинуться в области фармацевтики и биохимии.

Степень разработанности темы исследования

Физическое взаимодействие между молекулами белков имеет давнюю историю изучения многочисленными экспериментальными и вычислительными методами (Chothia et al., 1975; Archakov et al., 2003), включая методы биоинформатики (Shi et al, 2005). Одной из главных характеристик взаимодействия является константа диссоциации комплексов белок-белок (K_D), которая может быть выражена через энергию связывания $\Delta G = RT\ln K_D$.

На протяжении многих лет предлагались различные вычислительные методы предсказания аффинности связывания, резко различающиеся с точки зрения точности,

вычислительных затрат и физической правдоподобности (Siebenmorgen et al., 2019; Zheng et al., 2023).

В зависимости от постановки задачи используются различные метрики определения качества работы предсказательных алгоритмов. В случае задачи классификации наиболее часто применяемой метрикой является точность (англ. Accuracy), отражающая долю верно проклассифицированных объектов. В регрессионных задачах (предсказание значения энергии связывания), как правило, используется одновременно несколько метрик. Во-первых, для оценки способности алгоритма находить закономерности часто используется корреляция Пирсона, которая отражает степень линейной зависимости между экспериментально полученными значениями энергии связывания и предсказанными. Во-вторых, для оценки значения ошибки алгоритма, как правило, используется MAE и RMSE. Таким образом, используя разные метрики, можно с разных сторон оценить возможности и ограничения предсказательных алгоритмов.

Существуют достаточно сложные методы предсказания энергии связывания, такие как возмущение свободной энергии (Free Energy Perturbation, FEP) (Wang et al., 2012) и термодинамическое интегрирование (Bhati et al., 2017), подходы молекулярной механики с с расчетом уравнений Пуассона-Больцмана для площади поверхности (Molecular Mechanics Poisson-Boltzmann Surface Area, MMPBSA) (Rastelli et al., 2010; Panday et al., 2022). Эти методы обладают достаточно высокой точностью, однако, при этом в них используется обширная МД или конформационный поиск методом Монте-Карло, что делает данные подходы крайне требовательными к вычислительным ресурсам, обладая при этом ограниченной сферой применения. Например, в случаях, когда мутации неконтактных остатков значительно меняют аффинность связывания за счет существенного изменения конформации. Считается, что такого рода конформационные изменения выходят за рамки применимости FEP (Sampson et al., 2024). Были предложены альтернативные упрощенные эмпирические функции энергии для значительного снижения вычислительных затрат. Одним из таких методов является использование статистических потенциалов, которые используют наблюдаемые относительные положения атомов или остатков в экспериментальных структурах для определения потенциала взаимодействия (ROSSETTADOCK (Lyskov et al., 2008), DFIRE (Zhang et al., 2004), CP_PIE (Ravikant et al., 2010), FoldX (Schymkowitz et al., 2005) и др.). Также в последнее десятилетие в биоинформатике для решения подобных задач становятся популярными подходы, основанные на классическом машинном обучении и нейронных сетях (Zheng et al., 2023).

В настоящий момент реализованы алгоритмы, использующие данные о белковых комплексах в двух форматах: аминокислотная последовательность или пространственная структура. Наибольшая часть разработок данного направления сконцентрирована на изучении комплексов «белок-лиганд», и для этой задачи достигнуто достаточно высокое качество предсказания. В 2017 году был реализован Pafnucy – алгоритм предсказания связывания в комплексах «белок-лиганд», основанный на глубоких сверточных нейронных сетях и использующий в качестве обучающих данных PDB-структуры комплексов (Stepniewska-Dziubinska, 2017). Так, для тестового набора было достигнуто значение

корреляции Пирсона между предсказанными и экспериментально рассчитанными значениями, равное 0,78. В 2019 году был реализован алгоритм DeepAtom, также основанный на глубоких сверточных нейронных сетях, решающий эту же задачу со значением корреляции 0,83 (Li et al., 2019). Помимо этого, выходили алгоритмы, обученные на других наборах данных, обеспечивающие достаточно высокое качество предсказания на внутренних тестовых данных (Zhang et al, 2019). Однако, при отсутствии внешнего общепринятого репрезентативного тестового набора, объективное сравнение алгоритмов вызывает затруднения, а в ряде случаев не предоставляется возможным.

Что касается предсказания связывания в комплексах «белок-белок», здесь ситуация гораздо более сложная в связи с тем, что обе молекулы в комплексе обладают большим числом атомов и, как следствие, степеней свободы. В таком случае осложняется анализ особенностей конформационных состояний, оказывающих значительный вклад в сродство связывания между молекулами. Актуальные алгоритмы делятся на две группы: осуществляющие бинарную классификацию по наличию связывания (Asim el al., 2022), и решающие регрессионную задачу, обучаясь на данных об аминокислотной последовательности (ISLAND) (Abbasi et al., 2020) или структуре. В первом случае удалось добиться достаточно высокого качества предсказания (accuracy = 0,93), но результат недостаточно информативен, а во втором точность предсказания достаточно низкая (корреляция Пирсона = 0,44). Качество прогнозирования с использованием пространственных структур (PRODIGY (Xue et al., 2016), PPI-Affinity (Romero-Molina et al., 2022), AREA-AFFINITY (Yang et al., 2023)) выше (значение корреляции 0,5–0,6) на различных наборах тестовых данных.

В настоящее время разработано большое число методов, предсказывающих аффинность связывания в комплексах белок-белок и белок-пептид, однако до сих пор не удалось выявить метод, осуществляющий предсказание с высокой точностью для комплексов различной природы. Данное явление может быть связано со следующими ограничениями (Kastritis and Bonvin, 2012):

- Неоднозначность и нехватка экспериментальных данных;
- Отсутствие учета конформационных изменений или наличия кофакторов;
- Сложная кинетика комплекса и др.

На основании вышеизложенного можно утверждать, что остаётся достаточно большое поле для исследования белок-белковых комплексов, и создания алгоритмов, предсказывающих энергию связывания между белками с более высокой точностью.

Цель и задачи работы

Целью данной работы является разработка нейросетевого алгоритма, способного предсказывать аффинность связывания между белками в комплексах по их пространственным структурам. Для достижения поставленной цели были сформулированы следующие **задачи**:

1. Собрать набор данных из пространственных структур белок-белковых комплексов с известными характеристиками связывания и взаимодействующими цепями.
2. Проанализировать интерфейс белок-белковых взаимодействий для независимого набора комплексов, выявить взаимодействия конкретных аминокислот, включая опосредованные молекулами воды.
3. Разработать метод предобработки пространственных структур белок-белковых комплексов для их дальнейшего использования в обучении предсказательной модели.
4. Разработать, оптимизировать и обучить нейросетевой алгоритм, предсказывающий значение K_D для белок-белковых комплексов.
5. Апробировать новый алгоритм на репрезентативных тестовых наборах комплексов и провести анализ и сравнение получившихся результатов с другими подходами.
6. Провести анализ интерфейса взаимодействия в белок-белковых комплексах ACE2-RBD. Оценить аффинность связывания для набора комплексов ACE2-RBD с использованием разработанной модели, проанализировать результаты и сравнить с альтернативными методами.
7. С использованием разработанного алгоритма произвести анализ влияния разных вариантов гистонов H2A, H2B и H3 на стабильность образуемых ими димеров (для H2A-H2B), тетрамеров (H3-H4), а также комплексов между димерами и тетрамерами.

Объект и предмет исследования

Объектом исследования являются белок-белковые и белок-пептидные комплексы с известными характеристиками связывания. Предметом исследования являются пространственные структуры белковых комплексов, полученные с помощью экспериментальных методов, таких как рентгеновская кристаллография, ЯМР-спектроскопия и криоэлектронная микроскопия.

Научная новизна

Разработан новый подход прогнозирования аффинности связывания в белок-белковых комплексах, основанный на глубокой сверточной нейронной сети, позволяющий с высокой точностью предсказывать K_D и ΔG для белок-белковых и белок-пептидных комплексов разной природы. Полученные результаты апробации и сравнение с существующими аналогами указывают на стабильную качественную работу разработанной модели как на внутренних, так и на внешних тестах, содержащих белок-белковые комплексы различной природы.

Предложенная методология представления пространственной структуры комплексов в формате 4D-тензора, включающего информацию о расположении атомов и их способности участвовать в различных типах взаимодействий, является авторской и новой.

Практическая значимость работы

Собранный и предобработанный набор данных белок-белковых комплексов может в дальнейшем использоваться для изучения особенностей взаимодействия белковых молекул и для обучения различных предсказательных моделей. Разработанный и обученный нейросетевой алгоритм в дальнейшем может использоваться на ранних стадиях процессов разработки лекарственных препаратов, которые фокусируются на скрининге и оптимизации белок/пептид связывающих агентов для белка-мишени.

Данные об обучающем наборе, а также исходный код обученного алгоритма представлены в репозитории https://github.com/EABogdanova/ProBAN.

Методология и методы исследования

Для локализации интерфейса связывания были использованы методы машинного обучения (логистическая регрессия). Для разработки предсказательного алгоритма были использованы методы глубинного обучения (трехмерная сверточная нейронная сеть). Разработанный алгоритм был реализован на языке программирования Python 3 с использованием принципов объектно-ориентированного программирования (ООП). Изучаемые структуры белков были получены из базы данных PDB. Составление выборок для обучения и тестирования осуществлялось с использованием баз данных PDBBind v.2020 (Wang et al., 2020) и SKEMPI v.2.0 (Jankauskaitė et al., 2019).

Степень достоверности

Разработанная модель предсказания аффинности связывания в белок-белковых комплексах была апробирована и показала свою состоятельность на внутреннем тестовом наборе данных, содержащем комплексы, состоящие из трех и более молекул, так и на внешнем тесте, а также на наборе из комплексов ACE2-RBD и комплексов, образованных каноничными и замещающими формами гистонов. Анализ значимости признаков показал, что наиболее важными являются признаки, характеризующие некоторые наиболее важные взаимодействия в белках, что согласуется с известными данными о строении белковых молекул и белок-белковых взаимодействиях. В результате удалось добиться лучшего качества прогнозирования на тестовых наборах данных среди всех анализируемых моделей.

Личный вклад автора

Личный вклад автора заключается в: 1) анализе литературных источников; 2) разработке новых методов выявления и анализа структурных паттернов; 3) имплементации разработанных методов в качестве программного кода; 4) апробации разработанных методов; 5) анализе полученных результатов; 6) подготовке научных статей и представлении результатов на научных конференциях.

Положения, выносимые на защиту

1. Разработан новый алгоритм, основанный на трехмерной сверточной нейронной сети, предсказывающий значение аффинности связывания (константа диссоциации и свободная энергия Гиббса) для белок-белковых и белок-пептидных комплексов по их пространственным структурам.
2. Предложен новый метод предобработки пространственных структур белок-белковых комплексов, учитывающий различные типы контактов между молекулами, а также позволяющий сохранить информацию об их пространственных характеристиках.
3. В результате апробации на нескольких разнородных наборах комплексов (высоко- и низкоаффинные, нативные и мутантные формы комплексов) достигнуто лучшее качество предсказания энергии связывания в белок-белковых комплексах по сравнению со всеми существующими альтернативными подходами.
4. Предположено и в ходе тестирования алгоритма показано, что разработанная предсказательная модель способна оценивать влияние точечных мутаций на белок-белковые взаимодействия, а также на стабильность образуемых белковыми молекулами комплексов.

Публикации по теме работы

По материалам работы опубликованы 4 статьи в рецензируемых журналах, индексируемых в наукометрических базах данных Web of Science и/или Scopus (3 статьи в международных журналах и 1 статья в российском журнале из списка ВАК).

Апробация работы

Результаты исследования были представлены на 6-и конференциях: «OpenBio-2022», «OpenBio-2023» (Кольцово, Россия, 2022 и 2023 гг.), «Moscow Conference on Computational Molecular Biology» (MCCMB’23, Москва, 2023 г.), XXV Международная научно-техническая конференция "Нейроинформатика-2023" (Москва, Россия, 2023 г.), I Междисциплинарная всероссийская молодежная научная школа-конференция с международным участием «Молекулярный дизайн биологически активных веществ: биохимические и медицинские аспекты» (Казань, Россия, 2023 г.), 14-й Международной мультиконференции (Новосибирск, Россия, 2024 г.).

Структура и объем диссертации

Диссертационная работа состоит из следующих разделов: оглавление, список сокращений, введение, обзор литературы, методы, результаты и обсуждение, заключение, основные результаты и выводы, список литературы. Работа изложена на 141 странице, содержит 44 иллюстрации, 10 таблиц, 2 приложения и цитирует 187 литературных источников.

РЕЗУЛЬТАТЫ И ИХ ОБСУЖДЕНИЕ

Новый подход к преобразованию пространственных структур белок-белковых комплексов

В работе предложена новая методология предобработки структурных файлов белок-белковых комплексов, которая позволяет в автоматическом режиме локализовывать для разных комплексов интерфейс взаимодействия внутри ограничительной ячейки. И после отбора атомов происходит добавление каналов, отражающих различные типы взаимодействий, играющих ключевую роль в формировании белок-белковых комплексов. Предложенный подход состоит из нескольких этапов.

Этап 1. локализации интерфейса связывания молекул в ячейке универсального размера, подходящей для большинства комплексов.

На первом этапе обработки данных была решена задача автоматизации подбора ограничивающей ячейки. Сначала из структурных файлов была извлечена информация об атомах белковых молекул (координаты, аминокислотные остатки, цепи).

Далее для каждого комплекса цепи молекул, участвующих в связывании, были разделены на два класса. Чтобы определить оптимальное расположение ограничивающей ячейки для каждого комплекса, с использованием логистического метода рассчитывалась разделяющая плоскость (Рис. 1) между связывающими цепями. В качестве признаков использовались координаты атомов, а в качестве целевой метки был выбран класс цепи (0 или 1). Поскольку данный метод машинного обучения является линейным, в результате его обучения как классификатора цепей можно получить уравнение разделяющей плоскости вида:

$$D = W_1 x + W_2 y + W_3 z + \beta,$$

где W_1, W_2, W_3 – веса, β – смещение.

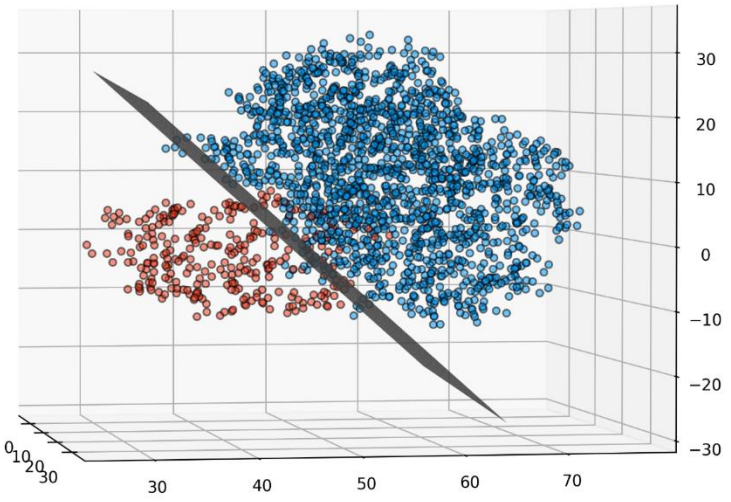

Рисунок 1. Точечное представление комплекса ингибитора триптазы с трипсином 1an1. Рассчитанная разделяющая плоскость выделена серым цветом, атомы трипсина — синим, а ингибитора — красным.

Далее были рассчитаны центры масс скоплений атомов, расположенных по обе стороны от разделяющей плоскости (в границах ± 10 Å). Из полученных точек строился вектор, а его пересечение с разделяющей плоскостью принималось за центр ячейки (Рис. 2).

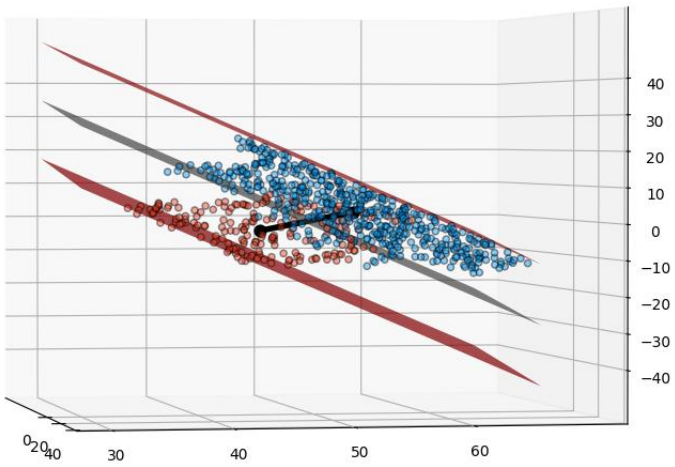

Рисунок 2. Точечное изображение комплекса 1an1 и вектор (черный), соединяющий центры масс областей, ограниченных разделяющей плоскостью и параллельных ей на расстоянии 10 Å (отмечено красным). Точка пересечения вектора и разделяющей плоскости принимается за центр ограничивающей ячейки.

Для определения положения ячейки относительно разделяющей плоскости были заданы три новые прямые. Для этого вблизи заданных заранее параллельных плоскостей, расположенных на расстоянии 10 Å от центра, были выбраны два наиболее удаленных атома. С их помощью была найдена прямая, определяющая наибольший разброс атомов в области взаимодействия. Проекция этой линии на разделяющую плоскость использовалась в качестве оси OX в новой системе координат, нормаль к плоскости стала осью OZ, а ортогональный им вектор стал осью OY. Затем центр ячейки смещался в начало координат и осуществлялся переход к новому базису по полученным векторам. После этого закреплялись границы ячейки размером 41x81x81: высота, ширина и длина в Å соответственно (Рис. 3).

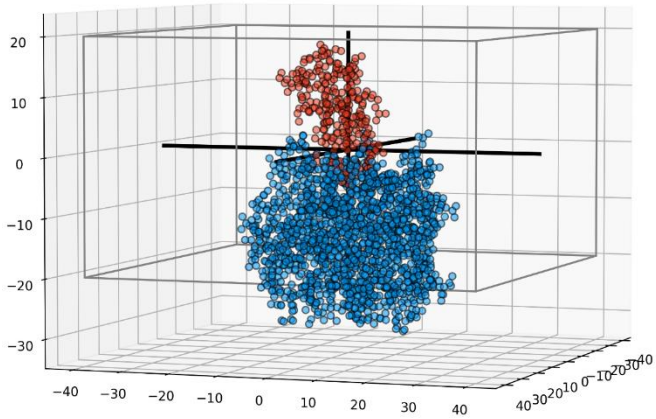

Рисунок 3. Точечное представление комплекса 1an1, координаты которого были преобразованы так, что центр ограничивающей ячейки (серый цвет) находится в начале координат. Оси OX, OY, OZ в новой системе координат выделены черным цветом.

Выбор такого размера ограничивающей ячейки был сделан на основе анализа размеров интерфейса связывания в изученных белок-белковых комплексах. Ширина и высота выбраны исходя из возможности включения в анализ максимального количества крупных

комплексов. Таким образом, области заданного размера было достаточно, чтобы вместить весь интерфейс взаимодействия для более чем 93% всех комплексов.

Этап 2. Отбор атомов.

Из атомов, попавших в ограничительную ячейку, для последующей работы отбирались те, которые находились на расстоянии не более 10 Å от ближайшего атома взаимодействующей цепи. Координаты выбранных атомов были сохранены в трехмерном массиве размером 41x81x81 с разрешением 1 Å.

Этап 3. Добавление каналов с информацией о химических свойствах молекул

Для успешного обучения нейронной сети, помимо самого расположения атомов взаимодействующих цепей, необходимо добавить дополнительную информацию о свойствах, влияющих на аффинность связывания белков и пептидов. При подборе таких признаков внимание было сосредоточено на контактах, которые вносят существенный вклад в формирование белок-белковых взаимодействий (водородные, ионные, гидрофобные, стэкинг-взаимодействия и т.д.) и в структуры основной цепи (расположение карбонильных групп). Распределение атомов по группам было основано на реализованном ранее алгоритме анализа белок-лигандных взаимодействий Arpeggio (Jubb et al., 2017). Итоговое распределение атомов по каналам показано в Таблице 1.

В результате к каждому массиву с атомами было добавлено по 10 каналов. В каждом канале атомы белка 1 обозначались цифрой 1, атомы второго белка -1. Такое представление особенностей позволяет подчеркнуть важные взаимодействия, происходящие на разных расстояниях между атомами. Таким образом, каждый белок-белковый комплекс был преобразован в 4D массив размером 10x41x81x81 (каналы, высота, ширина и длина соответственно).

Таблица 1. Распределение разных типов атомов по каналам.

Номер канала	Белок 1	Белок 2
0	Акцепторы водородных связей	Доноры водородных связей
1	Доноры водородных связей	Акцепторы водородных связей
2	Акцепторы водородных связей	Слабые доноры водородных связей
3	Слабые доноры водородных связей	Акцепторы водородных связей
4	Положительно заряженные атомы	Отрицательно заряженные атомы
5	Отрицательно заряженные атомы	Положительно заряженные атомы
6	Атомы гидрофобных групп	Атомы гидрофобных групп
7	Карбонильные углероды	Карбонильные углероды
8	Карбонильные кислороды	Карбонильные кислороды
9	Атомы ароматических групп	Атомы ароматических групп

Этап 4. Расширение обучающей выборки

Из-за ограниченного объема данных и сложного представления каждого объекта высока вероятность быстрого переобучения нейронной сети. Чтобы замедлить его возникновение и улучшить способность алгоритма к обобщению, было использовано несколько методов аугментации. Так, в обучающую выборку были добавлены новые конформации, полученные методом МД для 142 комплексов. Таким образом, с помощью этого метода была расширена обучающая выборка, за счет включения в нее информации о конформационной подвижности белков.

Помимо расширения обучающего набора данных, были использованы дополнительные подходы аугментации. В частности, перед каждой эпохой обучения с вероятностью 0,5 по каждой из осей (x, y, z) атомы внутри ячейки независимо поворачивались на 180°. Другим методом трансформации была замена с вероятностью 0,5 обозначений атомов одного белка на другой (1 заменялась на -1 и наоборот). Дополнительно для стабилизации значений функции потерь были стандартизированы значения pK_D.

Описание разработанного предсказательного алгоритма

В диссертационной работе был создан предсказательный алгоритм ProBAN (Protein Binding Affinity Network) на основе глубокой сверточной нейронной сети.

В итоговой архитектуре сети последовательно расположены четыре сверточных слоя Conv3D (3D Convolution Layer) с уменьшением размера ядра свертки (7, 5, 3, 3) и увеличением количества каналов (32, 64, 128, 256). Благодаря этому подходу анализируются сложные нелинейные зависимости на основе расстояний между атомами, участвующими в разных типах взаимодействий. После сверточных слоев данные преобразуются в одномерный массив и отправляются на последующие полносвязные слои. Последний полносвязный слой напрямую выводит стандартизированное значение pK_D, и поэтому после него нет функции активации (остальные слои содержат нелинейную функцию активации ReLU). Для обучения использовался оптимизатор AdamW, для расчета ошибки использовалась функция потерь MSELoss, подходящая для решения задачи регрессии. В качестве показателей качества прогнозирования использовались корреляция Пирсона и RMSE (ккал/моль). Для сравнения с другими алгоритмами для ΔG также рассчитывалось значение MAE (ккал/моль).

Общая блок-схема процесса прогнозирования значения pK_D на основе пространственной структуры показана на Рисунке 4.

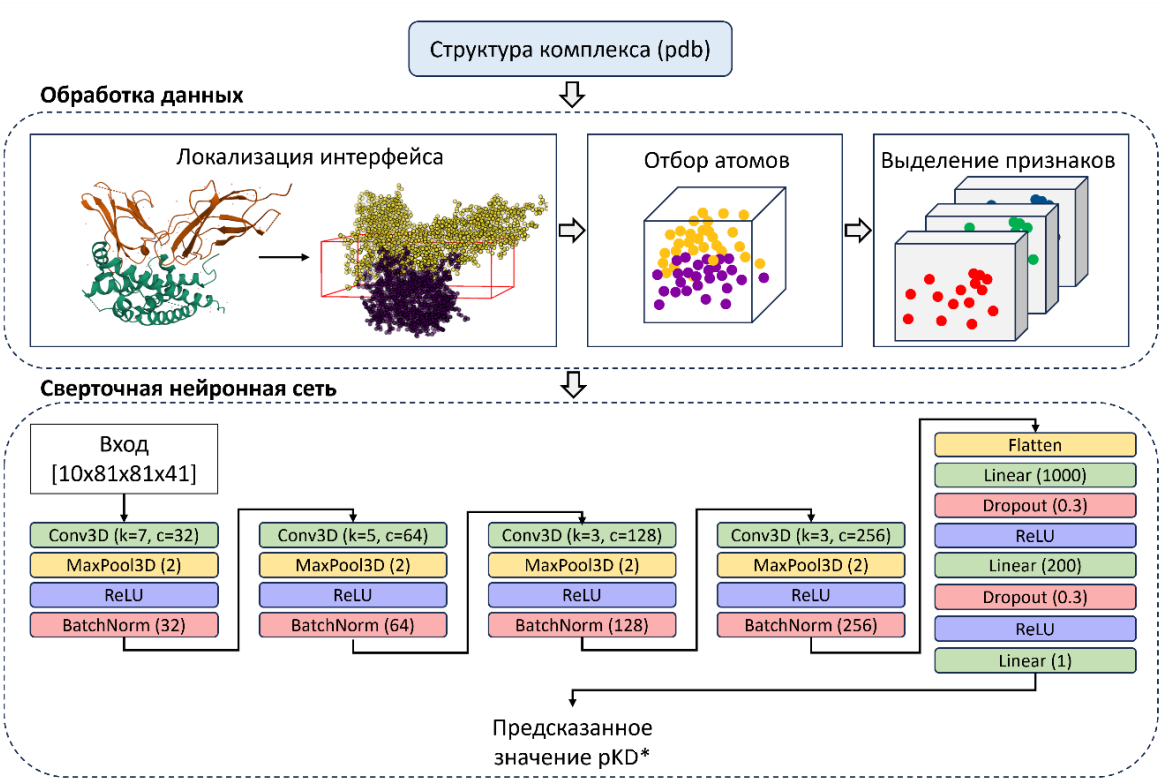

Рисунок 4. Полная схема обработки комплекса и прогнозирования его pK_D. Во-первых, интерфейс связывания локализуется внутри ограничивающей ячейки, а атомы внутри ячейки используются дальше. На следующем этапе происходит отбор атомов, важных для связывания. Затем к полученной трехмерной структуре добавляются каналы, в которые попадают атомы, участвующие в разных типах взаимодействий, и строится 4D-массив. Он отправляется на вход нейронной сети, состоящей из четырех сверточных и трех полносвязных слоев. На выходе последнего слоя выводится значение $pK_D = -\log(K_D)$.

Модель ProBAN обучалась в два этапа. На первом этапе обучение проводилось в течение 20 эпох со скоростью обучения (learning rate) = 0,0001 и Weight_decay = 0,001. В результате лучшая модель была сохранена и отправлена на дополнительное обучение на 10 эпох с learning rate = 0,00001, Weight_decay = 0,00001, и лучшая модель была сохранена (корреляция Пирсона 0,61, MSELoss = 0,45). Благодаря такому подходу удалось провести основную оптимизацию параметров на первом этапе и частично улучшить качество на втором за счет ослабления регуляризации (позволяет еще больше увеличить веса признаков) и уменьшения шага обучения. Остальные гиперпараметры были выбраны на этапе оптимизации нейронной сети и оставались постоянными на протяжении всего процесса обучения.

Апробация разработанного алгоритма на тестовых выборках

Обученная модель была апробирована на сформированных тестовых наборах комплексов. Для внутреннего тестового набора удалось получить значение корреляции Пирсона 0,6 (p-значение = 8e-14) и RMSE = 1,43. Наилучшие прогнозы были получены для комплексов с $pK_D > 8$, тогда как наибольшая ошибка наблюдалась для комплексов с $pK_D < 4$ (Рис. 5A), что

связано с их недостаточной представленностью в обучающем наборе данных. Для внешнего теста удалось получить значение корреляции Пирсона 0,55 (p-значение = 5e-08) и RMSE = 1,67 (Рис. 5B). При этом лучшие предсказания наблюдаются для комплексов со значениями pK_D от 4 до 6. Наибольшая ошибка характерна для комплексов с наиболее сильно отклоняющимися pK_D (менее 4 и более 10). Из-за большого разброса значений pK_D и отсутствия отбора комплексов по разрешению структуры метрики качества во внешнем тестовом наборе уступают внутреннему. В то же время внутренний тестовый набор содержит в основном комплексы, состоящие более чем из двух цепей, и полученная метрика значения указывает на стабильное качество предсказания аффинности для таких структур в диапазоне pK_D от 4 до 10.

Для итоговой модели был проведен анализ важности каналов, которые в данной задаче играют роль признаков. Разработанный алгоритм был обучен с добавлением регуляризации L2, которая ограничивает максимальные значения весов, поэтому можно оценить важность признаков, просматривая распределения весов, связанные со сверточными фильтрами на первом слое (Рис. 5C).

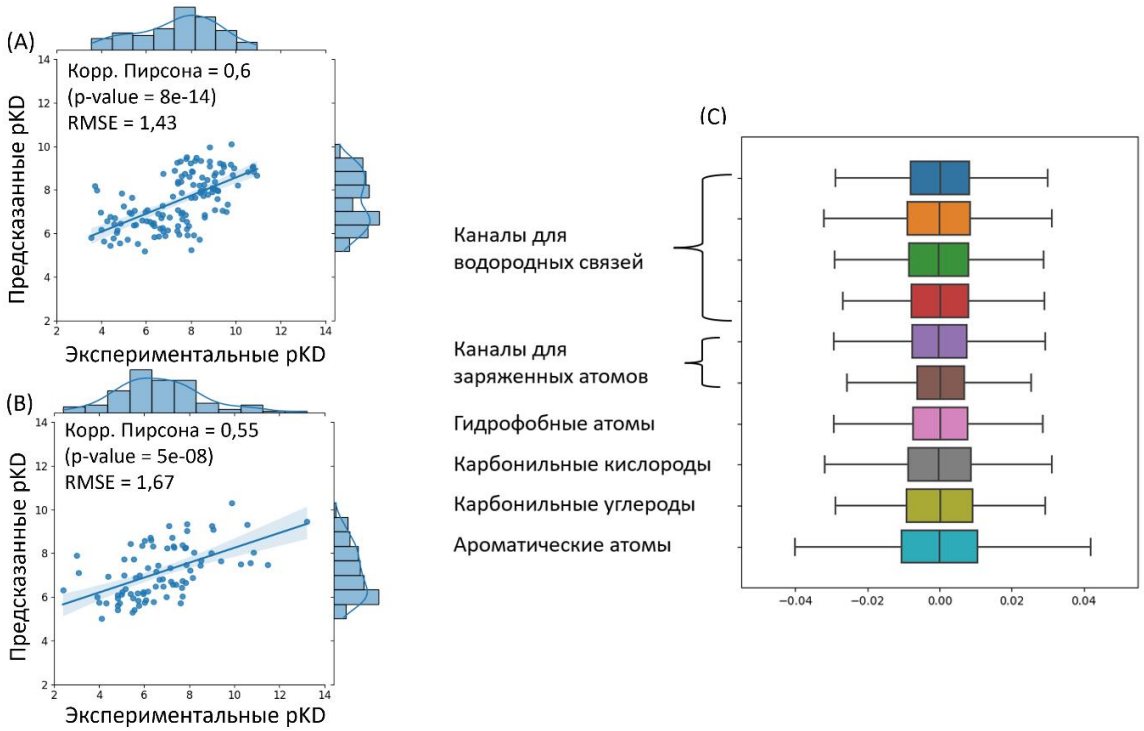

Рисунок 5. Результат тестирования обученной нейронной сети и анализа значимости признаков. (A) Диаграмма рассеяния комплексов из внутреннего тестового набора (тест 1). Ось X содержит истинные значения pK_D, а ось Y содержит прогнозируемые значения. (B) Диаграмма рассеяния для внешнего набора тестов (тест 2). (C) Коробчатая диаграмма, отражающая разброс весов, присвоенных каналам в первом слое нейронной сети. Чем больше разброс, тем выше значимость признака, отраженного в канале. Ось Y указывает на типы атомов, расположенных в каждом канале.

Основная идея заключается в том, что веса каналов, которые оказывают большее влияние на результаты, имеют более высокие абсолютные значения. Это происходит потому, что во

время обучения алгоритм распределяет веса таким образом, чтобы передать больше информации на более глубокие уровни сети. Однако благодаря наличию регуляризации L2 только самые важные каналы имеют такие высокие веса.

В целом можно сказать, что все каналы вносят существенный вклад в предсказание, поскольку нет каналов с критически малым разбросом весов. При этом наиболее широкий диапазон характерен для каналов с ароматическими атомами, заряженными ионами и атомами карбонильной группы. Следовательно, модель делает большую часть своих предсказаний исходя из этих особенностей, что согласуется с известными закономерностями в связывании молекул.

Чтобы сравнить ProBAN с другими прогностическими моделями, рассчитали значение ΔG на основе предсказанных констант диссоциации (Таблица 2). Результаты прогнозирования для внутреннего тестового набора не удалось сравнить с другими алгоритмами из-за наличия комплексов с тремя и более молекулами, для которых другие алгоритмы не делают прогнозы аффинности связывания. Результаты для обоих тестов достаточно высокие и стабильные, что указывает на возможность анализа белок-белковых и белок-пептидных комплексов, интерфейс связывания которых может быть локализован в пределах ячейки размером 41x81x81 Å.

Разработанный в 2022 году веб-сервис PPI-Affinity показал гораздо более высокую производительность, чем другие современные методы, на двух наборах тестов, один из которых в этой работе был собран непосредственно из данных PDBBind (v.2020). В данной работе этот набор тестов также использовался для оценки производительности ProBAN (Таблица 2, Тест 2) в сравнении с другими доступными в настоящее время инструментами по предсказанию аффинности связывания в белок-белковых комплексах.

В результате оценки эффективность разработанной модели на данном наборе данных получены следующие значения метрик качества: коэффициент корреляции R = 0,55, MAE = 1,75 ккал/моль и RMSE = 2,28 ккал/моль, что ставит ProBAN на первое место по всем показателям.

Таблица 2. Оценка ProBAN и других предикторов на двух тестовых наборах данных по предсказанию аффинности связывания в комплексах белок-белок

Метод	Корреляция Пирсона	MAE (ккал/моль)	RMSE (ккал/моль)
Тест 1			
ProBAN	0,60	1,6 ±0,1	2±0,1
Тест 2			
PRODIGY	0,28	2,5±0,3	3,5 ±0,4
DFIRE	0,08	25 ±1,6	29,2 ±2,1
CP_PIE	-0,10	10,9 ±0,3	11,3 ±0,3
ISLAND	0,28	2,3 ±0,2	2,9 ±0,3
PPI-Affinity	0,49	1,8 ±0,2	2,4 ±0,3
ProBAN	0,55	1,8±0,2	2,3 ±0,3

Эффективность работы алгоритма также оценивалась на наборе комплексов дикого типа, взятых из набора данных SKEMPI v2.0. Было отобрано подмножество из этого набора данных (только комплексы с известной пространственной структурой), применив следующие шаги фильтрации: (1) удаление комплексов, которые перекрывались между наборами данных SKEMPI и PDBbind (v.2020), которые использовались для обучения и тестирование моделей; и (2) удаление комплексов с более чем одним значением аффинности связывания. Использованные фильтры сократили набор данных до 117 комплексов дикого типа. Для пяти комплексов не удалось локализовать интерфейс взаимодействия в ограничивающей ячейке, поэтому они были удалены из набора. Таким образом, окончательный тестовый набор содержал 112 структур дикого типа. В связи с наличием более двух цепей в структуре большинства выбранных комплексов для сравнения с PPI-Affinity был выделен отдельный набор комплексов, состоящий всего из двух цепей. Дополнительный набор включал 26 комплексов дикого типа. Результаты прогнозирования для этого набора данных (R = 0,78 и MAE = 1,1 ккал/моль) были сопоставимы с результатами прогнозирования PPI-Affinity (R = 0,77 и MAE = 1,1 ккал/моль). Этот результат свидетельствует о стабильности ProBAN при работе со структурами, состоящими из двух цепочек. Однако показатели ProBAN для полного набора данных (R = 0,47 и MAE = 2 ккал/моль) уступают показателям, полученным на основе других наборов данных.

Полученный результат может быть связан с большим разбросом энергий связи в мультимолекулярных комплексах из этого набора данных, и, следовательно, для наиболее отклоняющихся значений прогнозы были более низкого качества.

Оценка влияния точечных мутаций на изменение энергии связывания в комплексах ACE2-RBD

Помимо апробации алгоритма на разнородных тестовых наборах данных, он также был протестирован на отдельно собранном наборе из комплексов RBD-ACE2. Данный набор состоит из комплексов белков, различающихся несколькими аминокислотными позициями, но при этом с разными значениями энергии связывания. Анализ работы алгоритма в таких условиях позволит оценить возможность его применимости для оценки влияния точечных мутаций на характеристики белок-белковых взаимодействий.

Рассматриваемые в настоящей работе комплексы RBD-ACE2 образованы RBD S-белка коронавирусов SARS-CoV и SARS-CoV-2 и молекулой ACE2 (Рис. 6A). Взаимная ориентация взаимодействующих белков и конформация интерфейса весьма консервативны (Рис. 6B): совмещение структур комплексов по Cα-атомам указанных 48 остатков даёт СКО<1 Å для всех рассмотренных структур. Поверхность непосредственного контакта имеет сложную форму, но может быть заключена в параллелепипед с размерами 45 Å x 15 Å x 15 Å. Анализ межмолекулярных взаимодействий показывает, что основными из них являются гидрофобные контакты и водородные связи.

В первом приближении интерфейс взаимодействия RBD-ACE2 состоит из двух макрообластей плотного примыкания молекул и полости между ними (Рис. 6C). Будем для определённости называть эти макрообласти большой (Рис. 6C, слева) и малой (Рис. 6C, справа), поскольку число вовлеченных в их образование аминокислотных остатков со стороны RBD составляет 15 и 11, а со стороны ACE2 11 и 11, соответственно. Для обеих макрообластей характерно наличие многочисленных гидрофобных контактов и водородных связей. Считается, что наиболее значимыми для связывания являются три области (hot spot), первая из которых соответствуют малой макрообласти, а две другие большой макрообласти.

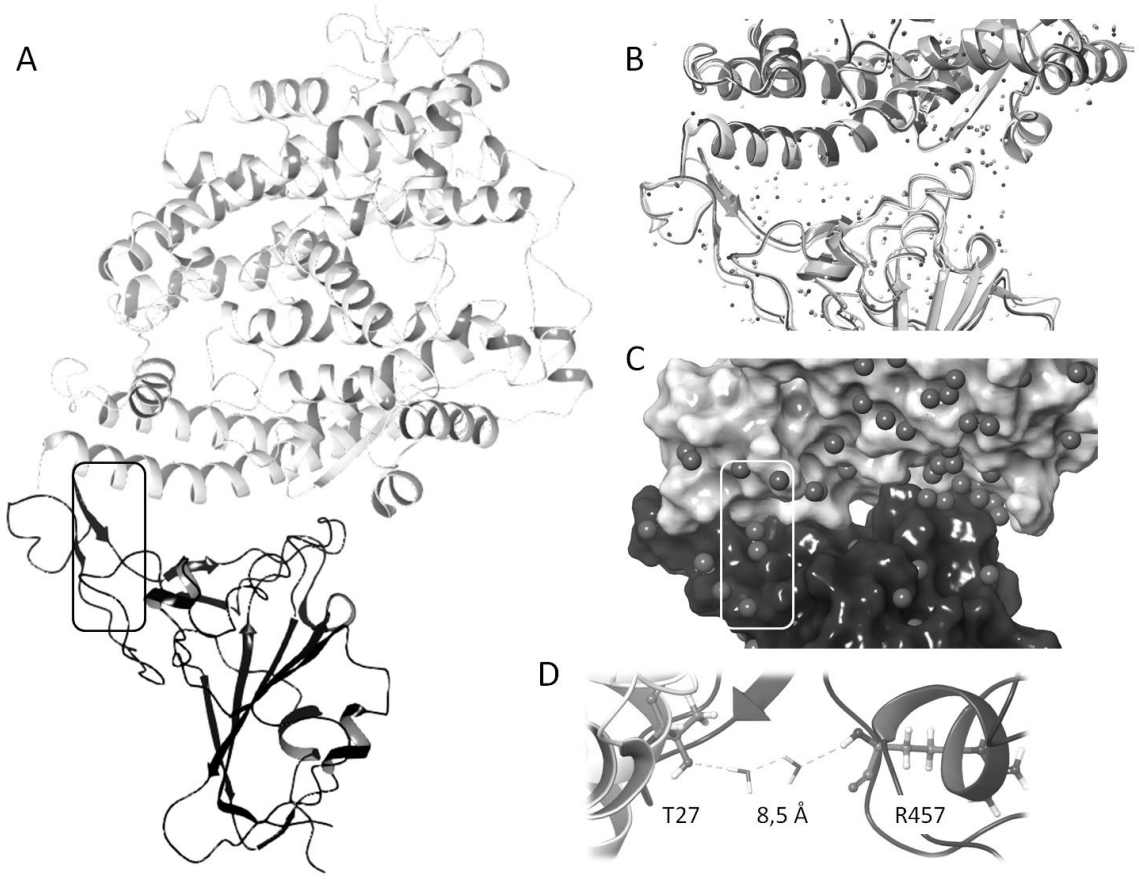

Рисунок. 6. A. Общий вид комплекса ACE2 (показан светлым) и-RBD (показан тёмным) (pdb код 6lzg) в ленточном представлении. Рамка соответствует области расположения остатков T27 (ACE2) и R457 (RBD). B. Совмещение интерфейсов взаимодействия ряда структур комплекса ACE2-RBD: 6lzg (показана самым тёмным), 7ekh (показана тёмным), 7lo4 (показана светлым) и 8df5 (показана самым светлым). Молекулы кристаллизационной воды показаны шариками соответствующих цветов (масштаб не соблюдён). C. Интерфейс взаимодействия ACE2-RBD (pdb код 6lzg). Молекулярные поверхности субъединиц, соответствующих ACE2 и RBD, показаны светлым и тёмным, соответственно. Атомы кислорода, соответствующие молекулам воды, показаны промежуточным серым. Рамка соответствует области расположения остатков T27 (ACE2) и R457 (RBD). D. Две молекулы воды и цепочка водородных связей, обеспечивающие взаимодействие остатка T27 ACE2 (слева) и остатка R457 RBD (справа) Расстояние между соответствующими атомами белка 8,5 Å.

Обращает на себя внимание полость, расположенная между макрообластями (Рис. 6C). В естественных условиях она, очевидно, заполнена молекулами воды и ионами, однако в известных кристаллографических структурах молекулы кристаллизационной воды в этой полости отсутствуют. Ключевой причиной этого явления может являться высокая подвижность молекул воды в этой области, вызванная несоответствием гидрофобных свойств поверхностей молекул ACE2 и RBD в этой области пространства.

Несмотря на влияние молекул воды и ионов, расположенных на интерфейсе взаимодействия или в его окрестности, на организацию белок-белковых комплексов (Reichmann et al., 2008), при описании экспериментальных структур комплексов RBD-ACE2 этим молекулам практически не уделяется внимание. Между тем, структуры, полученные методом рентгеновского структурного анализа с высоким разрешением (как правило, 2,5 Å или лучше), содержат большое число молекул кристаллизационной воды. Так, в структурах 6lzg, 7ekh, 7lo4 и 8df5 в непосредственной близости от ACE2 и RBD содержится 322, 250, 132 и 163 молекулы воды, соответственно, а структура 8df5 содержит ещё и один ион хлора. Большая часть этих молекул расположена в карманах на поверхности белка hACE2, однако заметное число находится и в окрестности интерфейса взаимодействия этого белка с RBD (Рис. 6C).

Для выявления молекул кристаллизационной воды, которые могут опосредовать белок-белковое взаимодействие, было выполнено добавление атомов водорода и оптимизация сети водородных связей для рассматриваемых структур высокого разрешения (6lzg, 7ekh, 7lo4 и 8df5) в программе Maestro (Schrodinger, LLC), причем положение тяжелых атомов не подвергалось изменению. Для рассмотренных структур высокого разрешение характерно наличие как минимум нескольких цепочек водородных связей. Рассмотрение таких цепочек показывает, что во взаимодействии белков в комплексе играют роль не только остатки, непосредственно образующие нековалентные взаимодействия, но и образующие такие взаимодействия посредством молекул воды. Это предположение находит подтверждение и в литературе (Schweke et al., 2020). В частности, показано, что поверхность взаимодействующих белков, примыкающих к интерфейсу их взаимодействия, но не вовлеченная в него непосредственно, обогащена полярными атомами (т.е. атомами N и O). Очевидно, полярные атомы, расположенные недостаточно близко для образования непосредственной водородной связи, могут сделать это, образовав связи посредством молекул воды (Рис. 6D).

Однако прямой учет подобных связей в белок-белковых комплексах по ряду причин является затруднительным. Прежде всего, для такого учета необходимо наличие молекул воды в явном виде и в достаточном количестве, что далеко не всегда наблюдается даже для структур высокого разрешения. Это делает необходимым обращение к методам молекулярного моделирования для создания и оптимизации водного окружения интерфейса.

Рассмотрение комплексов RBD-ACE2 с известными пространственной структурой и значением K_D выявило 48 комплексов. Большая часть этих комплексов имеют значение pK_D от 7 до 9, в то время как низкая аффинность связывания ($pK_D<7$) характерна для 5

комплексов, два из которых образованы RBD вируса SARS CoV (3sci, 6cs2), а три оставшихся содержат RBD SARS CoV-2 с мутациями F486L (7eke (ACE2 человека), 7wa1 (ACE2 норки)) или Y453F (7w8s (ACE2 норки)). Комплексы с самой высокой аффинностью связывания содержат hACE2 и RBD SARS CoV-2 с мутациями D614G, N501Y, E484K, K417N (7sy4, 7sy8) или RBD SARS CoV-2 Omicron BA.2.75 (8asy), BQ.1.1 (8if2) и рекомбинантный вариант XBB.1 (8iov).

В результате предсказания константы диссоциации моделью ProBAN удалось получить значение корреляции Пирсона между экспериментальными и рассчитанными значениями pK_D равное 0,56 и MAE = 0,5 (Рис. 7). Среди с комплексов с наибольшей абсолютной ошибкой (больше 1) подавляющая часть имела разрешение хуже 3 Å (6cs2, 7wk6, 7tex) и один имел разрешение 2,85 Å (8asy). Наблюдаемая закономерность свидетельствует о негативном вкладе нечетко разрешенного положения атомов в качество предсказания аффинности связывания, так как искажается информация об межатомных расстояниях, играющих ключевую роль во взаимодействии между белковыми молекулами. При этом стоит отметить, что, не считая комплексы 6cs2, 7wk6, для которых структуры получены с низким разрешением (4,4 Å и 3,7 Å соответственно), наилучшие предсказания характерны для комплексов с более низкой аффинностью связывания (pK_D <8), что ранее было замечено в работе, посвященной оценке других алгоритмов (Ozden et al., 2024). Данная закономерность может быть связана с тем, что мутации, дестабилизирующие интерфейс связывания, приводят к более крупным конформационным перестройкам, которые более эффективно могут учитываться предсказательными алгоритмами.

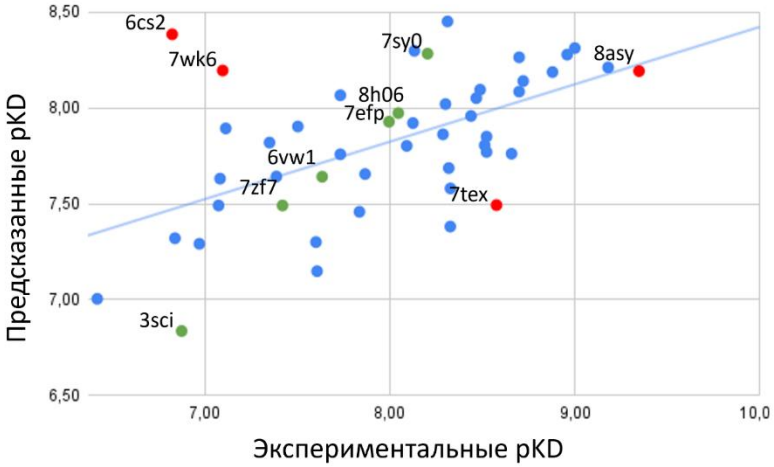

Рисунок 7. Результаты предсказаний pK_D для комплексов RBD-ACE2 алгоритмом ProBAN. (красным выделены предсказания для комплексов с абсолютной ошибкой больше 1, зеленым - с ошибкой меньше 0,1, синим – остальные). В качестве ярлыков добавлены pdb коды.

Для более полного анализа предсказания константы диссоциации алгоритмом ProBAN было проведено его сравнение с предсказаниями, полученными веб-сервисом Prodigy. Данный метод осуществляет оценку аффинности связывания функцией, основанной на межмолекулярных контактах и признаках непосредственно на интерфейсе и полученных из анализа поверхности, не относящейся к интерфейсу взаимодействия. Метрики, полученные в результате оценки данного алгоритма, находятся в Таблице 3. Хорошо видно, что ProBAN показывает более высокое качество предсказания по сравнению с Prodigy. Предположительно, причиной этого является использование как более полной информации

о взаимодействиях между атомами, так и большего порогового значения расстояния между атомами (10 Å), которое классифицирует пары атомов на взаимодействующие и нет. Используемое в Prodigy аналогичное пороговое значение расстояния между атомами (5,5 Å), по-видимому, отсеивает часть важных атомов, вносящих вклад в связывание.

Большое число алгоритмов, используемых для оценки аффинности связывания в белок-белковых комплексах, предсказывают не значение константы диссоциации, а свободную энергию Гиббса связывания. Для оценки работы данных алгоритмов (FoldX, DFIRE, ROSETTADOCK) на исследуемом наборе данных из полученных значений K_D были рассчитаны значения ΔG и проведено сравнение с ΔG полученной с использованием данных алгоритмов Рассчитанные значения метрик качества для разных алгоритмов представлены в Таблице 3.

Таблица 3. Метрики качества предсказания аффинности связывания для комплексов ACE2-RBD для отобранных алгоритмов.

Алгоритм	Корр. Пирсона для ΔG	p-value*	MAE для ΔG (ккал/моль)	MAE для pKD
ProBAN	**0,56**	3,3e-05	**0,7±0,1**	**0,5±0,1**
Prodigy	-0,38	7,2e-03	1,2±0,2	0,9±0,1
FoldX	0,41	4e-03	8,1±0,7	-
DFIRE (все комплексы)	-0,04	0,74	12,3±3,4	-
DFIRE (без 7u0n)	0,14	0,36	9,5±2,9	-
ROSETTADOCK	-0,11	0,46	5±0,4	-

*****расчет p-value осуществляется с использованием Z-преобразования Фишера

ProBAN оказался наиболее эффективным среди всех проанализированных алгоритмов. На втором месте по размеру средней ошибки находится Prodigy, однако, корреляция принимает отрицательное значение, что свидетельствует о неспособности алгоритма оценивать влияние мутаций на направление изменения аффинности связывания в изучаемом наборе данных. Таким образом, учитывая рассчитанное значение MAE, используемая в Prodigy оценочная функция, может использоваться для оценки ΔG с погрешностью в 1,2 ккал/моль. В свою очередь для определения вклада мутаций в аффинность относительно нативной структуры RBD-ACE2 более успешно может быть использован FoldX, который по значению корреляции (0,41) на исследуемом наборе данных находится на втором месте после ProBAN.

Таким образом, результаты предсказания аффинности связывания для комплексов ACE2-RBD подчеркивают важность выбора подходящих алгоритмов для предсказания аффинности связывания и их адаптации к специфике исследуемых белок-белковых

комплексов. Разработанный в данной работе алгоритм ProBAN может в дальнейшем использоваться для оценки влияния точечных мутаций в изменение энергии связывания в белок-белковых комплексах.

Анализ стабильности комплексов, образованных каноничными и замещающими вариантами гистонов

Помимо оценки вклада мутаций в изменение энергии связывания к областям применения разработанного алгоритма может также относиться оценка стабильности белок-белковых комплексов, для которых такие показатели связывания как константа диссоциации и свободная энергия Гиббса экспериментально, как правило, не измеряются, уступая оценке термостабильности (Darzynkiewicz et al., 1989; Berryhill et al., 2024), а также расчету изменения в эффективности FRET (Tóth et al., 2014) и др. В частности, такие подходы используются и для оценки стабильности нуклеосомы, вклад в которую будут вносить как белок-белковые взаимодействия, так и взаимодействия белок-ДНК и посттрансляционные модификации. Так как ProBAN концентрируется на анализе белок-белковых взаимодействий, его использование в данном случае может помочь оценить роль межгистоновых взаимодействий в поддержании стабильности нуклеосомы. Таким образом, объектом для данного анализа послужили димеры и тетрамеры, образуемые разными вариантами гистонов.

Существует четыре класса основных гистонов H2A, H2B, H3 и H4, составляющих октамер. Так, эукариотические нуклеосомы состоят из тетрамера $(H3\text{-}H4)_2$ и двух димеров H2A-H2B, вокруг которых 147 пар оснований ДНК намотаны в 1,7 витка левозакрученной спирали (Luger et al., 1997). Изменения в стабильности ядра нуклеосомы могут привести к изменению уровня экспрессии генов, что, в свою очередь, может повлиять на клеточные функции и процессы, такие как дифференцировка и ответ на стресс.

Канонические варианты гистонов преобладают в нуклеосомах и синтезируются и встраиваются в зависимости от репликации. Дополнительное разнообразие обеспечивается встраиванием в хроматин замещающих вариантов гистонов (Talbert et al., 2021). То, как различные варианты гистонов влияют на стабильность взаимодействий внутри октамера и ядра с ДНК, является предметом современных исследований в области молекулярной биологии (Szenker et al., 2011; Hirano et al., 2021; Kniazeva et al., 2022; El Kennani et al., 2018; Klein et al., 2023).

Перед анализом комплексов, образуемых между молекулами гистонов разных вариантов, производилась проверка алгоритма на способность идентифицировать и оценивать взаимодействия, оказывающие влияние на связывание гистонов с другими белками хроматина (шапероны, импортины и др.). Данное решение связано с отсутствием достаточного количества данных о свободной энергии Гиббса для димеров и тетрамеров гистонов. Таким образом, оценка производилась для комплексов, образованных гистонами и другими белками хроматина с известными значениями аффинности связывания.

По итогам тестирования алгоритма удалось добиться достаточно высокого качества предсказания (корреляция Пирсона = 0,53, MAE = 0,86 ккал/моль), соответствующего

значениям, полученным для ранее проанализированных тестовых выборок (Раздел 3.3 и 3.4). Также можно отметить присутствие в тестовом наборе данных не только каноничных форм гистонов, но и замещающих вариантов (H2A.Z, H2A.6, H3.3 и др.), высокое качество предсказания для которых также свидетельствует о хорошей обобщающей и предсказательной способности разработанного алгоритма. Полученный результат свидетельствует о возможности анализа стабильности белок-белковых комплексов, образуемых разными вариантами гистонов с использованием ProBAN.

Для анализа были отобраны несколько вариантов гистонов H2A, H2B и H3. К примеру, H2A.Z, вариант гистона H2A, необходимый для приспособлений у дрожжей и жизнеспособности многоклеточных организмов (Guillemette and Gaudreau, 2006), играет важнейшую роль в транскрипции генов, репликации ДНК, восстановлении ДНК и поддержании целостности генома (Henikoff et al., 2015; Venkatesh et al., 2015). Биологическая значимость измененной динамики H2A.Z-нуклеосомы плохо изучена, поскольку влияние H2A.Z на стабильность нуклеосомы было спорным (Abbott et al., 2001; Chen et al., 2013; Kim et al., 2016; Osakabe et al., 2018; Rudnizky et al., 2016), что оставляет этот вопрос открытым для исследования.

Из предсказанных значений ΔG можно заметить, что контакты между H2A.Z и H2B оказались немного более стабильными (-13,3 ккал/моль), чем для каноничной формы H2A (-13 ккал/моль), что согласуется с ранее приведенными исследованиями по изучению термостабильности и динамики димеров гистонов с каноничной формой H2A и измененной (Dai et al., 2021).

Другой вариант этого гистона – H2A.J накапливается в фибробластах человека *in vitro*, а также в тканях кожи мышей и человека *in vivo* во время репликативного, онкогенного и радиационно-индуцированного старения и влияет на экспрессию воспалительных генов в стареющих клетках (Contrepois et al., 2017; Isermann et al., 2020; Rube et al., 2021). Ранее в исследованиях утверждалось, что нуклеосома с H2A.J продемонстрировала аналогичный каноничному профиль тепловой денатурации, но первый шаг (отсоединение димеров H2A-H2B) был явно смещен в сторону более высокой температуры (Tanaka et al., 2020). Однако, предсказания энергии связывания между димером H2A-H2B и тетрамером H3-H4 для каноничной формы и варианта H2A.J практически не отличаются (-11,7 ккал/моль), что может свидетельствовать о повышении стабильности нуклеосомы с вариантом H2A.J за счет более прочных контактов гистонов с ДНК, а не путем изменения белок-белковых взаимодействий непосредственно между гистонами.

Также были проанализированы специфичные для семенников варианты гистонов TSH2A.1 и TSH2B.1, которые экспрессируются исключительно во время сперматогенеза (Tanaka et al., 2004; Luger et al., 1999; Cheung et al., 2003) и в ооцитах (Nusinow et al., 2007). В результате полученного предсказания можно заметить, что димеры H2A-H2B содержащие только один из специфичных для семенников вариантов гистонов менее стабильны (-10,5 ккал/моль и -11,9 ккал/моль), чем каноничный вариант (-13 ккал/моль), однако димер TSH2A.1-TSH2B.1 является даже более стабильным (-13,5 ккал/моль), чем в

каноничной форме. Полученные результаты согласуются с ранее проведенными исследованиями (Shinagawa et al., 2014), при этом можно заметить, что наибольший вклад в усиление взаимодействий вносит вариант TSH2B.1. Однако, присутствие варианта TSH2B.1 в нуклеосоме ослабляет взаимодействие между димером H2A-H2B и тетрамером H3-H4 (-11,2 ккал/моль). Данное явление может быть связано со специфичным для TSH2B.1 аминокислотным остатком Ser85. Остаток Ser85 TSH2B.1 не взаимодействует с H4 в нуклеосоме, но в канонической нуклеосоме остаток Asn84 H2B (соответствующий остатку Ser85 TSH2B.1) образует водородные связи с остатком Arg78 H4, опосредованные водой (Urahama et al., 2014).

Помимо вариантов гистонов H2A и H2B оценивалось взаимодействие двух вариантов H3 (H.3.3 и H.3.6) с H4. H3.3 — консервативный вариант гистона, который структурно очень близок к каноническому гистону H3 — связан с активной транскрипцией (Szenker et al., 2011). Кроме того, его роль в замещении гистонов в активных генах и промоторах очень консервативна, и было высказано предположение, что он участвует в эпигенетической регуляции активных состояний хроматина. В результате оценки взаимодействия между H3 и H4 было выявлено, что вариант H3.3 имеет небольшое снижение аффинности связывания (-12,7 ккал/моль) относительно каноничного варианта (-12,9 ккал/моль), при этом для варианта H3.6 это снижение является гораздо более значимым (-11,3 ккал/моль). Из литературных данных известно, что нуклеосома H3.6 менее термически стабильна по сравнению с нуклеосомой H3.3, что связано с остатком Val62 в H3.6, который, как видимо, полностью отвечает за нестабильность нуклеосомы H3.6, вероятно, из-за ослабленного гидрофобного взаимодействия с H4.

Полученные результаты открывают возможность изучения различных вариантов гистонов и вклада образуемых ими белок-белковых взаимодействий в общую стабильность и динамику нуклеосомы методами машинного обучения, в частности, с использованием разработанного нейросетевого алгоритма.

ЗАКЛЮЧЕНИЕ

Изучение механизмов и особенностей белок-белковых взаимодействий является одной из ключевых задач как биоинформатики, так и молекулярной биологии. Энергия связывания характеризует сродство молекул, вступающих во взаимодействие. Определение данной характеристики в белок-белковых комплексах является сложной задачей, которая напрямую влияет на разработку многих пептидных и белковых лекарственных препаратов (противоопухолевые, противовирусные и др.).

На основе проанализированной информации об особенностях белок-белковых взаимодействиях и альтернативных подходов предсказания энергии связывания был предложен новый метод предобработки пространственных структур, позволяющий в автоматическом режиме локализовывать интерфейс взаимодействия внутри ограничительной ячейки. И далее, с использованием подходов искусственного интеллекта был разработан новый алгоритм прогнозирования аффинности связывания в белок–белковых комплексах. Предсказательная модель основана на глубокой свёрточной

нейронной сети, архитектура которой позволяет выделять важные для связывания взаимодействия и свойства. По результатам тестирования на разнородных наборах данных, разработанная модель превосходит все существующие альтернативные методы предсказания аффинности. Использование подходов глубинного обучения в данном исследовании позволило учесть как пространственные характеристики, так и химических свойства контактирующих молекул.

В рамках апробации разработанного алгоритма были проанализированы особенности интерфейса взаимодействия в разнородных группах белковых комплексов, в частности в комплексах ACE2-RBD спайкового белка коронавирусов. В результате были выделены важные для связывания взаимодействия, в частности, опосредованные молекулами воды и сделаны предсказания энергии связывания для различных мутантных форм, превосходящие по точности альтернативные подходы.

Помимо комплексов с экспериментально рассчитанными значениями энергии связывания также была произведена оценка взаимодействий между различными вариантами гистонов, для которых нет такой информации, что позволило сопоставить известные характеристики термостабильности нуклеосом с разными замещающими вариантами гистонов с предсказанными значениями свободной энергии Гиббса. Также по полученным результатам были сделаны предположения о вкладе белок-белковых взаимодействий с участием замещающих вариантов гистонов в стабильность нуклеосомы в целом.

Таким образом, разработанный в диссертационном исследовании предсказательный алгоритм может применяться в различных областях молекулярной биологии, биоинформатики и фармакологии в частности для решения задач оценки влияния точечных мутаций на стабильность комплексов, а также для подбора новых терапевтических белковых мишеней и фармакологически активных пептидных соединений, что в дальнейшем может значительно ускорить ранние этапы разработки лекарственных препаратов, основанных на воздействии на белок-белковые взаимодействия или на создании новых белок-белковых или белок-пептидных комплексов.

ВЫВОДЫ

1. Собранный набор данных из пространственных структур белок-белковых комплексов с известными характеристиками связывания, расширенный конформациями, полученными методами МД, обладает репрезентативностью в широком диапазоне значений аффинности. Однако, для анализа особенностей взаимодействия в белок-белковых комплексах со значениями K_D меньше 4 и больше 10 необходимо получение новых экспериментальных данных о структуре и характеристиках связывания.
2. В результате анализа интерфейса взаимодействия в комплексах ACE2-RBD, данные о которых включали в себя как нативные, так и мутантные формы, были выявлены особенности структуры низко- и высокоаффинных комплексов, свидетельствующие о значительном вкладе в сродство связывания контактов, опосредованных молекулами воды.

3. Предложенный метод предобработки пространственных структур белок-белковых комплексов позволяет учитывать различные типы контактов (водородные, гидрофобные, ионные и т.д.), важных для формирования белок-белковых взаимодействий, а также позволяет сохранить информацию о пространственном расположении атомных групп, участвующих в образовании данных контактов.
4. Разработанный предсказательный алгоритм на основе трехмерной сверточной нейронной сети позволяет предсказывать значение константы диссоциации и свободной энергии Гиббса для белок-белковых комплексов, интерфейс взаимодействия в которых возможно локализовать в ограничительной ячейке размера 41x81x81 Å.
5. В результате оценки эффективности на внутреннем и внешнем тестовых наборах, разработанный алгоритм показал лучшее качество предсказания среди всех проанализированных подходов. Учитывая разнородность тестовых наборов данных, можно сделать вывод о возможности применения разработанного алгоритма для разных типов белок-белковых комплексов: белок-белковые, белок-пептидные, с моно-и мультидоменными взаимодействиями.
6. По результатам оценки аффинности для набора комплексов ACE2-RBD с использованием разработанной модели было достигнуто наиболее высокое качество предсказания по сравнению с альтернативными методами. Полученный результат свидетельствует о высокой чувствительности предсказательного алгоритма к структурным изменениям белковых молекул, обусловленных точечными аминокислотными заменами.
7. Проведенный анализ энергии связывания в комплексах, образованных каноничными и замещающими вариантами гистонов, показал, что варианты H2A.Z, TSH2A.1 и TSH2B.1 (при наличии обоих вариантов) оказывают стабилизирующее воздействие на белок-белковые взаимодействия в ядре нуклеосомы, а вариант H3.6 наоборот, дестабилизирует межгистоновые взаимодействия.

НАУЧНЫЕ СТАТЬИ ПО ТЕМЕ ДИССЕРТАЦИИ, ОПУБЛИКОВАННЫЕ В ЖУРНАЛАХ SCOPUS, WOS, RSCI[1]

- **Bogdanova E. A.**, Novoseletsky V. N. ProBAN: Neural network algorithm for predicting binding affinity in protein–protein complexes // Proteins: Structure, Function and Bioinformatics. — 2024. — V. 92, № 9, P. 1127–1136, JIF (для WoS) = 3,2, Q1 - (1,2/1,1), DOI: 10.1002/prot.26700.
- **Bogdanova E. A.**, Novoseletsky V. N., Shaitan K. V. Binding affinity prediction in protein-protein complexes using convolutional neural network // Advances in Neural Computation, Machine Learning, and Cognitive Research VII. NEUROINFORMATICS 2023. — Vol. 1120 of Studies in Computational Intelligence. — Springer Cham: 2023. — P. 389–397, SJR (для Scopus)=0,21, Q4 - (1/0,85), DOI: 10.1007/978-3-031-44865-2_42.

[1] В скобках приведен объем публикации в печатных листах и вклад автора в печатных листах

- **Богданова Е. А.**, Чернухин А. В., Шайтан К. В., Новоселецкий В. Н. Оценка аффинности связывания в комплексах ACE2-RBD S-белка коронавирусов с использованием сверточных нейронных сетей // Биофизика. – 2024. – Т. 69, № 5, Р. 979–989, РИНЦ (для RSCI и ВАК/МГУ)=0,58, (1,6/0,8), DOI: 10.31857/S0006302924050053.

ДРУГИЕ НАУЧНЫЕ РАБОТЫ ПО ТЕМЕ ДИССЕРТАЦИИ

- **Богданова Е. А.**, Тычинин Д. И., Новоселецкий В. Н. Анализ влияния мутаций на аффинность связывания в комплексах ACE2 и RBD S-белка коронавирусов // Journal of Bioinformatics and Genomics. — 2023. — Т. 4, № 22, (0,8/0,55), DOI: 10.18454/jbg.2023.22.8.