Санкт–Петербургский государственный университет

БАГДАСАРЯН Эрик Эдгарович

Выпускная квалификационная работа

Глубокое обучение для предсказания энергии связывания двух белков

Уровень образования: бакалавриат

Направление: 02.03.01 «Математика и компьютерные науки»

Основная образовательная программа:

СВ.5189.2021 «Науки о данных»

Научный руководитель:
старший преподаватель,
Факультет математики и
компьютерных наук СПбГУ,
Ершов Василий Алексеевич

Рецензент:
старший научный сотрудник,
Кафедра генетики и биотехнологии,
Рубель Александр Анатольевич

Санкт-Петербург

2025 г.

Содержание

Введение

1.1 Основные понятия

Белки представляют собой высокомолекулярные органические соединения, образованные цепочками аминокислот, соединённых пептидными связями. Эти биополимеры выполняют фундаментальные функции в живых организмах: формируют структурную основу клеток, выступают в роли биокатализаторов (ферменты), участвуют в передаче сигналов (рецепторы и гормоны) и осуществляют множество других жизненно важных процессов.

Свободная энергия связывания (ΔG) является ключевой термодинамической характеристикой, количественно описывающей силу взаимодействия между молекулами, в частности между белками и их лигандами или между белковыми молекулами. Данный параметр определяет термодинамическую выгодность образования молекулярного комплекса в стандартных условиях, что имеет принципиальное значение для понимания молекулярных механизмов биологических процессов.
Прогнозирование величины свободной энергии связывания между белковыми молекулами представляет собой сложную многопараметрическую задачу. Для её решения необходимо комплексно учитывать ряд критически важных факторов: трехмерную структуру взаимодействующих белков, особенности электростатических и вандерваальсовых взаимодействий, проявления гидрофобного эффекта, а также возможные конформационные изменения молекул при образовании комплекса. Современные вычислительные методы позволяют с различной степенью точности моделировать эти процессы, что открывает новые возможности для исследований в области молекулярной биологии и разработки лекарственных препаратов.

$\Delta G = G_{complex} - (G_{protein1} + G_{protein2})$, где G_x - свободная энергия Гиббса для x. Также свободную энергию связывания можно вычислить по следующей формуле: $\Delta G = RTlnK_d$, где R — универсальная газовая постоянная, T-температура, а $K_d = \frac{[R][L]}{[RL]}$ - константа диссоциации, равная отношению произведения концентраций белков R и L к концентрации их комплекса RL в равновесном состоянии

В настоящий момент существует 2 способа измерения ΔG:
Экспериментальный подход, основанный на измерении константы диссоциации (Kd) с последующим расчетом ΔG по фундаментальному термодинамическому уравнению. Данный метод предполагает проведение лабораторных исследований с использованием таких методик, как изотермическая титрационная калориметрия (ITC) или поверхностный плазмонный резонанс (SPR).
Вычислительные методы, в частности подходы машинного обучения, которые позволяют прогнозировать значения ΔG на основе анализа структурных и физико-химических характеристик белковых молекул. Эти методы развиваются как перспективная альтернатива трудоемким экспериментальным исследованиям.

1.2 Актуальность

Предсказание ΔG— одна из ключевых задач структурной биоинформатики и вычислительной биологии. Эта величина используется в разработке лекарств, изучении белковых взаимодействий и рациональном конструировании биомолекул.

Хотя современные экспериментальные методы позволяют измерить ΔG с высочайшей точностью, они остаются дорогими, трудоемкими и

медленными. Представьте: чтобы протестировать всего несколько сотен молекул, требуется месяцы работы и десятки тысяч долларов.Вот почему мир так отчаянно нуждается в вычислительных методах, способных предсказывать ΔG быстро, дёшево и — самое главное — точно.

Современные модели машинного обучения, от графовых нейросетей до трансформеров, уже научились предсказывать ΔG с разумной погрешностью. Но проблема в том, что "разумно" — недостаточно. Ошибка всего в 1–2 ккал/моль может превратить перспективный лекарственный кандидат в бесполезную молекулу.
Поэтому моя работа связана с попыткой улучшить уже имеющиеся результаты. И стать на шаг ближе к идеальному предсказанию ΔG

Обзор литературы

2.1 Известные модели

Есть всего 3 подхода к задаче: методы на основе ML, сверточные нейронные сети (3D-CNN), графовые нейронные сети (GNN)

2.1.1 Методы на основе ML

Методы машинного обучения используют заранее подготовленные характеристики, такие как:

- Геометрические свойства (расстояния между атомами, углы и т.д.).
- Энергетические параметры (например, вклад Ван-дер-Ваальсовых взаимодействий, электростатических взаимодействий, водородных связей).
- Химические свойства (типы атомов, заряды атомов, типы связей и т.д.).
  Примеры ML-базированных подходов: *RF-score*[1], *X-Score*[2]

2.1.2 Сверточные нейронные сети (3D-CNN)

Пространство вокруг белка и лиганда представляется в виде вокселей (трехмерных пикселей), где каждый воксель кодирует химическую информацию (например, тип атома, заряд, гидрофобность и т.д.).
Примеры:

- *AtomNet*[3]: Одна из первых сетей, использующих 3D-CNN для предсказания сродства связывания. AtomNet анализирует воксельные представления белка и лиганда.
- *Pufnucy*[4]: Улучшенный подход, который использует 3D-CNN для оценки энергии связывания, учитывая химические свойства атомов.
- *OnionNet*[5]: Пространство вокруг лиганда разбивается на концентрические сферические слои, где внутри каждого слоя агрегируется информация о взаимодействиях между атомами лиганда и ближайшими атомами белка.

2.1.3 Графовые нейронные сети (GNN)

Белки и лиганды можно представить в виде графов, где вершины — это атомы,а ребра — химические связи. Также некоторые модели проводят ребра между достаточно близкими атомами. Затем на полученных графах обучаются нейронные сети.
Примеры:

- *CMPNN*[6] (Communicative Message Passing Neural Network):
    - Не использует информацию о координатах атомов в пространстве.
    - CMPNN делает фокус на рёбрах, а не только на вершинах. Это позволяет модели лучше учитывать типы химических связей.

- *DimeNet*[7] (Directional Message Passing Neural Network):
    - Учитывает расстояния между атомами, а также углы между связями.
    - Обладает 3D-инвариантностью.
    - Передаёт сообщения через рёбра
- *SIGNN*[8] (Structure-aware Interactive Graph Neural Networks):
    - Как и DimeNet учитывает расстояния между атомами и углы между связями.
    - Использует механизм передачи сообщений для обмена информацией между атомами белка и лиганда, а также внутри каждого из них.

2.1.4 PAMNet

PAMNet[9] (Protein–Ligand Affinity Prediction Model Network)

- Тоже учитывает расстояния между атомами и углы между связями.
- Механизм передачи сообщений состоит из 2ух этапов: внутримолекулярная передача сообщений и Межмолекулярная передача сообщений
- Совместное представление для белка и лиганда, что позволяет эффективно моделировать их взаимодействия
- Обладает 3D-инвариантностью

Как показано в таблице 1 PAMNet показало наилучшие метрики на датасете PDBbind[10], содержащем почти 20000 взаимодействий вида белок-лиганд.

Model		RMSE ↓	MAE ↓	SD ↓	R ↑
ML-based	LR	1.675 (0.000)	1.358 (0.000)	1.612 (0.000)	0.671 (0.000)
	SVR	1.555 (0.000)	1.264 (0.000)	1.493 (0.000)	0.727 (0.000)
	RF-Score	1.446 (0.008)	1.161 (0.007)	1.335 (0.010)	0.789 (0.003)
CNN-based	Pafnucy	1.585 (0.013)	1.284 (0.021)	1.563 (0.022)	0.695 (0.011)
	OnionNet	1.407 (0.034)	1.078 (0.028)	1.391 (0.038)	0.768 (0.014)
GNN-based	GraphDTA	1.562 (0.022)	1.191 (0.016)	1.558 (0.018)	0.697 (0.008)
	SGCN	1.583 (0.033)	1.250 (0.036)	1.582 (0.320)	0.686 (0.015)
	GNN-DTI	1.492 (0.025)	1.192 (0.032)	1.471 (0.051)	0.736 (0.021)
	D-MPNN	1.493 (0.016)	1.188 (0.009)	1.489 (0.014)	0.729 (0.006)
	MAT	1.457 (0.037)	1.154 (0.037)	1.445 (0.033)	0.747 (0.013)
	DimeNet	1.453 (0.027)	1.138 (0.026)	1.434 (0.023)	0.752 (0.010)
	CMPNN	1.408 (0.028)	1.117 (0.031)	1.399 (0.025)	0.765 (0.009)
	SIGN	*1.316* (0.031)	*1.027* (0.025)	*1.312* (0.035)	*0.797* (0.012)
Ours	**PAMNet**	**1.263 (0.017)**	**0.987 (0.013)**	**1.261 (0.015)**	**0.815 (0.005)**

Таблица 1: метрики качества известных моделей на PAMNet

2.2 База данных

Свое исследование я проводил на датасете PPB-Affinity[11], так как это одна из самых больших, на данный момент, баз данных, содержащих белково-белковые взаимодействия(свыше 12000).

Особенность PPB-Affinity заключается в том, что создатели датасета взяли несколько уже известных датасетов (SKEMPI v2.0[12], SAbDab[13–15], PDBbind v2020 [16-21], Affinity Benchmark v5.5[22-24], and ATLAS[25]) и добавили в них эксперименты с мутациями в белках(в которых некоторые аминокислоты заменили на другие). Вследствии чего расширили датасет почти вдвое. Но обучить модель, предсказывающую энергию связи для белков с мутациями оказалось достаточно сложно.

Все данные о файлах и необходимых мутациях можно найти в таблице[32].

Мы взяли за тестовую выборку антитело-антиген комплексы (они указаны в последнем столбце xlsx файла), с которыми часто приходится работать на практике. Эта выборка составила около 11% от размера всего датасета.

В настоящее время в научной литературе представлено единственное

исследование, в котором приводится как описание используемого датасета, так и результаты прогнозирования свободной энергии связывания для данного набора данных. Согласно опубликованным результатам, модель демонстрирует следующие показатели точности:
RMSE: 1.994, SRCC: 0.298, PCC:0.329.
В указанной работе также представлен график 1, демонстрирующий зависимости предсказанных значений от экспериментально измеренных величин свободной энергии связывания, где по оси абсцисс отложены экспериментальные данные, а по оси ординат - результаты моделирования. Данная визуализация позволяет наглядно оценить степень соответствия между расчетными и фактическими значениями.

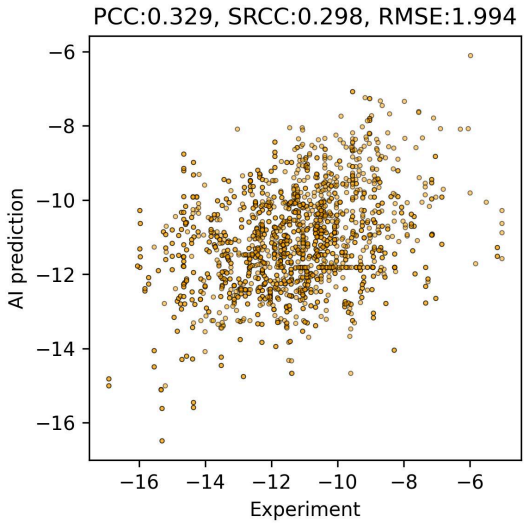

График 1: сравнение предсказанных значений с экспериментальными для базовой модели из статьи

2.3 Целевые метрики

Основные метрики для предсказания свободной энергии связывания:

- RMSE(среднеквадратическая ошибка)
  Чувствительна к большим ошибкам, что важно для ΔG, где отклонение в 1–2 ккал/моль уже значимо.
- MAE(средняя абсолютная ошибка)
  Устойчива к выбросам и прямо показывает среднюю ошибку предсказания.

Дополнительные метрики:

- PCC(коэффициент корреляции Пирсона)
  Пирсон показывает, насколько предсказания линейно связаны с экспериментальными данными.
- SRCC(коэффициент корреляции Спирмана)
  Спирман показывает, насколько хорошо модель сохраняет порядок значений (ранжирование).

Постановка цели и задач

Задачи:

1. Собрать датасет PPB-Affinity.
2. Рассчитать метрики для значений, предсказанных FoldX
3. Препроцессинг PPB-Affinity для обучения PAMNet.
4. Обучить различными способами на нашем датасете модель PAMNet, показавшую наилучшие результаты на PDBbind
5. Сравнить результаты с результатами базовой модели из статьи про PPB-Affinity и с результатами FoldX

Цель:

Превзойти результаты предыдущих моделей и стать на шаг ближе к идеальному предсказанию связывания двух белков

Исследование

4.1 Изучение прошлой модели

Изучив подробнее модель из статьи про PPB-Affinity появилось подозрение, что у них комплексы из трейн выборки, также попадают и в тестовую выборку, вследствии чего метрики, показанные в статье гораздо выше, чем должны быть на самом деле.
Тогда я решил связаться с авторами статьи, создав issue на github[33]. Они подтвердили мои подозрения.
Вследствии чего значение метрик из статьи не валидны.

4.2 FoldX

FoldX[26] представляет собой современный программный комплекс, нашедший широкое применение в области вычислительной структурной биологии. Данный инструментарий позволяет решать ключевые задачи молекулярного моделирования.
Методологическая основа FoldX базируется на эмпирическом силовом поле, что обеспечивает оптимальное сочетание вычислительной эффективности и удовлетворительной точности получаемых результатов. Это делает данный инструмент особенно востребованным при проведении масштабных исследований, требующих анализа множества структурных вариантов.
Основные возможности FoldX:

1. Расчет энергии стабильности белка
    - Оценка изменения свободной энергии сворачивания при мутациях.
    - Анализ влияния точечных мутаций на стабильность структуры.

2. Моделирование мутаций
    - Замена аминокислот в PDB-структуре.
    - Предсказание структурных изменений после мутации.
3. Анализ взаимодействий
    - Расчет энергии связывания (protein-protein, protein-ligand).
    - Выявление ключевых аминокислот для взаимодействий.

Для установки FoldX необходимо получить лицензию на официальном сайте[27].

4.2.1 BuildModel

FoldX BuildModel[28] — это ключевая функция пакета FoldX, предназначенная для предсказания и оптимизации трехмерных структур белков, включая Моделирование точечных мутаций (аминокислотных замен)
Входные данные:

- PDB-файл с исходной структурой белка
- Список мутаций или инструкции для сборки новой структуры
  В формате:
  [исходная аминокислота][id цепи][порядковый номер][новая аминокислота]

Основные шаги алгоритма:

- Разбор структуры: FoldX анализирует геометрию белка (связи, углы и прочее).
- Оптимизация боковых цепей: Перебирает возможные ротамеры аминокислот, выбирая наиболее энергетически выгодные.
- Расчет энергии: Оценивает стабильность структуры с помощью эмпирического силового поля FoldX (учитывает вандерваальсовы взаимодействия, водородные связи, электростатику и др.).

- Вывод структуры: Генерирует новый PDB-файл с предсказанной моделью.

Пример запуска:

```
FoldX --command=BuildModel --pdb=input.pdb --mutant-file=mutations.txt
--output=output.pdb
```

4.2.2 AnalyseComplex

FoldX AnalyseComplex[29] — специализированный инструмент пакета FoldX для:

- Расчета свободной энергии связывания между белками или белком и лигандом
- Определения вклада отдельных аминокислот в стабильность комплекса
- Выявления "горячих точек" (hot spots) межмолекулярного взаимодействия
- Сравнения стабильности различных конформаций комплекса

Сценарии применения:

1. Идентификация ключевых остатков для белково-белкового взаимодействия
2. Валидация докинговых комплексов
3. Оптимизация специфических антител
4. Исследование мутационных эффектов на аффинность связывания

Пример запуска:

```
FoldX --command=AnalyseComplex --pdb=AC.pdb --analyseComplexChains=A,B
```

4.3 RDKit

RDKit[30] представляет собой современную open-source платформу для вычислительной химии и хемоинформатики, реализованную на языке Python. Данная библиотека предоставляет комплексный инструментарий для решения широкого круга задач молекулярного анализа

Ключевые возможности RDKit:

1. Работа с молекулярными структурами
    - Чтение/запись химических форматов
    - Генерация 2D/3D структур
2. Химические дескрипторы и фичи
    - Расчет физико-химических свойств
    - Fingerprints (числовые представления молекулярных структур)
3. Молекулярные модификации
    - Генерация производных структур
4. Интеграция с ML (Scikit-learn, TensorFlow)

Благодаря своей гибкости и производительности, RDKit стала стандартным инструментом в современных исследованиях на стыке химии и data science, что подтверждается ее активным использованием как в академических работах, так и в промышленных проектах.

4.4 Open Babel

Open Babel[31] — это кроссплатформенная *open-source* программа и библиотека, предназначенная для обработки, конвертации и анализа химических данных. Она поддерживает более 150 химических форматов и предоставляет инструменты для:

- Конвертации структур между форматами
- Расчетов молекулярных свойств
- Фильтрации и поиска в химических базах данных
- 3D-оптимизации структур
- Интеграции с другими пакетами (RDKit, PyMOL, AutoDock)

4.5 Моделирование мутаций

Взяв данные из файла PPB-Affinity.xlsx с помощью FoldX BuildModel я начал проводить все необходимые мутации. Около 30 мутаций FoldX не смог произвести из-за неподходящих для него pdb файлов в нашем датасете. Также порядка 80 мутаций мы не стали проводить, так как файл, по которому строились мутации(‘1KBH.pdb’) был гораздо больше остальных и получен путем ядерно-магнитного резонанса, в отличие от других комплексов.

4.6 Предсказания FoldX

В ходе исследования была проведена оценка точности предсказаний модуля AnalyseComplex пакета FoldX на тестовой выборке. Сравнение расчетных значений свободной энергии связывания с экспериментальными данными позволило получить следующие показатели:
RMSE: 9.672, MAE: 6.932, SRCC: 0.139, PCC: 0.138

Как правило FoldX лучше справляется с предсказанием аффинности связывания. Такое расхождение может быть объяснено особенностями анализируемого датасета, который содержит: белковые комплексы со сложной динамикой взаимодействия и структуры, содержащие неканонические взаимодействия

4.7 Препроцессинг данных

Обычно Open Babel способен сам конвертировать pdb файлы в mol2 формат, но после мутаций файлы меняются и Open Babel не способен корректно с ними работать. Поэтому для начала с помощью RDKit я считал файлы(`Chem.MolFromPDBFile()`). Затем оставил только указанные в таблице цепи

для обоих белков (`atom.GetPDBResidueInfo().GetChainId()`). Только после этого восстановил полную химическую структуру (`Chem.SanitizeMol(mol); mol.UpdatePropertyCache()`) и сохранил ее во временный pdb(`Chem.MolToPDBFile()`). Дальше с помощью Open Babel стандартным образом преобразовал эту структуру в mol2 файл. Но во время работы программы возник ряд ошибок:

1. При мутациях атомы MG, MN и ZN мутировали неудачно и пришлось самостоятельно поправить файлы, содержащие их
2. Некоторые pdb содержали атомы, имеющие большую валентность, чем возможно, что вызвало ошибку в работе RDKit.
3. Некоторые комплексы содержали атомы-заглушки, которые не смог прочитать RDKit.

В итоге порядка 150 комплексов пришлось исключить.

4.8 Обучение

4.8.1 Взрыв градиентов

Я начал обучение PAMNet на нашем датасете. Но через несколько эпох обучения в предсказанных значениях появилось Nan. Это было следствие невероятно большого роста градиентов. Изучив подробнее проблему я осознал, что причиной этого стало нереалистично маленькие расстояния между атомами в некоторых комплексах. Модель PAMNet опиралась на знание, что два атома не могут быть друг к другу ближе, чем примерно 1.2 ангстрема (это длина ковалентной связи). Но из-за неточности экспериментальных данных это условие не выполнялось и начался взрыв градиентов.

Тогда я решил удалить все файлы в которых минимальное расстояние между атомами было меньше 1 ангстрема. Таких файлов оказалось порядка 720.

4.8.2 Предсказания PAMNet

На оставшихся 8918 комплексах в трейн выборке я обучил модель и предсказал 1326 значений для теста. Сравнение расчетных значений с экспериментальными данными показало следующие метрики:
Train RMSE: 4.240, Train MAE: 3.090, Train SRCC: 0.552, Train PCC: 0.464
Test RMSE: 5.006, Test MAE: 4.040, Test SRCC: 0.118, Test PCC: 0.115

При дальнейшем обучении, метрики на трейне падали, а на тесте сначала практически не менялись, а затем начали расти, что говорит о переобучении модели.

После не особо удачного эксперимента я решил попробовать предобучить модель на датасете PDBbind и только после этого обучать ее на нашем датасете, предварительно удалив из тестовой выборки все наблюдения обучающей выборки для датасета PDBbind. Метрики остались практически такими же:
Train RMSE: 2.914, Train MAE: 1.482, Train SRCC: 0.790, Train PCC: 0.663
Test RMSE: 5.095, Test MAE: 4.138, Test SRCC: 0.126, Test PCC: 0.120

И снова модель очень рано начала переобучаться.
Итоговые метрики можно увидеть в таблице 2. На графиках 2 и 3 представлены зависимости предсказанных данных от экспериментальных для PAMNet без предобучения и с ним, соответственно.

модель	RMSE	MAE	SRCC	PCC
Базовая модель из статьи	1.994	-	0.298	0.329
FoldX	9.672	6.932	**0.139**	**0.138**
PAMNet	**5.006**	**4.040**	0.118	0.115
предобученный PAMNet	5.095	4.138	0.126	0.120

Таблица 2: метрики качества моделей на антитело-антиген комплексах

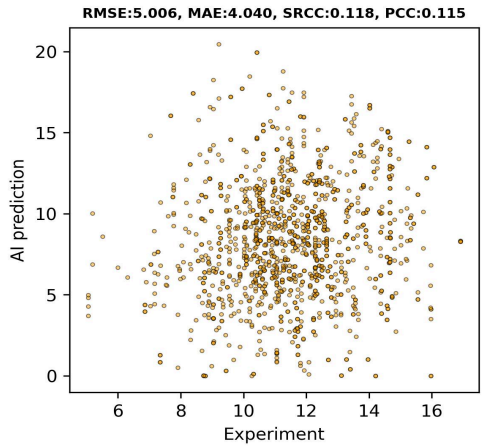

График 2: сравнение предсказанных значений с экспериментальными для PAMNet

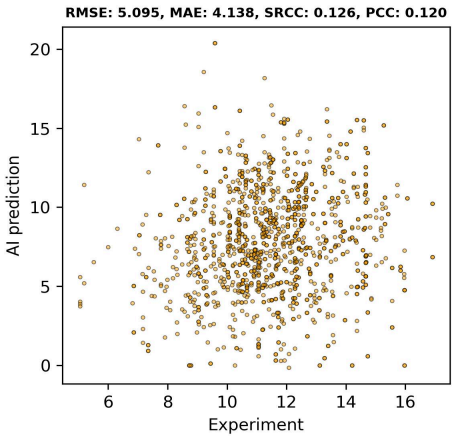

График 3: сравнение предсказанных значений с экспериментальными для предобученного PAMNet

4.8.3 Новое разбиение

В связи с ранним переобучением модели появилась теория, что антитело-антиген комплексы в тестовой выборке сильнее отличаются от оставшегося датасета, чем я ожидал, делая изначальное разбиение на трейн-тест выборки. Поэтому я решил сделать новое разбиение. Теперь я взял все pdb и 10% из них определил в тест выборку(как и все мутации, полученные из них). В итоге получили 9504 комплексов в трейне и 1191 - в тесте.

В таблице 3 можно увидеть, что для нового разбиения все метрики стали гораздо лучше. А на графиках 4 и 5 представлена зависимость предсказанных значений от экспериментальных на новой тестовой выборке для PAMNet без предобучения и с ним, соответственно.

модель	RMSE	MAE	SRCC	PCC
FoldX	10.730	7.439	-0.041	-0.037
PAMNet	**3.990**	**2.981**	**0.320**	**0.270**
предобученный PAMNet	4.166	2.937	0.191	0.149

Таблица 3: метрики качества моделей на случайной тестовой выборке.

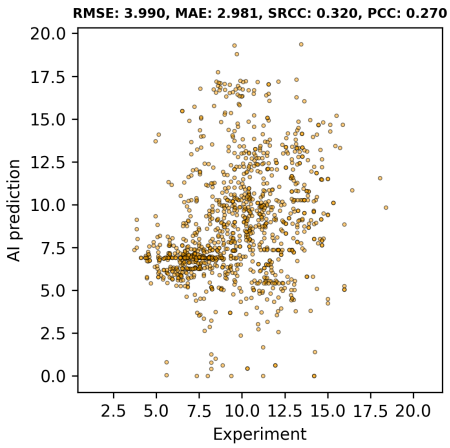

График 4: сравнение предсказанных значений с экспериментальными для PAMNet на новой тестовой выборке

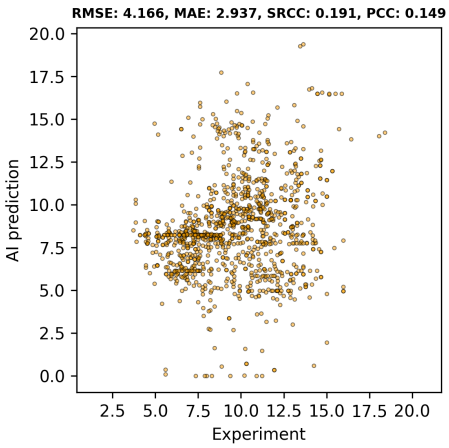

График 5: сравнение предсказанных значений с экспериментальными для предобученного PAMNet на новой тестовой выборке

К сожалению при новом разбиении также присутствует ярко выраженное переобучение модели. Видимо, многие комплексы обладают уникальными свойствами, и поэтому обучение на других комплексах слабо помогает в предсказании свободной энергии связывания на тесте.

Заключение

В ходе работы я пришел к следующим важным выводам:

- Предсказание свободной энергии связывания - важнейший инструмент в разработке лекарств и фундаментальных исследованиях. Достижения в этой области способны существенно повлиять на качество жизни людей и способствовать прогрессу в медицине и науке.
- Несмотря на то, что результаты обучения PAMNet на ppb-affinity далеко не идеальны, мы получили лучшие значения метрик на данный момент, на одном из самых крупных открытых датасетов содержащих белково-белковые взаимодействия.
- Мы показали, что сильное обучение на одних белково-белковых комплексах не гарантирует даже примерно таких же метрик на других комплексах.
- Модель PAMNet, которая очень хорошо предсказывала свободную энергию связи для белка и лиганда, показала результаты гораздо хуже для белково-белковых комплексов, что показывает насколько вторая задача сложнее первой.
- Многие стандартные инструментарии для работы с молекулярными комплексами не всегда работают корректно, из-за чего работа становится гораздо сложнее, чем могла бы быть.
- Нахождение критической ошибки в статье про PPB-Affinity.
- Загрузка почти всего датасета PPB-Affinity в открытый доступ[34], хотя раньше нужно было долго предобрабатывать данные, чтоб получить датасет
- Посмотреть код и более подробное пояснение к нему можно на моем github репозитории[35]

Список использованных литературных источников и информационных материалов

[1] A machine learning approach to predicting protein–ligand binding affinity with applications to molecular docking
ссылка
[2] XLPFE: A Simple and Effective Machine Learning Scoring Function for Protein−Ligand Scoring and Ranking
ссылка
[3] AtomNet: A Deep Convolutional Neural Network for Bioactivity Prediction in Structure-based Drug Discovery
ссылка
[4] Development and evaluation of a deep learning model for protein-ligand binding affinity prediction
ссылка
[5] OnionNet: a multiple-layer inter-molecular contact based convolutional neural network for protein-ligand binding affinity prediction
ссылка
[6] Learning Attributed Graph Representations with Communicative Message Passing Transformer
ссылка
[7] DIRECTIONAL MESSAGE PASSING FOR MOLECULAR GRAPHS
ссылка
[8] Structure-aware Interactive Graph Neural Networks for the Prediction of Protein-Ligand Binding Affinity
ссылка
[9] A Universal Framework for Accurate and Efficient Geometric Deep Learning of Molecular Systems
ссылка
[10] The PDBbind Database: Collection of Binding Affinities for Protein−Ligand Complexes with Known Three-Dimensional Structures
ссылка
[11] PPB-Affinity: Protein-Protein Binding Affinity dataset for AI-based protein drug discovery
ссылка
[12] Jankauskaite, J., Jiménez-García, B., Dapkunas, 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 35, 462–469 (2019).

[13] Schneider, C., Raybould, M. I. J. & Deane, C. M. SAbDab in the age of biotherapeutics: updates including SAbDab-nano, the nanobody structure tracker. Nucleic Acids Res 50, D1368–d1372 (2022).
[14] Dunbar, J. et al. SAbDab: the structural antibody database. Nucleic Acids Research 42, D1140–D1146 (2013).
[15] Raybould, M. I. J. et al. Tera-SAbDab: the Terapeutic Structural Antibody Database. Nucleic Acids Research 48, D383–D388 (2019).
[16] Wang, R., Fang, X., Lu, Y. & Wang, S. Te PDBbind database: collection of binding afnities for protein-ligand complexes with known three-dimensional structures. Journal of medicinal chemistry 47 (2004).
[17] Liu, Z. et al. Forging the Basis for Developing Protein-Ligand Interaction Scoring Functions. Accounts of chemical research 50, 302–309 (2017).
[18] Liu, Z. et al. PDB-wide collection of binding data: current status of the PDBbind database. Bioinformatics (Oxford, England) 31, 405–412 (2015).
[19] Li, Y. et al. Comparative assessment of scoring functions on an updated benchmark: 1. Compilation of the test set. Journal of chemical information and modeling 54, 1700–1716 (2014).
[20] Cheng, T., Li, X., Li, Y., Liu, Z. & Wang, R. Comparative assessment of scoring functions on a diverse test set. Journal of chemical information and modeling 49 (2009).
[21] Wang, R., Fang, X., Lu, Y., Yang, C. Y. & Wang, S. Te PDBbind database: methodologies and updates. Journal of medicinal chemistry 48, 4111–4119 (2005).
[22] Guest, J. D. et al. An expanded benchmark for antibody-antigen docking and afnity prediction reveals insights into antibody recognition determinants. Structure 29, 606–621.e605 (2021).
[23] Kastritis, P. L. et al. A structure‐based benchmark for protein–protein binding afnity. Protein Science 20, 482–491 (2011).
[24] Vreven, T. et al. Updates to the Integrated Protein-Protein Interaction Benchmarks: Docking Benchmark Version 5 and Afnity Benchmark Version 2. J Mol Biol 427 (2015).
[25] Borrman, T. et al. ATLAS: A database linking binding afnities with structures for wild-type and mutant TCR-pMHC complexes. Proteins 85, 908–916 (2017).
[26] Selective amplification and sequencing of cyclic phosphate–containing RNAs by the cP-RNA-seq method
ссылка
[27] https://foldxsuite.crg.eu/
[28] BuildModel | FoldX
[29] AnalyseComplex | FoldX
[30] Landrum G (2016) Rdkit: Open-source cheminformatics software
официальная документация
[31] Open Babel: An open chemical toolbox
ссылка
официальная документация
[32] PPB-Affinity.xlsx
[33] Training sample for 'checkpoints/results.csv' · Issue #6 · ChenPy00/PPB-Affinity
[34] PPB-Affinity
[35] https://github.com/ErikBag/diplom