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

ДИССЕРТАЦИЯ
на соискание ученой степени
кандидата биологических наук

Научный руководитель:
к.ф-м.н. Новоселецкий Валерий Николаевич

Москва – 2025

Оглавление

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

ВВЕДЕНИЕ

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 существующими 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 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 осуществляли 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)[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.

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

- **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.
- **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
- **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.

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.

ГЛАВА 1. ОБЗОР ЛИТЕРАТУРЫ

1.1. Белок-белковые взаимодействия

Перед дальнейшим анализом стоит заметить, что широко используемые в данной области исследований термины «белок», «полипептид» и «пептид» неоднозначны и могут перекрываться по смыслу. Термин «белок», как правило, используется для обозначения целостной биологической молекулы в стабильной конформации, тогда как под пептидом чаще всего подразумевается короткий аминокислотный олигомер, не имеющий стабильной пространственной структуры. Граница между пептидом и полипептидом четко не определена и расположена в диапазоне 20–30 остатков (Lodish et al., 2000). Соответственно, термин «полипептид» может применяться к любой одиночной достаточно длинной линейной цепи аминокислот, и также часто подразумевает отсутствие определенной конформации.

Белок-белковые взаимодействия лежат в основе интерактома каждой живой системы и регулируют сложные биологические процессы, такие как эндоцитоз, посттрансляционные модификации, сигнальные пути, иммунные ответы и т.д. Кроме того, белок-опосредованные взаимодействия играют важную роль в развитии ряда заболеваний человека, включая некоторые формы рака и вирусные инфекции. Из-за высокой медицинской ценности белок-белковых взаимодействий было проведено большое количество исследований для подбора и синтеза идеальных пептидов в терапевтических и косметических целях. Также было показано, что белок-пептидные взаимодействия можно регулировать с помощью малых молекул (Johansson-Åkhe et al., 2019), что делает их кандидатами на роль мишеней для лекарств. Таким образом, понимание структурных особенностей белок-белковых взаимодействий важно для понимания сложных клеточных процессов и многих заболеваний, а также может служить основой для разработки

лекарственных препаратов, способных к модификации данных взаимодействий.

В настоящее время в биоинформатике реализовано большое число предсказательных алгоритмов, причем наиболее мощными являются методы, основанные на искусственных нейронных сетях. Их широкое распространение в последнее время связано с увеличением вычислительных мощностей и быстрым ростом баз данных, что позволяет реализовывать более точные, но ресурсозатратные алгоритмы для предсказания связывания в комплексах пептид-белок. Экспериментальные методы определения аффинности связывания являются наиболее точными, однако проведение лабораторных экспериментов требует наличия дорогостоящего оборудования, а также значительных временных затрат. В связи с этим, данные методы не позволяют оценить сродство к терапевтической мишени всех потенциальных лекарственных кандидатов, так как во многих случаях их число выходит за несколько сотен тысяч. Следовательно, необходимо развитие новых предсказательных методов, способных ускорить отбор лекарственных кандидатов за счет предварительной оценки их связывания с мишенью.

Десятилетия исследований в области клеточной биологии, молекулярной биологии, биохимии, структурной биологии и биофизики позволили собрать и структурировать обширные данные о функциях и молекулярных свойствах отдельных белков. Однако белки редко выполняют свою функцию в одиночку (De Las Rivas et al., 2010). Нередко они объединяются в так называемые "молекулярные машины", вступая в сложные физико-химические динамические связи с другими белками для выполнения биологических функций как на клеточном, так и на более высоких уровнях. Важнейшим шагом на пути к раскрытию сложных молекулярных отношений в живых системах является изучение механизмов и особенностей белок-белковых взаимодействий в различных комплексах.

Первым необходимым шагом является точное определение того, что такое белок–белковые взаимодействия. Обычно под этим термином понимают высокоспецифичные физические контакты между двумя и более белковыми молекулам, которые происходят в клетке или в живом организме *in vivo* (De Las Rivas et al., 2010). Вопрос о том, имеют ли два белка общий "функциональный контакт", совершенно отличается от вопроса о том, взаимодействуют ли одни и те же два белка непосредственно друг с другом. Любой белок в рибосоме или в транскрипционном аппарате имеет функциональный контакт с другими белками в комплексе, однако, не все белки комплекса непосредственно взаимодействуют. Следовательно, функциональные связи между биомолекулярными образованиями (генами, белками, метаболитами и т. д.) в живых организмах не следует путать с белковыми физическими взаимодействиями (Mackay et al., 2007; De Las Rivas et al., 2010).

Хотя многие взаимодействия между белками включают классическое, хорошо характеризуемое связывание между двумя глобулярными доменами, в последнее время все большее число взаимодействий связывают с пептидно-белковыми взаимодействиями, где короткие линейные пептиды связываются с глобулярными белковыми рецепторами (Wu et al., 2022, Caporale et al., 2021). Так, значительная часть белково-белковых взаимодействий (15-40 %) представляет собой пептид-опосредованные взаимодействия (Raghavender et al., 2019; Petsalaki and Russell, 2008), при которых короткий фрагмент одного белка взаимодействует с более крупным фрагментом другого. Подобные короткие фрагменты часто неупорядочены в несвязанном состоянии, но приобретают стабильную структуру в комплексе (Nesterov et al., 2024; Uversky, 2024).

Пептиды служат перспективными лекарственными кандидатами с высокой специфичностью и относительно низкой токсичностью (Caporale et al., 2021; Wu et al., 2022). В связи с этим в последние годы увеличивается

количество реализуемых на рынке препаратов на основе пептидов (Caporale et al., 2021; Nielsen et al., 2024; Otvos, 2024). При этом мишенями для пептидных препаратов могут служить как липиды, так и белки (Otvos, 2024).

Белок-белковые взаимодействия встречаются в нескольких формах (Рис. 1) (Scott et al., 2016):

• Пары глобулярных белков, которые взаимодействуют через прерывистый эпитоп без существенных структурных изменений при связывании.

• Взаимодействия между парой глобулярных белков, в которых один или оба белка претерпевают существенные конформационные изменения при связывании.

• Комплексы, включающие глобулярный белок, взаимодействующий с пептидом.

• Взаимодействия между двумя пептидными цепями.

Последние два класса можно далее дифференцировать в зависимости от того, претерпевают ли пептиды существенные конформационные изменения при связывании. В некоторых случаях пептид представляет собой внутренне неупорядоченный пептид или участок белка, который сворачивается в специфическую конформацию при связывании, тогда как в других случаях предварительно свернутый участок белка взаимодействует с белком-партнером. Во многом характеристики и механизмы образования связей в данных комплексах будут схожи, однако, можно выделить некоторые особенности и различия, о который дальше и пойдет речь.

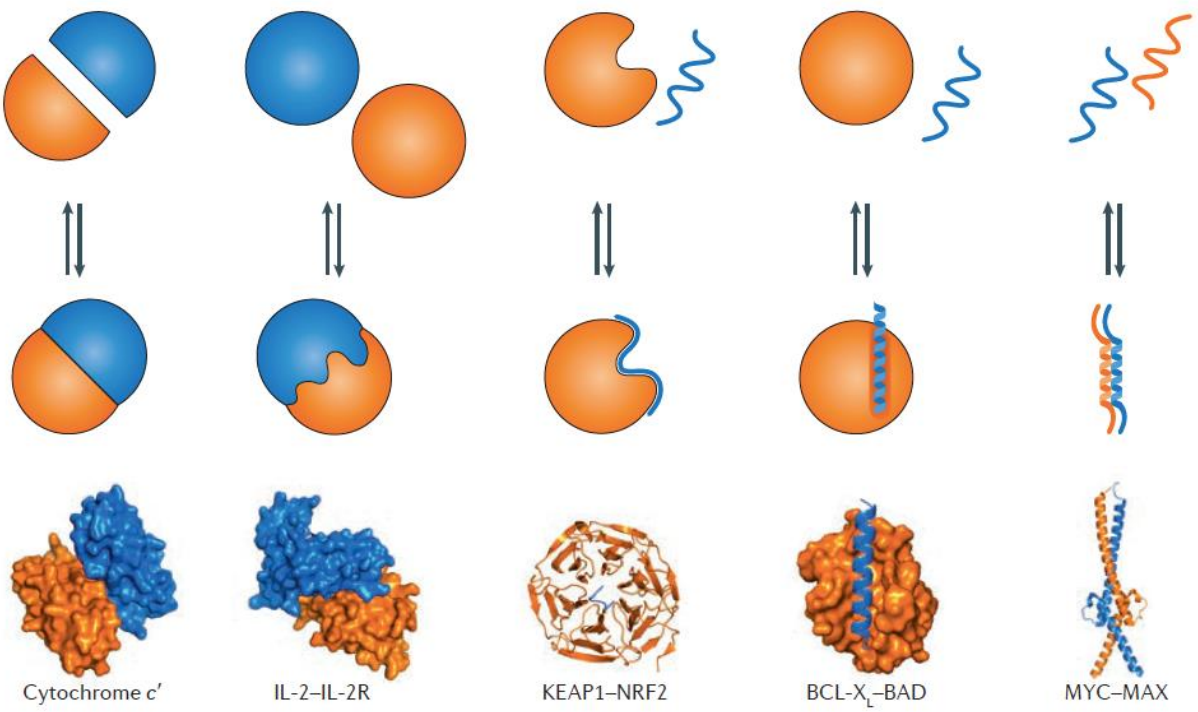


Рисунок 1. Различные варианты комплексов, образованных белок-белковыми взаимодействиями. В верхней части рисунка используются упрощенные иллюстрации для изображения партнеров по белкам и/или пептидам, а в нижней части рисунка показаны примеры кристаллических структур для каждого типа взаимодействия. А) Взаимодействие между двумя глобулярными белками с предварительно сформированными поверхностями (ID банка данных белков (PDB): 2ccy). Б) Взаимодействие между двумя глобулярными белками с индуцированной связывающей поверхностью (PDB id: 1z92). В) Взаимодействие жесткого глобулярного белка с пептидом (PDB id: 2dyh). Г) Взаимодействие гибкого глобулярного белка с пептидом (PDB id: 2xa0). Д) Взаимодействие двух пептидов (PDB id: 1nkp). BAD — BCL-2-ассоциированный агонист клеточной гибели; BCL, В-клеточная лимфома; ИЛ-2, интерлейкин-2; IL-2R, рецептор IL-2; KEAP1, kelch-подобный ECH-ассоциированный белок 1; MAX, фактор X, связанный с MYC; NRF2, ядерный фактор, связанный с эритроидом 2, фактор 2 (Scott et al., 2016).

Хотя и пептиды, и белки со внутренней неупорядоченностью (ВНБ) не имеют определенной структуры в нативном состоянии и приобретают

структуру при связывании, пептиды и неупорядоченные белки определяют два различных класса взаимодействия (Fuxreiter et al., 2007).

Так, например, средняя площадь поверхности белка, принимающей участие в связывании с пептидом, составляет около 500 $Å^2$, что в два раза меньше, чем в белок-белковых комплексах, и почти в три раза меньше, чем во взаимодействии с ВНБ. Однако, в пределах этой небольшой области, пептиды с большей частотой формируют различные связи, в частности водородные (London, 2010). На пептидно-белковых интерфейсах образуется в среднем 8,1 водородных связей в сравнении с 9,7 водородными связями в белково-белковых интерфейсах и 9,3 в ВНБ-опосредованных взаимодействиях (Meszaros, 2007). Учитывая меньший размер интерфейса, можно заметить, что пептиды образуют больше водородных связей на единицу площади интерфейса (примерно на 50% больше, чем при белок-белковых взаимодействиях, и более чем в два раза больше ВНБ-белковых взаимодействий на 100 $Å^2$) Это связано с высокой гибкостью пептидов и способностью подстраивать расположение основных доноров и акцепторов водородных связей под расположение таких групп на интерфейсе белка.

При этом α-спиральные пептиды образуют в среднем значительно меньшее количество водородных связей с белком (4,2 на пептид) и содержат гораздо больше неполярных атомов на интерфейсе связывания (53%). Анализ спиральных пептидов показал, что большинство этих пептидов образуют амфифильные спирали и связываются с гидрофобной поверхностью, что объясняет отчетливые свойства этого класса пептидов, в сравнение пептиды с β-тяжами образуют гораздо больше водородных связей (в среднем 12,5), причем значительную часть этих связей (примерно 32 %) составляют водородные связи между атомами основных цепей пептида и белка (London, 2010).

Достоверно установлено, что в белок-белковых интерфейсах наибольший вклад в энергию связывания обусловлен небольшим числом

остатков, называемых «активными точками» (Frank et al., 2024). Для пептидов также было показано, что небольшое число активных точек опосредует основную часть свободной энергии связывания: более 70% прогнозируемого снижения свободной энергии связывания обеспечивается аминокислотными остатками, лежащими в активных точках (London, 2010).

Часто пептид связывается в самом большом кармане, доступном на поверхности белка (Рис. 2.A). В случаях, когда пептид располагается в маленьком кармане, одна из боковых цепей пептида погружена в этот карман в виде ручки (Рис. 2.B). Так α-спиральные пептиды склонны связываться с использованием стратегии "ручка-отверстие", тогда как β-тяжевые пептиды предпочитают связывание в больших карманах.

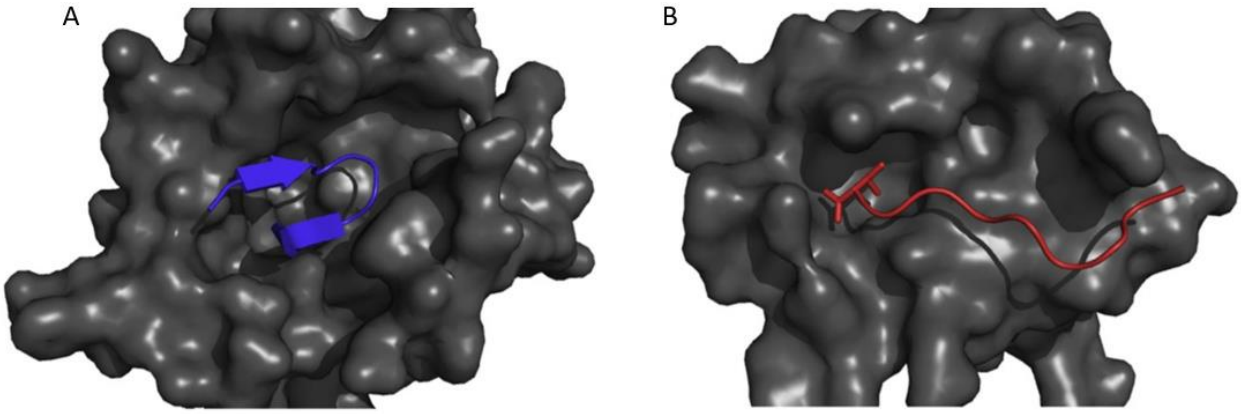


Рисунок 2. Область связывания пептидов на белковом интерфейсе. A) Белок комплемента C8 в комплексе с пептидом (код PDB: 2QOS). β-Тяжевой пептид (синий цвет) связывается в самом большом кармане на поверхности белка (серый цвет). B) PDZ-домен белка Erbin, связанный с субстратным пептидом (код PDB: 1MFG). Пептид (красный) прикрепляется к белку через его C-концевую валиновую ручку, которая входит в соответствующий карман в PDZ-домене (London, 2010).

Таким образом, комплексы белок-белок и белок-пептид различаются по ряду важных параметров, однако основные механизмы формирования

взаимодействий у них схожи, что позволяет экстраполировать закономерности в связывании одних комплексов на другие.

1.2. Характеристики связывания в белок-белковых комплексах

Важность оценки стабильности белок-белковых взаимодействий и выделения факторов, оказывающих на нее влияние приводит к задаче по предсказанию аффинности связывания в комплексах, образуемых белковыми и пептидными молекулами. Данная характеристика, определяемая как энергия взаимодействий между молекулами в комплексе, переводится в физико-химические термины как свободная энергия Гиббса. Изменение свободной энергии Гиббса, обозначаемое ΔG, представляет собой разницу в свободной энергии Гиббса между начальным и конечным состоянием реакции или процесса и даёт представление о её направлении и осуществимости. Отрицательное значение указывает на самопроизвольную реакцию, которая может протекать без внешнего источника энергии. Положительное значение означает, что реакция не является самопроизвольной и поэтому требует внешнего источника энергии.

В контексте белок-белковых взаимодействий изменение свободной энергии Гиббса при связывании одной белковой молекулы с другой может дать информацию о вероятности их эффективного взаимодействия и стабильности сформированного комплекса.

Согласно определению свободной энергии Гиббса, эта характеристика рассчитывается следующим образом:

$$G \equiv U + pV - TS,$$

где U – внутренняя энергия, p – давление, V – объем, T – абсолютная температура, S – энтропия.

Значение ΔG при постоянных значениях давления и температуры (изобарно изотермический потенциал) в стандартных условиях будет равно:

$$\Delta G = \Delta H - T\Delta S,$$

где ΔH — энтальпия системы, T — температура системы, ΔS — энтропия термодинамической системы.

Для короткоживущих белок-белковых комплексов характерно значение изменения энергии Гиббса больше, чем -8 ккал/моль, для долговременных это значение составляет -15 ккал/моль и ниже (Bashir et al., 2011; Wesley et al., 1997).

На практике свободная энергия Гиббса часто рассчитывается из экспериментально полученных значений константы равновесия реакции образования или диссоциации комплекса.

Все вещества (реагенты и продукты) в химической реакции могут находиться не в своей нормальной форме. В результате этой связи изменение энергии Гиббса реакции связано с изменением стандартной энергии Гиббса:

$$\Delta G = \Delta G^\circ + RT \ln Q,$$

где ΔG° — стандартное изменение энергии Гиббса (изменение энергии Гиббса, когда все вещества находятся в стандартном состоянии), Q — коэффициент реакции.

Выражение коэффициента реакции похоже на выражение константы равновесия, но между ними есть одно существенное различие: равновесные концентрации или парциальные давления продуктов и реагентов включены в константу равновесия. Тогда как Q выражается через начальные концентрации реагентов, парциальные давления и конечные концентрации или давления продуктов.

Когда реакция достигает равновесия, концентрации и парциальное давление достигают своих равновесных значений, и на этом этапе Q = K. При равновесии ΔG = 0 и Q = K, тогда стандартное уравнение энергии Гиббса становится таким:

$$0 = \Delta G^\circ + RT \ln K$$

Следовательно,

$$\Delta G^\circ = -RT \ln K$$

Это уравнение устанавливает связь между стандартным изменением энергии Гиббса для реакции и её константой равновесия.

Для белок-белковых комплексов наиболее часто рассматривается реакция диссоциации комплекса, и соответственно в качестве константа диссоциации (K_D), которая является экспериментальной мерой, определяющей, будет ли образован комплекс в растворе или нет. Помимо константы диссоциации используются также другие экспериментально определяемые характеристики, такие как IC_{50} (концентрация полумаксимального ингибирования) и K_i (константа ингибирования).

1.3. Базы данных, используемые для анализа белок-белковых комплексов

Для анализа особенностей белок-белковых взаимодействий, а также для обучения предсказательных моделей используется ряд общеизвестных баз данных, в которых хранится информация о структуре комплексов и характеристиках связывания.

Protein Data Bank (Berman et al., 2000) – это база данных для пространственных структурных данных крупных биологических молекул, таких как белки и нуклеиновые кислоты, которая контролируется Всемирным банком данных о белках. Эти структурные данные получены и депонированы биологами с помощью экспериментальных методологий, таких как рентгеновская кристаллография, ЯМР-спектроскопия и криоэлектронная микроскопия.

База данных PDBBind v. 2020 (Wang et al., 2020) представляет собой всеобъемлющую коллекцию экспериментально измеренных данных об аффинности связывания белковых комплексов, хранящихся в Protein Data Bank (Berman et al., 2000). Таким образом, она обеспечивает связь между энергетической и структурной информацией комплексов, что имеет большое значение для различных исследований молекулярного распознавания, проводимых в биологических системах.

SKEMPI представляет собой базу данных о свободной энергии связывания при введении мутаций, собранную из научной литературы, для гетеродимерных комплексов белок-белок с экспериментально определенной структурой (Moal et al., 2012). Версия SKEMPI v.2.0 (Jankauskaitė et al., 2019) содержит обработанные вручную данные о связывании для 7085 мутаций, включая изменения кинетики для 1844 мутаций, изменения энтальпии и энтропии для 443 мутаций и 440 мутаций.

1.4. Метрики оценивания качества предсказания аффинности связывания

Существует два подхода к решению задачи оценки аффинности белок-белковых комплексов. Во-первых, это решение задачи классификации. В базовом случае задается пороговое значение энергии связывания, разделяющее комплексы на два класса: Высокоаффинные и низкоаффинные. Второй и более предпочтительный подход – предсказание непосредственного значения энергии связывания или константы диссоциации. Для проверки качества работы алгоритма производится апробация на тестовых наборах данных с последующим расчетом метрик качества. И соответственно, для решения этих двух задач применяются различные метрики.

1.4.1. Метрики качества для задач классификации

Для рассмотрения метрик качества в задачах классификации необходимо ввести такое понятие как матрица ошибок (англ. Confusion Matrix). Рассмотрим наиболее простой вариант – бинарную классификацию. Тогда все предсказанные значения будут делиться на четыре вида: истинно положительные, ложно положительные, истинно отрицательные, ложно отрицательные (Табл. 1).

Таблица 1. Матрица ошибок.

	y=1	y=0
y'=1	Истинно положительное (ИП)	Ложно положительное (ЛП)
y'=0	Ложно отрицательное (ЛО)	Истинно отрицательное (ИО)

Где y' – предсказанное значение класса, y – истинное значение класса.

Так, на основании матрицы ошибок можно ввести некоторые метрики качества, в частности, применяемая в данной работе метрика Accuracy, она отражает долю правильных предсказаний:

$$Accuracy = \frac{\text{ИП} + \text{ИО}}{\text{ИП} + \text{ИО} + \text{ЛО} + \text{ЛП}}$$

Для оценки качества работы алгоритма на каждом из классов по отдельности используют метрики *precision* (точность) и *recall* (полнота). *Precision* можно интерпретировать как долю объектов, названных классификатором положительными и при этом действительно являющихся положительными, а *recall* показывает, какую долю объектов из всех объектов положительного класса нашел алгоритм.

$$Precision = \frac{\text{ИП}}{\text{ИП} + \text{ЛП}}$$

$$Recall = \frac{\text{ИП}}{\text{ИП} + \text{ЛО}}$$

Выбор метрики качества во многом зависит от сбалансированности классов, их количества и других параметров и подбирается индивидуально для каждой задачи.

1.4.2. Метрики качества для задач регрессии

В большинстве случаев решения регрессионных задач для оценки качества работы алгоритма на тестовой выборке используется корреляционный анализ. Для предсказанных и истинных значений попарно рассчитывается коэффициент корреляции Пирсона по формуле:

$$r_{xy} = \frac{\sum(y_i - M_{y_i})(\hat{y}_i - M_{\hat{y}_i})}{(n-1)\sigma_{y_i}\sigma_{\hat{y}_i}},$$

где y_i – истинное значение, $\hat{y}_i$ – предсказанное значение, M_{y_i}, $M_{\hat{y}_i}$ – средние значения наборов истинный и предсказанных значений соответственно, σ_{y_i}, $\sigma_{\hat{y}_i}$ – стандартные отклонения, n – количество объектов в наборе.

Таким образом, при помощи расчета значения корреляции Пирсона определяется, есть ли линейная зависимость между предсказанными и истинными значениями. По сути корреляция Пирсона отражает обобщающую способность модели, а также направление, в котором допускается ошибка.
Однако, для более полного анализа, нужно определить также величину ошибки, характерную для предсказаний. С этой целью часто используется метрика RMSE, которая рассчитывается по следующем формуле:

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

где n – количество объектов, y_i – истинное значение, $\hat{y}_i$ – предсказанное значение. Таким образом, RMSE отражает ошибку в абсолютных значениях (для ΔG – ккал/моль).

Также для определения ошибки в абсолютных значениях может рассчитываться средняя абсолютная ошибка (MAE) по формуле:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|(y_i - \hat{y}_i)|,$$

где n – количество точек, y_i – истинное значение, $\hat{y}_i$ – предсказанное значение.

1.5. Методы предсказания аффинности связывания, основанные на физических и статистических моделях

Все реализованные методы оценки аффинности связывания в той или иной степени сталкиваются с рядом значимых ограничений (Kastritis and Bonvin, 2012):

- Неоднозначность экспериментальных данных

- Отсутствие учета конформационных изменений, происходящих при связывании, или наличия кофакторов, которые могут потребоваться для связывания.
- Сложная кинетика комплекса
- Игнорирование влияния pH, температуры, концентрации растворителя и комплекса.
- Производительность (особенно для моделей прогнозирования аффинности) зависит от качества и размера набора экспериментальных данных, используемых для тестирования, а также от их разнообразия.
- Отсутствие учета вклада поверхности белка, не относящейся к интерфейсу взаимодействия, которая может играть значительную роль в модуляции аффинности.
- Сопоставление структуры, которая была определена в ее кристаллическом состоянии, с аффинностью, измеренной в состоянии раствора, может привести к некорректности полученных результатов из-за различной конформации этих двух состояний.

Каждый из реализуемых подходов при этом может учитывать одно или несколько вышеописанных ограничений. Далее подробнее будут описаны методы оценки аффинности связывания белок-белковых комплексов и их ограничения.

1.5.1. RosettaDock

Сервер RosettaDock (Lyskov et al., 2008) предназначен для определения стабильных конформаций при взаимодействиях белок–белок вблизи заданной начальной конфигурации путем оптимизации ориентации пептидного остова и конформаций боковых цепей. При этом помимо основного сервиса реализована библиотека PyRosetta, представляющая собой автономную реализацию пакета молекулярного моделирования Rosetta на основе Python. Данная библиотека позволяет пользователям писать алгоритмы

прогнозирования структуры и проектирования с использованием основных функций выборки и расчета значения оценочной функции Rosetta. При этом PyRosetta содержит привязки Python к библиотекам, которые определяют функции Rosetta, в том числе для доступа к структуре белка и манипулирования ею, вычисления энергий и запуска моделирования на основе метода Монте-Карло (Chaudhury et al., 2010).

Для оценки моделей белок-белковых комплексов исследуется изменение энергий как функция среднеквадратичного отклонения между остатками на интерфейсе связывания в каждой молекуле. Для расчета используются аминокислотные остатки в области интерфейса связывания, Cβ атомы которых расположены на расстоянии менее 8,0 Å от Cβ атома ближайшего остатка другой взаимодействующей молекулы. Оценка энергии связывания происходит за счет суммирования энергий различных взаимодействий с учетом эмпирически подобранных весов: энергия притяжения и отталкивания между атомами одного и разных остатков, энергии сольватации, энергия коротко- и дальнедействующий водородных связей и др. (Alford et al., 2017).

1.5.2. DFIRE

В 2004 году был разработан веб-сервис DFIRE, осуществляющий предсказание энергии взаимодействия в белок-белковых комплексах на основе расчета потенциала взаимодействия белковых молекул (Zhang et al., 2004).

Так, атом-атомный потенциал средней силы u(i,j,r) между типами атомов i и j, которые находятся на расстоянии r друг от друга, определяется выражением:

$$\bar{u}(i,j,r) = \begin{cases} -\eta RT \ln \dfrac{N_{obs}(i,j,r)}{\left(\dfrac{r}{r_{cut}}\right)^{\alpha}\left(\dfrac{\Delta r}{\Delta r_{cut}}\right)N_{obs}(i,j,r_{cut})}, & r < r_{cut}, \\ 0, & r \geq r_{cut}, \end{cases} \quad (1)$$

где η = 0,0157, R – газовая постоянная, T = 300 K, α = 1.61, N_{obs}(i,j,r) — количество пар (i,j) в пределах оболочки радиуса r, r_{cut} = 14,5 Å, а Δr(Δr_{cut}) — ширина интервала в точке r (r_{cut}). Параметр η определялся так, чтобы наклон линии тренда между прогнозируемыми и экспериментально измеренными изменениями был равен 1. Использовались типы атомов, специфичные для остатков (167 атомных типов) (Greer et al., 1980; Glaser et al., 2001). Число наблюдаемых пар атомов (i,j) на расстоянии r друг от друга [Nobs(i,j,r)] было получено из структурной базы данных, состоящей из 1011 негомологичных (менее 30% гомологии) белков с разрешением 2 Å.

Полный атом-атомный потенциал средней силы G для каждой структуры определяется выражением:

$$G = \frac{1}{2}\sum_{i,j} \bar{u}(_{i,j},r_{ij}) \ ,$$

где суммирование ведется по парам атомов, которые не входят в один и тот же остаток, и используется коэффициент 1/2, чтобы избежать двойного учета взаимодействий остаток-остаток и атом-атом. Свободная энергия связи димера AB расчитывается следующим образом:

$$\Delta G_{bind} = G_{complex} - (G_A + G_B) \ .$$

Таким образом, мономерный потенциал на основе DFIRE с определенным приближением обеспечивает описание энергетического и энтропийного вклада в стабильность связывания. При этом, важно учитывать, что при вычислении свободных энергий связи было сделано много допущений. К ним относятся приближение твердого тела (сводит количество степеней свободы твёрдого тела до 6) и отсутствие явного рассмотрения дальнодействующей электростатики и молекул воды.

1.5.3. CP_PIE

В 2010 году был разработали метод определения потенциала на основе контактов остатков и областей перекрытия CP_PIE (Ravikant et al., 2010). Для оптимизации параметров модели использовался комплексный обучающий набор из 640 двухцепочечных белковых комплексов. Данный метод на момент создания показал значительно лучшую способность определения оценочной функции связывания по сравнению с другими общедоступными потенциалами для докинга белков.

Целью создания алгоритма CP_PIE являлась разработка единого потенциала с физически значимой функциональной формой для отделения нативных способов связывания от ненативных. Такая оценочная функция может быть использована в статистико-механическом исследовании динамики и функционирования в дополнение к прогнозированию структур белковых комплексов. Для этой цели рассчитывается потенциал на основе остатков, поскольку он менее чувствителен к локальным конформационным изменениям и, как следствие, имеет больший радиус сходимости. В качестве энергетических характеристик, которые можно эффективно вычислить, применялись остаточные контакты и области перекрытия. Расчет площади перекрытия основан на вычислении для каждого атома интерфейса количества точек (у каждого атома задается 256 равномерно распределенных точек по его

сфере), потерянных при перекрытии атомами другой молекулы. Таким образом, используется дискриминационная структура для проектирования потенциала, поскольку она позволяет учиться на неправильных режимах связывания, в отличие от подходов, основанных только на положительном обучении (таких как статистические потенциалы). Результирующий потенциал достаточно хорошо показал себя в решении задачи распознавания нативных режимов связывания в большом наборе тестируемых комплексов.

1.5.4. FoldX

FoldX — это эмпирическое силовое поле, разработанное для быстрой оценки влияния мутаций на стабильность, сворачивание и динамику белков и нуклеиновых кислот. Основная функциональность FoldX, а именно расчет свободной энергии макромолекулы на основе ее трехмерной структуры с высоким разрешением, осуществляется через веб-сервер (Schymkowitzet al., 2005) или одноименную программу.

Для расчета свободной энергии (в ккал/моль) используется линейная комбинация эмпирических термов по формуле:

$$\begin{aligned}\Delta G = {} & a \cdot \Delta G_{\mathrm{vdw}} + b \cdot \Delta G_{\mathrm{solvH}} + c \cdot \Delta G_{\mathrm{solvP}} + d \cdot \Delta G_{\mathrm{wb}} \\ & + e \cdot \Delta G_{\mathrm{hbond}} + f \cdot \Delta G_{\mathrm{el}} + g \cdot \Delta G_{\mathrm{kon}} + h \cdot \mathrm{T}\Delta \mathrm{S}_{\mathrm{mc}} \\ & + k \cdot \mathrm{T}\Delta \mathrm{S}_{\mathrm{sc}} + l \cdot \Delta G_{\mathrm{clash}}.\end{aligned}$$

В этом выражении коэффициенты *a, b, ..., l* представляют собой относительные веса различных энергетических термов, используемых для расчета свободной энергии. Взаимодействие с растворителем рассматривается в два этапа: во-первых, учитываются вклады гидрофобных (ΔG_{solvH}) и полярных (ΔG_{solvP}) групп. Эти параметры сольватации были получены в результате экспериментов, в которых аминокислоты переносятся из воды в органический растворитель. Предполагается, что этот процесс имитирует

переход, который испытывает аминокислота во время сворачивания, от воздействия растворителя в развернутом состоянии к заглублению в гидрофобной среде в нативном состоянии. Во-вторых, те молекулы воды, которые имеют постоянное взаимодействие с группами белка, т.е. образуют более двух водородных связей с белком, явно учитываются в формуле за счет ΔG_{wb}. Вклад Ван-дер-Ваальсовых взаимодействий (ΔG_{vdw}) рассчитывается с учетом экспериментальных энергий при переходе от воды к пару. Водородные связи рассчитываются на основе простых геометрических соображений и их энергии, значение ΔG_{hbond} выводится из циклов двойных мутаций в белковой инженерии. Электростатический вклад в свободную энергию (ΔG_{el}) рассчитывается на основе простой реализации закона Кулона. Для белковых комплексов рассчитывается дополнительный электростатический вклад между атомами различных полипептидных цепей (ΔG_{kon}), основанный на эмпирическом уравнении Шрайбера (Selzer et al., 2000), который, как было показано, дает хорошую оценку скорости ассоциации (k_{on}) сложной формы. Важным отличием FoldX от других силовых полей является грубая оценка энтропии, которая используется для измерения свободной энергии. Расчеты энтропии обычно включают масштабное моделирование конформационной свободы боковых цепей и остова белка. В FoldX энтропийный штраф за фиксацию остова в заданной конформации (ΔS_{mc}) получен на основе статистического анализа распределения φ-ψ углов данной аминокислоты, наблюдаемого в наборе кристаллических структур с высоким разрешением. Энтропийные затраты на фиксацию боковой цепи в определенной конфигурации (ΔS_{sc}) определяются путем масштабирования набора параметров энтропии, рассчитанных Абагяном и его коллегами (Abagyan et al., 1994) для внедрения боковой цепи. Наконец, ΔG_{clash} позволяет учесть стерические перекрытия между атомами в структуре.

Взаимодействие между двумя молекулами определяется свободной энергией связывания (ΔG_{bind}), которая напрямую связана с термодинамической

константой диссоциации (K_D) следующим уравнением: $\Delta G_{bind} = -RT \ln(K_D)$, где R - газовая постоянная (1,9859 кал моль-1 К-1), а T - температура в кельвинах. Чтобы рассчитать свободную энергию связывания комплекса AB, FoldX вычисляет энергии Гиббса комплекса (ΔG_{AB}) и отдельно двух молекул A и B. Затем энергия взаимодействия определяется как:

$$\Delta G_{bind} = \Delta G_{AB} - (\Delta G_A + \Delta G_B).$$

1.6. Машинное обучение

Машинное обучение представляет собой ответвление искусственного интеллекта, при котором предсказательные алгоритмы используют закономерности во входных данных для прогнозирования новых значений (Mitchell et al., 1997).

В последнее десятилетие в связи с экспоненциальным ростом вычислительных мощностей (Schmidt et al., 2019), и доступностью больших массивов данных наблюдается всплеск интереса к машинному обучению. Алгоритмы машинного обучения можно разделить на три категории: обучение без учителя, обучение с учителем и обучение с подкреплением. Рассмотрим подробнее первые две категории.

1.6.1. Обучение с учителем

Алгоритмы, использующие обучение с учителем, основаны на тех же принципах, что и стандартная процедура подгонки: алгоритм пытается найти неизвестную функцию, которая соединяет известные входные данные с неизвестными выходами. Этот желаемый результат для неизвестных областей оценивается на основе экстраполяции шаблонов, найденных в размеченных обучающих данных.

Обычно в качестве обучающих данных создается набор из интересующих объектов, для которых известны значения целевого свойства (класса) (Schmidt et al., 2019). Значительная часть работы на этом этапе

заключается в создании, поиске и очистке данных для обеспечения их согласованности, точности и т.д (Shalev-Shwartz et al., 2014).

Дальше алгоритм оптимизирует параметры функции, связывающие объекты с их целевыми свойствами. Этот процесс называется обучением. В идеале для валидации гиперпараметров используется отдельный набор данных, отличный от тестового и обучающего наборов.

Алгоритмы, использующие обучение с учителем, в сравнение с методами, основанными на неконтролируемом обучении, способны решать задачи классификации и регрессии.

1.6.1.1. Алгоритмы обучения с учителем в классическом машинном обучении

Методы, основанные на подобном подходе, наиболее многочисленны среди алгоритмов машинного обучения. К одним из самых широко распространенных методов относятся:

- Линейная регрессия (англ. linear regression)

Данный метод используется для изучения линейной зависимости между зависимой переменной Y и одной или несколькими независимыми переменными X. Причем зависимая переменная должна быть непрерывной, в то время как независимые переменные могут быть либо непрерывными, двоичными, либо категориальными (Schneider et al., 2010).

Линейная регрессионная модель описывает зависимую переменную прямой линией, которая определяется уравнением:

$$Y = a \times X + b,$$

где a – коэффициент наклона прямой, b – точка пересечения с осью y. Параметры a и b регрессионной прямой оцениваются по значениям зависимой переменной Y и независимой переменной X статистическими методами. Прямая регрессии позволяет предсказать значение Y, основываясь на значении X (Fahrmeir et al., 2009). Таким образом, например, методом

линейной регрессии можно было бы оценить вес человека (зависимая переменная) по его росту (независимая переменная) (Рис. 3).

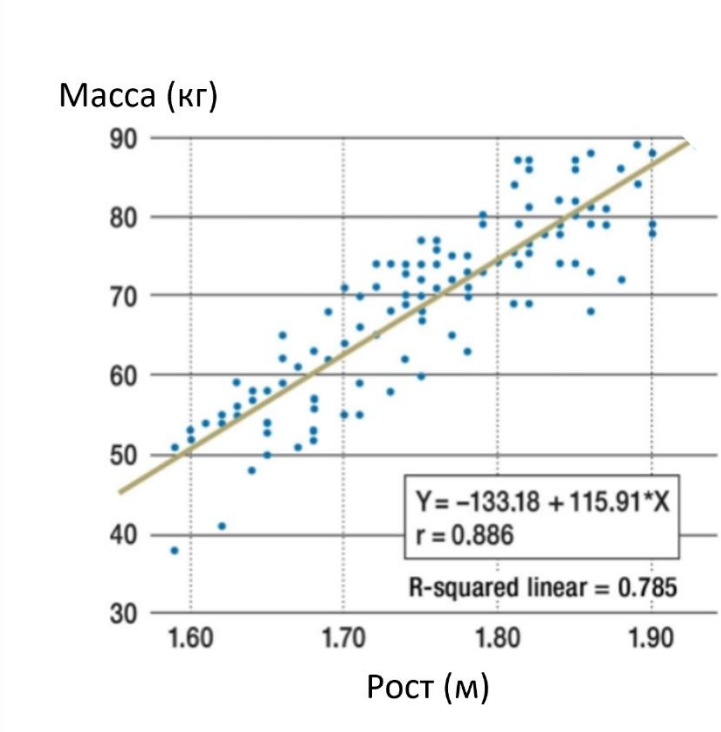


Рисунок 3. Точечная диаграмма и соответствующая ей линия регрессии и уравнение регрессии, отображающее связь между зависимой переменной (масса тела, кг) по оси ординат и независимой переменной (рост, м) по оси абсцисс (Schneider, 2010). R^2 = коэффициент детерминации; r = коэффициент корреляции Пирсона.

- Логистическая регрессия (англ. logistic regression)

Данный статистический метод широко используется в эмпирических исследованиях с участием категориальных зависимых переменных (Lian et al., 2018). Показано, что дихотомическая (то есть такая, которая может принимать два категориальных значения) зависимая переменная нарушает предположения о гомоскедастичности (однородность дисперсии случайной ошибки регрессионной модели). Следовательно, оценки стандартной ошибки не будут согласованными оценками истинных стандартных ошибок, и оценки коэффициентов линейной регрессионной модели больше не будут эффективными. Кроме того, оценка вероятности с помощью метода наименьших квадратов приведет к прогнозируемым значениям, которые

находятся за пределами диапазона вероятности (0,1). По этим причинам модель логистической регрессии используется для решения задач бинарной классификации, определяя вероятность принадлежности объекта к определенному классу (Lian et al., 2018). Модель логистической регрессии принимает следующую форму (Allison et al., 1999):

$$\log\left[\frac{p_i}{1-p_i}\right] = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik}$$

где i обозначает отдельное событие, p_i представляет вероятность возникновения события, $(1-p_i)$ представляет вероятность того, что событие не произошло, а отношение этих двух величин представляет собой шансы выбора события i; а выражение в левой части представляет логарифмические шансы или logit. В правой части уравнения α представляет точку пересечения, β представляет коэффициент регрессии, а x представляет независимую переменную (Rencher et al., 2000). Визуализация подхода логистической регрессии к решению задачи бинарной классификации представлена на Рис. 4.

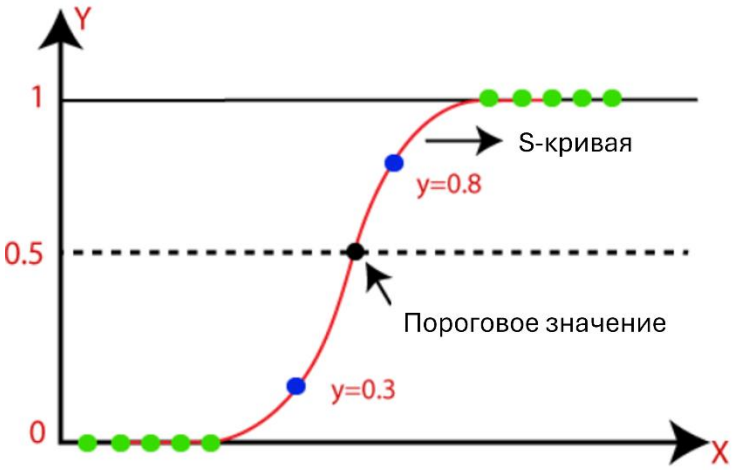


Рисунок 4. Визуализация работы логистической регрессии. Решение задачи бинарной классификации. По оси X обозначен признак объекта, по которому происходит предсказание, по оси Y – вероятность принадлежности объекта к классу 1. Зеленым обозначены объекты с известными классами; синим – объекты, для которых осуществляется предсказание.

Как и линейная регрессия, логистическая регрессия может обрабатывать как непрерывные, так и категориальные независимые переменные. Однако

интерпретация результатов является более сложной и менее интуитивной по сравнению с линейной регрессией. Это связано с тем, что в модели логистической регрессии связь между вероятностями и набором независимых переменных не является линейной (Perme et al., 2004). При этом несмотря на использование логистической регрессии для решения задачи классификации, данная модель предсказывает не сам класс, а вероятность принадлежности объекта к определенному классу.

- Метод К-ближайших соседей (англ. k-nearest neighbor algorithm, KNN)

Данный метод предназначен для классификации немаркированных объектов путем отнесения их к классу наиболее похожих меченых примеров, также он может применяться для регрессионной задачи (Sricharan et al., 2011).
Объект классифицируется на основании голосования множества его соседей, причем объекту присваивается класс, наиболее распространенный среди его ближайших k соседей (k – положительное целое число, обычно маленькое). Если k = 1, то объекту присваивается класс единственного ближайшего соседа (Рис. 5).

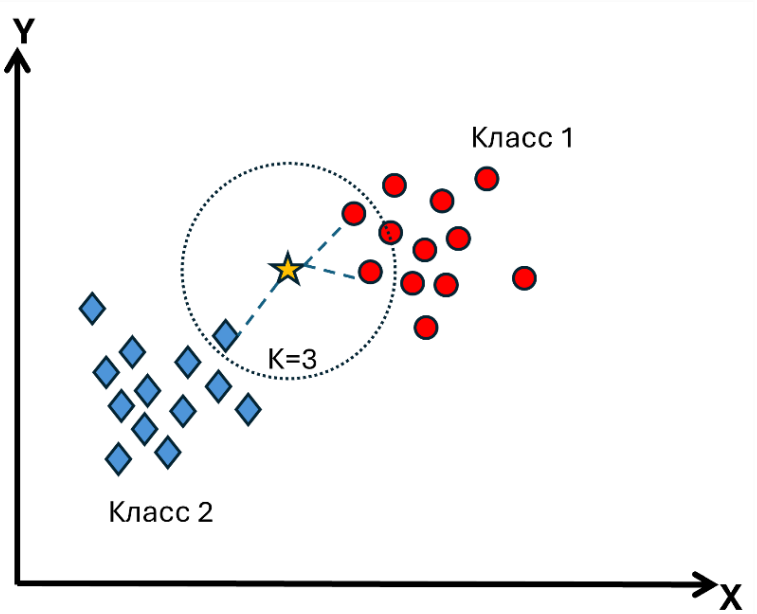


Рисунок 5. Визуализация работы метода К-ближайших соседей. Решение задачи бинарной классификации. По осям X и Y обозначены признаки

объектов, по которым происходит предсказание. Параметром K обозначено количество учитываемых соседей при классификации.

При решении задачи регрессии выходным значением алгоритма является значение целевого свойства объекта. Это число является средним среди значений k ближайших соседей (Altman et al., 1992).

- Метод опорных векторов (Support Vector Machines, или SVM)

Данный метод включает в себя непараметрическую статистическую технику обучения, следовательно, не делается никаких предположений относительно базового распределения данных. В своей первоначальной формулировке (Vapnik et al., 1979) метод представлен обучающим алгоритмом, который находит гиперплоскость для отдельных тестовых объектов (количество классов определяется заранее) таким образом, чтобы это соответствовало обучающим примерам. Термин «оптимальная гиперплоскость разделения» используется для обозначения границы решения, которая минимизирует ошибочные классификации, полученные на этапе обучения. Обучение относится к итеративному процессу поиска классификатора с оптимальной границей принятия решения для разделения обучающих объектов (в потенциально многомерном пространстве), а затем для классификации тестовых данных в том же пространстве (Рис. 6) (Zhu and Blumberg et al., 2002).

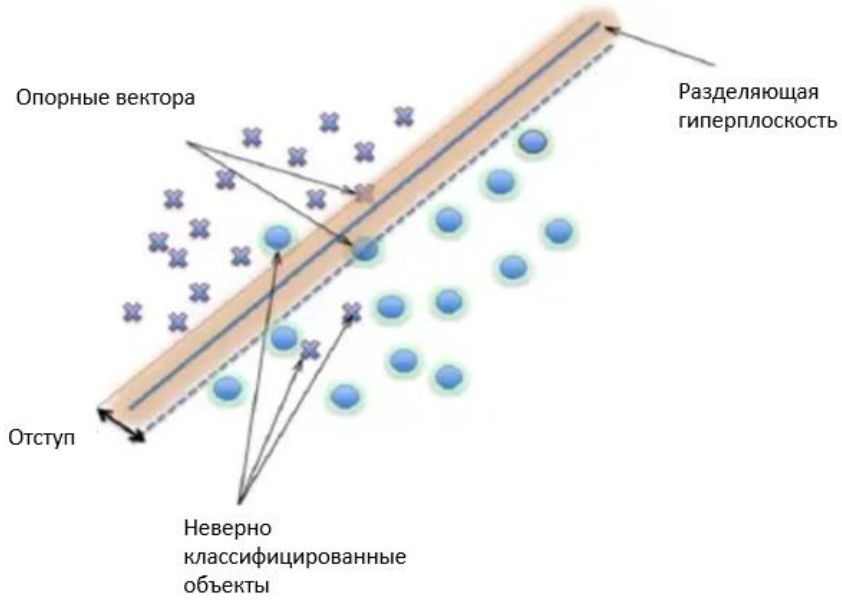


Рисунок 6. Визуализация метода опорных векторов. Задача классификации (Mountrakis et al., 2011).

- Решающие деревья (англ. decision trees)

Данный алгоритм решает задачу обобщения или поиска шаблонов в данных с целью их классификации. Это делается путем определения, какие тесты (вопросы) лучше всего разделяют объекты на отдельные классы, образуя дерево (Рис. 7). Затем на основе разветвления объекты распределяются, и эта процедура применяется рекурсивно, пока все экземпляры в узле не будут принадлежать одному и тому же классу (Kotsiantis et al., 2013). Как и в случае с другими парадигмами классификации шаблонов, более сложные модели (более крупные деревья решений), как правило, приводят к более низкой производительности обобщения.

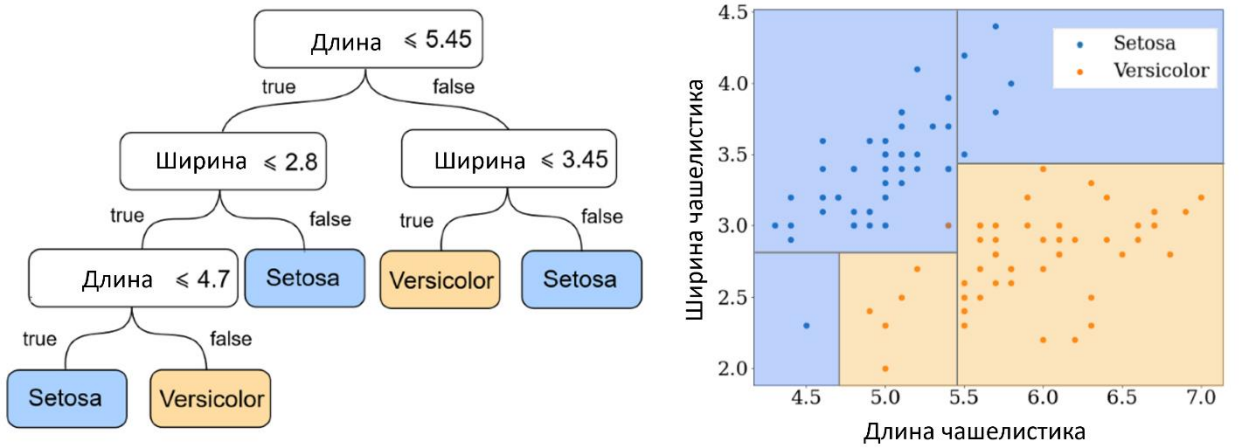


Рисунок 7. Дерево решений для классификации ирисов по морфологическим признакам (ширина и длина чашелистика). Каждый узел представляет атрибут или объект, а ветвь от каждого узла представляет результат этого узла. Наконец, листья дерева отражают окончательное решение (Costa et al., 2023).

1.6.1.2. Нейронные сети

Нейронные сети являются одним из наиболее универсальных алгоритмов машинного обучения, при этом они могут решать как задачи классификации, так и регрессии.

Искусственный нейрон обрабатывает и передает информацию как биологический нейрон. Он принимает данные на вход и выполняет необходимые вычисления, после этого генерирует результат.

Каждый вход в нейронной сети умножается на значение веса, и результат становится входом для следующего нейрона. Наряду с входным слоем и выходным слоем могут присутствовать скрытые слои для выполнения необходимых промежуточных вычислений (Arora et al., 2017). Особенностью многослойного перцептрона является наличие нескольких скрытых слоёв (Рис. 8).

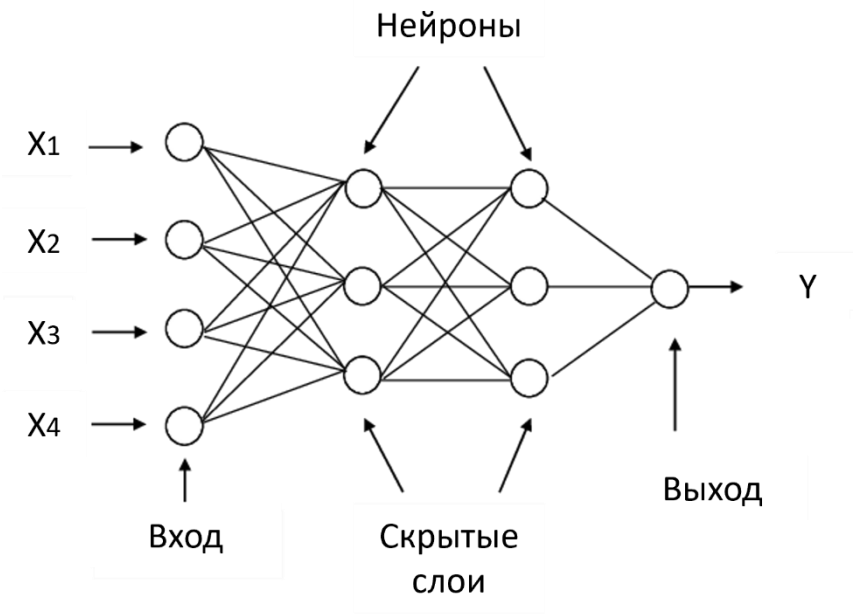


Рисунок 8. Базовая архитектура многослойного перцептрона. [X_1, X_2, X_3, X_4] – входные данные. Y – предсказанное значение (Hemalatha et al., 2017).

Нейронные сети представлены широким разнообразием простых и сложных структур, таких как нейронная сеть с прямой связью, рекуррентные, сверточные нейронные сети и другие. Существует множество методов обучения сети, подходящих под определенную архитектуру и задачу.

Наиболее значимыми элементами многослойного перцептрона (Multi-Layer Perceptron, MLP) являются веса связей и отклонения. Выход каждого узла рассчитывается в два этапа (Arora et al., 2017). На первом этапе взвешенное суммирование входных данных рассчитывается по формуле:

$$S_j = \sum_{i-1}^{n} w_{ij} I_i + \beta_j$$

где I_i - входная переменная, w_{ij} - вес связи между I_i и скрытым нейроном j, β_j - смещение.

На втором этапе используется функция активации для генерации выхода нейронов на основе рассчитанного значения взвешенного суммирования. В MLP могут использоваться различные типы функций активации, такие как логистическая, гиперболическая, экспоненциальная и сигмоидальная.

Сигмоидальная функция – наиболее часто применяемая функция активации в литературе. Уравнение для сигмоидальной функции:

$$f_i(x) = \frac{1}{1+e^{-s_j}}$$

Как только выход каждого скрытого нейрона вычислен, окончательный выход MLP рассчитывается по уравнению:

$$y_k = \sum_{i=1}^{m} w_{kj} f_i + \beta_k$$

Основной проблемой при использовании алгоритма, основанного на нейронных сетях, является оптимизация архитектуры под конкретную задачу. Однако, гибкость конфигурации также обеспечивает высокую вариативность решаемых задач.

1.6.1.2.1. Сверточные нейронные сети

Свёрточная нейронная сеть (англ. Convolutional Neural Network, CNN) – это разновидность многослойного перцептрона, особенность которой является наличие операции, называемой «свёртка». Для осознания этого процесса нужно ввести некоторые термины, использующиеся в рамках CNN. Ядро свёрточного слоя представляет из себя фильтр, содержащий систему разделяемых весов нейронов. Размер ядра обычно относительно мал, чаще всего это матрица 3x3 или 5x5. Фильтр скользит по так называемым картам, это области изображения или массива, которые принимает на вход ядро. Далее все значения карты поэлементно умножаются на ядро и суммируются, получившееся значение записывается в матрицу признаков следующего слоя. Процесс свёртки показан на Рисунке 9.

Помимо свёрточных слоев в CNN так же есть подвыборочный слой, цель которого уменьшить размерность карты признаков, выбрав максимальные значения в каждой ячейке размерности ядра. Данная операция называется MaxPooling.

Третьим идет полносвязный слой, в котором происходит оптимизация нелинейной функции с целью повышения качества предсказания. И после этого в выходном слое происходит окончательное присвоение классов.

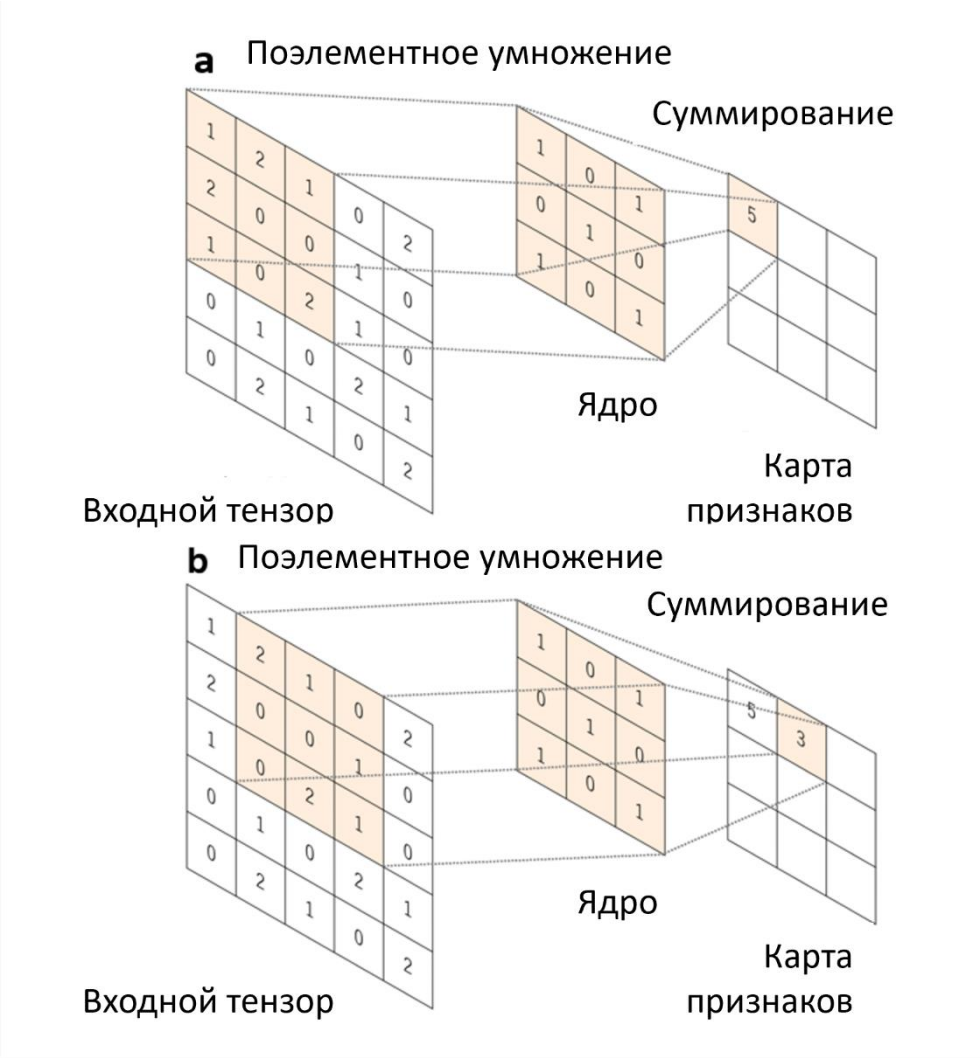


Рисунок 9. Операция свертки, происходящая в сверточном слое нейронной сети: a) свертка первой карты признаков; b) сдвиг ядра и свертка следующей карты признаков (Yamashita et al., 2018).

Свёрточные нейронные сети успешно показывали себя в ряде задач, связанных с выявлением шаблонных паттернов и определением закономерностей их расположения (Jiménez et al., 2018).

Одним из самых важных этапов в разработке алгоритма на основе нейронных сетей является процесс обучения, который заключается в

переопределении весов нейронов с целью минимизации функции потерь и максимизации точности.

За это отвечает большое число гиперпараметров:

- Функция активации.

Функция активации – это функция, используемая в нейронных сетях для вычисления взвешенной суммы входных данных и смещений, из которых можно определить, можно ли запустить нейрон или нет. Она манипулирует представленными данными посредством некоторой обработки градиента, обычно градиентного спуска, и затем создает выход для нейронной сети, который содержит параметры в данных. Функция активации может быть как линейной, так и нелинейной, в зависимости от функции, которую она представляет, и используется для управления выходами нейронных сетей в разных областях.

Одними из наиболее используемых функций активации являются Softmax и ReLU (Рис. 10). Функция Softmax выдает выходной сигнал, который представляет собой диапазон значений от 0 до 1 (распределение вероятностей по вектору действительных чисел), причем сумма вероятностей равна 1.

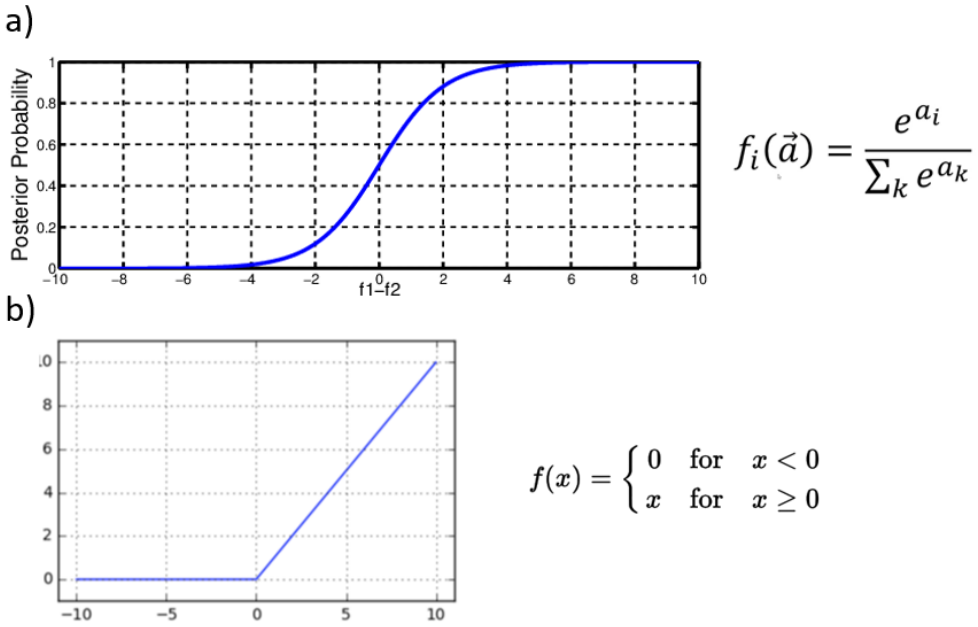


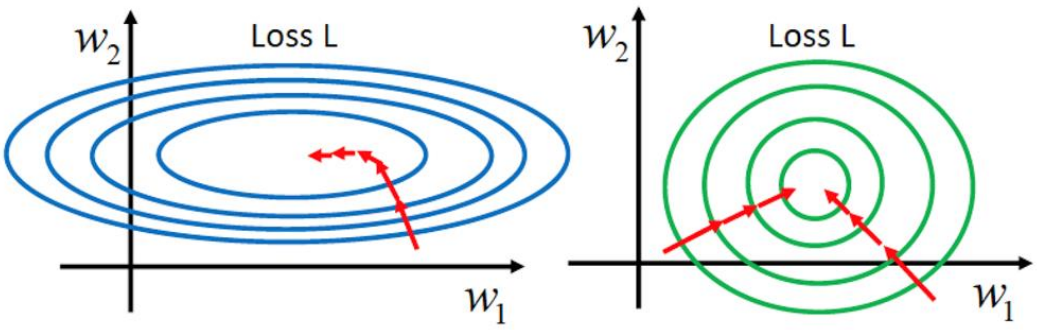
Рисунок 10. Применяемые в работе функции активации: a) Softmax; b) ReLU.

Данная активационная функция используется в мультиклассовых моделях, где она возвращает вероятности принадлежности объекта к каждому из классов, причем целевой класс имеет наибольшую вероятность.

Функция активации ReLU выполняет пороговую операцию для каждого элемента ввода, где выходное значение равно входному значению, если оно больше нуля, и нулю в противном случае. Данная функция активации способствует быстрому обучению (LeCun et al., 2015), так как представляет собой почти линейную функцию и поэтому сохраняет свойства линейных моделей, которые облегчают их оптимизацию с помощью методов градиентного спуска (Nwankpa et al., 2018).

- Нормализация

Нормализация входных данных нейронных сетей к нулевому среднему значению и постоянному стандартному отклонению известна в течение десятилетий (LeCun et al., 2018) как полезная для обучения нейронных сетей. При использовании в глубоких сетях аналогичная нормализация выполняется и для промежуточных слоёв с той лишь разницей, что для увеличения производительности она выполняется на каждой эпохе обучения не для всех нейронов слоя, а для их небольшой части (мини-пакета, англ. mini-batch), в связи с чем такой приём называется пакетной нормализацией (англ. Batch Normalization) (Bjorck et al., 2018). В своей оригинальной публикации (Ioffe, 2015) Иоффе и Сегеди выдвигают гипотезу, что пакетная нормализация может смягчать «внутренний ковариантный сдвиг» – явление, при котором распределения значений признаков в скрытых узлах слоёв имеют разные параметры (математическое ожидание, дисперсия и др.) во время обучения (Рис. 11).



$$x \;\rightarrow\; \hat{x} = \frac{x - \mu}{\sigma} \;\rightarrow\; y = \gamma\hat{x} + \beta$$

Рисунок 11. Минимизация функции потерь: слева – без пакетной нормализации, справа – с нормализацией. По осям отложены веса нейронов. Снизу – алгоритм пакетной нормализации, где x – входные значения, y – значения выхода слоя; μ - среднее значение x в мини-пакете; σ – среднее отклонение x в мини-пакете; γ и β – параметры, используемые для масштабирования и сдвига.

- Функция потерь

Для оптимизации значений весов связей необходимо видеть их влияние на точность предсказаний. Для этого вводится функция потерь, которая принимает большие значения при плохой классификации, а при улучшении качества классификации, её значение падает. Таким образом, при оптимизации параметров стоит задача минимизации функции потерь.

Для решения задач регрессии часто используют функцию средней квадратичной ошибки, которая задаётся уравнением:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Где n – количество объектов, y_i – вектор истинных значений, а $\hat{y}_i$ – вектор предсказаний.

Особенностью данной функции является высокая чувствительность к размеру ошибки, в связи с возведением ее в квадрат. Таким образом, данная функция может использоваться для увеличения значимости наиболее отклоняющихся от средних значений объектов и наиболее трудных для анализа.

- Алгоритм оптимизации

Как было выше сказано, обучение нейронной сети включает в себя подбор таких весов нейронов и сдвигов, чтобы предсказанные значения меток максимально совпадали с истинными. Для подбора оптимальных параметров необходимо минимизировать функцию потерь, для чего существует ряд алгоритмов.

Одним из самых используемых в глубоком обучении методов оптимизации является метод стохастической оптимизации Adam (Kingma et al., 2014), который был предложен в качестве адаптивного алгоритма оптимизации скорости обучения глубоких нейронных сетей. Суть метода заключается в отдельном подборе скоростей обучения для каждого параметра, при этом величины обновлений параметров не зависят от масштабирования градиента. Также алгоритм способен работать с разреженными градиентами. Реализация алгоритма показана на Рисунке 12.

$$\nu_t = \beta_1 * \nu_{t-1} - (1 - \beta_1) * g_t$$
$$s_t = \beta_2 * s_{t-1} - (1 - \beta_2) * g_t^2$$
$$\Delta\omega_t = -\eta \frac{\nu_t}{\sqrt{s_t + \epsilon}} * g_t$$
$$\omega_{t+1} = \omega_t + \Delta\omega_t$$

Рисунок 12. Алгоритм Adam для обновления параметров каждого w_j, где v_t и s_t – скользящие средние, рассчитанные по градиенту и квадратичному градиенту соответственно; β_1=0.9, β_2=0.999 (гиперпараметры алгоритма, обычно не варьируют, а оставляют стартовыми); g_t – временной градиент в

текущем мини-пакете; η – размер шага и w – вес нейрона. Таким образом на первом этапе происходит подсчет скользящих средних (1 и 2 формулы), дальше происходит обновление весов (3, 4 формулы) на основании, рассчитанных средних.

- Методы борьбы с переобучением

Одним из значительных недостатков глубоких нейронных сетей, является так называемое переобучение, возникающее из-за большого числа параметров, в связи с чем, алгоритм не выявляет закономерности, а запоминает значение класса для определенной комбинации признаков. Таким образом, можно наблюдать повышение точности предсказания на тренировочной выборке и плато (или спад) предсказания точности для тестовой выборки с течением процесса обучения.

Dropout (Hinton et al., 2012) – регуляризатор, используемый для борьбы с переобучением. Суть метода заключается в случайном обнулении ряда нейронов и соответственно их связей (Рис. 13), что предотвращает избыточное запоминание признаков.

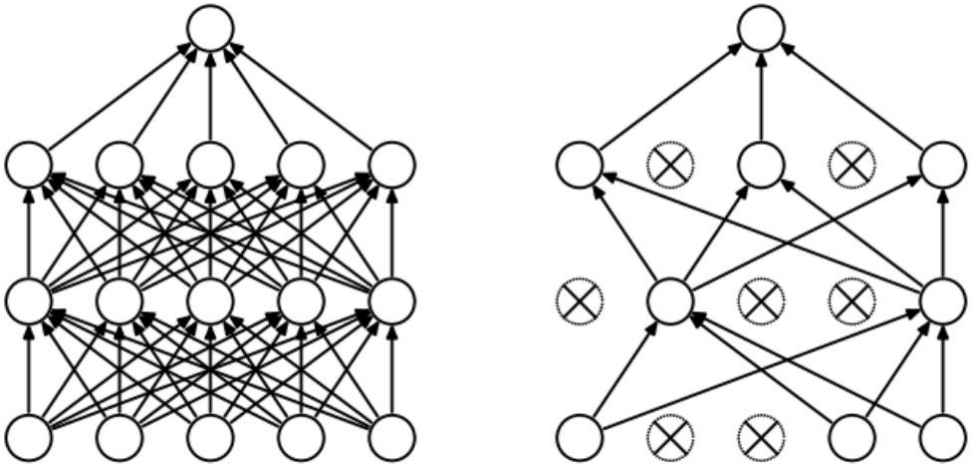


Рисунок 13. Схема работы Dropout. Слева – архитектура нейронной сети без прореживания, справа – та же нейронная сеть с добавлением Dropout (Srivastava et al., 2014).

Еще одним распространенным методом борьбы с переобучением является регуляризация, суть которой заключается в уменьшении слишком больших весов связей, что снижает вероятность запоминания предсказаний для конкретных наборов признаков. Регуляризация вводится как дополнительное слагаемое в уравнение оптимизационной функции:

$$J\left(w^{[1]}, b^{[1]}, \dots, w^{[L]}, b^{[L]}\right) = \frac{1}{m}\sum_{i=1}^{m} L\left(\hat{y}^{(i)}, y^{(i)}\right) + \frac{\lambda}{2m}\sum_{l=1}^{L}\left\|w^{[L]}\right\|_F^2$$

где L – функция потерь, w – вес нейрона, λ – параметр регуляризации.

Еще один существенный подход к борьбе с переобучением основан на введении шума. Данный способ позволяет вносить случайные отклонения, которые будут препятствовать запоминанию алгоритмом специфических индивидуальных признаков объектов.

1.7. Методы предсказания аффинности связывания, основанные на машинном обучении

Методы машинного обучения широко используются в биоинформатике, в частности в структурной биоинформатике. Увеличения набора свободно доступных данных в базе PDB (Berman et al., 2000)), и сложная пространственная организация делают данные о структуре белков идеальной областью применения современных методов машинного обучения, таких как глубокое обучение (Deep Learning) (LeCun et al., 2015). Тем не менее, существует ряд трудностей, которых следует избегать, чтобы правильно обучать и тестировать любой метод машинного обучения на данных о белковых структурах (Walsh et al., 2016).

1.7.1. Предсказание аффинности связывания в комплексах белок-пептид

Основными объектами для изучения белок-пептидных взаимодействий на данный момент являются белки, ассоциированные с иммунной системой, в

частности белки главного комплекса гистосовместимости (англ. MHC – Major histocompatibility complex). Это связано как с их высокой важностью для фармацевтики в синтезе новых пептидных лекарств, так и с наличием большого количества данных о связывании эти белков с пептидами, что позволяет собрать достаточную обучающую выборку.

В 2007 году была опубликована одна из наиболее ранних работ по предсказанию связывания в комплексах HLA-пептид (HLA – Human Leukocyte Antigen), основанная на искусственных нейронных сетях (Nielsen et al., 2007). Для реализации алгоритма, названного NetMHCpan, в обучающих данных молекулы HLA представлены в виде так называемых псевдопоследовательностей, состоящих из аминокислотных остатков последовательности белка, потенциально находящихся в контакте со связанным пептидом. Положения остатков, включенные в псевдопоследовательность, были получены из анализа последовательностей и структур HLA.

Архитектура использованной нейронной сети характеризуется наличием только прямых связей, то есть отсутствием петель и циклов (Kapp et al., 2003). Для каждого комплекса на вход в нейронную сеть подавались данные последовательностей для каждого комплекса, состоящей из 43 остатков пептид-HLA (9 из пептида и 34 из HLA), и в качестве выходных данных использовались соответствующие аффинности связывания, при этом аффинность связывания была переведена в диапазон от 0 до 1.

В качестве метрики качества использовалось значение AUC (Area Under the Curve), которое отражает площадь под ROC-кривой (Receiver Operating Characteristic, иногда её называют «кривая ошибок»). Часто результат работы алгоритма на фиксированной тестовой выборке визуализируют с помощью ROC-кривой, а качество оценивают как площадь под этой кривой – AUC (Рис. 14). Так, полнота (recall) метода составила 0,74, значение AUC составило 0,91 (Nielsen et al., 2007).

Рисунок 14. ROC-кривые для наилучшего (AUC=1), случайного (AUC=0.5) и наихудшего (AUC=0) алгоритма. Синим цветом обозначена ROC-кривая, розовым – площадь под ней, называемая AUC.

Упомянутая работа (Nielsen et al., 2007) послужила отправной точкой к использованию нейронных сетей для решения задачи. Так, в 2014 году была опубликована работа, посвященная реализации алгоритма NetMHCpan на расширенных данных, содержащих не только HLA, но также и молекулы MHC, отличные от человеческих (Pro et al., 2014). Структурная визуализация псевдопоследовательностей в NetMHCpan и в расширенном наборе показана на Рис.15. Даже учитывая расширение обучающих данных, точки контактов MHC с пептидами не изменились, что свидетельствует об их высокой консервативности в области интерфейса.

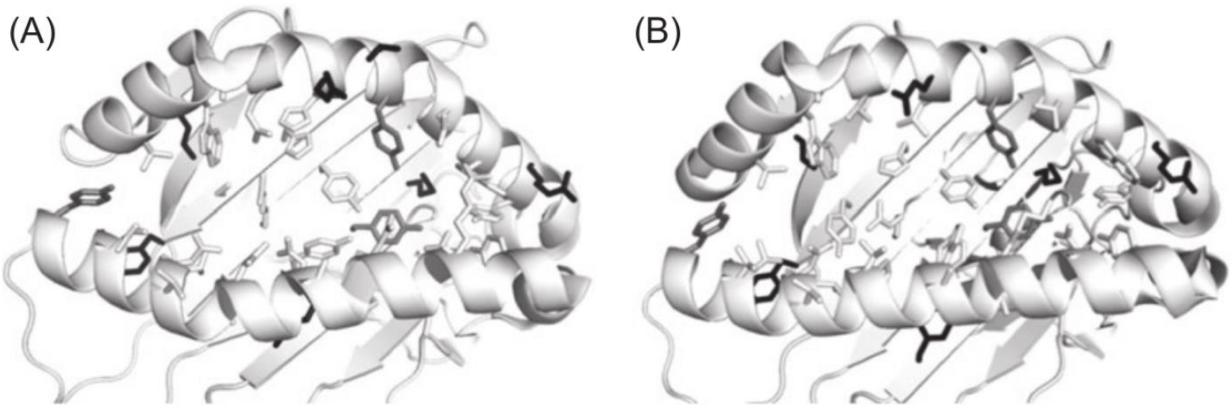


Рисунок 15. Структурная визуализация NetMHCpan и дополненные псевдопоследовательности в (A) Mamu-A*02 (код PDB: 3JTS) и (B) HLA A*

02:01 (код PDB: 3D25). Остатки, показанные белым, обнаружены в обеих конструкциях, остатки, присутствующие исключительно в NetMHCpan, показаны светло-серым, а остатки, включенные только в «расширенный обучающий набор», показаны черным (Pro et al., 2014).

Также было обнаружено, что учет контактов, опосредованных водой, позволил улучшить предсказательную точность, что указывает на важность непрямых контактов в связывание пептида. Новая псевдопоследовательность была проверена на трех различных тестовых наборах данных MHCI и дальше сравнивали ее эффективность с NetMHCpan. Точность предсказания повысилась для MHC, отличных от человеческих (Pro et al., 2014), было достигнуто значение AUC = 0,84, которое превосходит значение для NetMHCpan, равное 0,81.

В последние годы широкое применение также получили глубокие нейронные сети. Глубокое обучение – это совокупность методов и приёмов для эффективного обучения искусственных нейронных сетей различного типа с большим числом скрытых слоёв (LeCun et al., 2015). Эти слои содержат иерархические представления интересующих объектов, извлеченные из входных данных. В последнее время они широко применяются для решения многих биологических задач, включая оценку связывания в комплексах белок-пептид (Nielsen et al., 2024; Otvos et al., 2024).

В 2015 году была опубликована работа, в которой использование нелинейного алгоритма, основанного на функционально-связанных нейронных сетях, показало результат предсказания связывания в комплексах MHC-пептид более высокий, чем при использовании глубоких нейронных сетей (Kuksa et al., 2015).

В функционально-связанных нейросетях расширяется стандартная архитектура прямого и обратного распространения ошибки модификацией узлов на входном слое. Входы комбинируются математическим

преобразованием с помощью таких функций, как квадратичные, кубические или синусоидальные. Из названий функционально-связанных входов и вытекает название нейросетей (Gupta et al., 2013). Архитектуры глубокой и функционально-связанной нейронных сетей показаны на Рис. 16.

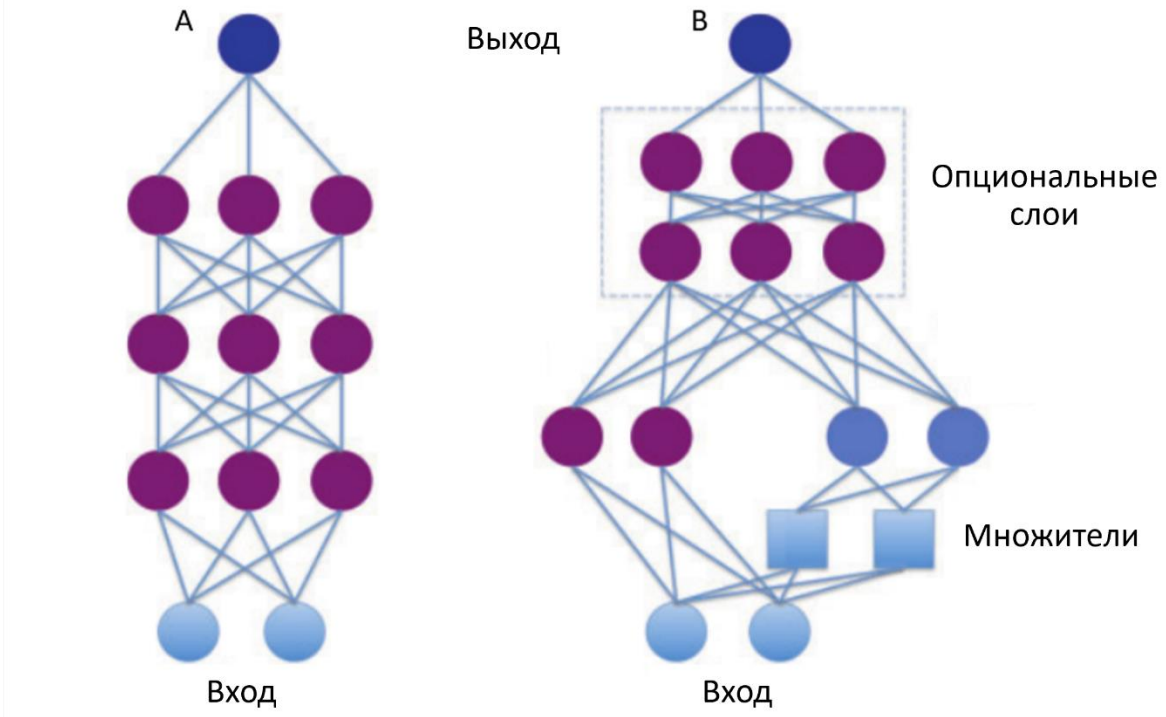


Рисунок 16. Архитектуры глубокой нейронной сети (A) и функционально-связанной нейронной сети (B). Голубым обозначены входные слой, фиолетовым – скрытые слои, синим – выходные слои. Квадратами обозначено математическое преобразование входных данных (Kuksa et al., 2015).

Авторы в своем исследовании в качестве обучающих данных использовали пептиды, состоящие из 9 остатков. В задаче прогнозирования связывания MHC-I было установлено пороговое значение IC_{50} = 500 нМ для разделения пептидов по силе связывания. Соответственно, пептиды с сильным связыванием ($IC_{50} < 500$) и пептиды со слабым связыванием ($IC_{50} > 500$). Было показано, что сочетание метода опорных векторов и функционально-связанных нейронных сетей позволяет достичь значения AUC = 0,95, что значительно выше, чем при использовании глубоких нейронных сетей (0,918) и NetMHCpan (0,91).

В свою очередь в работе 2017 года применялись методы машинного обучения из области обработки естественного языка для решения задачи прогнозирования связывания в комплексах HLA-пептид (Vang and Xie, 2017). Была предложена глубокая сверточная архитектура нейронной сети, названная HLA CNN, для задачи предсказания связывания HLA класса I с пептидами. В качестве обучающих данных использовалась преобразованная информация о последовательностях пептидов, в частности, учитывалось распределение различных остатков в определенном положении пептида.

Архитектура сверточной нейронной сети, используемая в работе Ванга показана на Рис. 17.

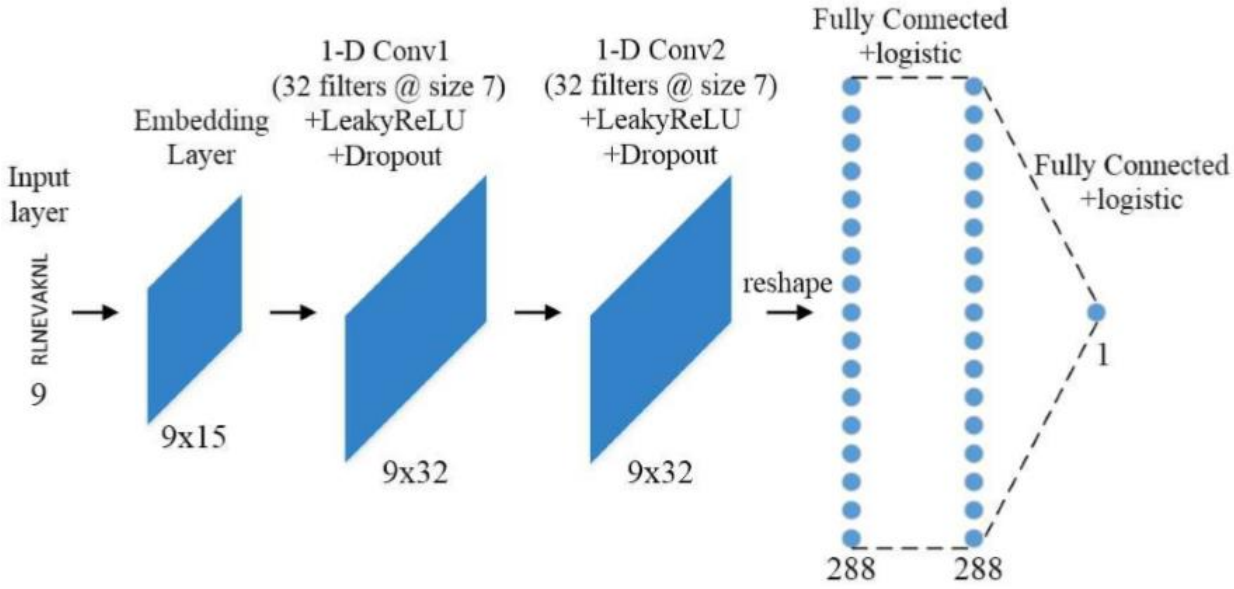


Рисунок 17. Архитектура сверточной нейронной сети для предсказания связывания в комплексе MHC-пептид (пептиды размером в 9 остатков).

На вход подаются данные о пептиде. Слой встраивания (Embedding layer) переводит данные в векторное представление. Далее идут два сверточных слоя сохраняют входную длину, используя 32 фильтра размера 7. Выход 2-го сверточного слоя преобразуется в одномерный вектор, который полностью связан со следующим полносвязным слоем того же размера. Затем этот слой

связывается со выходным (Vang and Xie, 2017). Для борьбы с переобучением использовался метод прореживания Dropout. Благодаря использованию сверточной нейронной сети, удалось достигнуть значения AUC = 0,836 на наборе данных, при предсказании которого вышеупомянутый алгоритм NetMHCpan достигает значения 0,778 (Vang and Xie, 2017).

В 2019 году была опубликована работа, в которой также использовался предсказательный алгоритм на базе глубоких сверточных нейронных сетей (Zhao et al., 2019). В методе предложенном Чжао для составления обучающих данных использовались комплексы MHCI-пептид, включающие пептиды разной длинны, а также в данные вводились дополнительные признаки пептидов. Архитектура данного алгоритма представлена на Рис. 18.

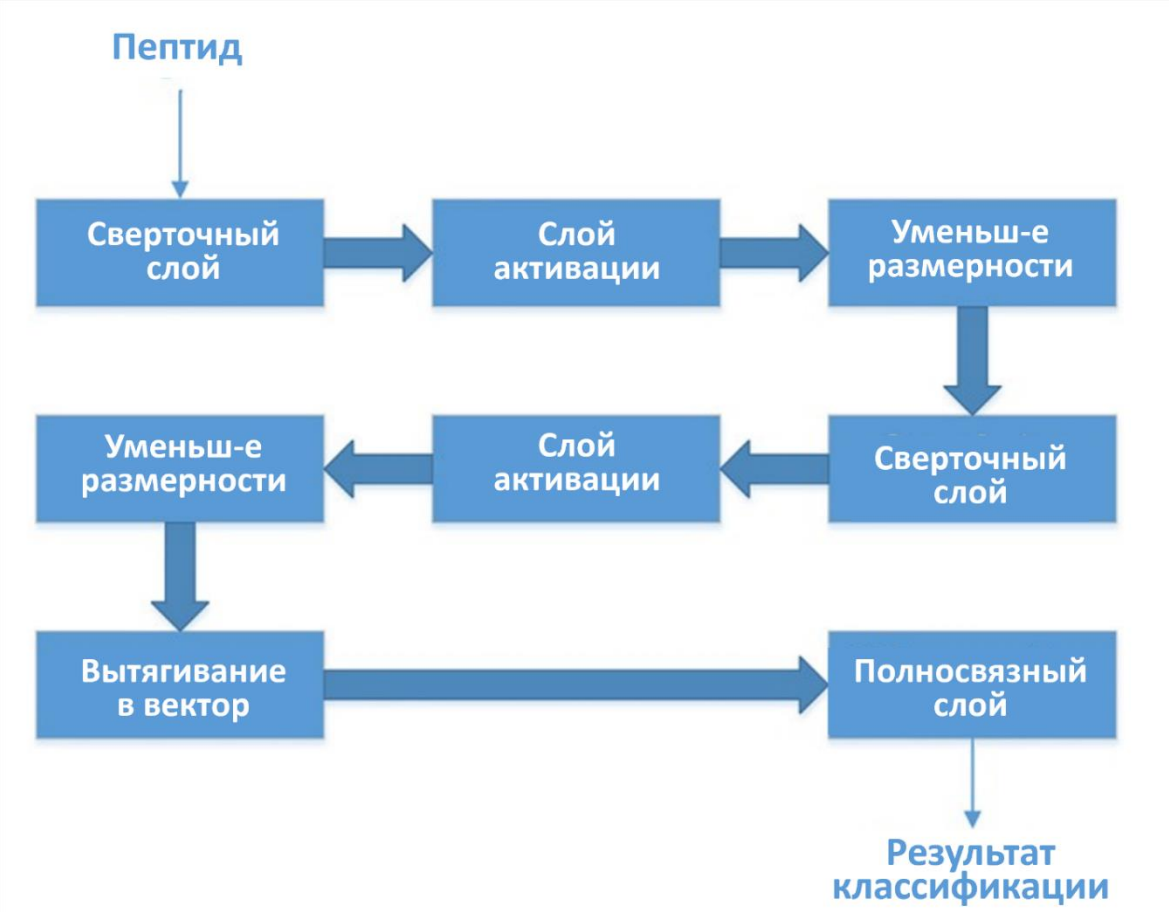


Рисунок 18. Архитектура свёрточной нейронной сети в работе Чжао (Zhao et al., 2019).

В результате для многих аллелей MHCI удалось достичь значения AUC >0,9, а также наблюдалось меньшее значение стандартного отклонения, чем в ряде других методов, что говорит о большей стабильности данного алгоритма.

Помимо методов, использующих данные об аминокислотной последовательности, для комплексов белок-лиганд реализованы также алгоритмы, работающие с пространственными структурами. В 2017 году был реализован алгоритм под названием Pafnucy, основанный на глубоких сверточных нейронных сетях и использующий в качестве обучающих данных PDB-структуры комплексов (Stepniewska-Dziubinska, 2017).

Для каждого комплекса задавалась кубическая ячейка, со стороной в 20 Å, центром ячейки назначался центр пептида (Рис. 19). Так, создавался 4D-тензор, 3 измерения – координаты, а четвертое – признаки.

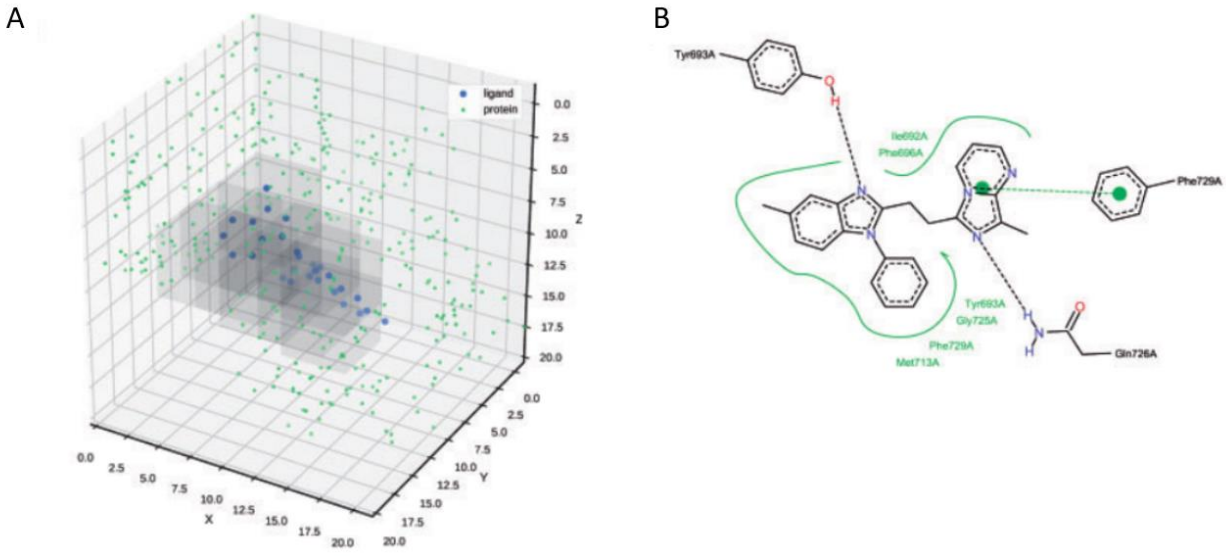


Рисунок 19. Входные данные для алгоритма Pafnucy. A) Кубическая ячейка, включающая координаты атомов пептида и белка в комплексе. B) Взаимодействия между белком и лигандом (Stepniewska-Dziubinska, 2017).

Преобразованные структурные данные подавались на вход алгоритму, основанному на глубокой сверточной нейронной сети, архитектура которой показана на Рис. 20.

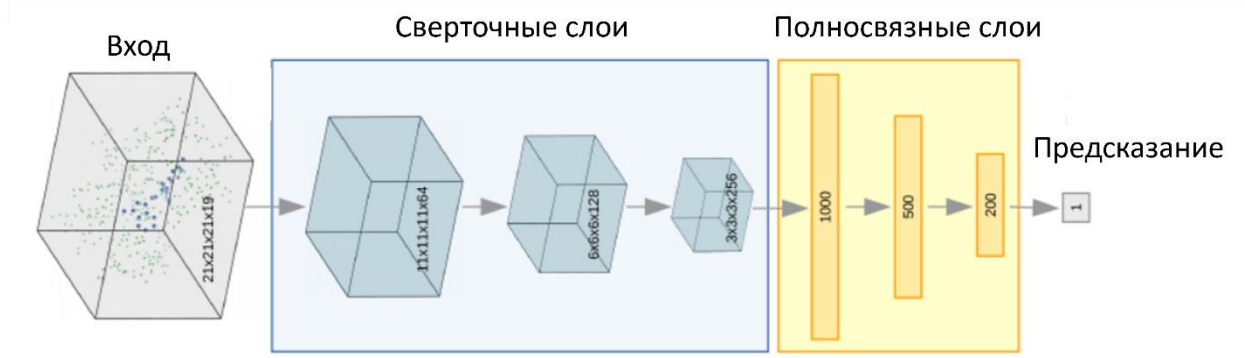


Рисунок 20. Архитектура глубокой нейронной сети Pafnucy (Stepniewska-Dziubinska, 2017).

В качестве целевой метрики было выбрано значение pK_D (равное $-\log K_D$ или –logKi). По итогу для тестового набора было достигнуто значение корреляции Пирсона, равное 0,78, между предсказанными значениями pKa и экспериментально рассчитанными.

В 2018 году было опубликовано похожее исследование по предсказанию константы связывания в комплексах белок-лиганд с использованием трехмерных глубоких сверточных нейронных сетей (Jiménez et al., 2018).

По сравнению с предыдущей работой в KDEEP используется другое представление данных, в частности воксельное отображение комплексов белка с лигандом (Рис. 21 (a)). При таком способе отображения объемное изображение записывается в двумерном пространстве с добавлением дополнительной оси, отвечающей за глубину, в данном случае на дополнительную ось записывались признаки атомов белка и лиганда, в общей сложности 16 каналов. Таким образом на вход алгоритму подавался не 4D-тензор, а 3-D тензор. Общая схема работы алгоритма показана на Рис. 21 (b).

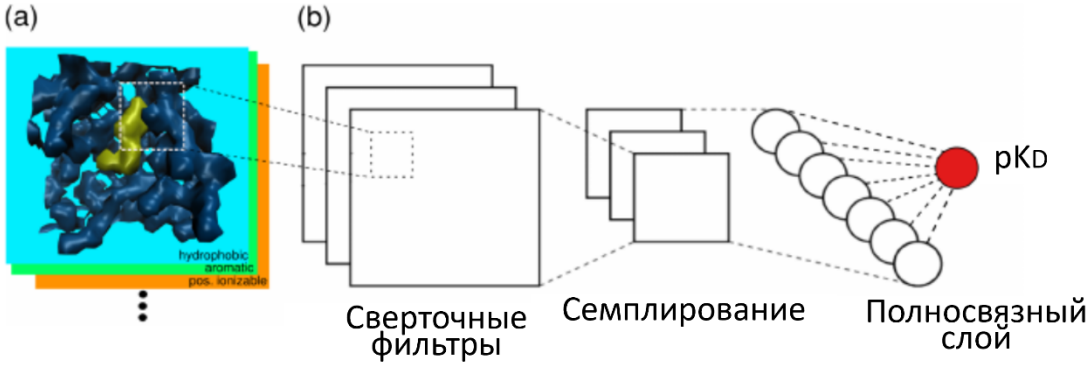


Рисунок 21. Схема работы модели KDEEP. (a) Воксельное представление комплекса белок-лиганд; (b) архитектура используемой сверточной нейронной сети (Jiménez et al., 2018).

В качестве целевой переменной так же, как и в предыдущей работе было выбрано значение pK_D. В результате для тестового набора было достигнуто значение корреляции Пирсона, равное 0,82 между предсказанным значением pK_D и полученным экспериментально (Jiménez et al., 2018).

Таким образом, помимо данных о первичной структуре комплексов белок-лиганд используются также данные о пространственной структуре для предсказания константы связывания. Однако, реализован такой подход на данное время в основном для более простых лигандов, чем пептидные. Несмотря на это, в перспективе использование пространственных структур может открыть новые возможности в решении задачи предсказания энергии связывания и в комплексах белок-пептид.

1.7.2. Предсказание аффинности связывания в комплексах белок-белок

Хотя предсказание аффинности связывания в комплексах белок-низкомолекулярный лиганд решается в настоящее время с достаточно высоких качеством, до сих пор не удалось разработать метод, позволяющий с высокой точностью предсказать эти характеристики для белковых комплексов различной природы. Это явление может быть связано с рядом ограничений, таких как недостаточный объем и разнообразие обучающих данных,

невозможность учитывать конформационную подвижность и другие (Kastritis et al., 2013). В частности, прогностические модели могут плохо работать в области высокой аффинности связывания. Это связано с тем, что аффинность связывания большинства белок-белковых комплексов из использованных наборов данных находится в диапазоне низкой аффинности (Guo et al., 2022). Поэтому для изучения структуры высокоаффинных комплексов в настоящее время недостаточно данных, чтобы сделать качественные прогнозы аффинности связывания. Кроме того, первоначально стандартные ошибки измерения K_D обычно составляют порядка 0,25 ккал/моль для ΔG. Эти оценки получены в результате повторных измерений с использованием того же оборудования, среды и протокола (Jankauskaitė et al., 2019). Однако реальная погрешность эксперимента с учетом условий окружающей среды, оборудования и т.п. для ΔG может составлять 1,4–2,3 ккал/моль (Tian et al., 2012).

Одними из наиболее перспективных методов прогнозирования в настоящее время являются методы, основанные на классическом машинном обучении и нейросетевые методы.

1.7.2.1. ISLAND

В работе 2020 года был реализован алгоритм ISLAND (Abbasi et al., 2020). Аббаси с коллегами использовали информацию о последовательности белка вместо структуры белка наряду с методами машинного обучения для точного прогнозирования аффинности связывания. Для преобразования данных ученые использовали несколько явных функций (особенности аминокислотного состава, физико-химические свойства белка и т.д.) и различные представления ядра для моделирования атрибутов белковых комплексов на основе последовательностей. Методы ядра представляют альтернативный способ представления последовательности путем

моделирования степени сходства между белковыми последовательностями вместо явного представления признаков (Рис. 22).

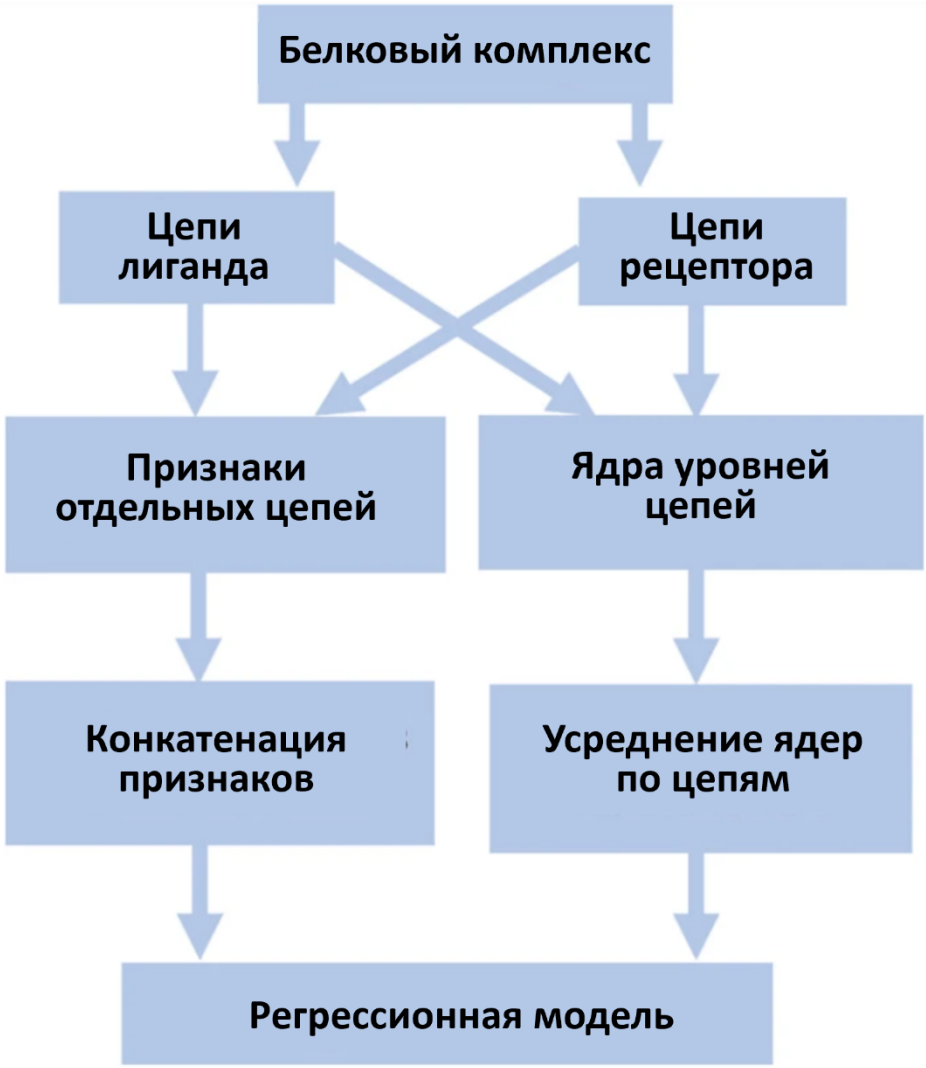


Рисунок 22. Методы, принятые для генерации представления признаков белкового комплекса на основе последовательности для разработки моделей прогнозирования аффинности связывания белков на основе машинного обучения (Abbasi et al., 2020).

В качестве предсказательных моделей использовались следующие методы: метод наименьших квадратов, метод опорных векторов, случайный лес. Лучший результат предсказания показал метод опорных векторов, использовании дескрипторов ядра получили максимальную корреляцию 0,44 с RMSE = 2,56 между прогнозируемыми и экспериментальными значениями

ΔG. Данный результат является одним из самых высоких для методов, использующих информацию о первичной структуре белка для анализа. Однако полученного качества предсказания недостаточно для дальнейшего применения данного метода в экспериментальной работе. Следовательно, на данный момент для оценки аффинности связывания в белок-белковых комплексах необходима информация не только об аминокислотной последовательности, но и о пространственной структуре.

1.7.2.2. PRODIGY

В 2016 году был реализован веб сервис PRODIGY (Xue et al., 2016), для прогнозирования сродства к связыванию белок-белковых комплексов на основе их трехмерной структуры. Сервер PRODIGY реализует прогностическую модель, основанную на межмолекулярных контактах и свойствах, полученных на основе поверхности, не являющейся границей раздела.

Прогностическая модель основана на простой линейной регрессии с учетом контактов в области интерфейса связывания (ICs – interfacial contacts) и некоторых свойствах не интерфейсных поверхностей (Nis – non-interacting surfaces), которые, как было показано, влияют на аффинность связывания (Kastritis et al., 2013):

$$predicted = -0{,}095\, ICs_{charged/charged} - 0{,}1\, ICs_{charged/apolar} + 0{,}196\, ICs_{polar/polar} - 0{,}227\, ICs_{polar/apolar} + 0{,}187\, \%NIS_{apolar} + 0{,}381\, \%NIS_{charged} - 15{,}943,$$

где ICxxx/yyy отражает количество контактов, обнаруженных на границе раздела между взаимодействующим агентом 1 и агентом 2, классифицировано в соответствии с полярной / неполярной / заряженной природой взаимодействующих остатков (т.е. $ICs_{charged/apolar}$ - количество контактов на интерфейсе между заряженными и неполярными остатками). Два остатка

считаются потенциально контактирующими, если какой-либо из их тяжелых атомов находится на расстоянии не более 5,5 Å (Рис 23).

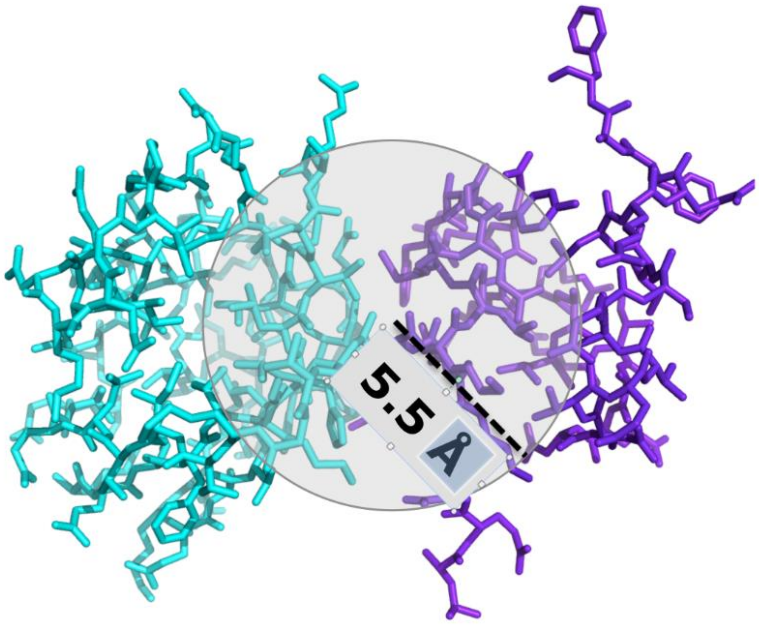


Рисунок 23. Пример ICs (Kastritis et al., 2013).

По заявлению авторов работы, разработанная модель предсказывает сродство к связыванию с беспрецедентной точностью для большого и разнородного набора данных комплексов (всего 81) (Vangone et al., 2015), так корреляция Пирсона между прогнозируемыми и экспериментальными значениями составила 0,73 (p-value<0,0001) и RMSE=1,89 ккал/моль.

1.7.2.3. PPI-Affinity

В 2022 году был предложен алгоритм PPI-Affinity (Romero-Molina et al., 2022), оценивающий аффинность связывания в белок-белковых комплексах с использованием методов классического машинного обучения. Данные о характеристиках пространственных структур комплексов были представлены в виде дескрипторов, реализованных в пакете кодификации белков ProtDCal (Romero-Molina et al., 2019). Данная программа вычисляет все комбинации определенных параметров, таким образом создавая один индивидуальный вектор признаков, называемый дескриптором, из каждой комбинации (Рис.

24). Комбинация выбранных индексов, операторов близости, групп и операторов агрегации приводит к большому набору дескрипторов для каждого белка. При данном подходе вся информация о комплексе записывается в виде одномерного вектора.

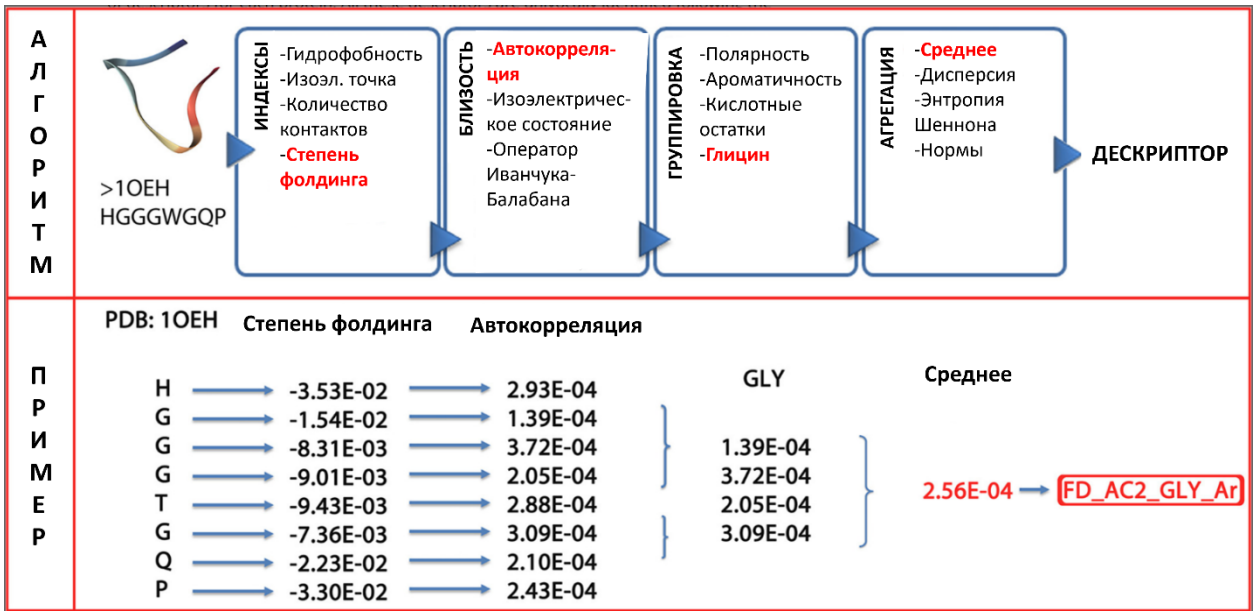


Рисунок 24. Основные шаги для вычисления дескриптора белка. Фрагмент человеческого прионного белка (верхняя панель, крайний слева) с идентификационным кодом PDB ID 1OEH используется в качестве примера кодируемого белка (Romero-Molina et al., 2019).

Для обучения подбирались структуры белок-белковых комплексов, состоящих из двух цепей из базы PDBBind v. 2020 (Wang et al., 2020). В качестве предсказательного алгоритма использовался метод опорных векторов. По итогам тестирования на двух наборах данных получился результат, превосходящий все существующие на тот момент алгоритмы (PRODIGY (Xue et al., 2016), DFIRE (Zhang et al., 2004), CP_PIE (Ravikant et al., 2010), ISLAND (Abbasi et al., 2020)).

1.7.2.4. AREA-AFFINITY

В 2023 году был реализован веб-сервис AREA-AFFINITY (Yang et al., 2023), решающий задачу предсказания аффинности связывания в белок-белковых комплексах с использованием ансамблевого объединения линейных и нелинейных алгоритмов. Для представления данных о комплексах используется также метод формирования одномерного дескриптора. Так, на основе структуры белок-белкового комплекса вычисляются 18 дескрипторов на основе данных о поверхности различных партнеров связывания в комплексе, в частности учитываются многочисленные взаимодействия как как на интерфейсе связывания, так и аминокислотных остатков с молекулами воды (Yang et al., 2022).

Аффинность связывания прогнозируется с использованием моделей, основанных на анализе поверхности интерфейса взаимодействия (60 репрезентативных моделей для прогнозирования сродства связывания белок-белок или 37 репрезентативных моделей, специфичных для прогнозирования сродства связывания антитело–белок-антиген). 60 моделей для прогнозирования сродства связывания белок –белок состоят из 12 линейных моделей, 17 построенных нелинейных моделей, 10 смешанных моделей на основе построенных нелинейных моделей, 18 нелинейных (нейросетевых) моделей и 3 смешанных моделей на основе нейросетевых моделей.

Для тестирования использовались два подмножества, содержащие 52 и 24 комплекса белок–белок с экспериментально рассчитанными значениями сродства связывания. Наилучший коэффициент корреляции Пирсона (R) между предсказанным и экспериментальным сродством связывания для этих 60 моделей составляет 0,87 для 52 комплексов. Алгоритм работы AREA-AFFINITY показан на Рис 25.

Рисунок 25. Блок-схема для прогнозирования сродства связывания на основе областей контактирующих соединений. Для анализируемой структуры белок–белкового комплекса, AREA-AFFINITY вычисляет дескрипторы на основе площадей в комплексе и предсказывает сродство связывания в соответствии с предварительно подготовленными моделями (Abbasi et al., 2023).

ГЛАВА 2. МАТЕРИАЛЫ И МЕТОДЫ

2.1. Базы данных, используемые для сборки обучающего и тестовых наборов данных

Обучающий набор комплексов был собран с использованием базы данных PDBBind v.2020 (Wang et al., 2020) и PDB (Berman et al., 2000) (версия 2023 года). Для создания тестовых выборок использовалась также база данных SKEMPI v.2.0 (Jankauskaitė et al., 2019).

2.2. Библиотеки, использованные для предобработки данных и обучения предсказательного алгоритма

Для создания предиктивной модели, оценивающей аффинность связывания в белок-белковых комплексах, был использован язык программирования Python3. Использованные библиотеки описаны в Таблице 2.

Таблица 2. Описание использованных в работе библиотек.

Название библиотеки	Версия	Назначение
numpy	1.22.4	Используется для работы с массивами
pandas	1.5.3	Используется для работы с табличными данными
matplotlib	3.7.1	Используется для визуализации атомов и процесса обучения
biopython	1.81	Используется для извлечения необходимой информации о комплексах из

		структурных файлов pdb
scikit-learn	1.2.2	Используется для обучения модели логистической регрессии, необходимой для локализации интерфейса связывания
torch	2.0.1 +cu118	Используется для создания и обучения предсказательной нейросетевой модели

2.3. Создание набора данных для обучения предсказательного алгоритма

Для обучения нейронной сети был создан набор данных из структурных файлов белково-белковых и белок-пептидных комплексов с известной аффинностью связывания. Структурные файлы были извлечены из PDB в формате pdb. Поскольку большая часть структур (1913 структуры) содержат в pdb-файле более двух цепей, для них были вручную размечены цепи, участвующие в связывании. После анализа данных часть комплексов (8% от исходного набора) не вошла в окончательный набор данных по следующим причинам: слишком большой размер интерфейса связывания (размер больше анализирующей ячейки), наличие небелковых молекул на интерфейсе (например, ДНК) или наличие в файле менее 2 цепей, также если невозможно отметить цепи, участвующие в связывании. В результате для обучения было выбрано 2397 комплексов из исходного набора. В связи с относительно небольшим набором доступных данных было решено расширить обучающую выборку конформациями белок-белковых комплексов, полученными методами МД. В результате для 142 комплексов были получены траектории со 100 конформациями молекул. Чтобы расширить набор обучающих данных,

для каждого комплекса было взято по 11 конформаций (каждая десятая от 0 до 100). В результате обучающая выборка состояла из 3959 структурных файлов. Для валидации случайным образом было выбрано 90 экспериментально полученных структур.

В качестве целевого значения использовалось значение pK_D. Для сравнения с существующими прогнозными моделями рассчитывалась свободная энергия Гиббса по полученным значениям по формуле:

$$\Delta G = TRln(K_D),$$

где T — температура (принимается за 300 К), R — универсальная газовая постоянная (8,314 Дж/(моль*К)).

2.4. Гиперпараметры обучения нейросетевого алгоритма

Для оптимизации процесса обучения в ходе работы подбирались гиперпараметры, описанные в Таблице 3.

Таблица 3. Гиперпараметры обученного нейросетевого алгоритма ProBAN

Название	Значение	Функция
Функция потерь	MSELoss	Используется для расчета ошибки предсказаний
Оптимизатор	AdamW	Осуществляет процесс минимизации функции потерь с возможностью внесения L2-регуляризации
Функция активации	ReLU	Создает выход для нейрона на основе поступивших данных
Нормализация	Batch Normalization	Используется для удержания выходных данных скрытых слоев в постоянном диапазоне
Регуляризация	L2, Dropout	Препятствует наступлению переобучения

Для апробации алгоритма на валидационной и тестовых выборках применялись следующие метрики качества: корреляция Пирсона и RMSE.

2.5. Создание тестовых выборок для апробации предсказательного алгоритма

Для проверки обученного алгоритма использовалось два набора тестов. Для оценки работоспособности модели на комплексах с тремя и более цепями собран новый набор тестов (тест 1). Для него были отобраны комплексы с высоким разрешением (более 2,8 Å), не содержащие белков из обучающего набора, в результаты тестовый набор данных состоял из 127 комплексов. Другой набор данных, взятый из работы по разработке PPI-Affinity (Romero-Molina et al., 2022), содержал комплексы из PDBBind, состоящие только из двух цепей (тест 2). Этот набор состоит из 90 комплексов, из которых оставлено 82 комплекса, так как 8 из исходного набора были исключены по указанным выше причинам.

Для дополнительного анализа аффинности связывания были отобраны комплексы RBD-ACE2 с известными пространственными структурами и значениями константы диссоциации. Итоговый набор данных содержит 48 комплексов ACE2 и RBD вирусов SARS CoV (3 структуры) и SARS CoV-2 (45 структур), включая исходный вариант, а также различные штаммы и инженерные варианты. Также для тестирования был собран набор комплексов с известными значениями K_D, состоящих из гистонов, взаимодействующих с различными белками хроматина.

Для анализа стабильности комплексов, образованных разными вариантами гистонов H2A, H2B и H3, был собран набор из 8 пространственных структур нуклеосом, содержащих анализируемые варианты, также были размечены взаимодействующие цепи.

2.6. P-оценка статистической значимости

P-оценка – это статистическая достоверность, наименьшее значение уровня значимости, при котором полученная проверочная статистика ведёт к отказу от основной (нулевой) гипотезы. Значения, близкие к 0, говорят о высокой значимости результатов. Математически P-оценка рассчитывается как площадь области под кривой распределения вероятностей, находящаяся правее наблюдаемого значения (Рис. 26).

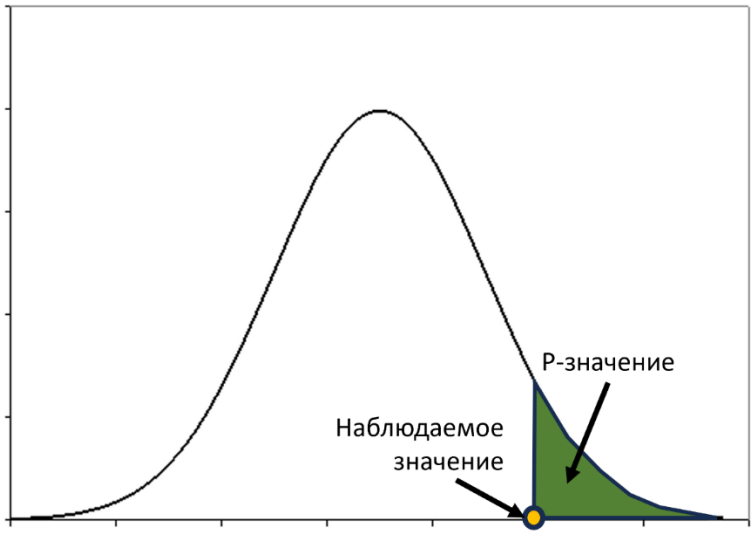


Рисунок 26. P-оценка, рассчитываемая как площадь области под кривой распределения вероятностей, находящееся правее наблюдаемого значения.

2.7. Анализ межмолекулярных взаимодействий и расчет траекторий МД

Визуализация комплексов и оптимизация сети водородных связей выполнена с помощью программ Maestro (Schrödinger, LLC) и VMD (Humphrey et al., 1996)).

Молекулярно-динамический расчет проводился с помощью программы Gromacs (Van Der Spoel et al., 2005) в полноатомном силовом поле CHARMM36m (Huang et al., 2017) в ансамбле NPT при температуре 300 К в течение 100 нс с использованием модели воды TIP3P.

2.8. Программы, используемые для расчета аффинности связывания в комплексах альтернативными методами.

Предсказание с использованием Prodigy осуществлялось на одноименном онлайн-сервисе, для FoldX использовалась одноименная программа FoldX Suite 4.0. Расчет с использованием остальных методов (RosettaDock, DFIRE, CP_PIE) осуществлялся на онлайн-сервисе CCharPPI (Moal et al., 2015) (версия 2024 г.).

ГЛАВА 3. РЕЗУЛЬТАТЫ И ИХ ОБСУЖДЕНИЕ

3.1. Новый подход к преобразованию пространственных структур белок-белковых комплексов

3.1.1. Анализ обучающей и тестовых выборок

Обучающая выборка состояла из 3959 структурных файлов, в состав которых вошли как экспериментально полученные структуры, так и конформации, полученные методами МД. Такое расширение позволило добавить данные для обучения со значениями константы диссоциации, которые встречаются реже относительно исходного распределения (Рис. 27A,B). Однако из-за небольшого количества комплексов с pK_D <4 и pK_D >12 трудности в прогнозировании энергии связывания в таких комплексах оказались неизбежны.

При распределении pK_D для наборов тестовых данных можно заметить, что внутренний тестовый набор содержит больше комплексов с pK_D >7 (Рис. 27C), а внешний содержит больше комплексов с pK_D <8 (Рис. 27D). Благодаря такому распределению комплексов на тестовые выборки работу алгоритма можно проверить как в области низкой аффинности, так и высокой.

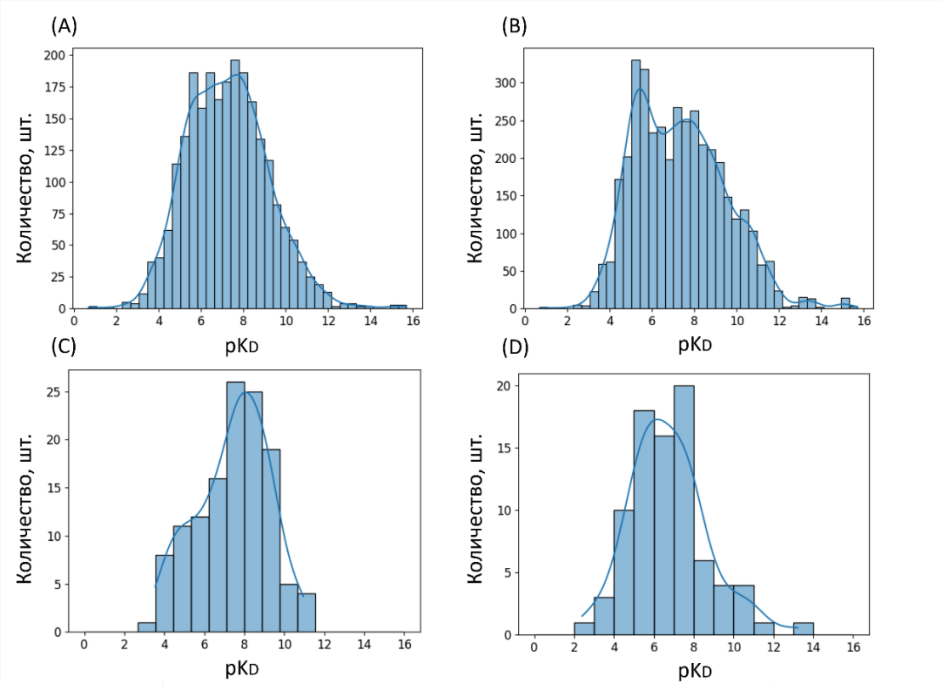


Рисунок 27. Распределение белок-белковых комплексов в использованных

наборах данных по целевым значениям (pK_D). (A) Распределение комплексов по целевым значениям для исходного обучающего набора. (B) Распределение комплексов в обучающем наборе после добавления дополнительных структур, полученных из МД. (C) Распределение комплексов во внутреннем тестовом наборе. (D) Распределение комплексов во внешнем тестовом наборе (Romero-Molina et al., 2022).

3.1.2. Локализация интерфейса связывания и формирование ограничительной ячейки

Ориентация молекул в пространстве относительно системы координат и расположение интерфейса связывания в разных комплексах сильно различаются. Поэтому на первом этапе обработки данных решалась задача локализации интерфейса связывания молекул в ячейке универсального размера, подходящей для большинства комплексов. Сначала из структурных файлов была извлечена информация об атомах белковых молекул (координаты, аминокислотные остатки, цепи) (Рис. 28).

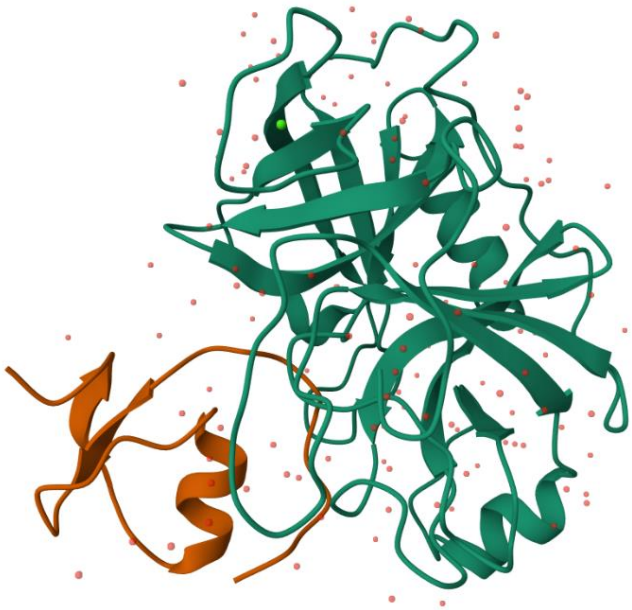


Рисунок 28. Ленточное отображение белок-белкового комплекса ингибитора триптазы с трипсином (pdb id 1an1), трипсин выделен зеленым цветом; его

ингибитор оранжевым.

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. 29) 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.

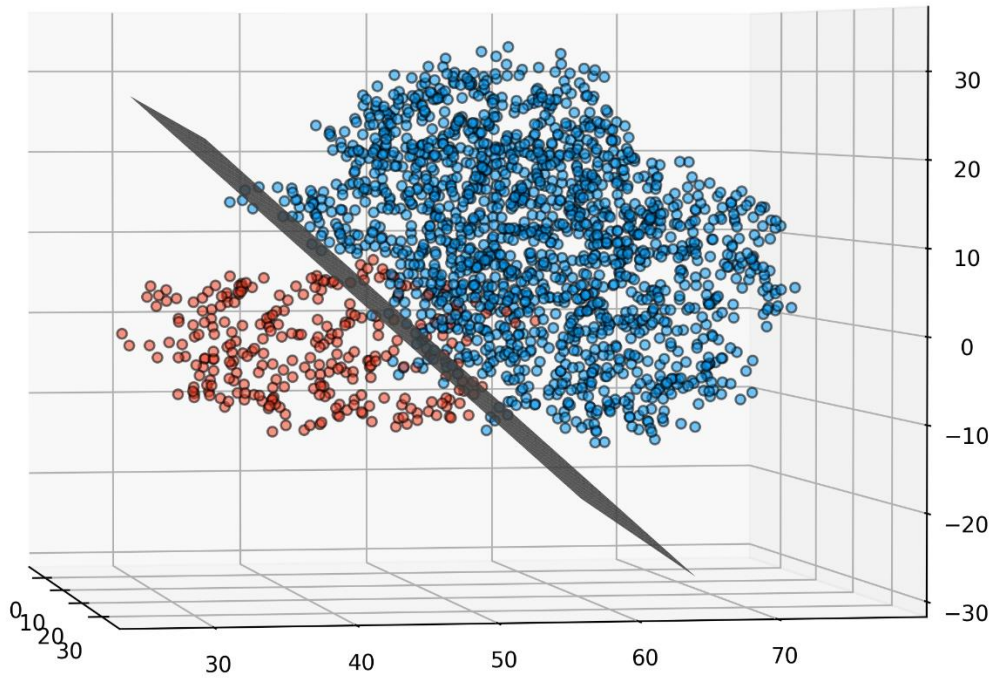


Рисунок 29. Точечное представление комплекса 1an1. Рассчитанная разделяющая плоскость выделена серым цветом, атомы трипсина — синим, а ингибитора — красным.

Следующим шагом был расчет центров масс в области интерфейса по обе стороны от разделяющей плоскости. Для этого определяли две плоскости, параллельные разделяющей, расположенные на расстоянии 10 Å от нее. Коэффициенты для координат остались такими же, как и для полученной плоскости, а смещения рассчитывались по формуле:

$$\beta_{1,2} = \pm 10\sqrt{W_1^2 + W_2^2 + W_3^2}$$

Далее были рассчитаны центры масс скоплений атомов, расположенных по обе стороны от разделяющей области, ограниченной дополнительными плоскостями. Из полученных точек строился вектор, а его пересечение с разделяющей плоскостью принималось за центр ячейки (Рис. 30).

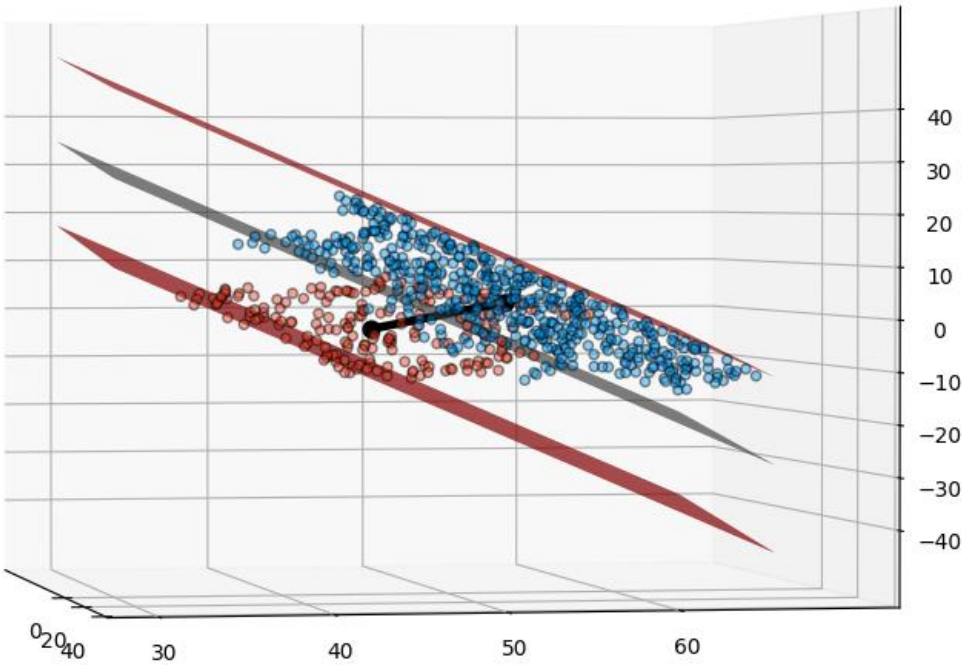


Рисунок 30. Точечное изображение комплекса 1an1 и вектор (черный), соединяющий центры масс областей, ограниченных разделяющей плоскостью и параллельных ей на расстоянии 10 Å (отмечено красным). Точка пересечения вектора и разделяющей плоскости принимается за центр ограничивающей ячейки.

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 перемещался 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. 31).

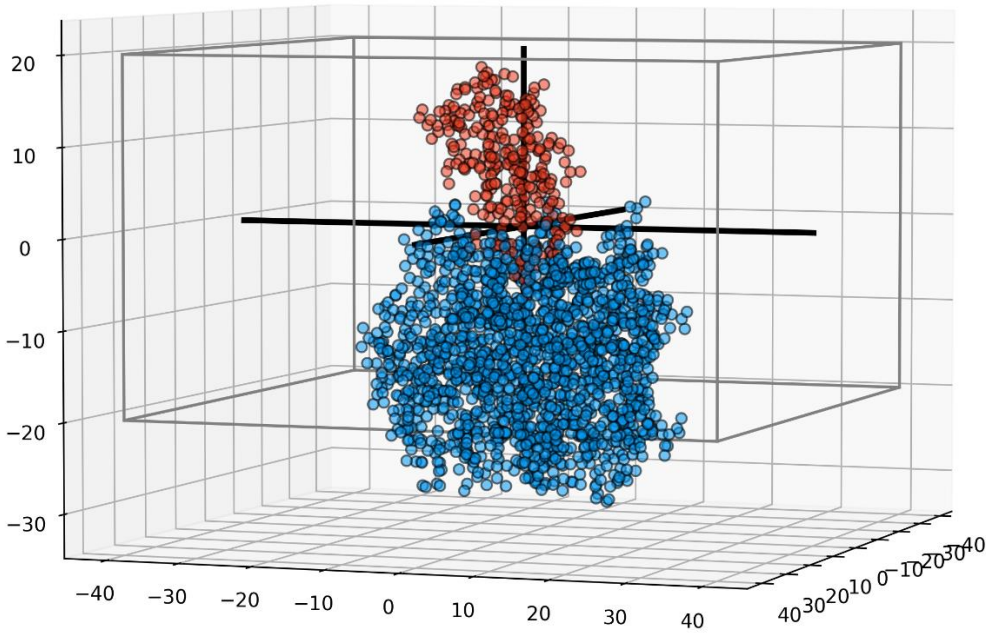


Рисунок 31. Точечное представление комплекса 1an1, координаты которого были преобразованы так, что центр ограничивающей ячейки (серый цвет) находится в начале координат. Оси OX, OY, OZ в новой системе координат выделены черным цветом.

Выбор такого размера ограничивающей ячейки был сделан на основе анализа размеров интерфейса связывания в изученных белок-белковых комплексах.

Подбор размера ячейки состоял из нескольких этапов. Сначала подбиралась высота, исходя из информации о поверхности связывания белка и максимального расстояния между атомами взаимодействующих молекул, необходимого для образования контактов, которое было принято равным 10 Å. Поверхность взаимодействия во многих случаях имеет неровности, впадины и т. д., соответственно, возникает необходимость увеличения высоты ячейки для учета всех возможных атомов, которые могут повлиять на сродство связывания в комплексе. Таким образом, от расчетного центра ячейки должно быть расстояние до границ 10 Å с каждой стороны, т.е. высота ячейки должна быть больше или равна 20 Å. Дальнейшее увеличение роста происходило исходя из результатов обучения модели (Таблица 4). Таким образом, оптимальным вариантом среди рассмотренных оказалось расстояние от центра ячейки до верхней и нижней границ 20 Å (высота = 41 Å, поскольку рассматривается и центральная точка). Ширина и высота выбраны исходя из возможности включения в анализ максимального количества крупных комплексов. Таким образом, были протестированы варианты от 41 до 91 Å. Подробная информация о размерах ячеек представлена в Таблице 4. Области заданного размера было достаточно, чтобы вместить весь интерфейс взаимодействия для более чем 93% всех комплексов.

Таблица 4. Результат подбора размеров ограничительной ячейки

Размер ограничительной ячейки	Доля подходящих комплексов, %	Значение функции потерь на валидационной выборке
21x41x41	85,7	0,87
31x41x41	87,6	0,81
41x41x41	88,9	0,75
51x41x41	88,9	0,82
41x61x61	92,2	0,66
41x81x81	93,4	0,5
41x91x91	93,6	0,5

Из атомов, попавших в ограничительную ячейку, для последующей работы отбирались те, которые находились на расстоянии не более 10 Å от ближайшего атома взаимодействующей молекулы. Координаты выбранных атомов были сохранены в трехмерном массиве размером 41x81x81 с разрешением 1 Å (Рис. 32).

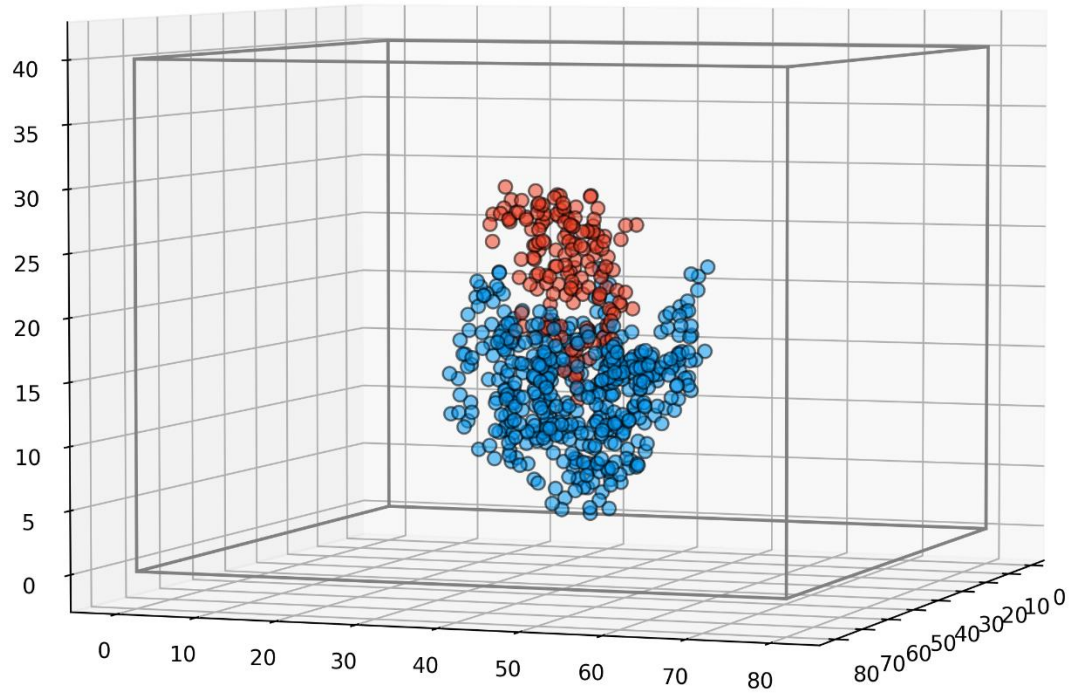


Рисунок 32. Точечное изображение выбранных сложных атомов комплекса 1an1 для дальнейшего анализа.

3.1.3. Выделение признаков из пространственных структур комплексов

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 взаимодействиях, которые вносят существенный вклад в белок-белковые взаимодействия (водородные, ионные, гидрофобные, стэкинг-взаимодействия и т.д.) и в структуру основной цепи (расположение карбонильной группы). Распределение атомов по группам было основано на реализованном ранее

алгоритме анализа белок-лигандных взаимодействий Arpeggio (Jubb et al., 2017). Подробная информация о распределении атомов по группам представлена в Приложении 1.

В исходном варианте атомы одной из групп помещались в каждый канал («one-hot» кодирование на 9 каналов), при этом все атомы одной из взаимодействующих молекул добавлялись в каналы 10 или 11. Такая организация пространства признаков не позволила добиться высокого качество прогнозирования (значение функции потерь 0,7), при этом процесс обучения был достаточно длительным из-за размера массивов (для каждого комплекса 11x41x81x81). Далее были апробированы различные варианты этого подхода, например, вместо 1 в каналах с донорами и акцепторами водородных связей использовалось число возможных одновременно образующихся водородных связей для атомов, в каналах с гидрофобными атомами использовалось значение гидрофобности и т.д. Однако эти изменения способствовали ухудшению обучения (значение функции потерь составляет 0,7–0,8).

Новый подход заключался в изменении структуры каналов и исключении каналов, содержащих все атомы молекул. Для ввода информации о принадлежности атомов к разным молекулам использовалось их кодирование в каждом канале (атомы одной молекулы во всех каналах обозначались как 1, другой как -1). Затем был изменен набор атомов в каждом канале, чтобы сосредоточиться на взаимодействиях, важных для связывания белков. Соответственно, четыре канала включали комбинации доноров (или слабых доноров) водородных связей одной молекулы и акцепторов другой молекулы, два канала содержали комбинации положительно и отрицательно заряженных атомов молекул. Остальные каналы не изменились; они содержали лишь информацию о свойствах атомов, принадлежащих взаимодействующим молекулам (Таблица 5).

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 тензора соответственно).

Таблица 5. 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

3.1.4. Аугментация данных

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. Аугментация данных — это процесс искусственного увеличения объема и разнообразия обучающих данных путем создания новых данных на основе существующих.

Сначала, как упоминалось ранее, в обучающую выборку добавлялись конформации, полученные методом МД для 142 комплексов. Благодаря этому в молекулы вводятся новые позиции атомов при сохранении сродства связывания, что позволяет расширить набор обучающих данных наиболее редкими со сложными для предсказания примерами. Первоначально моделью, обученной только на экспериментальных комплексах, были сделаны предсказания аффинности для полученных конформаций нескольких комплексов. В результате анализа было замечено, что прогнозы различаются для разных конформаций одного комплекса (Рис. 33).

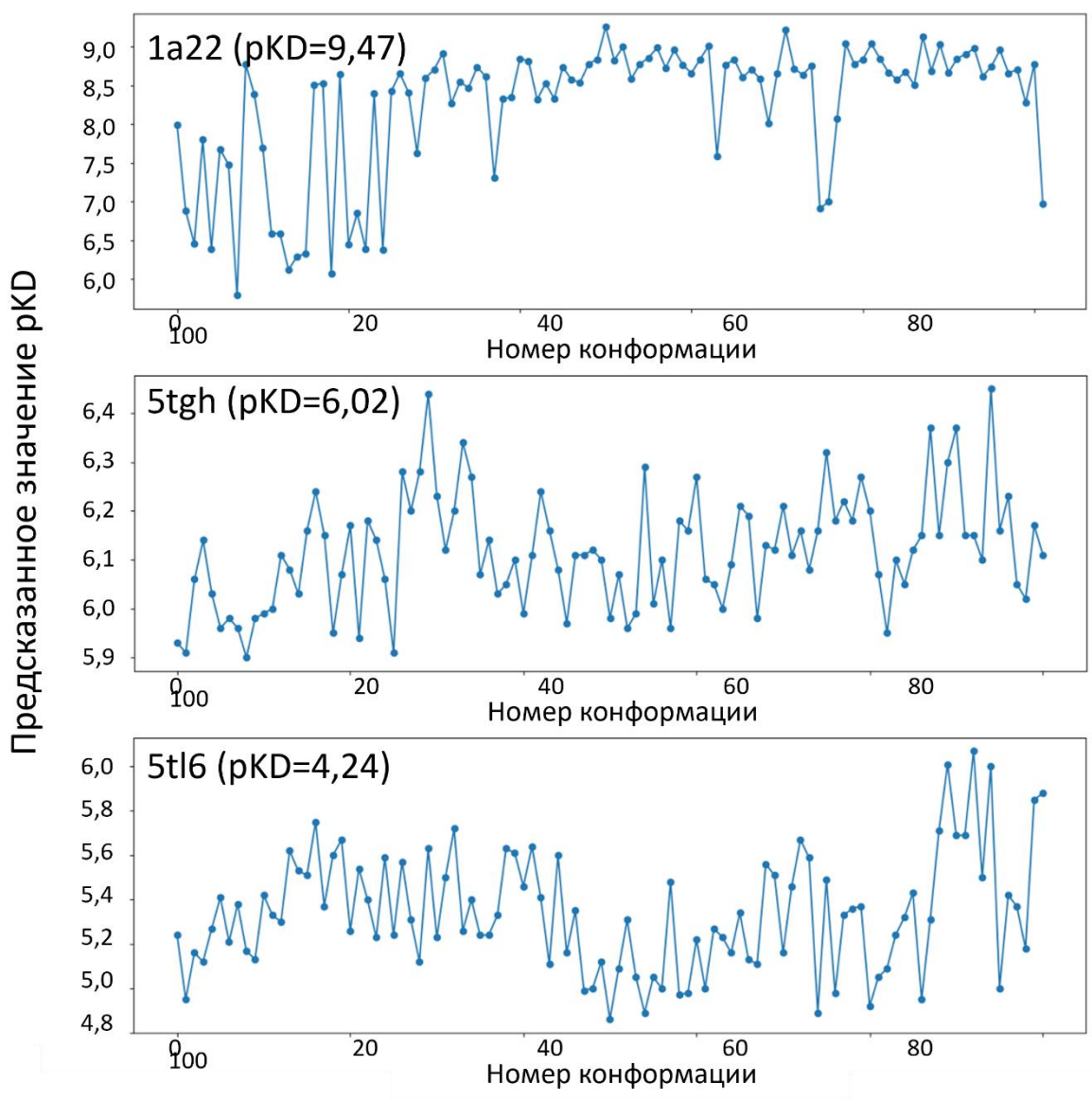


Рисунок 33. Результаты предсказания pK_D для конформаций комплексов, полученных методами МД.

Таким образом обучающая выборка была расширена с помощью использования метода МД, включив в нее информацию о конформационной подвижности белков. В результате использования расширенного набора данных удалось улучшить как общие метрики прогнозирования, так качество предсказания для комплексов с наиболее высоким и низким сродством (Рис. 34).

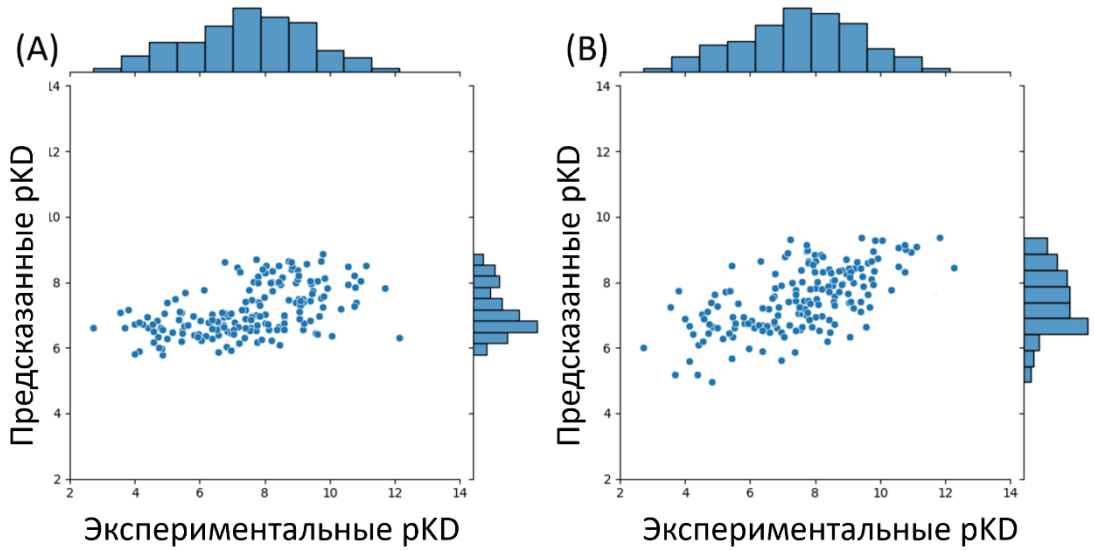


Рисунок 34. Результаты прогнозирования pK_D для валидационного набора. (A) Модель обучена только на экспериментальных комплексах (корреляция Пирсона равна 0,51). (B) Модель обучена на экспериментальных комплексах с добавлением структур, полученных методами МД (корреляция Пирсона 0,61).

Помимо расширения обучающего набора данных, были использованы и дополнительные методы 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.

3.2. Разработка предсказательного алгоритма

В данной работе был разработан предсказательный алгоритм ProBAN (**Pro**tein **B**inding **A**ffinity **N**etwork) на основе глубокой сверточной нейронной сети. Выбор этой архитектуры обусловлен форматом входных данных и их разреженностью.

На первом этапе была обучена нейронная сеть с одним сверточным слоем для извлечения пространственных признаков из структуры с дальнейшей обработкой сформированного одномерного массива полносвязными слоями. Модель с такой архитектурой имела множество параметров обучения, поэтому достаточно быстро переобучалась и давала низкое качество прогнозирования (функция потерь равна 0,9). При дальнейшем увеличении количества сверточных слоев и количества выходных каналов обучение происходило быстрее, а качество прогнозирования на валидационном наборе улучшалось (0,7 для сети с двумя сверточными слоями и 0,53 для трех слоев). Далее был добавлен еще один сверточный слой, который замедлил наступление переобучения и улучшил качество прогнозирования после оптимизации гиперпараметров (функция потерь равна 0,5) (Рис. 35). Дальнейшее увеличение слоев не привело к улучшению обучения, поэтому было решено остановиться на текущей архитектуре.

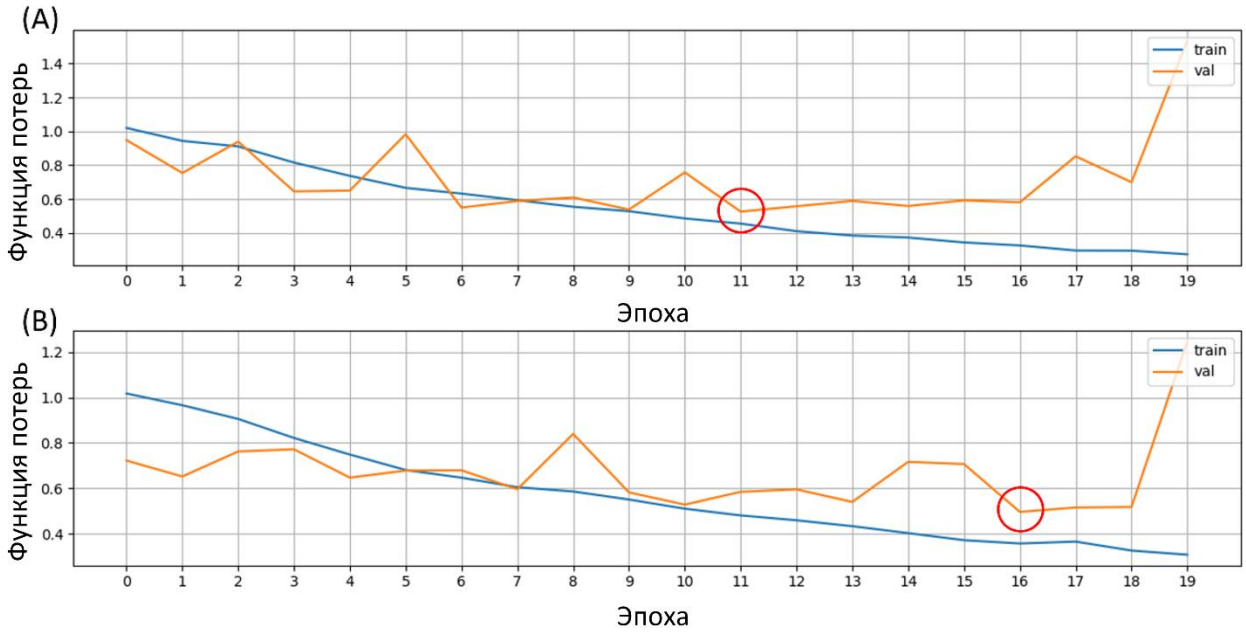


Рисунок 35. Процесс изменения функции потерь (MSELoss) в процессе обучения нейронной сети с двумя вариантами архитектуры: (A) три сверточных слоя, три полносвязных слоя. (B) Четыре сверточных слоя, три полносвязных слоя. Красный кружок отмечает эпоху с лучшим значением функции потерь на валидационном наборе.

Таким образом, в сети последовательно выполняются 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 выводит непосредственно стандартизированное значение pK_D, и поэтому после него нет функции активации (остальные слои содержат нелинейную функцию активации ReLU). Для обучения использовался оптимизатор AdamW, ввиду возможности добавления к нему регуляризации L2. Для расчета ошибки использовалась функция потерь MSELoss, подходящая для решения задачи регрессии. В качестве метрик качества прогнозирования использовались корреляция Пирсона и RMSE. 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 36.

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 (параметр, отражающий значение регуляризации). В результате сохранялась лучшая модель (значение корреляции на валидационной выборке 0,57, значение функции потерь 0,5) 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).

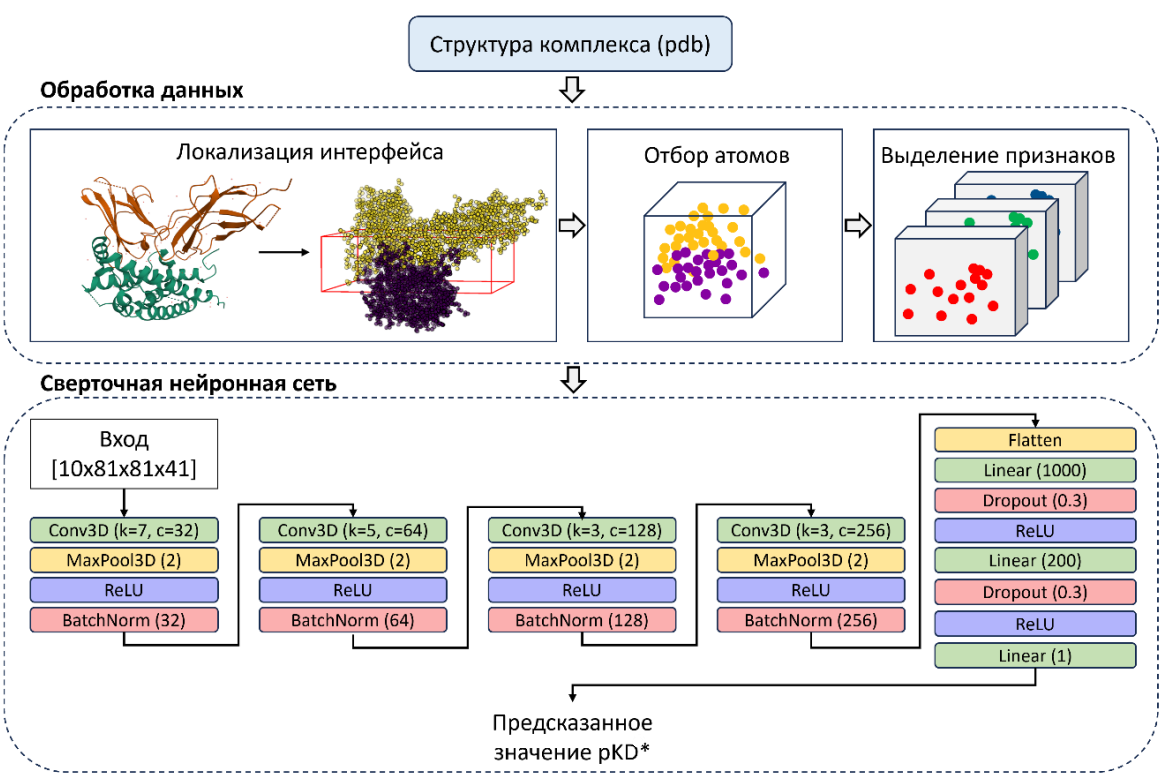


Рисунок 36. Полная схема обработки комплекса и прогнозирования его pK_D. Во-первых, интерфейс связывания локализуется внутри ограничивающей ячейки, и дальше анализируются атомы, оказавшиеся внутри ячейки. На следующем этапе происходит отбор атомов, важных для связывания. Затем к полученной трехмерной 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.

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 (Dropout = 0,3, batch size = 32 и другие). Также было замечено, что добавление в обучающую выборку большего количества

конформаций (более 10 для каждого комплекса), полученных с помощью молекулярно-динамического моделирования, привело к более быстрому началу переобучения. Этот факт может быть связан с тем, что в данном случае нейросеть скорректировала параметры для лучшего прогнозирования именно для этих примеров, при этом упустив обобщение и поиск закономерностей в комплексах, для которых не были получены траектории. Однако именно добавление к экспериментальным данным определенное количество новых смоделированных конформаций комплексов позволило повысить качество прогнозирования на тестовых данных.

3.3. 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. 37A), что связано с их недостаточной представленностью в обучающем наборе данных. Для внешнего тестового набора удалось получить значение корреляции Пирсона 0,55 (p-значение = 5e-08) и RMSE = 1,67 (Рис. 37B). 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 тесте 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. Этот 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. 37C). Этот подход ранее использовался для алгоритма Pafnucy (Stepniewska-Dziubinska et al., 2018).

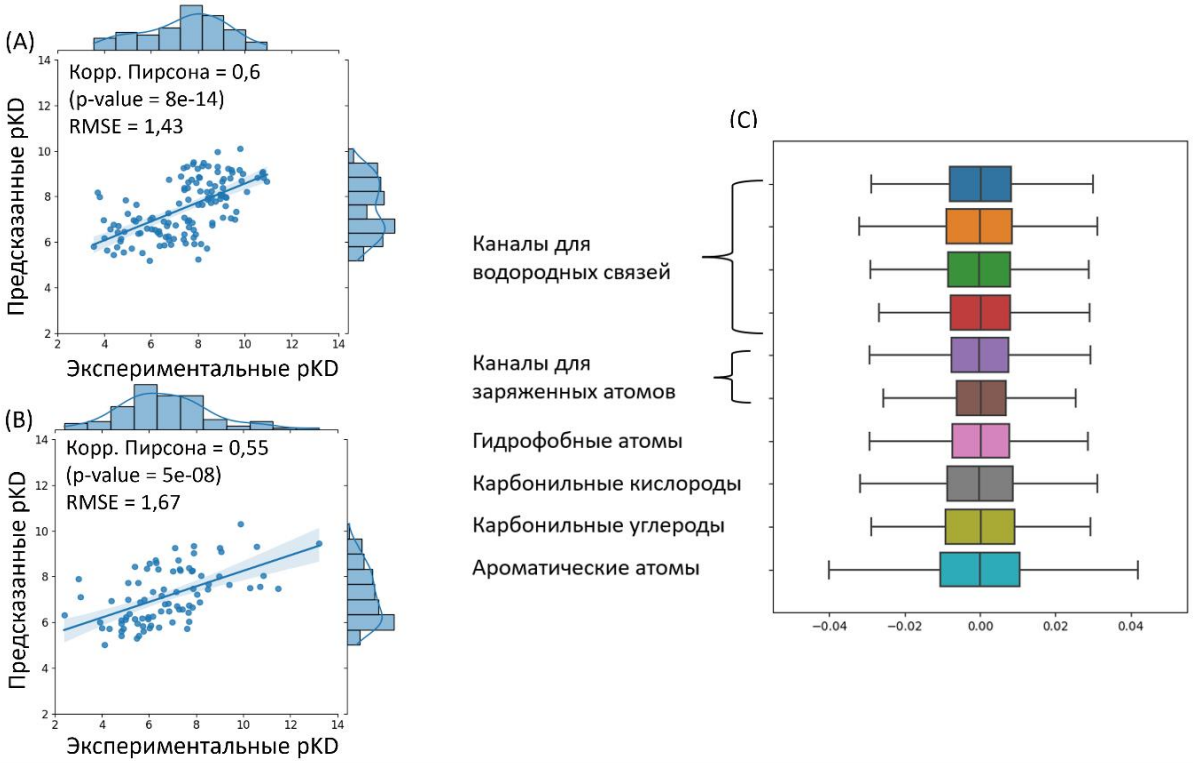


Рисунок 37. Результат тестирования обученной нейронной сети и анализа значимости признаков. (A) Диаграмма рассеяния комплексов из внутреннего тестового набора (тест 1). Ось X содержит истинные значения pK_D, а ось Y содержит прогнозируемые значения. (B) Диаграмма рассеяния для внешнего набора тестов (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 варианты 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 фокусирует большую часть своих предсказаний исходя из этих особенностей, что согласуется с известными закономерностями в связывании молекул. Так, ароматические соединения образуют стэкинг-взаимодействия, которые могут усиливать связывание между белками (Toulmé, 1985; McGaughey et al., 1998). Ионные связи также имеют решающее значение для формирования пространственной организации белков (Bošnjak et al., 2014; Batoulis et al., 2016; Furutani, 2018), что делает этот тип контакта одним из ключевых в межмолекулярных взаимодействиях. Атомы карбонильной группы участвуют в образовании карбонил-карбонильных контактов в белках, которые играют важную роль в формировании вторичной и третичной структуры белков (Esposito et al., 2000; Rahim et al., 2017; Sahariah et al., 2018), таким образом, можно дополнительно извлечь информацию о конформации молекулы. Описанные закономерности могут объяснить большой диапазон весов для этих каналов.

Чтобы сравнить ProBAN с другими прогностическими моделями, было рассчитано значение ΔG на основе предсказанных констант диссоциации. Рассчитанные метрики для обоих тестовых наборов представлены в Таблице 6. Результаты прогнозирования для внутреннего 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 стабильность разработанной модели и возможность анализа белок-

белковых и белок-пептидных комплексов, интерфейс связывания которых может быть локализован в пределах ограничительной 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). Этот тестовый набор также использовался для оценки производительности разработанного в данной работе метода (Таблица 6, Тест 2) в сравнении с другими доступными в настоящее время инструментами по предсказанию аффинности связывания в белок-белковых комплексах. Результаты прогнозирования этих алгоритмов были получены из материалов публикации, описывающей работу PPI-Affinity (Romero-Molina et al., 2022).

В результате оценки эффективность разработанной модели на данном наборе данных (тест 2) 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 6. 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

Значение метрик качества (корреляция Пирсона и MAE) PPI-Affinity уступает разработанной в диссертационной работе модели. Тем не менее, этот алгоритм показал высокие стабильные результаты, в целом превосходящие результаты, полученные другими методами на проанализированном наборе тестов. Другие предикторы (PRODIGY, DFIRE, CP_PIE, ILAND) показывают значительное снижение своей производительности по сравнению с результатами в тестовых наборах, первоначально использованных в исходных исследованиях. Такое резкое снижение качества прогнозирования предполагает переобучение по сравнению с предыдущим набором контрольных показателей.

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 (Рис. 38A) (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 (Рис. 38B) (R

= 0,47 и MAE = 2 ккал/моль) уступают показателям, полученным на основе других наборов данных.

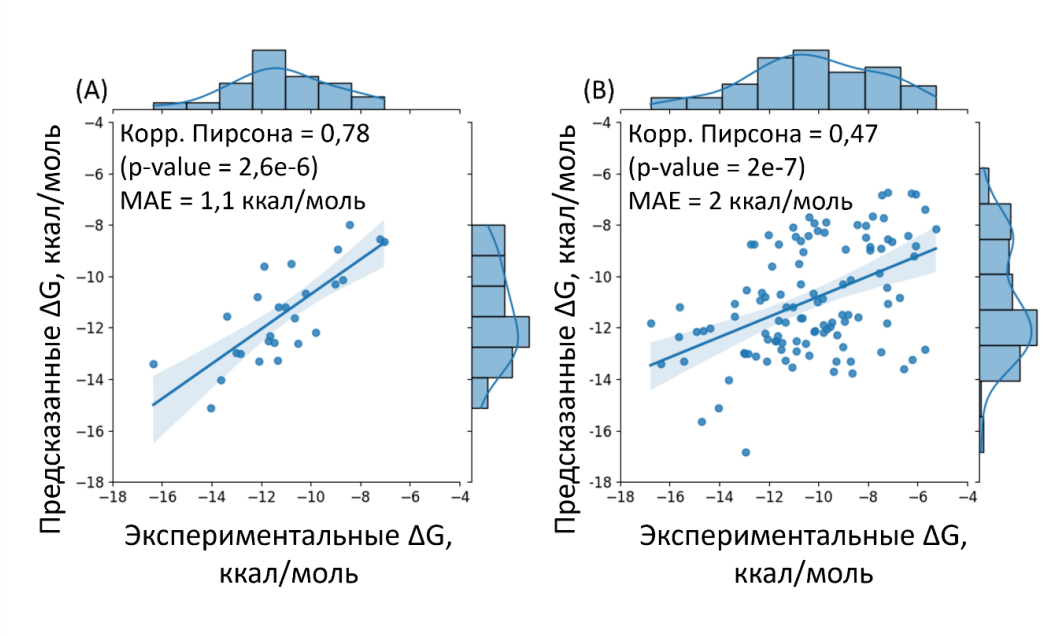


Рисунок 38. Результат тестирования ProBAN на комплексах дикого типа из SKEMPI v2.0. (A) Диаграмма рассеяния для дополнительного набора тестов (26 комплексов); (B) Диаграмма рассеяния комплексов из полного набора данных (112 комплексов).

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.

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

Помимо апробации алгоритма на разнородных тестовых наборах данных, производилось его тестирование 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.

3.4.1. Анализ интерфейса взаимодействия

Рассматриваемые в настоящей работе комплексы RBD-ACE2 образованы рецептор-связывающим доменом (receptor binding domain, RDB) S-белка коронавирусов SARS-CoV и SARS-CoV-2 и молекулой ангиотензин-превращающего фермента 2 (англ. angiotensin converting enzyme 2, ACE2) (Рис. 39A). В непосредственном контакте RBD-ACE2 принимают участие 26 остатков со стороны RBD (позиции 403, 417, 439, 446, 449, 453, 455, 456, 458, 475 – 477, 484 – 487, 489, 490, 493, 496, 498, 500 – 503, 505) и 22 остатка со стороны ACE2 (позиции 19, 24, 27, 28, 30, 31, 34, 35, 37, 38, 41, 42, 79, 82, 83, 329, 330, 352 – 355, 357). Взаимная ориентация взаимодействующих белков и конформация интерфейса весьма консервативны (Рис. 39B): 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. 39C). Будем для определённости называть эти макрообласти большой (Рис. 39C, слева) и малой (Рис. 39C, 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.

Межмолекулярные взаимодействия в комплексах RBD-ACE2 подробно освещаются практически в каждой работе, посвящённой той или иной

расшифрованной структуре (Han et al., 2021; Su et al., 2022). Считается, что наиболее значимыми для связывания являются три области (hot spot), первая из которых соответствуют малой макрообласти, а две другие большой макрообласти. Итак, первая область локализована вокруг остатка Lys353 ACE2 и характеризуется взаимодействиями, образованными остатками Lys353, Asp38, Tyr41, Gln42, Leu45 и Asn330 со стороны ACE2 и остатками Thr500, Asn501, Gln498 и Tyr505 со стороны RBD. Мутации в указанных позициях значимо влияют на аффинность связывания. Так, например, выявлено, что замена аспарагина на тирозин в 501 позиции RBD у альфа штамма SARS-CoV-2 значительно увеличивает аффинность (Lawad et al., 2021; Salleh et al., 2021).

Вторая область ассоциирована с остатком Lys31 ACE2 и характеризуется контактами, образованными остатками Leu455, Glu484, Lys417 со стороны RBD и Asp30 и Lys31 со стороны ACE2 (Wang et al., 2020). Также в мутантах дикого штамма выявлен остаток Gln493, который усиливает взаимодействие в данной области.

Третья область также расположена вблизи N-конца ACE2, с которым взаимодействует короткая подвижная петля RBD (Рис. 39A, слева). Эта петля охватывает спираль ACE2 (остатки 21–52) с почти противоположной стороны от основного интерфейса, что обеспечивает более обширную область взаимодействий (Geng et al., 2022) и стабильное относительное расположение RBD и ACE2. Критически важными для взаимодействия в этой области являются следующие остатки RBD: Ala475, Gly476 и Phe486, первые два взаимодействуют с Ser19, а третий с Met82 и Leu79 (Nelson-Sathi et al., 2022) (Рис. 39D). При этом ранее было показано, что мутации в позиции Gly476, а также Ala475 негативно сказываются на аффинности связывания (Yang et al., 2021).

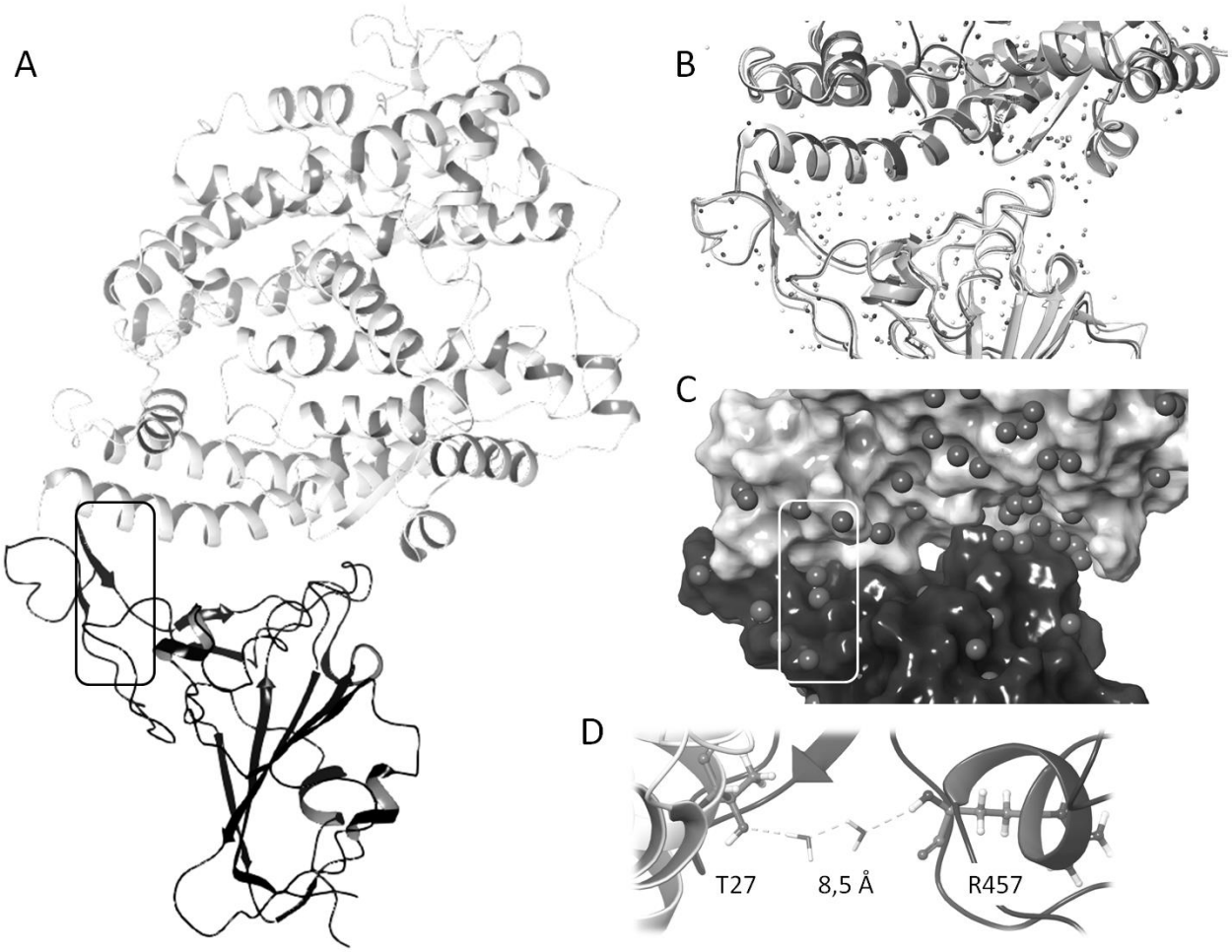


Рисунок. 39. A. Общий вид комплекса ACE2 (показан светлым) и-RBD (показан тёмным) (pdb код 6lzg) в ленточном представлении. Рамка соответствует области расположения остатков T27 (ACE2) и R457 (RBD). B. Совмещение интерфейсов взаимодействия ряда структур комплекса ACE2-RBD: 6lzg (показана самым тёмным), 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 заимодействие 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. 39C). В естественных условиях она, очевидно, заполнена молекулами воды и ионами, однако в известных кристаллографических структурах молекулы кристаллизационной воды в этой полости отсутствуют. Предположительно, ключевой причиной этого явления является высокая подвижность молекул воды в этой области, вызванная несоответствием гидрофобных свойств поверхностей молекул ACE2 и RBD в этой области пространства. Рассмотрение свойств поверхностей показывает, что поверхность ACE2 между остатками Lys353 и Lys31 обладает гидрофильными свойствами, в то время как соответствующая ей поверхность RBD между пятен контактов этих остатков обладает гидрофобными свойствами (результаты не приведены). Это наблюдение хорошо соотносится с результатами анализа структур комплексов «белок-белок», полученных с высоким разрешением, который выявил более стабильное состояние молекул кристаллизационной воды вблизи полярных незаряженных остатков белка по сравнению с заряженными или неполярными остатками (Kastritis et al., 2014).

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. 39C). Интересно отметить, что

множества молекул воды, находящиеся в этих структурах, пересекаются не полностью, что, с одной стороны, позволяет выявить наиболее консервативные сайты связывания воды, а с другой стороны, гипотетически, позволяет создать молекулярную модель интерфейса, содержащую в себе все возможные молекулы воды, включая подвижные молекулы воды, расположенные в вышеупомянутой полости. Однако создание такой модели лежит за рамками данной работы.

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. Собственно выявление молекул воды выполняли визуально. Итоговая информация представлена в Таблице 7. Видно, что 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. 39D).

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. Эти же методы должны быть использованы и для учёта тепловых колебаний молекул как белков, так и воды. Далее, даже будучи выявленной, такая цепочка водородных связей сложна в интерпретации с точки зрения собственной энергии и вклада в энергию взаимодействия белков. Наконец, зачастую ставится задача быстрой оценки энергии взаимодействия в комплексе (например, в задачах белок-белкового докинга), что вообще не позволяет рассматривать молекулы воды в явном виде.

Таблица 7. Наблюдаемые в структурах высокого разрешения цепочки водородных связей с участием молекул воды. Остаток ACE2 – молекулы воды – остаток RBD. В скобках дано расстояние между соответствующими атомами указанных остатков, Å

Код структуры, (разрешение, Å)	Непрямое взаимодействие посредством молекул воды		
	одной	двух	трех
6lzg (2,5)	K31-w-E484 (4,2) A386-w-Y505 (5,4)	T27-w-w-R457 (8,5) Q325-w-w-P499 (8,3) G354-w-w-V503 (5,4) G354-w-w-G504 (6,9)	G354-w-w-w-D405 (9,0)
7ekh (2,4)	T27-w-A475 (5,9) Y83-w-N487 (4,4)	T324-w-w-T500 (7,4) G354-w-w-G504 (6,9)	-
7lo4 (2,5)	-	G354-w-w-V503 (5,1) G326-w-w-V503 (7,0) Q325-w-w-V503 (7,3)	-
8df5 (2,7)	T27-w-A475 (5,6) D30-w-N417 (5,6) H34-w-N417 (4,2)	E35-w-w-F490 (7,6) G354-w-w-G504 (6,8)	E35-w-w-w-G485 (10,5)

3.4.2. Оценка связывания в комплексах

Рассмотрение комплексов RBD-ACE2 с известными пространственной структурой и значением K_D выявило 48 комплексов, полный набор представлен в Приложении 2.

Большая часть этих комплексов имеют значение pK_D от 7 до 9 (Рис. 40), 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).

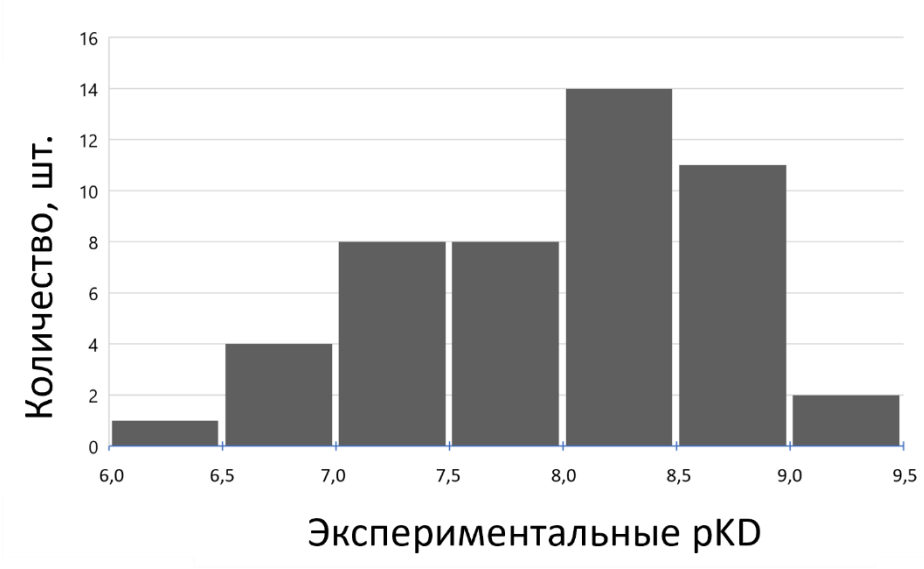


Рисунок. 40. Распределение комплексов RBD-ACE2 из исследуемого набора по значениям pK_D.

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.41). 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.

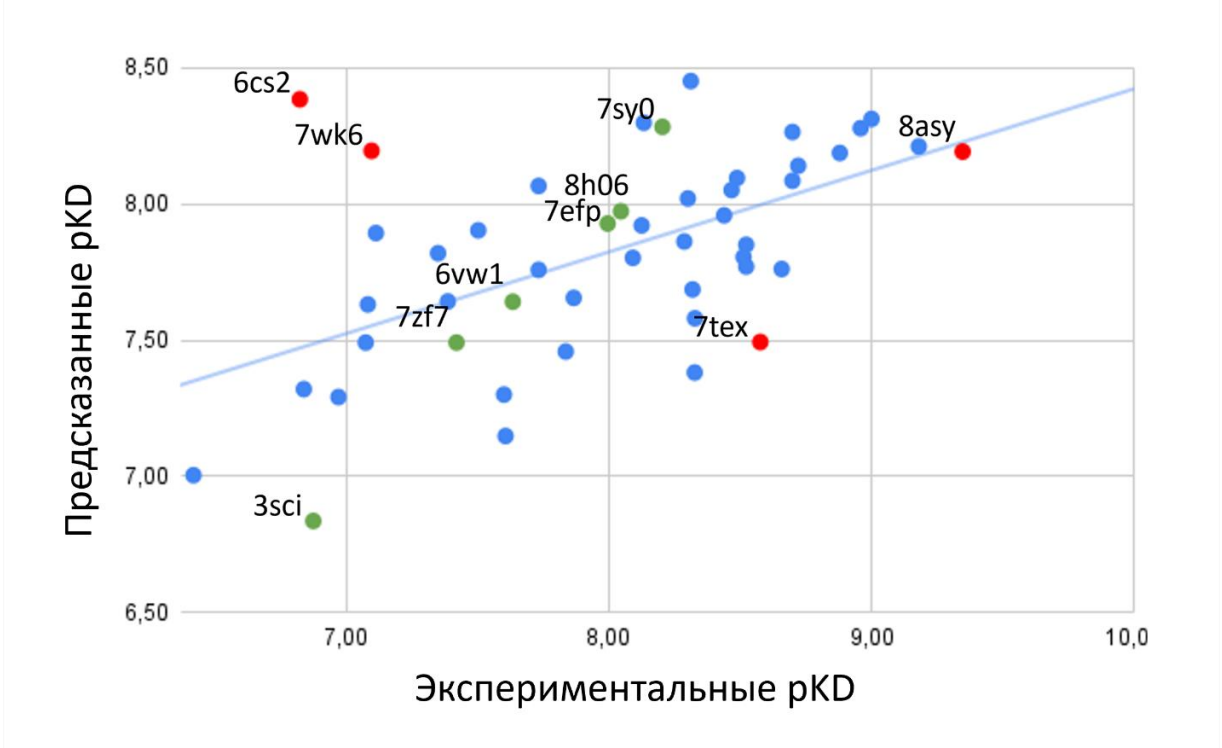


Рисунок 41. Результаты предсказаний pK_D для комплексов RBD-ACE2 алгоритмом ProBAN. (красным выделены предсказания для комплексов с абсолютной ошибкой больше 1, зеленым - с ошибкой меньше 0,1, синим – остальные). В качестве ярлыков добавлены pdb коды комплексов.

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 8.

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 проводилось сравнение с ΔG полученными с использованием данных алгоритмов (Рис. 42). Рассчитанные значения метрик качества для разных алгоритмов представлены в Таблице 8.

Таблица 8. 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 pK_D
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

Видно, что ProBAN оказывается 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. Данный вывод согласуется с более ранними исследованиями по предсказанию аффинности связывания в комплексах RBD-ACE2 (Ozden et al., 2024). Остальные алгоритмы (DFIRE, ROSETTADOCK) оказались менее успешными в предсказании свободной энергии Гиббса для изучаемых белок-белковых комплексов.

Также стоит заметить, что для одного из комплексов (7u0n) DFIRE предсказал значение ΔG = 137,6 ккал/моль, что явно является выбросом и свидетельствует о возможной нестабильности алгоритма. Для более объективной оценки работы данного метода метрики качества были пересчитаны для набора из 47 комплексов без 7u0n, что позволило их улучшить, но статистически значимой корреляции также не получилось достичь (Таблица 8), как и для ROSETTADOCK.

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. В дальнейшем развитие методов, основанных на машинном обучении, смогут способствовать более точному пониманию механизмов взаимодействия белков и разработке эффективных терапевтических стратегий лечения различных заболеваний.

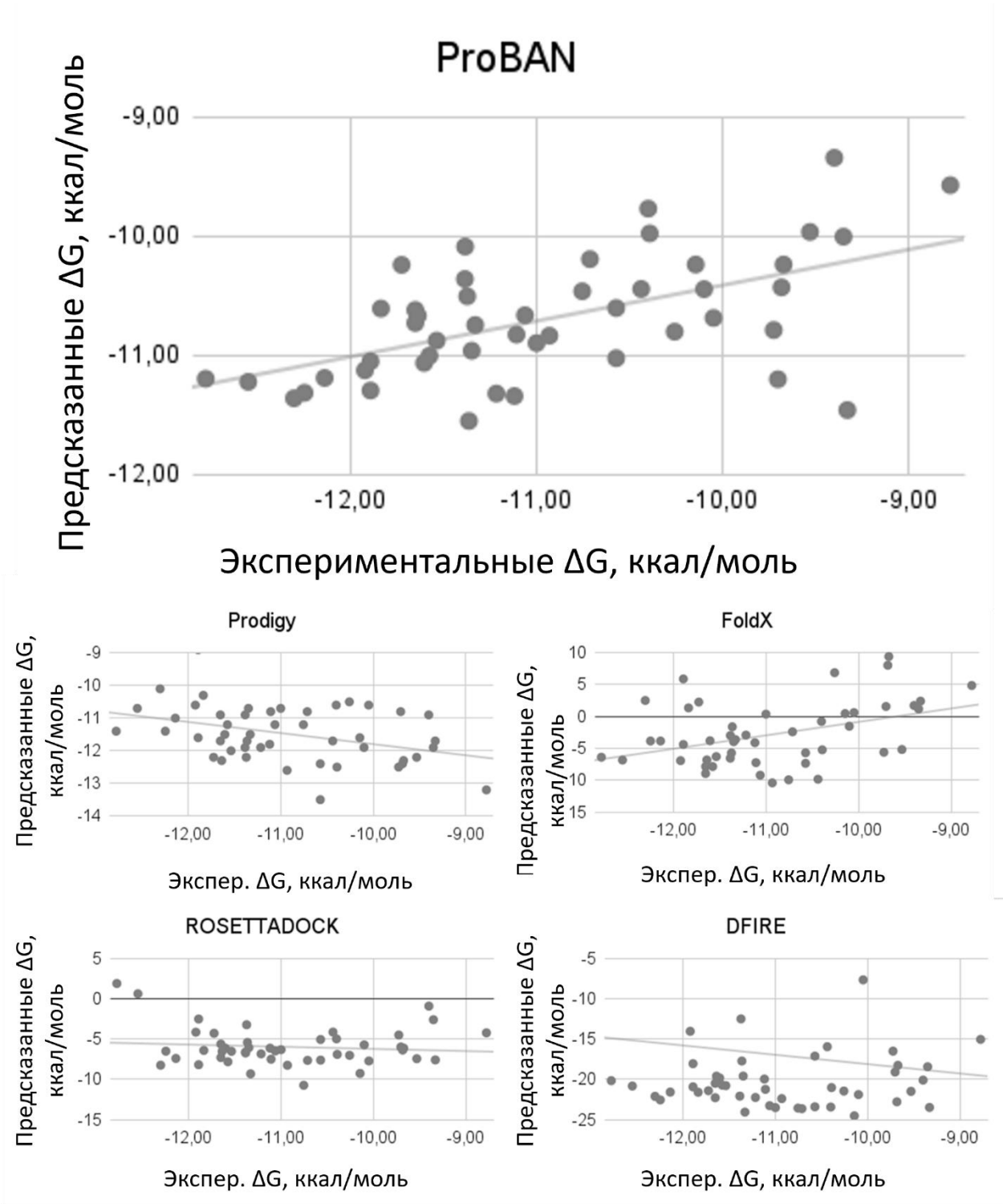


Рисунок 42. Результаты предсказаний **Δ**G для комплексов RBD-ACE2 моделей: ProBAN, Prodigy, FoldX, DFIRE (было исключено предсказание для комплекса 7u0n) и ROSETTADOCK.

3.5. Анализ стабильности комплексов, образованных разными вариантами гистонов

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.

Гистоны представляют собой группу структурных эукариотических белков, которые, в составе нуклеосомы играют ключевую роль в упаковке (Pepenella et al., 2014) и регуляции ДНК в клеточном ядре, в частности, в регуляции транскрипции. (Chang et al., 2022; Kulaeva et al., 2009). 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). Все четыре семейства основных гистонов имеют высокий положительный заряд, консервативный С-концевой домен гистоновой складки и уникальные N-концевые хвосты (Kamakaka et al., 2005). Домены гистоновой складки тесно взаимодействуют с другими основными гистонами внутри нуклеосомы, а также с нуклеосомной ДНК (Рис. 43).

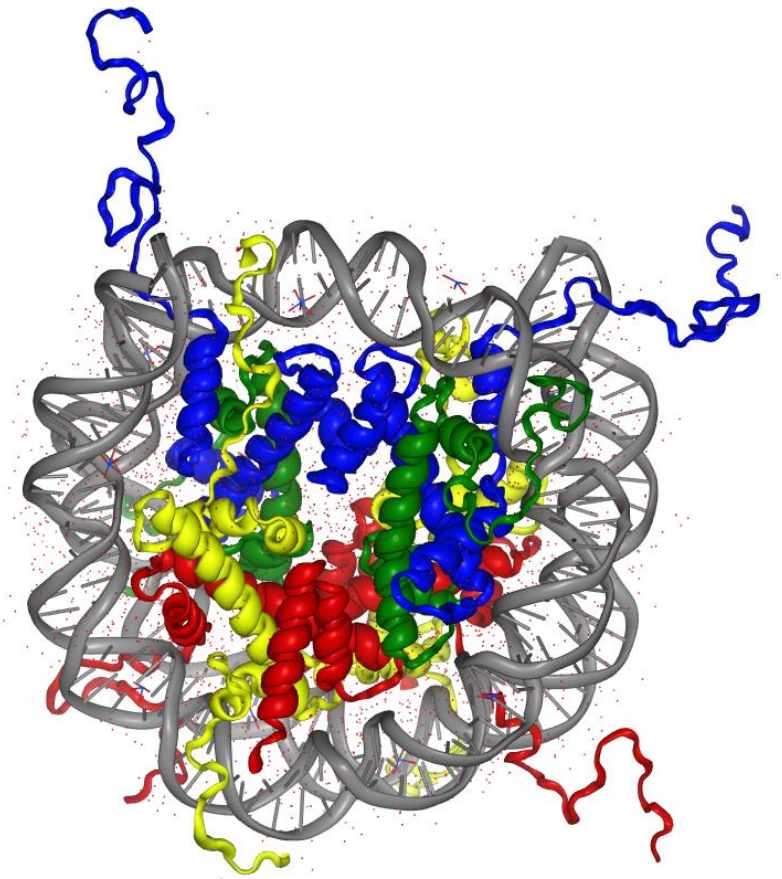


Рисунок 43. Ленточное отображение структуры нуклеосомы человека (PDB id: 1kx5). Желтым обозначены молекулы гистона H2A, красным – H2B, синим – H3, зеленым – H4, серым – фрагмент молекулы ДНК.

Динамические процессы связывания гистонов с ДНК непосредственно влияют на доступность генов для транскрипции (Shi et al., 2024). 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.

Нуклеосомы могут подвергаться не только большому разнообразию ковалентных посттрансляционных модификаций, в основном происходящих в N-концевых участках гистонов, но и встраиванию вариантов гистонов, соответствующих гистонам H3, H2A, H2B, но не H4, для которого на данный момент выявлена только одна форма. Канонические формы гистонов преобладают в нуклеосомах и синтезируются и встраиваются в зависимости от репликации. Дополнительное разнообразие обеспечивается встраиванием в

хроматин вариантов гистонов (Talbert et al., 2021). Варианты гистонов были обнаружены на основе различий в их аминокислотной последовательности, которые могут варьироваться от нескольких аминокислотных замен до крупных доменов. Эти варианты демонстрируют различные механизмы регуляции их экспрессии и накопления, которые потенциально могут придавать нуклеосомам особые свойства (Talbert et al., 2021). Так, замещающие варианты гистонов могут напрямую влиять на структуру и стабильность нуклеосом (Tachiwana et al., 2011). 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). Так, изменения в стабильности гистоновых комплексов могут быть связаны с развитием различных заболеваний, включая рак. Понимание этих изменений может помочь в разработке новых терапевтических стратегий, направленных на восстановление нормальной регуляции генов.

3.5.1. Тестирование алгоритма на комплексах, образованных гистонами с другими белками

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 с известными значениями аффинности связывания. Результаты предсказания для отобранных комплексов представлены в Таблице 9.

Таблица 9. Результат предсказания свободной энергии Гиббса для комплексов, образованных гистонами и молекулами других белков.

Код PDB	Гистоны	Связываемые белки	Экспер. Δ G, ккал/моль	Предск. Δ G, ккал/моль
5chl	H2A.Z	YL1	-9,6	-8,3
5fug	H2A.Z, H2B	YL1	-11,4	-9,5
5vey	H2B 1-J, H2A 1-B/E	RNF169 (653-708)	-8,9	-8,5
6kbb	H2A.Z-H2B	DEF/Y мотив Swc5	-10,7	-10,2
6n1z	H2A H2B 1.1	Импортин-9	-10,1	-10,7
7bp6	H2A.6 H2B.1	AtNRP1-CTAD	-7,8	-8,3
7c7x	H2A.6 H2B.1	AtNRP1	-9,3	-8,8
6ae8	H2A.Z, H2B	шаперон Chz1	-8,2	-9,7
7wlp	H2A-H2B	Белок вируса Эпштейна-Барра BKRF4	-8,1	-8,6
5wvo	H3	DNMT1 RFTS	-10,6	-9,4
6s1r	H4	Mis16	-10,3	-9,2
7ciz	H3.3-H4	DNAJC9	-9,9	-9,6

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.

3.5.2. Оценка влияния разных вариантов гистонов на стабильность комплексов

На первом этапе анализа стабильности комплексов гистонов производилась предобработка структурных файлов с выделением взаимодействующих цепей, описанная ранее. Также в связи со сложной неровной поверхностью интерфейса взаимодействия и его размерами оценивалась возможность локализации всех причастных к связыванию между молекулами атомов. Результат подбора расположения ограничивающей ячейки для комплексов с каноническими формами гистонов (pdb: 1kx5) представлен на Рис. 43. Как можно заметить, размеры используемой ячейки достаточны как для анализа взаимодействующих поверхностей между отдельными гистонами (Рис. 43A), так и между димером и тетрамером (Рис. 43B). Таким образом были обработаны комплексы всех анализируемых вариантов гистонов и отправлены на вход нейронной сети.

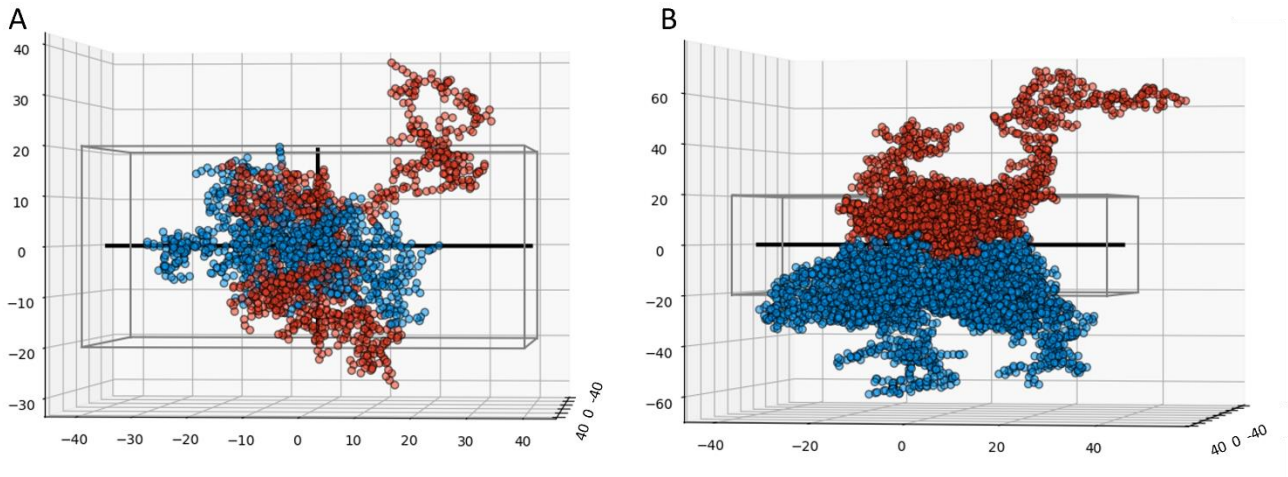


Рисунок 44. Локализация интерфейса связывания между гистонами в нуклеосоме (pdb: 1kx5). A – Между H2A (красный) и H2B (синий); B – между димером H2A-H2B (красный) и тетрамером $(H3\text{-}H4)_2$ (синий).

Для анализа были отобраны несколько вариантов гистонов H2A, H2B и H3. К примеру, H2A.Z, вариант гистона H2A, необходимый для приспособленности 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.

Результаты предсказания энергии связывания между гистоновыми молекулами каноничных форм и альтернативных вариантов представлены в Таблице 10.

Таблица 10. Результат предсказания энергии связывания между различными вариантами гистонов.

PDB_ID	Вариант	ΔG, ккал/моль
H2A-H2B		
1kx5	Каноничная форма	-13,0
1f66	H2A.Z	-13,3
6kvd	H2A.J	-10,0
5gt0	TSH2A.1	-10,5
5gt3	TSH2B.1	-11,9
5gsu	TSH2A.1, TSH2B.1	-13,5
H3-H4		
1kx5	Каноничная форма	-12,9
5x7x	H3.3	-12,7
5gxq	H3.6	-11,3
H2A-H2B с H3-H4		
1kx5	Каноничная форма	-11,7
1f66	H2A.Z	-12,0
6kvd	H2A.J	-11,6
5gt0	TSH2A.1	-11,6
5gt3	TSH2B.1	-11,1
5gsu	TSH2A.1, TSH2B.1	-11,2

Как видно из предсказанных значений ΔG контакты между H2A.Z и H2B оказались немного более стабильными, чем для каноничной формы H2A, что согласуется с ранее приведенными исследованиями по изучению термостабильности и динамики димеров гистонов с каноничной формой H2A и измененной (Dai et al., 2021). Ранее было показано, что замены аминокислот в H2A.Z значительно стабилизируют α-спиральную конформацию, что, вероятно, помогает формировать более стабильные контакты с ДНК. (Kniazeva et al., 2022), что также могло сказаться и на взаимодействиях с H2B.

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, 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, чем каноничный вариант, однако димер TSH2A.1-TSH2B.1 является даже более

стабильным, 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. 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 передаче активных состояний хроматина. В результате оценки взаимодействия между H3 и H4 было выявлено, что вариант H3.3 имеет небольшое снижение аффинности связывания относительно каноничного варианта, при этом для варианта H3.6 это снижение являетя гораздо более значимым. Это может быть связано с тем, что в нуклеосоме с H3.6 специфический для H3.6 остаток Val62 образует гидрофобный контакт с родственной молекулой H4, но площадь контакта меньше, чем у соответствующего остатка Ile62 в H3.3 (Thakar et al., 2009). Так же по литературным данным известо что нуклеосома H3.6 менее термически стабильна по сравнению с нуклеосомой H3.3, что также 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 факмакологически 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.

БЛАГОДАРНОСТИ

Автор выражает благодарность своему научному руководителю, Новоселецкому Валерию Николаевичу за направление исследования, ценные советы и наставления. Автор выражает благодарность некоммерческому фонду ИНТЕЛЛЕКТ и курсу для молодых ученых «Нейронные сети и их применение в научных исследованиях» за возможность углубить знания, необходимые для развития в научной деятельности и за поддержку данного исследования. Автор выражает благодарность за возможность расширения области применимости данного исследования Шайтану Алексею Константиновичу и поддержку исследования в рамках гранта на проведение крупных научных проектов по приоритетным направлениям научно-технологического развития No 075-15-2024-539 от 24.04.2024 по теме: «Эпигенетика как основа для разработки новых стратегий лечения болезней». Автор выражает благодарность Чернухину Артему Валерьевичу за поддержку на протяжении всей работы над диссертацией и предоставленные вычислительные ресурсы. Автор благодарит коллектив лаборатории молекулярного моделирования кафедры биоинженерии биологического факультета МГУ за добрую рабочую атмосферу, взаимовыручку и моральную поддержку. Также автор благодарен своей семье и близким людям за понимание, терпение и вдохновение.

СПИСОК ЛИТЕРАТУРЫ

1. Abagyan R.A., Totrov M.M. Biased probability Monte Carlo conformational searches and electrostatic calculations for peptides and proteins // J. Mol. Biol. 1994. V. 235. P. 983–1002.
2. Abbasi, W.A., Yaseen, A., Hassan, F.U. et al. ISLAND: in-silico proteins binding affinity prediction using sequence information. // BioData Mining. 2020. V. 13.
3. Abbott D.W., Ivanova V.S., Wang X., Bonner W.M., Ausió J. Characterization of the stability and folding of H2A.Z chromatin particles: implications for transcriptional activation // The Journal of biological chemistry. 2001. V. 276. N. 45. P. 41945–41949.
4. Ahn H., Calderon B.M., Fan X., Gao Y., Horgan N.L., Jiang N., Blohm D.S., Hossain J., Rayyan N.W.K., Osman S.H., Lin X., Currier M., Steel J., Wentworth D.E., Zhou B., Liang B. Structural basis of the American mink ACE2 binding by Y453F trimeric spike glycoproteins of SARS-CoV-2 // Journal of medical virology. 2023. V. 95. N. 10.
5. Alford R.F., Leaver-Fay A., Jeliazkov J.R., O'Meara M.J., DiMaio F.P., Park H., Shapovalov M.V., Renfrew P.D., Mulligan V.K., Kappel K., Labonte J.W., Pacella M.S., Bonneau R., Bradley P., Dunbrack R.L. Jr., Das R., Baker D., Kuhlman B., Kortemme T., Gray J.J. The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design // J Chem Theory Comput. 2017. V. 13. N. 6. P. 3031-3048.
6. Altman N.S. An introduction to kernel and nearest-neighbor nonparametric regression // The American Statistician. 1992. V. 46. N. 3. P. 175–185.
7. Archakov A. I., Govorun V. M., Dubanov A. V., Ivanov Y. D., Veselovsky A. V., Lewi P., Janssen P. Protein-protein interactions as a target for drugs in proteomics // Proteomics. 2003. V. 3. N. 4. P. 380–391.
8. Arora I., Saha A. Comparison of Back Propagation Training Algorithms for Software Defect Prediction // 2nd International Conference on Contemporary Computing and Informatics (IC3I). 2016. P. 51–58.
9. Asim M.N., Ibrahim M.A., Malik M.I., Dengel A., Ahmed S. ADH-PPI: An attention-based deep hybrid model for protein-protein interaction prediction. // iScience. 2022. V. 25.
10. Bashir Q., Scanu S., and Ubbink, M. Dynamics in electron transfer protein complexes // The FEBS journal. 2011. V. 278. N. 9. P. 1391–1400.

11. Batoulis H., Schmidt T., Weber P., et al. Concentration Dependent Ion-Protein Interaction Patterns Underlying Protein Oligomerization Behaviours // Sci Rep. 2016. V. 6.
12. Berman H.M., Westbrook J., Feng Z., Gilliland G., Bhat T.N., Weissig H., Shindyalov I.N., Bourne P.E., The Protein Data Bank // Nucleic Acids Research. 2000. V. 28. P. 235-242.
13. Berryhill C.A., Doud E.H., Hanquier J.N., Smith-Kinnaman W.R., McCourry D.L., Mosley A.L., Cornett E.M. Protein Thermal Stability Changes Induced by the Global Methylation Inhibitor 3-Deazaneplanocin A (DZNep) // Biomolecules. 2024. V. 14. N. 7. P. 817.
14. Bhati A.P., Wan S., Wright D.W., Coveney P.V. Rapid, accurate, precise, and reliable relative free energy prediction using ensemble based thermodynamic integration // J Chem Theory Comput. 2017. V. 13. P. 210–222.
15. Bjorck J., Gomes C., Selman B. Understanding Batch Normalization // 32nd Conference on Neural Information Processing Systems. 2018.
16. Boike L., Henning N. J., Nomura, D. K. Advances in covalent drug discovery // Nature reviews. Drug discovery. 2022. V. 21. N. 12. P. 881–898.
17. Bošnjak I. Occurrence of protein disulfide bonds in different domains of life: a comparison of proteins from the Protein Data Bank // Protein Engineering, Design & Selection. 2014. V. 27. N. 3. P. 65–72.
18. Bryant P., Pozzati G., Elofsson A. Improved prediction of protein-protein interactions using AlphaFold2 // Nat Commun. 2022. V. 13. N. 1.
19. Caporale A., Adorinni S., Lamba D., Saviano, M. Peptide-Protein Interactions: From Drug Design to Supramolecular Biomaterials // Molecules (Basel, Switzerland). 2021. V. 26. N.5.
20. Chang H.W., Feofanov A.V., Lyubitelev A.V., Armeev G.A., Kotova E.Y., Hsieh F.K., Kirpichnikov M.P., Shaytan A.K., Studitsky V.M. N-Terminal Tails of Histones H2A and H2B Differentially Affect Transcription by RNA Polymerase II In Vitro // Cells. 2022. V. 11. N. 16. P. 2475.
21. Chaudhury S., Lyskov S., Gray J.J. PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta // Bioinformatics. 2010. V. 26. N. 5. P. 689–691.
22. Chen P., Zhao J., Wang Y., Wang M., Long H., Liang D., Huang L., Wen Z., Li W., Li X., Feng H., Zhao H., Zhu P., Li M., Wang Q. F., Li G. H3.3 actively marks enhancers and primes gene transcription via opening higher-ordered chromatin // Genes & development. 2013. V. 27. N. 19. P. 2109–2124.
23. Cheung W.L., Ajiro K., Samejima K., Kloc M., Cheung P., Mizzen C.A., Beeser A., Etkin L.D., Chernoff J., Earnshaw W.C., Allis C.D. Apoptotic

phosphorylation of histone H2B is mediated by mammalian sterile twenty kinase // Cell. 2003. V. 113. N. 4. P. 507–517.
24. Chothia C., Janin J. Principles of protein–protein recognition // Nature. 1975. V. 256. P. 705–708.
25. Contrepois K., Coudereau C., Benayoun B.A., Schuler N., Roux P.F., Bischof O., Courbeyrette R., Carvalho C. Thuret J.Y., Ma Z. Histone variant H2A.J accumulates in senescent cells and promotes inflammatory gene expression // Nat. Commun. 2017. V. 8. N. P. 14995.
26. Costa V.G., Pedreira C.E. Recent advances in decision trees: an updated survey // Artif Intell Rev. 2023. V. 56. P. 4765–4800.
27. Dai L., Xiao X., Pan L., Shi L., Xu N., Zhang Z., Feng X., Ma L., Dou S., Wang P., Zhu B., Li W., Zhou Z. Recognition of the inherently unstable H2A nucleosome by Swc2 is a major determinant for unidirectional H2A.Z exchange // Cell reports. 2021. V. 35. N. 8. P. 109183.
28. Darzynkiewicz Z., and Carter S.P. Thermal stability of nucleosomes studied in situ by flow cytometry: effect of ionic strength and n-butyrate // Experimental cell research. 1989. V. 180. N. 2. P. 551–556.
29. De Las Rivas J., Fontanillo C. Protein-protein interactions essentials: key concepts to building and analyzing interactome networks // PLoS Comput Biol. 2010 V. 6. N. 6.
30. Dejnirattisai W., Zhou D., Supasa P., Liu C., Mentzer A. J., Ginn H.M., Zhao Y., Duyvesteyn H.M.E., Tuekprakhon A., Nutalai R., Wang B., López-Camacho C., Slon-Campos J., Walter T.S., Skelly D., Costa Clemens S.A., Naveca F.G., Nascimento V., Nascimento F., Fernandes da Costa C., Screaton G.R. Antibody evasion by the P.1 strain of SARS-CoV-2 // Cell. 2021. V. 184. N. 11. P. 2939–2954.
31. El Kennani S., Adrait A., Permiakova O., Hesse A.M., Ialy-Radio C., Ferro M., Brun V., Cocquet J., Govin J., Pflieger D. Systematic quantitative analysis of H2A and H2B variants by targeted proteomics // Epigenetics & chromatin. 2018. V. 11. N. 1.
32. Erausquin, E., Glaser, F., Fernández-Recio, J., López-Sagaseta, J. Structural bases for the higher adherence to ACE2 conferred by the SARS-CoV-2 spike Q498Y substitution // Acta crystallographica. Section D, Structural biology. 2022. V. 78. P. 1156–1170.
33. Esposito L., Vitagliano L., Zagari A., Mazzarella L. Pyramidalization of backbone carbonyl carbon atoms in proteins // Protein science: a publication of the Protein Society. 2000. V. 9. N. 10. P. 2038–2042.

34. Fahrmeir L., Kneib T., Lang S. Regression - Modelle, Methoden und Anwendungen // Statistik und ihre Anwendungen. 2009. 2 edn. Berlin, Heidelberg: Springer.
35. Frank Y., Unger R., Senderowitz H. Statistical analysis of sequential motifs at biologically relevant protein-protein interfaces // Computational and structural biotechnology journal. 2024. V. 23. P. 1244–1259.
36. Furutani Y. Ion–protein interactions of a potassium ion channel studied by attenuated total reflection Fourier transform infrared spectroscopy // Biophys Rev. 2018. V. 10. P. 235–239.
37. Fuxreiter M., Tompa P., Simon I. Local structural disorder imparts plasticity on linear motifs // Bioinformatics. 2007. V. 23. N. 8. P. 950–956.
38. Geng Q., Shi K., Ye G., Zhang W., Aihara H., Li, F. Structural Basis for Human Receptor Recognition by SARS-CoV-2 Omicron Variant BA.1 // Journal of virology. 2022. V. 96. N. 8.
39. Glaser F., Sternberg D., Vasker I., and Ben-Tal N. Residue frequencies and pairing preferences at protein–protein interfaces // Proteins. 2001. V. 43. P. 89–102.
40. Goncearenco A., Li M., Simonetti F.L., Shoemaker B.A., Panchenko A.R. Exploring Protein-Protein Interactions as Drug Targets for Anti-cancer Therapy with In Silico Workflows // Methods Mol Biol. 2017. N. 1647. P. 221–236.
41. Greer J. Model for haptoglobin heavy chain based upon structural homology // Proc. Natl. Acad. Sci. 1980. V. 77. P. 3393–3397.
42. Guo Z., Yamaguchi R. Machine learning methods for protein-protein binding affinity prediction inprotein design // Front. Bioinform. 2022. V. 2.
43. Gupta M.M., Bukovsky I., Homma N., Solo A.M., Hou Z. Fundamentals of Higher Order Neural Networks for Modeling and Simulation. 2013.
44. Han P., Su C., Zhang Y. *et al.* Molecular insights into receptor binding of recent emerging SARS-CoV-2 variants // Nat Commun. 2021. V. 12.
45. Hemalatha K. Advancements in Multi-Layer Perceptron Training to Improve Classification Accuracy // International Journal on Recent and Innovation Trends in Computing and Communication. 2017. V. 5. P. 353–357.
46. Henikoff S., and Smith, M. M. Histone variants and epigenetics // Cold Spring Harbor perspectives in biology. 2015. V. 7. N. 1. P. a019364.
47. Hinton G., Sejnowski T. Unsupervised Learning: Foundations of Neural Computation // MIT Press. 1999.
48. Hirano R., Arimura Y., Kujirai T., Shibata M., Okuda A., Morishima K., Inoue R., Sugiyama M., Kurumizaka, H. Histone variant H2A.B-H2B dimers are

spontaneously exchanged with canonical H2A-H2B in the nucleosome // Communications biology. 2021. V. 4. N. 1. P. 191.
49. Huang J., Rauscher S., Nawrocki G., Ran T., Feig M., de Groot B., Grubmüller H., MacKerell A.J. CHARMM36m: an improved force field for folded and intrinsically disordered proteins // Nat Methods. 2017. V. 14. N. 1. P. 71–73.
50. Humphrey W., Dalke A., Schulten K. VMD: visual molecular dynamics // Journal of molecular graphics. 1996. V. 14. N. 1. P. 33–28.
51. Isermann A., Mann C., Rube C.E. Histone Variant H2A.J Marks Persistent DNA Damage and Triggers the Secretory Phenotype in Radiation-Induced Senescence // Int. J. Mol. Sci. 2020. V. 21. P. 9130.
52. Ito J., Suzuki R., Uriu K. Convergent evolution of SARS-CoV-2 Omicron subvariants leading to the emergence of BQ.1.1 variant // Nat Commun. 2023. V. 14.
53. Jankauskaitė J., Jiménez-García B., Dapkūnas J., Fernández-Recio J., Moal I.H. SKEMPI 2.0: an updated benchmark of changes in protein–protein binding energy, kinetics and thermodynamics upon mutation // Bioinformatics. 2019. V. 35. P. 462–469.
54. Jawad B., Adhikari P., Podgornik R., Ching W.Y. Key Interacting Residues between RBD of SARS-CoV-2 and ACE2 Receptor: Combination of Molecular Dynamics Simulation and Density Functional Calculation // Journal of chemical information and modeling. 2021. V. 61. N. 9. P. 4425–4441.
55. Jiménez J., Skalic M., Martínez-Rosell G., De Fabritiis G. K DEEP: Protein–Ligand Absolute Binding Affinity Prediction via 3D-Convolutional Neural Networks // Journal of chemical information and modeling. 2018. V. 58. P. 287–296.
56. Jiménez J., Škalič M., Martínez-Rosell G., Fabritiis G.D. KDEEP: Protein-Ligand Absolute Binding Affinity Prediction via 3D-Convolutional Neural Networks // Journal of chemical information and modeling. 2018. V. 58. N. 2. P. 287-296.
57. Johansson-Åkhe I., Mirabello C., Wallner B. Predicting protein-peptide interaction sites using distant protein complexes as structural templates // Sci Rep. 2019. V. 9. N. 4267.
58. Jubb H.C., Higueruelo A.P., Ochoa-Montano B., Pitt B.W.R., Ascher D.B., Blundell T.L. Arpeggio: A web server for calculating and visualising interatomic interactions in protein structures // Journal of molecular biology. 2017. V. 429. N. 3. P. 365–371.
59. Kamakaka R.T., and Biggins S. Histone variants: deviants? // Genes & development. 2005. V. 19. N. 3. P. 295–310.

60. Kapp M.N., Freitas C.O., Nievola J.C., Sabourin R. Evaluating the conventional and class-modular architectures feedforward neural network for handwritten word recognition // 16th Brazilian Symposium on Computer Graphics and Image Processing (SIBGRAPI). 2003. P. 315-319.
61. Kastritis P.L., Bonvin A.M. On the binding affinity of macromolecular interactions: daring to ask why proteins interact // J R Soc Interface. 2012. V. 10. N. 79.
62. Kastritis P.L., Bonvin A.M. On the binding affinity of macromolecular interactions: daring to ask why proteins interact // J R Soc Interface. 2013.
63. Kim J., Wei S., Lee J., Yue H., Lee T.H. Single-Molecule Observation Reveals Spontaneous Protein Dynamics in the Nucleosome // The journal of physical chemistry. B. 2016. V. 120. N. 34. P. 8925–8931.
64. Kimura I., Yamasoba D., Tamura T., Nao N., Suzuki T., Oda Y., Mitoma S., Ito J., Nasser H., Zahradnik J., Uriu K., Fujita S., Kosugi Y., Wang L., Tsuda M., Kishimoto M., Ito H., Suzuki R., Shimizu R., Begum M.M., Sato K. Virological characteristics of the SARS-CoV-2 Omicron BA.2 subvariants, including BA.4 and BA.5 // Cell. 2022. V. 185. N. 21. P. 3992–4007.
65. Kingma D., Ba J.L. Adam : A method for stochastic optimization // arXiv:1412.6980v9. 2014.
66. Kirchdoerfer R.N., Wang N., Pallesen J. Stabilized coronavirus spikes are resistant to conformational changes induced by receptor recognition or proteolysis // Sci Rep. 2018. V. 8.
67. Klein R.H., Knoepfler P.S. Knockout tales: the versatile roles of histone H3.3 in development and disease // Epigenetics & Chromatin. 2023. V. 16. N. 38.
68. Kniazeva A.S., Armeev G.A., Shaytan, A.K. H2A-H2B Histone Dimer Plasticity and Its Functional Implications // Cells. 2022. V. 11. N. 18. P. 2837.
69. Kotsiantis S.B. Decision trees: A recent overview // Artificial Intelligence Review. 2013. V. 39. P. 261–283.
70. Kuksa P.P., Min M.R., Dugar R., Gerstein M.B. High-order neural networks and kernel methods for peptide-MHC binding prediction // Bioinformatics. 2015. V. 31. N. 22. P. 3600–3607.
71. Kulaeva O., Gaykalova D., Pestov N. et al. Mechanism of chromatin remodeling and recovery during passage of RNA polymerase II // Nat Struct Mol Biol. 2009. V. 16. P. 1272–1278.
72. Ladbury J.E., Chowdhry B.Z. Sensing the heat: the application of isothermal titration calorimetry to thermodynamic studies of biomolecular interactions // Chem Biol. 1996. V. 3. P. 791–801.

73. Lan J., Ge J., Yu J., Shan S., Zhou H., Fan S., Zhang Q., Shi X., Wang Q., Zhang L., Wang X. Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor // Nature. 2020. V. 581. P. 215–220.
74. LeCun Y., Bengio Y., Hinton G. Deep learning // Nature. 2015. V. 521 N. 7553. P. 436–444.
75. Li Y., Rezaei M.A., Li C., Li X., Wu, D.O. DeepAtom: A Framework for Protein-Ligand Binding Affinity Prediction // 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). 2019. P. 303–310.
76. Lian N. A review of the application of logistic regression in educational research: common issues, implications, and suggestions // Educational Review. 2018. P. 1–27.
77. Lodish H., Berk A., Zipursky S.L. Hierarchical Structure of Proteins // Molecular Cell Biology. 4th edition. 2000.
78. London N., Movshovitz-Attias D., Schueler-Furman O. The structural basis of peptide-protein binding strategies // Structure. 2010. V. 18. N. 2. P. 188–199.
79. Lu H., Zhou Q., He J., Jiang Z., Peng C., Tong R., Shi J. Recent advances in the development of protein-protein interactions modulators: mechanisms and clinical trials // Signal transduction and targeted therapy. 2020. V. *5*. N. 1.
80. Lucero B., Francisco K.R., Liu L.J., Caffrey C.R., Ballatore C. Protein-protein interactions: developing small-molecule inhibitors/stabilizers through covalent strategies // Trends Pharmacol Sci. 2023. V. 44. N. 7. P. 474–488.
81. Luger K., Mäder A. W., Richmond R. K., Sargent D. F., Richmond T. J. Crystal structure of the nucleosome core particle at 2.8 A resolution. Nature. 1997. V. 389. N. 6648. P. 251–260.
82. Luger K., Rechsteiner T.J., Richmond T.J. Preparation of nucleosome core particle from recombinant histones // Methods in enzymology. 1999. V. 304. P. 3–19.
83. Lyskov S., Gray J.J. The RosettaDock Server for Local Protein-Protein Docking // Nucleic Acids Research. 2008. V. 36. P. 233–238.
84. Mackay J.P., Sunde M., Lowry J.A., Crossley M., Matthews J.M. Protein interactions: is seeing believing? // Trends Biochem. 2007. V. 32. P. 530–531.
85. Mannar D., Saville J.W., Zhu X., Srivastava S.S., Berezuk A.M., Zhou S., Tuttle K.S., Kim A., Li W., Dimitrov D.S., Subramaniam S. Structural analysis of receptor binding domain mutations in SARS-CoV-2 variants of concern that modulate ACE2 and antibody binding // Cell reports. 2021. V. 37. N. 12.
86. Manu M. K-Means Clustering in Machine Learning // a Review. 2019. V. 1. P. 1–19.

87. McGaughey G.B., Gagné M., Rappé A.K. Pi-Stacking interactions. Alive and well in proteins // The Journal of biological chemistry. 1998. V. 273. N. 25. P. 15458–15463.
88. Mészáros B., Tompa P., Simon I., Dosztányi Z. Molecular principles of the interactions of disordered proteins // Journal of molecular biology. 2007. V. 372. N. 2. P. 549–561.
89. Mitchell T.M., McGraw H. Machine Learning Definition // Science/Engineering/Math. 1997.
90. Moal I. H., Fernández-Recio, J. SKEMPI: a Structural Kinetic and Energetic database of Mutant Protein Interactions and its use in empirical models // Bioinformatics. 2012. V. 28. N. 20. P. 2600–2607.
91. Moal I.H., Jiménez-García B., Fernández-Recio J. CCharPPI web server: computational characterization of protein-protein interactions from structure // Bioinformatics. 2015. V. 31. N. 1. P. 123–125.
92. Mountrakis G., Im J., Ogole C. Support vector machines in remote sensing: A review // ISPRS Journal of Photogrammetry and Remote Sensing. 2011. V. 66. N. 3. P. 247–259.
93. Nelson-Sathi S., Umasankar P. K., Sreekumar E., Nair R. R., Joseph I., Nori S.R.C., Philip J.S., Prasad R., Navyasree K.V., Ramesh S., Pillai H., Ghosh S., Santosh Kumar T.R., Pillai M.R. Mutational landscape and in silico structure models of SARS-CoV-2 spike receptor binding domain reveal key molecular determinants for virus-host interaction // BMC molecular and cell biology. 2022. V. 23. N. 1.
94. Nesterov S.V., Ilyinsky N.S., Plokhikh K.S., Manuylov V.D., Chesnokov Y.M., Vasilov R.G., Kuznetsova I.M., Turoverov, K.K., Gordeliy, V.I., Fonin, A.V., Uversky V.N. Order wrapped in chaos: On the roles of intrinsically disordered proteins and RNAs in the arrangement of the mitochondrial enzymatic machines // International journal of biological macromolecules. 2024. V. 267.
95. Ni D., Turelli P., Beckert B., Nazarov S., Uchikawa E., Myasnikov A., Pojer F., Trono D., Stahlberg H., Lau K. Cryo-EM structures and binding of mouse and human ACE2 to SARS-CoV-2 variants of concern indicate that mutations enabling immune escape could expand host range // PLoS pathogens. 2023. V. 19. N. 4.
96. Nielsen J.C., Hjo Rringgaard C., Nygaard M.M.R., Wester A., Elster L., Porsgaard T., Mikkelsen R.B., Rasmussen S., Madsen A.N., Schlein M., Vrang N., Rigbolt K., Dalbo Ge L.S. Machine-Learning-Guided Peptide Drug Discovery: Development of GLP-1 Receptor Agonists with Improved Drug

Properties // Journal of medicinal chemistry. 2024. V. 67. N. 14. P. 11814–11826.

97. Nielsen M., Lundegaard C., Blicher T., Lamberth K., Harndahl M., Justesen S., Røder G., Peters B., Sette A., Lund O., Buus, S. NetMHCpan, a Method for Quantitative Predictions of Peptide Binding to Any HLA-A and -B Locus Protein of Known Sequence // PLoS ONE. 2007. V. 2.
98. Nusinow D.A., Sharp J.A., Morris A., Salas S., Plath K., Panning B. The histone domain of macroH2A1 contains several dispersed elements that are each sufficient to direct enrichment on the inactive X chromosome // Journal of molecular biology. 2007. V. 371 N. 1. P. 11–18.
99. Nutalai R., Zhou D., Tuekprakhon A., Ginn H.M., Supasa P., Liu C., Huo J., Mentzer A. J., Duyvesteyn H.M.E., Dijokaite-Guraliuc A., Skelly D., Ritter T.G., Amini A., Bibi S., Adele S., Johnson S.A., Constantinides B., Webster H., Temperton N., Klenerman P., Screaton G.R. Potent cross-reactive antibodies following Omicron breakthrough in vaccinees // Cell. 2022. V. 185. N. 12. P. 2116–2131.
100. Nwankpa C., Ijomah W., Gachagan A., Marshall S. Activation Functions: Comparison of trends in Practice and Research for Deep Learning // arXiv:1811.03378v1. 2018.
101. Osakabe A., Lorkovic Z.J., Kobayashi W., Tachiwana H., Yelagandula R., Kurumizaka H., Berger, F. Histone H2A variants confer specific properties to nucleosomes and impact on chromatin accessibility // Nucleic acids research. 2018. V. 46. N. 15. P. 7675–7685.
102. Otvos L. The latest trends in peptide drug discovery and future challenges // Expert opinion on drug discovery. 2024. V. 19. N. 8. P. 869–872.
103. Ozden B., Şamiloğlu E., Özsan A., Erguven M., Yükrük C., Koşaca M., Oktayoğlu M., Menteş M., Arslan N., Karakülah G., Barlas A.B., Savaş B., Karaca E. Benchmarking the accuracy of structure-based binding affinity predictors on Spike-ACE2 deep mutational interaction set // Proteins. 2024. V. 92. N. 4. P. 529–539.
104. Panday S.K., Alexov E. Protein-Protein Binding Free Energy Predictions with the MM/PBSA Approach Complemented with the Gaussian-Based Method for Entropy Estimation // ACS Omega. 2022. V. 7. P. 11057–11067.
105. Patel A., Kumar S., Lai L., Chakravarthy C., Valanparambil R., Reddy E.S., Gottimukkala K., Bajpai P., Raju D.R., Edara V.V., Davis-Gardner M.E., Linderman S., Dixit K., Sharma P., Mantus G., Cheedarla N., Verkerke H.P., Frank F., Neish A.S., Roback J.D., Ortlund E.A. Molecular basis of SARS-CoV-

2 Omicron variant evasion from shared neutralizing antibody response // Structure. 2023. V. 31. N. 7. P. 801–811.

106. Pepenella S., Murphy K.J. and Hayes, J.J. Intra- and inter-nucleosome interactions of the core histone tail domains in higher-order chromatin structure // Chromosoma. 2014. V. 123. P. 3–13.
107. Perme M.P., Blas M., Turk S. Comparison of logistic regression and linear discriminant analysis // Advances in Methodology and Statistics. 2004.
108. Petsalaki E., Russell R.B. Peptide-mediated interactions in biological systems: new discoveries and applications // Current opinion in biotechnology. 2008. V. 19. N. 4. P. 344–350.
109. Phillip Y., Kiss V., Schreiber G. Protein-binding dynamics imaged in a living cell // Proc Natl Acad Sci. 2012. V. 109. P. 1461–1466.
110. Pro S.C., Zimic M., Nielsen M. Improved pan-specific MHC class I peptide-binding predictions using a novel representation of the MHC-binding cleft environment // Tissue antigens. 2014. V. 83. N. 2. P. 94–100.
111. Raghavender U.S., Rathore, R.S. Protein-Peptide Interactions in Regulatory Events // Encyclopedia of Bioinformatics and Computational Biology. 2019.
112. Rahim A., Saha P., Jha K.K., Sukumar N., Sarma B.K. Reciprocal carbonyl-carbonyl interactions in small molecules and proteins // Nature communications. 2017. V. 8. N. 1.
113. Rastelli G., Del Rio A., Degliesposti G., Sgobba M. Fast and accurate predictions of binding free energies using MM-PBSA and MM-GBSA // J Comput Chem. 2010. V. 31. P. 797–810.
114. Ravikant D.V.S., Elber R. PIE-efficient filters and coarse grained potentials for unbound protein-protein docking // Proteins. 2010. V. 78. P. 400–419.
115. Reichmann D., Phillip Y., Carmi A., Schreiber G. On the contribution of water-mediated interactions to protein-complex stability // Biochemistry. 2008. V. 47. N. 3. P. 1051–1060.
116. Romero-Molina S., Ruiz-Blanco Y.B., Green J.R., Sanchez-Garcia E. ProtDCal-Suite: A web server for the numerical codification and functional analysis of proteins // Protein science : a publication of the Protein Society. 2019. V. 28. N. 9. P. 1734–1743.
117. Romero-Molina S., Ruiz-Blanco Y.B., Mieres-Perez J., Harms M., Münch J., Ehrmann M., Sánchez-García E. PPI-Affinity: A Web Tool for the Prediction and Optimization of Protein–Peptide and Protein–Protein Binding Affinity // Journal of Proteome Research. 2022. V. 21. P. 1829–1841.
118. Rube C.E., Baumert C., Schuler N., Isermann A., Schmal Z., Glanemann M., Mann C., Scherthan H. Human skin aging is associated with increased

expression of the histone variant H2A.J in the epidermis // NPJ Aging Mech. Dis. 2021. V. 7. N. 7.

119. Rudnizky S., Bavly A., Malik O., Pnueli L., Melamed P., Kaplan A. H2A.Z controls the stability and mobility of nucleosomes to regulate expression of the LH genes // Nature communications. 2016. V. 7. P. 12958.
120. Sahariah B., Sarma B.K. Relative orientation of the carbonyl groups determines the nature of orbital interactions in carbonyl-carbonyl short contacts // Chemical science. 2018. V. 10. N. 3. P. 909–917.
121. Salleh M.Z., Derrick J.P., Deris Z.Z. Structural Evaluation of the Spike Glycoprotein Variants on SARS-CoV-2 Transmission and Immune Evasion // International Journal of Molecular Sciences. 2021. V. 22. N. 14.
122. Sampson J.M., Cannon D.A., Duan J., Epstein J.C.K., Sergeeva A.P., Katsamba P.S., Mannepalli S.M., Bahna F.A., Adihou H., Guéret S.M., Gopalakrishnan R., Geschwindner S., Rees D.G., Sigurdardottir A., Wilkinson T., Dodd R.B., De Maria L., Mobarec J. C., Shapiro L., Honig B., Wang, L. Robust prediction of relative binding energies for protein-protein complex mutations using free energy perturbation calculations // bioRxiv : the preprint server for biology. 2024.
123. Saville J.W., Mannar D., Zhu X., Berezuk A.M., Cholak S., Tuttle K.S., Vahdatihassani F., Subramaniam S. Structural analysis of receptor engagement and antigenic drift within the BA.2 spike protein // Cell reports. 2023. V. 42. N. 1.
124. Saville J.W., Mannar D., Zhu X. Structural and biochemical rationale for enhanced spike protein fitness in delta and kappa SARS-CoV-2 variants // Nat Commun. 2022. V. 13. N. 742.
125. Schmidt J., Marques M.R.G., Botti S. Recent advances and applications of machine learning in solid-state materials science // Comput Mater 5. 2019. V. 83.
126. Schneider A., Hommel G., Blettner M. Linear regression analysis: part 14 of a series on evaluation of scientific publications // Dtsch Arztebl Int. 2010. V. 107. N. 44. P. 776–782.
127. Schweke H., Mucchielli M.H., Sacquin-Mora S., Bei W., Lopes A. Protein Interaction Energy Landscapes are Shaped by Functional and also Non-functional Partners // Journal of molecular biology. 2020. V. 432. N. 4. P. 1183–1198.
128. Schymkowitz J., Borg J., Stricher F., Nys R., Rousseau F., Serrano L. The FoldX web server: an online force field // Nucleic acids research. 2005. V. 33. P. W382–W388.

129. Scott D.E., Bayly A.R., Abell C., Skidmore J. Small molecules, big targets: drug discovery faces the protein-protein interaction challenge // Nature reviews. Drug discovery. 2016. V. 15. N. 8. P. 533–550.
130. Selzer T., Albeck S., Schreiber G. Rational design of faster associating and tighter binding protein complexes. Nature Struct. Biol. 2000. V. 7. P. 537–541.
131. Shalev-Shwartz S., Shai B.D. Understanding-Machine-Learning. 2014.
132. Shang J., Ye G., Shi K. Structural basis of receptor recognition by SARS-CoV-2 // Nature. 2020. V. 581. P. 221–224.
133. Shi T.L., Li Y.X., Cai Y.D., Chou K.C. Computational methods for protein–protein interaction and their application // Curr Protein Pept Sci. 2005. V. 6. N. 5. P. 443–449.
134. Shi X., Fedulova A.S., Kotova E.Y., Maluchenko N.V., Armeev G.A., Chen Q., Prasanna C., Sivkina A.L., Feofanov A.V., Kirpichnikov M.P., Nordensköld L., Shaytan A. K., Studitsky V.M. Histone Tetrasome Dynamics Affects Chromatin Transcription // bioRxiv : the preprint server for biology. 2024. 2024.07.18.604164.
135. Shinagawa T., Takagi T., Tsukamoto D., Tomaru C., Huynh L. M., Sivaraman P., Kumarevel T., Inoue K., Nakato R., Katou Y., Sado T., Takahashi S., Ogura A., Shirahige K., Ishii S. Histone variants enriched in oocytes enhance reprogramming to induced pluripotent stem cells // Cell stem cell. 2014. V. 14. N. 2. P. 217–227.
136. Siebenmorgen T., Zacharias M. Computational prediction of protein–protein binding affinities // WIREs Comput Mol Sci. 2019. V. 10.
137. Soleymani F., Paquet E., Viktor H., Michalowski W., Spinello D. Protein-protein interaction prediction with deep learning: A comprehensive review // Comput Struct Biotechnol J. 2022. V. 20. P. 5316–5341.
138. Sricharan K., Raich R., Hero A.O. K-nearest neighbor estimation of entropies with confidence. 2011 IEEE International Symposium on Information Theory Proceedings. 2011. P. 1205–1209.
139. Srivastava N., Hinton G., Krizhevsky A., Sutskever I., Salakhutdinov R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting // Journal of Machine Learning Research. 2014. V. 15. P. 1929–1958.
140. Stepniewska-Dziubinska M.M., Zielenkiewicz P., Siedlecki P. Development and evaluation of a deep learning model for protein-ligand binding affinity prediction. 2017.
141. Su C., He J., Han P., Bai B., Li D., Cao J., Tian M., Hu Y., Zheng A., Niu S., Chen Q., Rong X., Zhang Y., Li W., Qi J., Zhao X., Yang M., Wang Q., Gao

G.F. Molecular basis of mink ACE2 binding to SARS-CoV-2 and its mink-derived variants // J Virol. 2022. V. 96.

142. Szenker E., Ray-Gallet D., and Almouzni G. The double face of the histone variant H3.3 // Cell Res. 2011. V. 21. P. 421–434.
143. Talbert P.B., and Henikoff S. Histone variants at a glance // J. Cell Sci. 2021. V. 134.
144. Talbert P.B., and Henikoff, S. Histone variants--ancient wrap artists of the epigenome. Nature reviews. Molecular cell biology. 2010. V. 11. N. 4. P. 264–275.
145. Tamura T., Ito J., Uriu K. Virological characteristics of the SARS-CoV-2 XBB variant derived from recombination of two Omicron subvariants // Nat Commun. 2023. V. 14.
146. Tanaka H., Sato S., Koyama M., Kujirai T., Kurumizaka H. Biochemical and structural analyses of the nucleosome containing human histone H2A.J // Journal of biochemistry. 2020. V. 167. N. 4. P. 419–427.
147. Tanaka Y., Tawaramoto-Sasanuma M., Kawaguchi S., Ohta T., Yoda K., Kurumizaka H., Yokoyama, S. Expression and purification of recombinant human histones // Methods (San Diego, Calif.). 2004. V. 33. N. 1. P. 3–11.
148. Thakar A., Gupta P., Ishibashi T., Finn R., Silva-Moreno B., Uchiyama S., Fukui K., Tomschik M., Ausio J., Zlatanova J. H2A.Z and H3.3 histone variants affect nucleosome structure: biochemical and biophysical studies // Biochemistry. 2009. V. 48. N. 46. P. 10852–10857.
149. Tian F., Lv Y., Yang L. Structure-based prediction of protein-protein binding affinity with consideration of allosteric effect // Amino acids. 2012. V. 43. N. 2. P. 531–543.
150. Tóth K., Gansen A., Hetey S., Székvölgyi L., Nordenskiöld L., Langowski J. How Histone Modifications Change Nucleosome Stability – FRET Studies on Single Molecules and in Bulk // Microscopy and Microanalysis. 2014. V. 20. N. S3. P. 1204–1205.
151. Toulmé J.J. Stacking Interactions: The Key Mechanism for Binding of Proteins to Single-Stranded Regions of Native and Damaged Nucleic Acids? // Chromosomal Proteins and Gene Expression. 1985. V. 101.
152. Urahama T., Horikoshi N., Osakabe A., Tachiwana H., Kurumizaka H. Structure of human nucleosome containing the testis-specific histone variant TSH2B // Acta crystallographica. Section F, Structural biology communications. 2014. V. 70. N. 4. P. 444–449.

153. Uversky V.N. Functional unfoldomics: Roles of intrinsic disorder in protein (multi)functionality // Advances in protein chemistry and structural biology. 2024. V. 138. P. 179–210.
154. Van Der Spoel D., Lindahl E., Hess B., Groenhof G., Mark A. E., Berendsen H. J. GROMACS: fast, flexible, and free // Journal of computational chemistry. 2005. V. 26. N. 16. P. 1701–1718.
155. Vang Y.S., Xie, X. HLA class I binding prediction via convolutional neural networks // Bioinformatics. 2017. V. 33. P. 2658–2665.
156. Vangone A., Bonvin A. M. Contacts-based prediction of binding affinity in protein-protein complexes // eLife. 2015. V. 4.
157. Vapnik V. Estimation of Dependences Based on Empirical Data // Nauka (Moscowz). 1982. V. 27. P. 5165–5184.
158. Venkatesh S., and Workman J.L. Histone exchange, chromatin structure and the regulation of transcription // Nature reviews. Molecular cell biology. 2015. V. 16. N. 3. P. 178–189.
159. Walsh I., Pollastri G., Tosatto S.C. Correct machine learning on protein sequences: a peerreviewing perspective // Brief. Bioinform. 2016. V. 17. N. 5. P. 831–840.
160. Wang L., Berne B.J., Friesner R.A. On achieving high accuracy and reliability in the calculation of relative protein-ligand binding affinities // Proc Natl Acad Sci USA. 2012. V. 109. P. 1937–1942.
161. Wang Q., Zhang Y., Wu L., Niu S., Song C., Zhang Z., Lu G., Qiao C., Hu Y., Yuen K. Y., Wang Q., Zhou H., Yan J., Qi J. Structural and Functional Basis of SARS-CoV-2 Entry by Using Human ACE2 // Cell. 2020. V. 181. N. 4. P. 894–904.
162. Wang R., Fang X., Lu Y., Yang C. Y., Wang, S. The PDBbind database: methodologies and updates // Journal of medicinal chemistry. 2005. V. 48. N.12. P. 4111–4119. PDBbind-CN v. 2020. Database. Accessed May 16, 2024.
163. Wang S., Peng J., Ma J., Xu J. Protein secondary structure prediction using deep convolutional neural fields // Scientific Reports. 2016. V. 6. N. 1. P. 18962.
164. Wang Y., Liu C., Zhang C. Structural basis for SARS-CoV-2 Delta variant recognition of ACE2 receptor and broadly neutralizing antibodies // Nat Commun. 2022. V. 13. N. 871.
165. Wang Y., Xu C., Wang Y. Conformational dynamics of the Beta and Kappa SARS-CoV-2 spike proteins and their complexes with ACE2 receptor revealed by cryo-EM // Nat Commun. 2021. V. 12.

166. Wesley E. Protein−Protein Interactions Interface Structure Binding Thermodynamics and Mutational Analysis. Chem. Rev. 1997. V. 97. P. 1233–1250.
167. Willander M., Al-Hilli S. Analysis of biomolecules using surface plasmons // Methods Mol Biol. 2009. N. 544. P. 201–229.
168. Wrobel A.G., Benton D.J., Roustan C. Evolution of the SARS-CoV-2 spike protein in the human host. Nat Commun. 2022. V. 13.
169. Wu K., Peng G., Wilken M., Geraghty R. J., Li F. Mechanisms of host receptor adaptation by severe acute respiratory syndrome coronavirus // The Journal of biological chemistry. 2012. V. 287. N. 12. P. 8904–8911.
170. Wu W. H., Guo J., Zhang L., Zhang W. B., Gao W. Peptide/protein-based macrocycles: from biological synthesis to biomedical applications // RSC chemical biology. 2022. V. 3. N. 7. P. 815–829.
171. Xiao T., Lu J., Zhang J. A trimeric human angiotensin-converting enzyme 2 as an anti-SARS-CoV-2 agent // Nat Struct Mol Biol. 2021. V. 28. P. 202–209.
172. Xue L.C., Rodrigues J.P., Kastritis P.L., Bonvin A.M., Vangone A. PRODIGY: a web server for predicting the binding affinity of protein–protein complexes // Bioinformatics. 2016. V. 32.
173. Yamashita R., Nishio M., Do R.K.G. Convolutional neural networks: an overview and application in radiology // Insights Imaging. 2018. V. 9. P. 611–629.
174. Yang Y.X., Huang J.Y., Wang P., Zhu B.T. AREA-AFFINITY: A Web Server for Machine Learning-Based Prediction of Protein-Protein and Antibody-Protein Antigen Binding Affinities // Journal of chemical information and modeling. 2023. V. 63. N. 11. P. 3230–3237.
175. Yang Y.X., Wang P., Zhu B.T. Importance of interface and surface areas in protein-protein binding affinity prediction: A machine learning analysis based on linear regression and artificial neural network // Biophysical chemistry. 2022. V. 283.
176. Ye F., Lin X., Chen Z. S19W, T27W, and N330Y mutations in ACE2 enhance SARS-CoV-2 S-RBD binding toward both wild-type and antibody-resistant viruses and its molecular basis // Sig Transduct Target Ther. 2021. V. 6.
177. Zhang C., Liu S., Zhou Y. Accurate and efficient loop selections by the DFIRE-based all-atom statistical potential // Protein science: a publication of the Protein Society. 2004. V. 13. N. 2. P. 391–399.
178. Zhang H., Liao L., Saravanan K.M., Yin P., Wei Y. DeepBindRG: a deep learning based method for estimating effective protein–ligand affinity. // PeerJ. 2019. V. 7.

179. Zhang N. PremPRI: predicting the effects of missense mutations on protein-RNA interactions // Int J Mol Sci. 2020. V. 21.
180. Zhang N., Chen Y., Zhao F., Yang Q., Simonetti F.L., Li M. PremPDI estimates and interprets the effects of missense mutations on protein-DNA interactions // PLoS Comput Biol. 2018. V.14.
181. Zhang W., Shi K., Geng Q., Ye G., Aihara H., Li F. Structural basis for mouse receptor recognition by SARS-CoV-2 omicron variant // Proceedings of the National Academy of Sciences of the United States of America. 2022. V. 119. N. 44.
182. Zhang Z., Zhang Y., Liu K. The molecular basis for SARS-CoV-2 binding to dog ACE2 // Nat Commun. 2021. V. 12.
183. Zhao T., Cheng L., Zang T., Hu Y. Peptide-Major Histocompatibility Complex Class I Binding Prediction Based on Deep Learning With Novel Feature // Frontiers in Genetics. 2019. V. 10.
184. Zhao Z., Xie Y., Bai B. Structural basis for receptor binding and broader interspecies receptor recognition of currently circulating Omicron sub-variants // Nat Commun. 2023. V. 14.
185. Zheng F., Jiang X., Wen Y., Yang Y., Li M. Systematic investigation of machine learning on limited data: A study on predicting protein-protein binding strength // Comput Struct Biotechnol J. 2023. V. 23. P. 460–472.
186. Zheng A., Wu L., Ma R. A binding-enhanced but enzymatic activity-eliminated human ACE2 efficiently neutralizes SARS-CoV-2 variants // Sig Transduct Target Ther. 2022. V. 7. N. 10.
187. Zhu G., Blumberg D.G. Classification using ASTER data and SVM algorithms // The case study of Beer Sheva, Israel. Remote Sensing of Environment. 2002. V. 80. N.2. P. 233–240.

ПРИЛОЖЕНИЕ

Приложение 1. Распределение атомов по функциональным группам

Аминокислоты записаны в трехбуквенном коде, обозначения атомов совпадают с таковыми в базе данных PDB.

Название группы	Аминокислоты																			
	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
Hbond acceptor	O, OXT	O, OXT	ND2, O, OD1, OXT	OD1, OD2, O, OXT	SG, O, OXT	NE2, O, OE1, OXT	OE1, OE2, O, OXT	O, OXT	ND1, NE2, CE1, CD2, O, OXT	O, OXT	O, OXT	O, OXT	SD, O, OXT	O, OXT	O, OXT	OG, O, OXT	OG1, O, OXT	O, OXT	OH, O, OXT	O, OXT
Hbond donor	N	N, NE, NH1, NH2	N, ND2, OD1	N	N, SG	N, NE2, OE1	N	N	N, ND1, CE1, NE2, CD2	N	N	N, NZ	N	N	N, OG	N, OG	N, OG1	N, NE1	N, OH	N
Weak hbond donor	CA, CB	CA, CB, CG, CD	CA, CB	CA, CB	CA, CB	CA, CB, CG	CA, CB, CG	CA	CA, CB	CA, CB, CG1, CD1, CG2	CA, CB, CG, CD1, CD2	CA, CB, CG, CD, CE	CA, CB, CG, CE	CA, CB, CG, CD1, CD2, CE1, CE2, CZ	CA, CB, CG, CD	CA, CB	CA, CB, CG2	CA, CB, CD1, CE3, CZ3, CH2, CZ2	CA, CB, CG, CD1, CD2, CE1, CE2, CZ	CA, CB, CG1, CG2

Pos ionisable		NE, CZ, NH1, NH2							CG, ND1, CE1, NE2, CD2			NZ							
Neg ionisable				OD1, OD2			OE1, OE2												
Hydrophobe	CB	CB, CG	CB	CB	CB	CB, CG	CB, CG		CB	CB, CG1, CD1, CG2	CB, CG, CD1, CD2	CB, CG, CD	CB, CG, CE, SD	CB, CG, CD1, CD2, CE1, CE2, CZ	CB, CG		CG2	CB, CG, CD2, CE3, CZ3, CH2, CZ2	CB, CG, CD1, CD2, CE1, CE2	CB, CG1, CG2
Carbonyl oxygen	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
Carbonyl carbon	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C
Aromatic									CG, ND1, CE1, NE2, CD2					CG, CD1, CD2, CE1, CE2, CZ				CG, CD1, CD2, NE1, CE2, CE3, CZ2, CZ3, CH2	CG, CD1, CD2, CE1, CE2, CZ	

Приложение 2. Набор комплексов ACE2-RBD спайкового белка коронавирусов, используемый в исследовании интерфейса и для тестирования предсказательного алгоритма.

PDB ID	K_D,nM	Цепь ACE2	Цепь RBD	Источник	Описание
6lzg	18,5	A	B	Wang et al., 2020	Комплекс Omicron ACE2-RBD
6m0j	4,7	A	E	Lan et al., 2020	SARS-CoV-2 RBD спайкового белка в комплексе с ACE2
7e3j	18,5	A	B	Zhang et al., 2021	SARS-CoV-2 связанный с ACE2 собаки
7ekc	5,16	A	B	Han et al., 2021	Структура SARS-CoV-2 **Гамма** варианта RBD спайкового белка в комплексе с ACE2 человека
7eke	106,94	A	B	Han et al., 2021	Структура of SARS-CoV-2 RBD спайкового белка с мутацией **F486L** в комплексе с ACE2 человека
7ekf	3,64	A	B	Han et al., 2021	Структура SARS-CoV-2 **Альфа** варианта RBD спайкового белка в комплексе с ACE2 человека
7ekg	8,1	A	B	Han et al., 2021	Структура SARS-CoV-2 **Бета** варианта RBD спайкового белка в комплексе с ACE2 человека
7ekh	3,07	A	B	Han et al., 2021	Структура SARS-CoV-2 RBD спайкового белка с мутацией **Y453F** в комплексе с ACE2 человека
7efp	10,1	A	B	Ye et al., 2021	Структура SARS-CoV-2 RBD спайкового белка в комплексе с мутантным ACE2 (S19W,N330Y)
7efr	13,6	A	B	Ye et al., 2021	Структура SARS-CoV-2 RBD спайкового белка в комплексе с мутантным ACE2 (T27W,N330Y)
7sxy	7,36	E	B	Mannar et al., 2021	Структура SARS-CoV-2 RBD спайкового белка с мутацией **D614G** в комплексе с ACE2 человека
7sy0	6,25	E	B	Mannar et al., 2021	Структура SARS-CoV-2 RBD спайкового белка с мутацией **D614G,L452R** в комплексе с ACE2 человека

7sy2	3,41	E	B	Mannar et al., 2021	Структура SARS-CoV-2 RBD спайкового белка с мутацией **D614G,N501Y** в комплексе с ACE2 человека
7sy4	1,1	E	B	Mannar et al., 2021	Структура SARS-CoV-2 RBD спайкового белка с мутацией **D614G,N501Y,E484K** в комплексе с ACE2 человека
7sy6	3,25	E	B	Mannar et al., 2021	Структура SARS-CoV-2 RBD спайкового белка с мутацией **D614G,N501Y,E484K, K417N** в комплексе с ACE2 человека
7sy8	1,32	E	B	Mannar et al., 2021	Структура SARS-CoV-2 RBD спайкового белка с мутацией **D614G,N501Y,E484K, K417T** в комплексе с ACE2 человека
7nxc	4,8	A	B	Dejnirattisai et al., 2021	Структура RBD SARS-CoV-2 P.1 (**K417T, E484K,** и **N501Y**) варианта спайкового белка в комплексе с ACE2
3sci	133,33	A	E	Wu et al., 2012	Структура RBD домена спайкового белка SARS коронавируса в комплексе с ACE2 человека
7u0n	4,71	A	E	Gegg et al., 2022	Структура SARS-CoV-2 Omicron RBD (BA.1) в комплексе с ACE2 человека
6cs2	150	B	D	Kirchdoerfer et al., 2018	Структура RBD домена спайкового белка SARS-CoV S 2P в комплексе с ACE2 человека
6vw1	23,2	A	E	Shang et al., 2020	Структура SARS-CoV-2 химерного RBD в комплексе с ACE2 человека
7wk6	80	A	E	Patel et al., 2023	ACE2 с Omicron-S RBD
7vx4	82,66	A	E	Wang et al., 2021	ACE2-RBD SARS-CoV-2 Бета варианта
7vx5	84,22	A	E	Wang et al., 2021	ACE2-RBD SARS-CoV-2 Kappa варианта
7kj2	77	B	D	Xiao et al., 2021	SARS-CoV-2 RBD в комплексе с ACE2 человека
7w9i	41	A	E	Wang et al., 2022	SARS-CoV-2 Delta RBD в комплексе с ACE2 человека
7tex	2,65	B	E	Saville et al., 2022	Структура SARS-CoV-2 Delta

					(B.1.617.2) RBD в комплексе с ACE2 человека
7zf7	38	A	B	Nutalai et al., 2022	SARS-CoV-2 Omicron BA.2 RBD в комплексе с ACE2 человека
8t23	2	A	B	Ahn et al., 2023	RBD-ACE2 интерфейс SARS-CoV-2 тримера спайкового белка в комплексе с ACE2
7yhw	24,7	A	B	Zhao et al., 2023	SARS-CoV-2 Omicron BA.2.12.1 RBD RBD в комплексе с ACE2 человека
7yj3	14,6	A	B	Zhao et al., 2023	SARS-CoV-2 Omicron BA.2 RBD в комплексе с ACE2 человека
8h06	9	A	B	Zhao et al., 2023	SARS-CoV-2 Omicron BA.4/5 RBD в комплексе с ACE2 человека
8h5c	7,5	A	B	Zhao et al., 2023	SARS-CoV-2 Omicron BA.2.75 RBD в комплексе с ACE2 человека
8iov	1	A	B	Tamura et al., 2023	SARS-CoV-2 XBB.1 RBD в комплексе с ACE2 человека
8if2	0,66	A	B	Ito et al., 2023	SARS-CoV-2 Omicron BQ.1.1 RBD в комплексе с ACE2 человека
8aqs	2,2	B	A	Ni et al., 2023	BA.4/5 SARS-CoV-2 RBD в комплексе с ACE2 человека
8dm6	5	D	A	Saville et al., 2023	SARS-CoV-2 Omicron BA.2 RBD в комплексе с ACE2 человека
8asy	0,45	A	B	Saville et al., 2023	SARS-CoV-2 Omicron BA.2.75 RBD в комплексе с ACE2 человека
7wnm	3	B	A	Zheng et al., 2023	SARS-CoV-2 Gamma RBD в комплексе с мутантным ACE2 (T27F,R273Q) человека
7ufk	2	A	E	Zhang et al., 2022	SARS-CoV-2 BA.2 RBD в комплексе с ACE2 человека
7xwa	1,9	A	B	Kimura et al., 2022	SARS-CoV-2 BA.4/5 RBD в комплексе с ACE2 человека
7w8s	145	A	B	Su et al., 2022	SARS-CoV-2 RBD **Y453F** в комплексе с ACE2 американской норки
7wa1	381	A	B	Su et al., 2022	SARS-CoV-2 RBD **F486L** в комплексе с ACE2 американской норки
7p19	3	A	E	Erausquin et al.,	SARS-CoV-2 RBD **Q498Y** в

				2022	комплексе с ACE2 человека
7tez	4,88	E	B	Saville et al., 2022	SARS-CoV-2 Kappa (B.1.617.1) RBD в комплексе с ACE2 человека
7r11	44,7	D	A	Wrobel et al., 2022	SARS-CoV-2 Beta RBD в комплексе с ACE2 человека
7wbl	31,4	A	B	Han et al., 2022	SARS-CoV-2 Omicron RBD в комплексе с ACE2 человека
7wbq	25,07	A	B	Han et al., 2022	SARS-CoV-2 Delta RBD в комплексе с ACE2 человека