МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ

БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

ФАКУЛЬТЕТ ПРИКЛАДНОЙ МАТЕМАТИКИ И ИНФОРМАТИКИ

Кафедра биомедицинской информатики

**АНАЛИЗ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ПРЕДСКАЗАНИЯ СВОЙСТВ МАЛЫХ МОЛЕКУЛ**

Курсовая работа

Снежко Льва Владимировича

студента 3 курса, специальность

«информатика»

Научный руководитель:

Карпенко А.Д.

Оглавление

[Оглавление 2](#_Toc198379889)

[ВВЕДЕНИЕ 3](#_Toc198379890)

[Глава 1. Постановка задачи. Основные определения и понятия. Краткий обзор основных свойств малых молекул 5](#_Toc198379891)

[1.1. Основные определения и понятия 5](#_Toc198379892)

[1.2. Свойства малых молекул 5](#_Toc198379893)

[1.3. ADMET 6](#_Toc198379894)

[1.4. Липофильность 7](#_Toc198379895)

[1.5. Обоснование актуальности задачи предсказания свойств малых молекул. 8](#_Toc198379896)

[Глава 2. Краткий обзор методов машинного обучения для решения задачи предсказания свойств малых молекул. 9](#_Toc198379897)

[2.1 Графовые нейронные сети (GNN) 9](#_Toc198379898)

[2.2 Transformers 9](#_Toc198379899)

[2.3 QSAR 9](#_Toc198379900)

[2.4. 3D GNN 9](#_Toc198379901)

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 МО 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. Основные определения и понятия

Чёткого международного юридического определения малой молекулы не существует. В настоящей работе под малой молекулой будет пониматься следующее.

Малая молекула – химическое соединение со сравнительно малой молекулярной массой (не более 900-1000 дальтон), обладающее той или иной биологической активностью и размером порядка 1 нм. Определяющим свойством малых молекул является её способность проникать в клетку и покидать её без необходимости взаимодействия с другими молекулами.

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. 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. 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. 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:

где - растворимость соединения в октаноле

а – растворимость в воде.

Часто на практике используют не сам коэффициент растворимости, а его логарфим: .

При logP > 0 соединение считается липофильным, а при logP < 0 – гидрофильным.

Оптимальным диапазоном коэффициента липофильности для лекарственных препаратов (по Липински) считается .

**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. 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. Под неперспективными молекулами подразумеваются молекулы с неудовлетворительной фармакинетикой, низкой биодоступностью и высокой токсичностью.

Благодаря стремительному развитию методов машинного обучения и росту уровня доступности данных задача появилась возможность решения этой задачи методами МО.

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

2.1 Графовые нейронные сети (GNN)

2.2 Transformers

2.3 QSAR

2.4. 3D GNN