Разработка модели машинного обучения для оценки способности молекул проникать через гематоэнцефалический барьер

Работу выполнил: Пронин Михаил Васильевич

Санкт-Петербург 2025 год

Методические материалы

Проблематика:

Гематоэнцефалический барьер (ГЭБ) - это полупроницаемый барьер между кровью и нервной тканью, который регулирует транспорт биологически активных веществ в мозг, а также препятствует проникновению в ЦНС различных чужеродных агентов (микроорганизмов, токсинов, антибиотиков). Способность проникать через ГЭБ является одной из важнейших характеристик потенциальных лекарственных молекул. Препараты для лечения заболеваний нервной системы должны проходить через ГЭБ легко, тогда как среди лекарств для лечения заболеваний желудка предпочтение отдается тем, которые преодолевают ГЭБ хуже (во избежание различных побочных эффектов). Численной мерой преодоления ГЭБ обычно выступает параметр logBB. Существует большое количество экспериментальных методов для оценки способности молекул проникать через ГЭБ, однако все они являются очень трудоемкими. Сейчас ученые возлагают большие надежды на машинное обучение (ML). Однако эта научная область находится в активном развитии, создаются и улучшаются различные типы моделей. Поэтому нет общепринятого стандарта их создания, и проверка различных инструментов для извлечения признаков из молекул, алгоритмов ML, количества данных является значимой задачей.

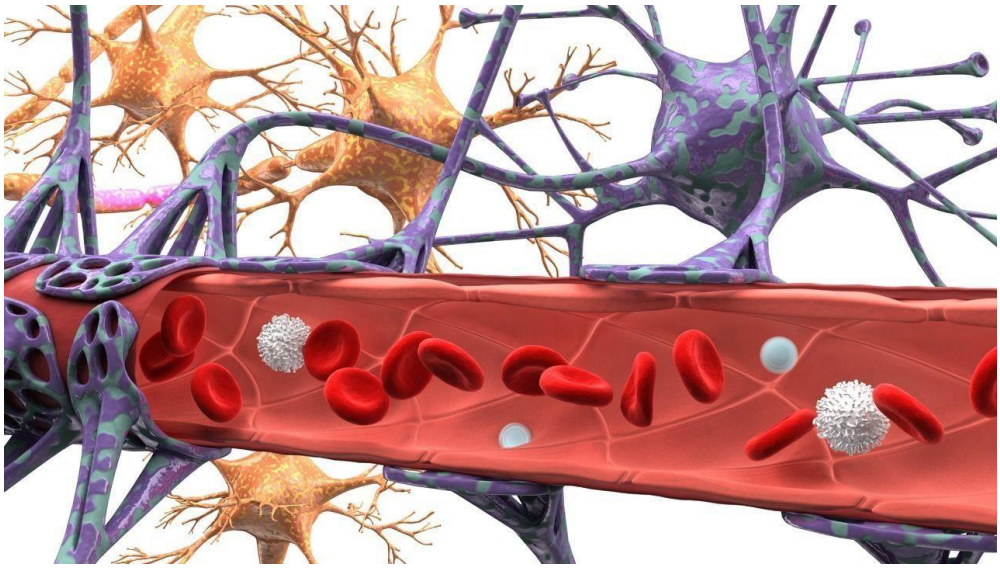

Рис.1 3d моделирование ГЭБ

Цель моего исследования: разработать надежные, общедоступные модели машинного обучения, для предсказания вероятности прохождения молекулы лекарственного препарата через гематоэнцефалический барьер(ГЭБ), обученные на большом наборе данных. Проверить эффективность различных алгоритмов и библиотек с параметрами молекул.

Модели, полученные в ходе работы, можно использовать, как один из способов проверки прохождения молекулой ГЭБ. Также данные о том, какие характеристики, и в каком количестве, показали лучший результат обучения, какие модели оказались точнее и надежнее, можно использовать в дальнейших исследованиях на эту тему.

Задачи:

Сбор и очистка данных для обучения (молекулы в формате SMILES). Вычисление параметров молекул (фингерпринты), которые можно использовать для машинного обучения. Отбор 25%, 50%, 75% лучших параметров, чтобы избежать нехватки данных и переобучения. Проверка эффективности различных моделей. Сравнение различных моделей и формулировка выводов, публикация работы.

Интердисциплинарность проекта выражается в использовании программирования, современных способов обработки и анализа информации для решения задачи, связанной с химией и биологией. Также, изначальные данные о прохождении молекул через ГЭБ были получены экспериментальным способом.

Ход работы:

В исследовании были использованы сведения открытой базы B3DB, данные о молекулах собраны Tevosyan et. al. Затем, все молекулы приводились к каноническому виду SMILES (такой вид облегчает обработку и анализ химических данных с использованием компьютерных программ). Были удалены повторяющиеся, отсутствующие значения, неорганические соединения. Так как задачей было предсказывать численное значение logBB, удалялись молекулы значения logBB, для которых не входило в два стандартных отклонения от среднего, т.е. выбросы. В итоговой таблице осталось 1130 молекул.

logBB	SMILES
1.19	CN(C)CCCN1C2=CC=CC=C2SC2=CC=CC=C21
1.20	CCC1=CC=CC(N=C(C)NC2=CC=CC(CC)=C2)=C1
1.25	OC1=CC=C2C(OC3=CC=C(OCCN4CCCCC4)C=C3)=C(C3=CC=...
1.25	CN(C)CCOC(C1=CC=CC=C1)C1=CC=CC=C1
1.25	CC(NC(C)(C)C)C(=O)C1=CC=CC(Cl)=C1
1.26	COC1=CC=C(C2=NN(C)C(C3=CC=CC(CN(C)C)=C3)=N2)C=C1
1.26	CN(C)COC(C1=CC=CC=C1)C1=CC=CC=C1
1.26	COC1=CC2=C(C=C1CNC1CCCNC1C1=CC=CC=C1)N=C(C1=CC...
1.30	CNCCCC12CCC(C3=CC=CC=C31)C1=CC=CC=C12
1.30	CC(CN1C2=CC=CC=C2SC2=CC=CC=C21)N(C)C

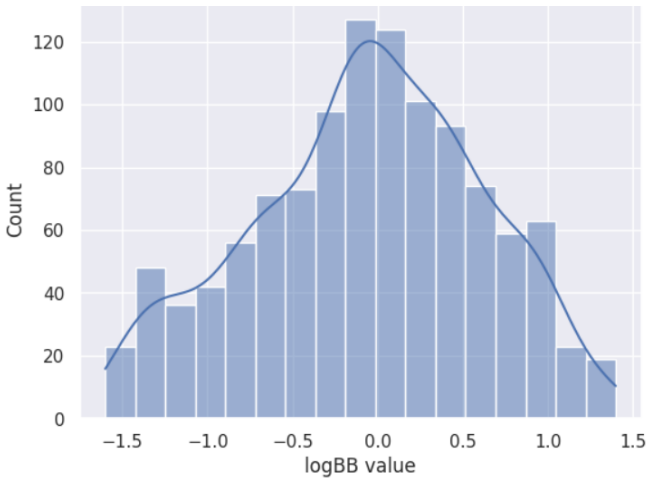

Рис. 2 Некоторые из исследуемых молекул (в формате SMILES), их значение logBB и распределение этого значения

После предобработки, используя такие библиотеки python, как: rdkit, Avalon, PubChem, были сгенерированы характерные признаки каждой молекулы. Затем, создавались разные наборы данных. При этом библиотека Avalon определяет фингерпринты молекулы, её “отпечатки пальцев”, т.е. наличие или отсутствие определенных структурных элементов. Признаки, основанные на базе PubChem, включает в себя информацию не только о фингерпринтах, но и физико-химических параметрах, таких как масса молекулы, максимальный частичный заряд, кол-во атомов, отличных от углерода и водорода и т.д. В перспективе это может дать больше полезной информации.

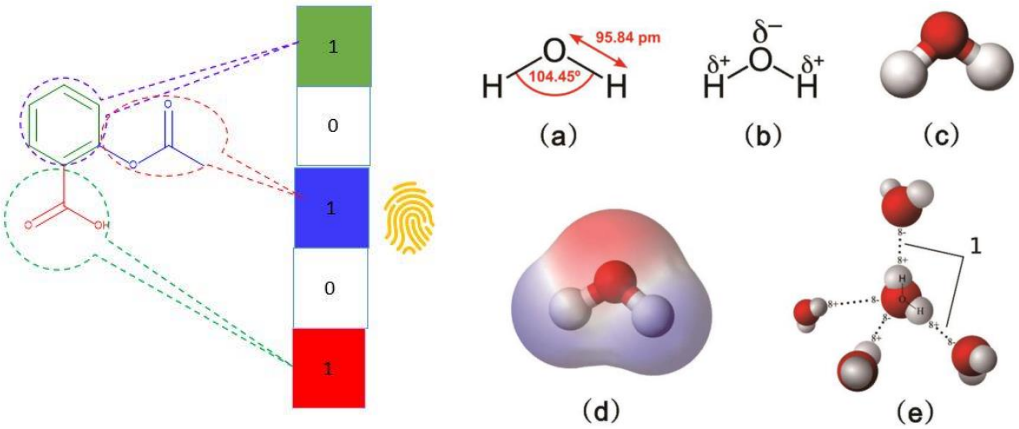

Рис. 3 Иллюстрации создания фингерпринтов и определения физико-химических параметров. Наличие определенных элементов в молекуле определяет 0 или 1 будет в соответствующем поле

Метод SelectKBest из библиотеки sklearn выбирал лучшие 25, 50, 75% от всех признаков молекул, по показателю взаимной информации, который демонстрирует зависимость между двумя множествами. Таким образом, для обучения и проверки моделей использовались характеристики с наибольшим влиянием на параметр logBB. Всего было проверено 6 моделей: LR (линейная регрессия), KNN (К ближайших соседей), DT (дерево решений), RF (случайный лес), LGBM (градиентный бустинг), ANN (искусственная нейронная сеть). Каждую модель обучали на 9 вариантах данных: 25,50,75 процентов признаков Avalon/PubChem и 50% признаков обеих библиотек.

Чтобы проверить работу модели, использовались три метрики. Среднеквадратичная ошибка ($\boldsymbol{RMSE_{cv}}$) – это одна из функций потерь, в моем случае она демонстрирует разницу

между предсказанными значениями logBB и экспериментально проверенными. Идеальное значение среднеквадратичной ошибки – 0. Другая метрика – коэффициент детерминации **R²**. Она показывает насколько хорошо модель работает на разных наборах данных. Лучшее значение метрики – 1. В моем случае для подсчёта R² проводилась кросс валидация 5 folds. То есть, весь набор данных делился на 5 частей. На 4 модель обучалась, на последней проводился тест. На тесте подсчитывались значения $\mathrm{RMSE_{cv}}$ и R². Это повторялось для каждой из 5 частей, среднее значение метрики - итоговое. **Q²** – значение R², но при обучении и проверке на полном датасете, без валидации.

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

$$R^2 = 1 - \frac{\Sigma(y_{pred} - y_{mean})^2}{\Sigma(y_{actual} - y_{mean})^2}$$

Рис.4 Формулы для расчёта $RMSE_{cv}$ и R², где $\hat{y}_i$ и y_{pred} – предсказанные значения, y_i и y_{actual} – верные значения

После обучения и проверки моделей с дескрипторами Avalon и PubChem, к данным для обучения были добавлены метки классов, предсказанные моделью классификации. То есть, значения, которые были получены моделью при предсказании класса (пройдет молекула через ГЭБ или нет), добавились к данным из библиотек. Модель классификации разработана моим коллегой по учебной группе проекта Сириус.Лето и по SMILES предсказывает класс молекулы. Таким образом, её предсказание можно применить в работе над моими моделями. Добавление этих данных улучшило показатели метрик в среднем на 11%.

По итогам исследования, лучшие результаты показали три модели: RF, LGBM, ANN. Значения их метрик схожи, но лучшие у третьей модели: **R² = 0.98, Q² = 0.694, $RMSE_{cv}$ = 0.332.**

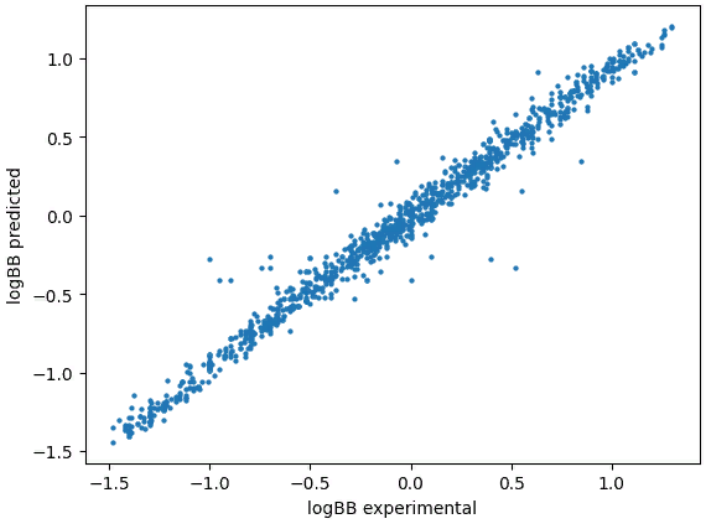

Рис. 5 График показывающий различия в предсказанном и экспериментальном значениях logBB для молекул

Напомним, что ANN по строению схожа с нейронами головного мозга. Такая модель хорошо работает с неструктурированными данными, некоторые переменные в которых могут быть связаны. Эти же характеристики относятся к информации о молекулах. Другие модели: RF с беггингом, в котором каждое отдельное дерево решений обучается на своей части датасета, а итоговое решение – среднее из всех решений деревьев или LGBM, в котором при построении деревьев программа стремится уменьшить значение функции ошибки – справляются с предсказанием в данных со сложной структурой и взаимосвязью параметров.

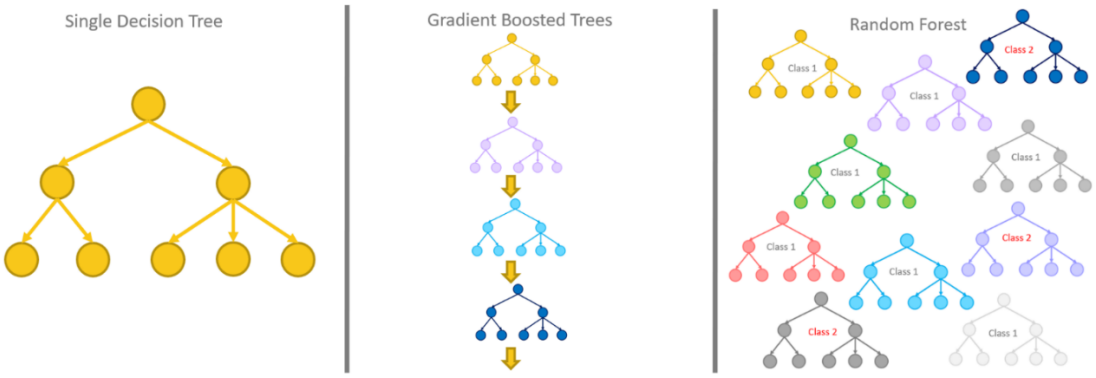

Рис.6 Отличие LGBM – каждое следующее дерево с меньшим значением функции потерь, от RF – множества деревьев с разными данными для обучения

Также, было определено, что использование параметров PubChem даёт результаты значительно хуже, чем параметров Avalon. При их смешении, качество работы не растёт. В зависимости от модели, использование различного количества лучших параметров даёт небольшие различия в качестве. Принимая главным показателем $RMSE_{cv}$, лучшие значения

метрик получаются при использовании 75% признаков Avalon (50% для LGBM). Как упомянуто выше, добавление метки класса значительно повышает качество моделей.

Модель	Значение метрики		
	$RMSE_{cv}$	Q^2	R^2
KNN Avalon 50%	0,49	0,44	0,78
KNN PubChem 25%	0,65	0,10	0,19
DT Avalon 25%	0,50	0,40	0,93
DT PubChem 50%	0,81	0,11	0,24
RF Avalon 75%	0,39	0,64	0,95
RF PubChem 75%	0,61	0,08	0,87
LGBM Avalon 75%	0,35	0,71	0,98
LGBM PubChem 25%	0,64	0,01	0,94
ANN PubChem 75%	0,35	0,04	0,05
ANN Avalon 75%	0,33	0,70	0,98

Рис.7 Значения метрик для разных моделей. Для каждой модели указаны значения при добавлении к признакам Rdkit, признаков Avalon или PubChem

Сравнивая результаты этой работы с другими подобными, можно выделить такие характеристики, как: большой набор исследуемых молекул, общедоступность, модели выложены в открытое GitHub-хранилище. При этом итоговые метрики моей работы сравнимы с другими исследованиями (См.рис.8)

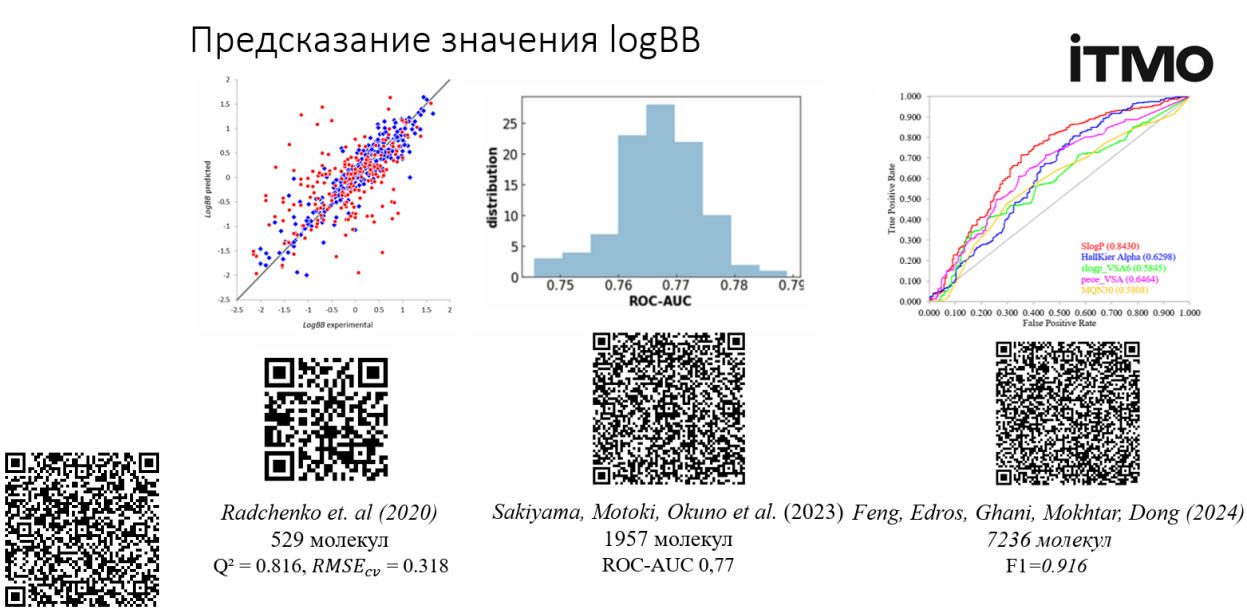

Рис.8 QR-код с ссылкой на GitHub-репозиторий и инфографика о подобных исследованиях, информация о метриках, размере изучаемых данных, графики предсказанных и экспериментальных значений logBB, QR -коды с ссылками на статьи

Таким образом, цель проекта выполнена. Созданы и обучены модели, получена достоверная информация о том, какие параметры и алгоритмы дают преимущество. По сравнению с подобными работами, данное исследование опирается на базу данных достаточного объема, общедоступно и при этом не уступает в точности и надежности другим.

Федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет ИТМО»

Научный руководитель: Исакова Анастасия Михайловна, магистрант 2-ого года обучения НОЦ инфохимии