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

ПРИМЕНЕНИЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ПРЕДСКАЗАНИЯ РАСТВОРИМОСТИ МАЛЫХ МОЛЕКУЛ

Курсовая работа

Снежко Льва Владимировича
студента 3 курса, специальность «информатика»
Научный руководитель:
Карпенко А.Д.

Минск 2025

Оглавление

Введение 3

Глава 1. Постановка задачи. Основные определения и понятия. Краткий обзор основных свойств малых молекул 4

1.1. Основные определения и понятия 4

1.2. Свойства малых молекул 4

1.3. ADMET 4

1.4. Липофильность 6

1.5. Обоснование актуальности задачи предсказания свойств малых молекул. 7

Глава 2. Краткий обзор методов машинного обучения для решения задачи предсказания свойств малых молекул. 9

2.1 Графовые нейронные сети (GNN) 9

2.2 Стандартные трансформеры 9

2.3. BERT 10

2.4. Классические методы 11

Глава 3. Постановка целей и задач практической работы. Обоснование выбора средств для достижения поставленных целей. Выбор оцениваемых моделей и датасета 12

3.1 Постановка задачи практической части 12

3.2. Выбор моделей для сравнительного анализа и его обоснование. 13

3.2.1. BERT 13

3.2.2. GCN 15

3.2.3. Random Forest 16

3.3. Выбор датасета и его обоснование. 17

3.4. Обоснование выбора метрики оценки эффективности 19

Глава 4. Анализ полученных результатов. 21

4.1. BERT 21

4.2. GCN 22

4.3. Random Forest 22

4.4. Проведение сравнительного анализа 23

Заключение 25

Список литературы 27

Введение

Предсказание свойств малых молекул – крайне актуальная на данный момент задача, поскольку на основании этих свойств можно отыскать закономерность, характерную для лекарственных препаратов. Следовательно, решив эту задачу, мы получим возможность определять потенциальные лекарственные препараты, определять их назначение и эффективность. А, как известно, методы машинного обучения (МО) достаточно эффективно решают задачу поиска закономерностей, неочевидных для человека. Таким образом, применение методов машинного обучения для предсказания свойств малых молекул и последующего определения эффективности применения молекулярного соединения в области создания лекарственных препаратов является естественным и, как будет показано далее, эффективным шагом.

Цель работы – провести сравнительный анализ эффективности различных методов машинного обучения в решении задачи предсказания свойств малых молекул и объяснить полученные результаты.

Задачи работы:

- Обоснование актуальности задачи предсказания свойств малых молекул
- Поиск и выборка различных методов МО для сравнительного анализа
- Проверка эффективности моделей для предсказания свойств малых молекул (например, растворимости) на едином датасете.
- Сравнительный анализ результатов различных моделей.
- Трактовка результатов и подведение итогов.

Суть проделанной работы заключается в анализе проблем создания лекарственных препаратов, установлении важности решения задачи предсказания свойств малых молекул, выборке моделей МО различных архитектур, сравнении их эффективности в решении задач предсказания свойств малых молекул, сравнительном анализе и подведении итогов, трактовке и объяснению полученного результата.

В процессе работы возникли трудности, связанные с недостатком ресурсов, поскольку некоторые модели, реально использующиеся для решения этой задачи либо не находятся в открытом доступе, либо требуют внушительных вычислительных ресурсов для функционирования. В целях получения наиболее полной и реалистичной статистики для сравнительного анализа выборка экспериментальных моделей была составлена из открытых аналогов моделей, реально использующихся для решения задачи предсказания свойств малых молекул, и реализаций классических методов МО, не предназначенных специально для решения указанной задачи.

Глава 1. Постановка задачи. Основные определения и понятия. Краткий обзор основных свойств малых молекул

1.1. Основные определения и понятия

Научный и практический интерес к малым молекулам и их свойствам обусловлен тем фактом, что большинство лекарств являются малыми молекулами.

Главная особенность малых молекул – достаточно низкий верхний предел молекулярной массы, что позволяет этим молекулам быстро проникать сквозь липидный бислой клеточной мембраны и достигать внутриклеточных мишеней. По сравнению с белками они обладают большей биодоступностью (т.е. большее количество молекул способно пройти сквозь клеточную мембрану к мишени). Малые молекулы могут служить *передатчиками сигнала,* лекарствами, удобрениями, пестицидами и проч.

1.2. Свойства малых молекул

В этом пункте проведён краткий обзор основных и наиболее интересных свойств малых молекул.

Молекулярная масса (Molecular Weight, MW) – меньшая масса способствует увеличению биодоступности.

Растворимость в воде (aqua solubility) – важна для абсорбации и доставки в кровоток.

Липофильность (Lipophilicity, жирорастворимость, гидрофобность) – влияет на биодоступность и растворимость. Измеряется как логарифм коэффициента распределения между октанолом и водой.

Мембранная проницаемость (permeability) – характеризует способность молекулы проходить через биологические барьеры.

ADMET (Absorption, Distribution, Metabolism, Excretion and Toxicity, абсорбация, распределение, метаболизм, выведение и токсичность) – комплекс критериев, описывающих положение соединения в организме.

Полярность (PSA) – число водородных связей. Высокое значение этого параметра может снизить биодоступность.

Избирательность (selectivity) – параметр, определяющий активность взаимодействия молекулы с другими соединениями, помимо мишени. Влияет на эффективность и наличие побочных эффектов

В этой главе рассмотрены важнейшие из этих свойств.

1.3. ADMET

ADMET – комплекс критериев, описывающих положение лекарственного препарата в организме. Все четыре критерия влияют на кинетику и уровень

воздействия лекарства на организм, а значит, влияют на эффективность и активность выбранного соединения как лекарственного препарата.

Абсорбация (absorbation)

Абсорбация – это характеристика процесса перемещения лекарственного препарата от места введения к месту действия. Это важное свойство лекарственного препарата, поскольку оно непосредственно влияет на скорость воздействия препарата на организм. Такие факторы, как плохая растворимость соединения, время опорожнения желудка, время прохождения через кишечник, химическая нестабильность в желудке и неспособность проникать через кишечную стенку, могут снизить степень всасывания препарата после перорального приема. Абсорбция в решающей степени определяет биодоступность соединения.

Распределение (distribution)

Распределение описывает скорость переноса лекарственного препарата из одних органов в другие. Каждый орган и ткань получат различные дозы препарата и препарат может находиться в органах или тканях различное время. Распределение зависит от проницаемости сосудов, от способности препарата связывать белки с плазмой, его растворимости в липидах и pH. Препарат способен перемещаться из плазмы в ткань до тех пор, пока между ними не установится равновесие.

Некоторые факторы, влияющие на распределение лекарств, включают скорость регионарного кровотока, размер молекулы, полярность и связывание с белками сыворотки, образующими комплекс. Распределение может стать серьёзной проблемой для некоторых естественных барьеров, таких как гематоэнцефалический барьер.

Метаболизм (metabolism)

Метаболизм (обмен веществ) – процесс расщепления и синтеза молекул для получения энергии. Особенности метаболизма определяют, будет ли использована данная молекула для получения энергии. В процессе метаболизма исходное (родительское) соединение преобразуется в новые соединения, называемые метаболитами. Когда метаболиты фармакологически инертны, с помощью метаболизма дезактивируется введенная доза исходного препарата, что обычно снижает его воздействие на организм. Метаболиты также могут быть фармакологически активными, иногда более активными, чем исходный препарат

Выведение (excretion)

Выведение – характеристика процесса экскрекции, в результате которого лекарственные препарат выводится из организма. Если выведение не

завершено, накопление чужеродных веществ может негативно повлиять на нормальный обмен веществ.
Существует три основных органа, через которые происходит выведение лекарств. Почки являются наиболее важным органом, с помощью которого продукты метаболизма выводятся с мочой. Билиарная экскреция или фекальная экскреция — это процесс, который начинается в печени и проходит через кишечник, пока препараты окончательно не выводятся вместе с отходами жизнедеятельности или калом. Последний основной способ выделения — через легкие (например, анестезирующие газы).

Токсичность (Toxity)

Токсичность – характеристика, обратно пропорциональная абсолютному значению среднесмертельной дозы или концентрации ядовитого вещества. Параметры, используемые для характеристики токсичности, включают среднюю летальную дозу и терапевтический индекс.

1.4. Липофильность

Липофильность (гидрофобность) – свойство вещества, характеризующее химическое сродство с органическими веществами. Это свойство крайне важно, поскольку оно может влиять на все пять критериев ADMET [6]
Характеризуется коэффициентом липофильности, который вычисляется по формуле:

$$P = \frac{[solute]_{octanol}}{[solute]_{water}}$$

где $[solute]_{octanol}$ - растворимость соединения в октаноле
а $[solute]_{water}$ – растворимость в воде.
Часто на практике используют не сам коэффициент растворимости, а его логарфим: $logP = log_{10}P$.
При logP > 0 соединение считается липофильным, а при logP < 0 – гидрофильным.
Оптимальным диапазоном коэффициента липофильности для лекарственных препаратов (по Липински) считается $logP \in [0,3]$.

Методы определения logP

Величина logP определяется двумя видами методов: экспериментальными и вычислительными.
Существует большое разнообразие экспериментальных методов [6], анализ которых выходит за рамки тематики данной работы. Однако стоит отметить, что экспериментальные методы отличаются от вычислительных повышенной

точностью хотя и проигрывают в универсальности применения и эффективности.
Наибольший интерес в рамках этой работы представляют вычислительные методы (in silico), а именно определение коэффициента липофильности методами машинного обучения. На ранних стадиях открытия лекарств методы in silico очень полезны для отбора соединений, похожих на лекарства. Однако как можно скорее прогнозируемые значения должны быть заменены более точными измеренными величинами.

1.5. Обоснование актуальности задачи предсказания свойств малых молекул.

Актуальность задачи предсказания свойств малых молекул (small molecules property prediction) обусловлена её центральным значением в фармацевтике, медицине, химии и смежных науках.
Решение этой задачи позволяет ускорить разработку лекарственных препаратов, сократить число синтезируемых соединений, а также исключить неперспективные молекулы на ранних этапах исследования [10]. Под неперспективными молекулами подразумеваются молекулы с неудовлетворительной фармакинетикой, низкой биодоступностью и высокой токсичностью. Благодаря возможности применения методов машинного обучения стало возможно исключить неперспективные соединения на самых ранних этапах разработки и молекулярного дизайна.
В результате стремительного развития методов машинного обучения и взрывного роста уровня доступности данных появилась возможность решения этой задачи и многих других задач смежных областей методами МО.
Также не следует забывать об этической стороне вопроса. In silico методы позволяют уменьшить потребность в in vitro и in vivo испытаниях, что способствует соблюдению принципа 3Rs (Replacement, Reduction, Refinement) и снижению этической нагрузки [11].

Решение задачи предсказания свойств малых молекул имеет широкое практическое применение. Помимо разработки лекарственных препаратов оно может быть использовано для ускорения разработки новых материалов (полимеров), пестицидов и инсектоцидов, а также электролитов и новых видов топливных элементов. Подобная технология может принести пользу в области экотоксикологии для оценки воздействия на окружающую среду [12].
Модели могут быть адаптированы для предсказания того, как конкретные молекулы взаимодействуют с биологическими мишенями у разных пациентов, что способствует развитию персонализированных подходов к терапии [13]. Для болезней с малым коммерческим интересом (например, редкие заболевания или

устойчивые инфекции), in silico подходы помогают снизить порог входа и сделать разработку более доступной и быстрой[14].
Модели предсказания свойств полезны не только для новых молекул, но и для модификации известных соединений с целью улучшения их профиля (например, повышение растворимости или снижение токсичности) [15].
Предсказание свойств становится важной частью автоматизированного drug discovery pipeline, где синтез, тестирование и отбор соединений происходят с участием роботов, ИИ и предсказательных моделей.

Глава 2. Краткий обзор методов машинного обучения для решения задачи предсказания свойств малых молекул.

2.1 Графовые нейронные сети (GNN)

Графовые нейронные сети – нейронные сети, разработанные для решения задач, входными данными для которых являются графы.

В случае задачи молекулярного дизайна лекарств каждый входной образец представляет собой графическое представление молекулы, в котором каждый атом соответствует вершине, а химическая связь – ребру. В дополнение к графическому представлению входные данные также включают известные химические свойства для каждого из атомов. Таким образом, образцы наборов данных могут отличаться по длине, отражая различное количество атомов в молекулах и различное количество связей между ними. Задача состоит в том, чтобы предсказать эффективность данной молекулы для конкретного медицинского применения, например, устранения бактерий. Ключевым элементом дизайна GNN является использование попарной передачи сообщений , так что узлы графа итеративно обновляют свои представления, обмениваясь информацией со своими соседями.

По состоянию на 2022 год остается открытым вопрос, возможно ли определить архитектуры GNN, «выходящие за рамки» передачи сообщений, или вместо этого каждая GNN может быть построена на передаче сообщений по соответствующим образом определенным графам [7].

2.2 Стандартные трансформеры

Трансформер – архитектура глубоких нейросетей, предназначенных для обработки последовательностей (например, текста). Преимущество трансформеров заключается в отсутствии необходимости обработки последовательности по порядку, что облегчает параллельную обработку вычислений и увеличивает скорость обучения.

Архитектура трансформера состоит из кодировщика и декодировщика. Кодировщик получает на вход векторизованую последовательность с позиционной информацией. Декодировщик получает на вход часть этой последовательности и выход кодировщика. Кодировщик и декодировщик состоят из слоев. Слои кодировщика последовательно передают результат следующему слою в качестве его входа. Слои декодировщика последовательно передают результат следующему слою вместе с результатом кодировщика в качестве его входа.

Каждый кодировщик состоит из механизма самовнимания (вход из предыдущего слоя) и нейронной сети с прямой связью (вход из механизма самовнимания). Каждый декодировщик состоит из механизма самовнимания

(вход из предыдущего слоя), механизма внимания к результатам кодирования (вход из механизма самовнимания и кодировщика) и нейронной сети с прямой связью (вход из механизма внимания).
Для использования трансформеров в целях решения задач молекулярного дизайна лучшей практикой считается использование строк в формате SMILES. Упрощенная система ввода строк молекулярных данных (SMILES) — это спецификация в форме линейной нотации для описания структуры химических веществ с использованием коротких строк ASCII .
С точки зрения графовой вычислительной процедуры SMILES — это строка, полученная путем печати узлов символов, встречающихся при обходе дерева в глубину химического графа . Химический граф сначала обрезается для удаления атомов водорода, а циклы разрываются, чтобы превратить его в остовное дерево . Там, где циклы были разорваны, включаются числовые суффиксные метки для указания связанных узлов. Скобки используются для указания точек ветвления на дереве.
С точки зрения формальной теории языка, SMILES — это слово. SMILES можно анализировать с помощью контекстно-свободного анализатора. Использование этого представления было в прогнозировании биохимических свойств (включая токсичность и биоразлагаемость) на основе основного принципа хемоинформатики, что схожие молекулы имеют схожие свойства.

2.3. BERT

Отдельного внимания в рамках настоящей работы заслуживает одна из разновидностей трансформера.
BERT (Bidirectional Encoder Representations from Transformers) — модель на основе энкодерной части архитектуры трансформера**,** обучаемая в маскированной языковой модели. Основная идея: предсказание замаскированных токенов на основе контекста с обеих сторон.
BERT (Bidirectional Encoder Representations from Transformers) — модель на основе энкодерной части архитектуры трансформера**,** обучаемая в маскированной языковой модели. Основная идея: предсказание замаскированных токенов на основе контекста с обеих сторон [15].
На высоком уровне архитектура BERT состоит из следующих компонентов:

1. Tokenizer – преобразует исходный текст в токены
2. Embedding - преобразует последовательность токенов в массив векторов с действительными значениями, представляющих токены.
3. Encoder
4. Task head - преобразует окончательные векторы представления в one-hot закодированные токены, создавая предсказанное распределение вероятностей по типам токенов.

Использование BERT для решения задач молекулярного дизайна и предсказания свойств малых молекул имеет ряд преимуществ. BERT учитывает контекст с обеих сторон, что улучшает захват химической семантики. Предобучение позволяет модели обобщать между различными корректными представлениями одной молекулы. Механизм self-attention может использоваться для анализа фрагментов молекулы, важных для предсказания. Однако даже для этой архитектуры характерен ряд недостатков. Например, разные SMILES-версии одной молекулы могут затруднять обобщение без аугментаций. BERT работает только со SMILES и не учитывает пространственную конфигурацию молекулы. Обучение трансформера с нуля требует значительных ресурсов. Также нельзя исключать риск того, что химически значимые фрагменты могут быть разделены на неестественные токены.

2.4. Классические методы

В некоторых случаях задачи молекулярного дизайна решаются при помощи классических методов МО, таких как многослойный перцептрон, случайный лес и XGBoost.

Многослойный перцептрон(MLP) состоит из формальных нейронов или узлов и связей (весов) между ними. В архитектуре MLP нейроны организованы в слои (входной слой, один или несколько скрытых слоев и выходной слой), и связи являются однонаправленными от входа к выходу. Смежные слои полностью связаны, но между нейронами внутри одного слоя не существует связей. Эта архитектура вычисляет числовое выходное значение $f(x)$ для заданного числового входного вектора x , который является строкой матрицы X, соответствующей заданному объекту (молекуле, виду и т. д.). Формальный нейрон суммирует входящие сигналы, умноженные на веса связей, вычитает пороговое значение (или смещение θ) и вычисляет выходной сигнал, используя так называемую передаточную функцию. Нейроны могут иметь разные передаточные функции. Входные нейроны просто распределяют данные дескриптора по нейронам скрытого слоя без каких-либо дальнейших вычислений [9].

Алгоритм градиентного бустинга (XGBoost) — это интегрированный алгоритм машинного обучения, основанный на деревьях решений, использующий структуру градиентного подъема, подходящий для задач классификации и регрессии, и используемый для решения задач контролируемого обучения. Ансамблевое обучение относится к построению нескольких слабых классификаторов для прогнозирования набора данных, а затем использования определенной стратегии для интеграции ожидаемых результатов нескольких классификаторов в качестве окончательного результата прогнозирования.

Глава 3. Постановка целей и задач практической работы. Обоснование выбора средств для достижения поставленных целей. Выбор оцениваемых моделей и датасета.

3.1 Постановка задачи практической части.

Целью практической части настоящей работы является оценка эффективности различных методов машинного обучения в решении задачи предсказания свойств малых молекул.

В рамках практической части настоящей работы были выполнены следующие задачи:

1. Составлена выборка из моделей - представителей популярных архитектур машинного обучения. Главными критериями отбора были:
    a. разнообразность архитектур (для усиления объективности сравнительного анализа)
    b. актуальность конкретной архитектуры (при отборе больший приоритет имели модели, реально использующиеся для решения задачи предсказания свойств малых молекул или их открытые аналоги)
    c. открытость
2. Выбран датасет для независимой и объективной оценки эффективности выбранных моделей. В процессе выполнения этой задачи составлена выборка подходящих датасетов, основными критериями которых являлись:
    a. открытость (свободный доступ)
    b. универсальность (работу над датасетом могла осуществлять любая модель вне зависимости от архитектуры)
    c. соответствие тематике настоящей работы (датасет должен содержать параметры молекул и их свойства)

    Над полученной выборкой проведено дополнительное исследование в формате сравнительного анализа и в результате этого исследования выбран единый датасет для оценки эффективности всех моделей, полученных в задаче 1.
3. Проведен анализ датасета, выбранного в результате выполнения задачи 2 и дополнительное исследование в формате сравнительного анализа для выбора метрик оценки эффективности моделей.
4. Составлен механизм дообучения, оценки эффективности модели по метрике, выбранной в результате выполнения задачи 3, на датасете, выбранном в результате выполнения задачи 2, и добавления результатов этой оценки в общую статистику.

5. Осуществлено применение механизма, полученного в результате выполнения задачи 4, к каждой модели из выборки, полученной в результате выполнения задачи 1. Сохранение общей статистики.
6. Проведен анализ результатов, выраженных статистикой, полученной в результате выполнения задачи 5. Выполнено подведение итогов, проведение сравнительного анализа оценок различных моделей и объяснение полученных результатов в контексте сравнения различных архитектур и методик машинного обучения.

3.2. Выбор моделей для сравнительного анализа и его обоснование.

Для проведения объективного сравнительного анализа различных методов машинного обучения крайне важным условием является составление широкой выборки моделей, отражающих полноту разнообразия архитектур и подходов машинного обучения.

При отборе моделей автор настоящей работы руководствовался следующими критериями:

1. Модели должны быть различных архитектур. Составление выборки из моделей одной и той же (или близких) архитектур даёт повод усомниться в объективности проведенного сравнительного анализа, а значит, и всего исследования, проведенного в рамках настоящей работы. От подобных ограничений следовало избавиться на самых ранних этапах проведения исследования.
2. Доступность. В рамках настоящей работы в сравнительном анализе участвуют лишь модели, находящиеся в свободном доступе и не имеющие чрезвычайно высоких требований к объему вычислительных ресурсов. Это связано с ограничениями, в рамках которых находится автор настоящей работы.
3. Актуальность. При отборе моделей больший приоритет уделялся тем кандидатам, которые имели наибольшее сходство с моделями, реально использующимися в науке и промышленности для решения задачи предсказания свойств малых молекул, но только если их выбор не противоречил критериям 1 и 2.

В результате проведенного исследования были отобраны следующие архитектуры:

3.2.1. BERT

В качестве представителя архитектуры BERT была выбрана модель ChemBERTa.

ChemBERTa - это модель, действующая по принципу архитектуры Transformer и адаптированная для обработки химических молекул, представленных в виде строк SMILES (Simplified Molecular Input Line Entry System). Модель основана на архитектуре RoBERTa (Robustly Optimized BERT Pretraining Approach), модифицированной для задач хемоинформатики [17].
Модель ChemBERTa использует энкодерную часть трансформера, подобно оригинальному BERT, и обучается по задаче Masked Language Modeling (MLM). Основные компоненты модели включают:

- Tokenizer: модифицированный byte-level BPE-токенизатор, специально адаптированный для химической нотации SMILES.
- Энкодер: несколько слоёв самовнимания (multi-head self-attention), реализованных в рамках архитектуры RoBERTa.

Целевая функция обучения ChemBERTa формулируется как:

$$\mathcal{L}_{\mathcal{MLM}} = -\sum_{i \in \mathcal{M}} \log P\left(x_i \mid x_{\setminus i}\right)$$

где $\mathcal{M}$ - множество маскированных токенов
x_i - истинный токен
$x_{\setminus i}$ - последовательность с маскированным токеном x_i
$P\left(x_i \mid x_{\setminus i}\right)$ – вероятность восстановить маскированный токен по контексту
Модель обучается на крупной выборке молекул из базы ZINC15, содержащей около 10 миллионов молекул. Во время обучения случайные токены маскируются, и задача модели — предсказать их на основе контекста.
После предобучения модель может быть дообучена (fine-tuned) на конкретных задачах предсказания молекулярных свойств — как регрессионных, так и классификационных.
Эта модель была рассмотрена в рамках сравнительного анализа по следующим причинам:

- Модель активно применяется в научных целях, продемонстрировала конкурентоспособность на задачах регрессии и классификации свойств малых молекул
- Способствует архитектурной полноте выборки и вносит требуемое разнообразие
- Модель и обученные веса находятся в свободном доступе на Hugging Face.
- Применяется к молекулам в формате SMILES, что исключает необходимость в 3D-геометрии или квантово-химических расчётах.

- Может быть дообучена и применена на стандартной CPU-машине, что критически важно в условиях ограниченных вычислительных ресурсов.

Модель ChemBERTa представляет собой современное, доступное и архитектурно уникальное решение для предсказания свойств малых молекул. Её включение в выборку моделей для сравнительного анализа усиливает научную объективность исследования за счёт архитектурной разнородности, повышает прикладную значимость за счёт актуальности подхода в научной и промышленной практике и обеспечивает воспроизводимость экспериментов в условиях ограниченных ресурсов.

3.2.2. GCN

GCNPredictor — это модель графовой нейронной сети (GNN), реализованная в библиотеке GDL-LifeSci, специально разработанной для биоинформатики и хемоинформатики. Она основана на Graph Convolutional Network (GCN) [18]. GCNPredictor моделирует молекулы как неориентированные графы, где узлы (nodes) соответствуют атомам, а рёбра (edges) — химическим связям. Модель обучается предсказывать молекулярные свойства, обобщая информацию, проходящую через структуру молекулярного графа.
Основные компоненты GCNPredictor:

1. Graph Convolutional Layers: множественные слои GCN осуществляют агрегацию информации о соседях каждого узла:

$$h_i^{(l+1)} = \sigma\left(\sum_{j \in \mathcal{N}(i) \cup \{i\}} \frac{1}{\sqrt{d_i d_j}} W^{(l)} h_j^{(l)}\right)$$

где $h_i^{(l)}$ - вектор признаков узла i на слое l,
$\mathcal{N}(i)$ - множество соседей узла i,
d_i – степень узла i,
$W^{(l)}$ - обучаемая матрица весов слоя l,
σ – нелинейная функция активации (например, ReLU

2. Readout-функция. После свёрток производится агрегация узловых векторов в вектор молекулы (глобальное представление графа), с помощью mean/sum/max pooling или learnable attention pooling.
3. MLP head. Глобальный вектор подаётся на полносвязную сеть (MLP), которая и производит финальное предсказание.

В зависимости от задачи (регрессия или классификация), используется:

Регрессия (например, предсказание растворимости):

$$\mathcal{L}_{MSE} = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2$$

Классификация (например, токсичность):

$$\mathcal{L}_{BCE} = -\sum_{i=1}^{N}(y_i \log \hat{y}_i + (1 - y_i)\log(1 - \hat{y}_i))$$

GCNPredictor является лучшим кандидатом для нашей выборки в своём классе по ряду причин. GCNPredictor основан на графовой нейросетевой архитектуре, принципиально отличной от трансформеров (как ChemBERTa) и моделей на основе SMILES (например, RNN или CNN). Это делает модель ценной для сравнения с иными подходами, так как она опирается на структурное представление молекулы**.**

GCN-модели активно используются в современных публикациях для предсказания биологических и физико-химических свойств [18]. Подход графовых нейросетей показывает высокую точность на стандартных датасетах (Tox21, ESOL, HIV и др.). GCN лежит в основе более сложных GNN-моделей (GAT, MPNN, DMPNN), поэтому её включение позволяет заложить базовый уровень сравнения в классе GNN.

Модель dgllife.model.GCNPredictor входит в открытое ПО DGL-LifeSci. Код и примеры обучения доступны на GitHub. Не требует особых вычислительных ресурсов - обучение возможно на CPU при работе с небольшими датасетами. Модель широко применяется в исследовательских статьях и соревнованиях по химическому ML. Обучение и воспроизведение результатов может быть проведено с использованием общедоступных молекулярных баз.

Таким образом включение GCNPredictor в выборку моделей усиливает архитектурное разнообразие, обеспечивает реалистичное сравнение с GNN-классом моделей, соответствует современным научным и прикладным практикам в предсказании свойств малых молекул и при этом не нарушает критерии доступности и воспроизводимости, что делает эту модель оптимальным кандидатом.

3.2.3. Random Forest

Random Forest (RF) — ансамблевый алгоритм машинного обучения, представляющий собой комбинацию решающих деревьев, обученных на случайных подмножествах признаков и объектов. Он остается одним из

наиболее распространённых и надёжных методов для решения задач регрессии и классификации в химоинформатике и смежных областях.
Random Forest не относится к нейросетевым архитектурам и тем более к графовым моделям. Это классический алгоритм из семейства ансамблевых деревьев решений, что позволяет расширить архитектурное разнообразие выборки. Его наличие важно для избежания архитектурной однородности.
Эта модель активно применяется для предсказания ADMET-свойств, токсичности, биоактивности молекул и других свойств малых молекул [19], применима к дескрипторам молекул (например, ECFP, MACCS, RDKit-фингерпринты), которые можно извлекать без графовой структуры, что даёт возможность проводить аналитически интерпретируемые эксперименты.
Random Forest не требует обучения на GPU или большого объема оперативной памяти. Это делает его удобным выбором в условиях ограниченных вычислительных ресурсов, соответствуя критерию доступности. RF имеет небольшое число гиперпараметров и предоставляет удобные средства анализа важности признаков. Это полезно при анализе влияния различных молекулярных дескрипторов на результат.
RF часто используется как базовая модель, с которой сравниваются более сложные архитектуры. Включение её в исследование обеспечивает точку отсчета и позволяет оценить, насколько архитектуры типа GNN или трансформеров действительно превосходят классические методы. Благодаря бутстрапированию и усреднению предсказаний, RF демонстрирует хорошую обобщающую способность, особенно при небольших объемах данных.
Также использование RF позволяет выполнять воспроизводимые эксперименты, что особенно важно в рамках практической части настоящей работы. RF устойчив к несбалансированным классам (в задачах классификации), особенно с применением техник взвешивания классов.
Для оценки эффективности и проведения сравнительного анализа будет использоваться реализация Random Forest из библиотеки sklearn.

3.3. Выбор датасета и его обоснование.

В результате дополнительного исследования и проведенного анализа было принято решение в рамках практической части настоящей работы использовать датасет ESOL.
Датасет ESOL (Estimated Solubility) представляет собой широко используемую открыто доступную коллекцию экспериментальных данных, отражающих водную растворимость малых органических молекул. Впервые упомянут в работе «ESOL: Estimating Aqueous Solubility Directly from Molecular Structure» J.S.Delaney в 2004 г.и с тех пор занял прочное место в хемоинформатике как эталонный набор данных для решения задач регрессионного типа [4]. Основной

предсказываемой характеристикой в данном датасете является логарифм растворимости молекул в воде (logS), выраженный в логарифмической шкале концентрации (молярность). Значения logS были получены на основе экспериментальных измерений и охватывают широкий диапазон растворимостей, что обеспечивает репрезентативность выборки. [4]
В датасет входит 1128 малых молекул, каждая из которых представлена в формате SMILES, что позволяет эффективно конвертировать данные как для классических методов машинного обучения (например, через векторизацию дескрипторов), так и для современных подходов, основанных на графовых нейронных сетях и трансформерах. Такой формат делает датасет универсальным: он может быть использован с моделями любых архитектур — от случайных лесов до BERT-подобных моделей и графовых нейросетей. Дополнительно к строкам SMILES предоставляются вычисленные молекулярные дескрипторы, включая молекулярную массу, число водородных доноров и акцепторов, площадь полярной поверхности (TPSA) и другие параметры, что позволяет расширить пространство признаков для моделей, не использующих явное представление структуры молекулы.
Таким образом, ESOL отвечает ряду ключевых требований. Он входит в состав MoleculeNet и поддерживается библиотеками вроде DeepChem), может быть использован с различными типами моделей) и представляет актуальное физико-химическое свойство, напрямую влияющее на биодоступность молекул. Благодаря умеренному объёму данных (около одной тысячи наблюдений) датасет также может использоваться в условиях ограниченных вычислительных ресурсов, что делает его особенно удобным для практических экспериментов и сравнительных исследований.
Таким образом выбор датасета ESOL полностью соответствует методологическим требованиям настоящего исследования, а именно гарантирует архитектурную независимость применения, предоставляет реальные и значимые молекулярные свойства, обладает открытым и стандартизированным форматом и обеспечивает воспроизводимость экспериментов и сравнимость с результатами предыдущих экспериментов.
Достоверность данных в датасете ESOL подтверждается как его происхождением, так и методом составления, основанным на проверенных источниках и научной методологии. Растворимость молекул (в логарифмической шкале — logS) была собрана автором из экспериментальных публикаций, химических справочников и научной литературы, например, «Handbook of Aqueous Solubility Data»[20] и экспериментальные базы данных в фармацевтической и агрохимической промышленности. Молекулы, содержащие нестабильные фрагменты, сильно ионизированные формы или редкие атомы, были исключены. В датасет вошли только органические малые

молекулы, пригодные для моделирования. Все молекулы были приведены к формату SMILES для универсального описания структуры. Это дало возможность использовать данные для самых разных подходов: от QSAR-моделей до графовых и трансформерных архитектур. Дополнительно к экспериментальным значениям logS, Delaney рассчитал набор простых молекулярных дескрипторов (молекулярный вес, число акцепторов/доноров водородных связей, TPSA и т.д.), что позволило создать базовую регрессионную модель на основе линейной комбинации признаков (также приведённую в статье как бенчмарк). Все значения logS были перепроверены вручную, чтобы исключить ошибки ввода или несоответствие структур. Таким образом, данные прошли экспертную валидацию на этапе подготовки.
Таким образом в настоящей работе для сравнительного анализа эффективности различных подходов машинного обучения будет использоваться современный датасет, отвечающий тематике исследования и подходящий для использования моделями любой архитектуры в условиях ограниченных вычислительных ресурсов.

3.4. Обоснование выбора метрики оценки эффективности

В рамках настоящего исследования для количественной оценки качества моделей, решающих задачу предсказания свойств малых молекул, используются две широко признанные регрессионные метрики: среднеквадратичная ошибка (RMSE) и коэффициент детерминации (R^2). Обоснование выбора именно этих метрик базируется как на их теоретических свойствах, так и на практике их использования в современной научной литературе по хемоинформатике и машинному обучению.
Во-первых, задача предсказания свойств молекул формулируется как задача регрессии, поскольку целью является аппроксимация непрерывной количественной переменной, такой как растворимость, липофильность или активность соединения. RMSE и R^2 являются стандартными метриками качества для задач такого типа и позволяют получить объективную численную оценку точности предсказаний.
Среднеквадратичная ошибка (RMSE) определяется как квадратный корень из средней квадратичной разности между истинными и предсказанными значениями. Данная метрика чувствительна к большим отклонениям, что особенно важно в контексте предсказания молекулярных свойств, где крупные ошибки могут привести к выбору неподходящих соединений для дальнейших этапов разработки. Кроме того, RMSE выражается в тех же единицах измерения, что и предсказываемое свойство, что упрощает интерпретацию результатов и обеспечивает их прикладную значимость.

Коэффициент детерминации (R^2), в свою очередь, характеризует долю дисперсии зависимой переменной, объясняемую моделью. Его значения лежат в диапазоне $[-\infty; 1]$, где значение, близкое к 1, указывает на высокую предсказательную силу модели. В отличие от RMSE, R^2 является относительной метрикой, позволяющей оценить эффективность модели по сравнению с базовой моделью, предсказывающей среднее значение по выборке. Совместное использование обеих метрик позволяет получить всестороннюю оценку качества модели: RMSE дает представление об абсолютной точности, тогда как R^2 оценивает способность модели захватывать структуру данных. Такая комплементарность делает их особенно полезными в задачах сравнительного анализа моделей различных архитектур.
Следует отметить, что применение RMSE и R^2 является общепринятой практикой в ведущих работах в области предсказания свойств молекул [21], что дополнительно подтверждает их актуальность и обоснованность. Более того, обе метрики инвариантны к архитектуре модели, что обеспечивает их применимость как к классическим алгоритмам машинного обучения (например, Random Forest), так и к современным архитектурам, основанным на графовых нейронных сетях или трансформерах.

Таким образом, выбор метрик RMSE и R^2 обусловлен необходимостью обеспечить объективность, интерпретируемость и сопоставимость результатов эксперимента в рамках данного исследования, а также соответствует признанным стандартам оценки моделей в смежных научных дисциплинах.

Глава 4. Анализ полученных результатов.

4.1. BERT

Полученные значения метрик RMSE = 0.63 и R^2 = 0.78 на датасете ESOL свидетельствуют о высоком качестве предсказаний модели ChemBERTa и могут быть объяснены её архитектурными особенностями.

ChemBERTa является представителем трансформерных моделей, адаптированных для обработки химических структур в виде SMILES-строк. Основной архитектурной базой модели служит RoBERTa — модифицированный вариант оригинального трансформера BERT, который отличается улучшенным обучением на больших корпусах данных с использованием динамического маскирования и более длительного предобучения. В контексте химии предобучение на обширных наборах SMILES позволяет модели усвоить глубокие закономерности химической структуры и взаимодействий, что критично для точного предсказания физико-химических свойств молекул.

Ключевым элементом трансформерной архитектуры является механизм внимания (self-attention), который обеспечивает эффективное моделирование сложных зависимостей между атомами и функциональными группами в молекуле, отражающимися в SMILES-последовательности. В отличие от традиционных моделей на основе дескрипторов или сверточных нейронных сетей (GCN), ChemBERTa напрямую оперирует с исходным химическим кодированием, что снижает потери информации и повышает выразительность признакового пространства.

Достигнутый RMSE = 0.63 указывает на низкую среднеквадратичную ошибку в абсолютных значениях логарифма растворимости, что свидетельствует о высокой точности количественного предсказания. Значение R^2 = 0.78 отражает, что модель объясняет порядка 78% вариабельности целевого свойства, что подтверждает её способность к обобщению на тестовых данных.

Таким образом, успешность ChemBERTa в задаче предсказания свойств малых молекул обусловлена сочетанием предобучения на масштабных химических корпусах, обеспечивающего глубокое представление структуры, архитектуры трансформера с механизмом внимания, позволяющим моделировать сложные внутренние взаимосвязи и способности работать с последовательной химической информацией без промежуточного снижения размерности или потери важных структурных деталей.

Эти архитектурные особенности делают ChemBERTa конкурентоспособной по сравнению с другими современными подходами в области биоинформатики и способствуют получению высококачественных прогнозов для задач, подобных ESOL.

4.2. GCN

Результаты модели графовой сверточной сети (GCN) на датасете ESOL, характеризующиеся значениями RMSE около 1.04 и коэффициента детерминации R^2 порядка 0.77, демонстрируют как сильные, так и ограниченные стороны данной архитектуры применительно к задаче предсказания физико-химических свойств малых молекул.

Ключевым аспектом архитектуры GCN является принцип локальной агрегации признаков вершин графа, который реализуется посредством суммирования или усреднения информации от соседних узлов на каждом слое сети. Такая локальная агрегация обеспечивает эффективное кодирование локальной топологии молекулы и её непосредственного окружения, что способствует улавливанию значимых химических закономерностей. Однако ограниченная «глубина восприятия» модели, обусловленная небольшим числом слоев и эффектом переусреднения (over-smoothing) при увеличении глубины, снижает способность GCN учитывать отдалённые связи и комплексные глобальные зависимости внутри молекулы. Это, в свою очередь, ограничивает точность численного предсказания и проявляется в сравнительно высоком значении RMSE.

Кроме того, фиксированная структура молекулярного графа в GCN не позволяет моделировать гибкость и динамичность конформационных изменений, которые могут влиять на свойства молекул. Отсутствие механизма глобального внимания, присущего трансформерным моделям, ограничивает способность GCN выделять наиболее информативные химические подструктуры и взвешивать их значимость в процессе предсказания.

4.3. Random Forest

Результаты модели случайного леса (Random Forest) на датасете ESOL с RMSE примерно 0.586 и коэффициентом детерминации R^2 около 0.619 свидетельствуют о хорошей способности модели к аппроксимации зависимости между структурными признаками молекул и их растворимостью, но при этом с более ограниченной объяснительной мощностью по сравнению с глубокими нейронными архитектурами.

Random Forest представляет собой ансамбль решающих деревьев, построенных с использованием бутстрапа и случайного подмножества признаков, что обеспечивает высокую устойчивость к переобучению и способность выявлять нелинейные зависимости в данных без предположений о форме этих зависимостей. Данный подход хорошо работает на табличных признаках, которые могут быть извлечены из молекул, и не требует сложных архитектурных решений. Тем не менее, отсутствие встроенного механизма обработки структурных особенностей молекул, таких как топология и

пространственное расположение атомов, ограничивает глубину анализа химической информации. Это сказывается на сравнительно низком значении R^2, что указывает на меньшую долю объяснённой дисперсии по сравнению с нейросетевыми моделями, особенно с ChemBERTa, которая использует трансформерную архитектуру и учитывает контекст SMILES-последовательностей.

В итоге, Random Forest демонстрирует конкурентоспособную точность предсказания с относительно низкой вычислительной сложностью и простотой настройки, что делает её полезным инструментом для быстрой и интерпретируемой оценки свойств малых молекул, однако уступающим в выразительности более специализированным глубоким моделям.

4.4. Проведение сравнительного анализа

Представленные результаты моделей ChemBERTa, GCN и Random Forest на контрольном датасете ESOL демонстрируют различные уровни эффективности в задаче предсказания физико-химических свойств малых молекул, что обусловлено их архитектурными особенностями и способностью моделировать химическую информацию.

Модель ChemBERTa, основанная на архитектуре трансформера RoBERTa, продемонстрировала высокое качество предсказаний с RMSE = 0.63 и $R^2 = 0.78$. Эти показатели отражают её способность точно количественно оценивать логарифм растворимости и объяснять значительную часть вариабельности целевого признака. Архитектурные преимущества ChemBERTa связаны с механизмом self-attention, который позволяет эффективно моделировать сложные и дальнодействующие взаимосвязи между атомами и функциональными группами, представленными в виде последовательностей SMILES. Предобучение на больших химических корпусах обеспечивает глубокое усвоение химических закономерностей и способствует устойчивости модели к различиям в структуре молекул. Отсутствие необходимости промежуточного снижения размерности и сохранение контекста последовательности делают ChemBERTa одной из наиболее выразительных моделей для обработки химических данных.

Графовая сверточная сеть (GCN), с результатами RMSE ≈ 1.04 и $R^2 \approx 0.77$, демонстрирует хорошие показатели, однако с определёнными ограничениями. Модель эффективно кодирует локальную топологию молекулы за счёт агрегации признаков соседних вершин графа, что позволяет захватывать локальные химические взаимодействия. Тем не менее, ограниченная глубина сети и эффект переусреднения ограничивают способность учитывать глобальные и более отдалённые зависимости, которые важны для точного описания молекулярных свойств. Кроме того, отсутствие механизма

глобального внимания затрудняет выделение наиболее релевантных химических подструктур, что отражается в более высокой ошибке предсказания по сравнению с ChemBERTa. Несмотря на это, GCN сохраняет значительную эффективность, учитывая структурный характер входных данных и относительно низкие требования к предварительной обработке.

Модель Random Forest достигла RMSE около 0.59 и R^2 около 0.62, показывая удовлетворительную точность предсказаний, однако с более низкой объяснительной способностью по сравнению с глубокими архитектурами. Этот ансамблевый метод, оперирующий табличными признаками молекул, успешно выявляет нелинейные зависимости и устойчив к переобучению, что делает его простым и надёжным инструментом для предварительного анализа данных. Тем не менее, отсутствие встроенного механизма обработки молекулярной топологии и пространственных отношений атомов ограничивает глубину химического анализа, что сказывается на меньшем значении R^2 и, соответственно, на способности модели к комплексному учёту химических особенностей.

Таким образом, ChemBERTa, благодаря своей трансформерной архитектуре и предобучению на масштабных химических корпусах, обеспечивает более глубокое и контекстно обоснованное представление молекулярных данных, что отражается в наилучших метриках качества. GCN эффективно использует структурную информацию молекул, однако его локальные механизмы агрегации и ограничения глубины снижают точность по сравнению с ChemBERTa. Random Forest служит полезным базовым методом с хорошей устойчивостью и простотой применения, но уступает глубоким моделям в способности моделировать сложные химические зависимости.

Заключение

Таким образом в рамках настоящей работы было проведено исследование важнейших свойств малых молекул, обоснование актуальности задачи предсказания этих свойств, рассмотрены различные методы и подходы машинного обучения для решения данной задачи, проведен сравнительный анализ таких методов на примере задачи предсказания липофильности.

В результате дополнительных исследований и тщательного отбора была составлена выборка моделей, в полной мере отражающая разнообразие архитектур машинного обучения. Выбранные модели были дообучены и протестированы на датасете ESOL [4].

Сравнительный анализ моделей ChemBERTa, GCN и Random Forest на задаче предсказания растворимости малых молекул позволяет выявить как сильные стороны, так и ограничения каждой архитектурной парадигмы в контексте молекулярного моделирования.

Модель	RMSE	R^2
ChemBERTa	0.63	0.78
GCN	1.04	0.77
RF	0.586	0.619

ChemBERTa показала наилучшее качество с точки зрения R^2, несмотря на несколько выше RMSE по сравнению с Random Forest. Это обусловлено тем, что R^2 отражает долю объяснённой дисперсии и лучше характеризует способность модели к обобщению.

GCN демонстрирует хорошую способность моделировать химическую структуру за счёт графового представления молекулы. Архитектурное преимущество GCN заключается в использовании локальной агрегации признаков, что делает модель особенно эффективной для задач, где важно учитывать ближайшее окружение атомов. Однако отсутствие механизма глобального внимания и ограниченная глубина затрудняют интеграцию более отдалённых зависимостей, что снижает точность предсказания по сравнению с трансформером.

Random Forest, несмотря на простоту, продемонстрировал наименьшее значение RMSE, но уступил другим моделям по коэффициенту детерминации. Это свидетельствует о хорошем приближении к среднему значению растворимости, но с меньшей способностью улавливать дисперсию, присущую данным. Данный метод ограничен в моделировании структурных особенностей молекул, поскольку работает исключительно с табличными дескрипторами. Его предсказательная способность базируется на агрегированных признаках, что лишает модель гибкости при описании тонких химических взаимодействий.

Сопоставление архитектур показывает, что наиболее перспективным направлением для задач предсказания молекулярных свойств являются глубокие модели, оперирующие с оригинальными химическими представлениями без необходимости ручной инженерии признаков. Среди них ChemBERTa обладает наилучшей способностью к обобщению за счёт self-attention механизма и обширного предобучения. GCN представляет собой мощный инструмент, эффективно использующий графовую природу молекул, но требующий дополнительных улучшений, таких как введение глобального внимания (например, GAT или GINE). Random Forest, несмотря на ограниченность в выразительности, остаётся надёжным и интерпретируемым базовым методом, особенно полезным в условиях ограниченных вычислительных ресурсов или для анализа важности признаков.
Таким образом, архитектурные преимущества глубоких моделей, особенно трансформеров, делают их наиболее предпочтительными для задач прогнозирования свойств малых молекул в условиях комплексных химических взаимосвязей.

Список литературы

1. Prediction of small-molecule compound solubility in organic solvents by machine learning algorithms, Zhuyifan Ye, Defang Ouyang
2. Trends in small molecule drug properties: A developability molecule assessment perspective, P. Agarwal, J. Huckle, J. Newman, D. L. Reid
3. Antibiotic Discovery and Development, T.J.Dougherty, M.J.Pucci
4. ESOL: Estimating Aqueous Solubility Directly from Molecular Structure, John S Delaney
5. Strategy of utilizing in vitro and in vivo ADME tools for lead optimization and drug candidate selection, Balani SK, Miwa GT, Gan LS, Wu JT, Lee FW
6. Liquid Chromatography on the Different Methods for the Determination of Lipophilicity: An Essential Analytical Tool in Medicinal Chemistry, José X. Soares,Álvaro Santos,Carla Fernandes, Madalena M. M. Pinto
7. Message passing all the way up, P. Veličković
8. Do we need different machine learning algorithms for QSAR modeling? A comprehensive assessment of 16 machine learning algorithms on 14 QSAR data sets Zhenxing Wu , Minfeng Zhu , Yu Kang , Elaine Lai-Han Leung , Tailong Lei , Chao Shen , Dejun Jiang , Zhe Wang , Dongsheng Cao , Tingjun Hou
9. Non-linear QSAR modeling by using multilayer perceptron feedforward neural networks trained by back-propagation, D González-Arjona, G López-Pérez, A Gustavo González
10. How to improve R&D productivity: the pharmaceutical industry's grand challenge, Steven M. Paul, Daniel S. Mytelka, Christopher T. Dunwiddie,Charles C. Persinger, Bernard H. Munos, Stacy R. Lindborg, Aaron L. Schacht
11. qiRNA is a new type of small interfering RNA induced by DNA damage, Heng-C. Lee, S. Chang, S. Choudhary, A. P. Aalto, M. Maiti, D. H. Bamford, Y. Liu
12. Reinforced Adversarial Neural Computer for de Novo Molecular Design, E. Putin, A. Asadulaev, Y. Ivanenkov, V. Aladinskiy, B. Sanchez-Lengeling, A. Aspuru-Guzik, A. Zhavoronkov
13. Applications of machine learning in drug discovery and development, J. Vamathevan, D. Clark, P. Czodrowski, I. Dunham, E. Ferran, G. Lee, B. Li, A. Madabhushi, P. Shah, M. Spitzer
14. Drug repurposing: progress, challenges and recommendations, Sudeep Pushpakom, Francesco Iorio, Patrick A. Eyers
15. Analyzing Learned Molecular Representations for Property Prediction, Kevin Yang, Kyle Swanson, Wengong Jin
16. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, J. Devlin, M. Chang, K. Lee, K. Toutanova
17. ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction, S. Chithrananda, G. Grand, B. Ramsundar

18. Semi-Supervised Classification with Graph Convolutional Networks, T. N. Kipf, M. Welling
19. Computational Modeling of β-Secretase 1 (BACE-1) Inhibitors Using Ligand Based Approaches, G. Subramanian, B. Ramsundar, V. Pande, R. A. Denny
20. Handbook of Aqueous Solubility Data, S. H. Yalkowsky, Y. He
21. MoleculeNet: a benchmark for molecular machine learning" Z. Wu, B. Ramsundar, E. N. Feinberg, J. Gomes, C. Geniesse, A. S. Pappu, K. Leswing, V. Pande