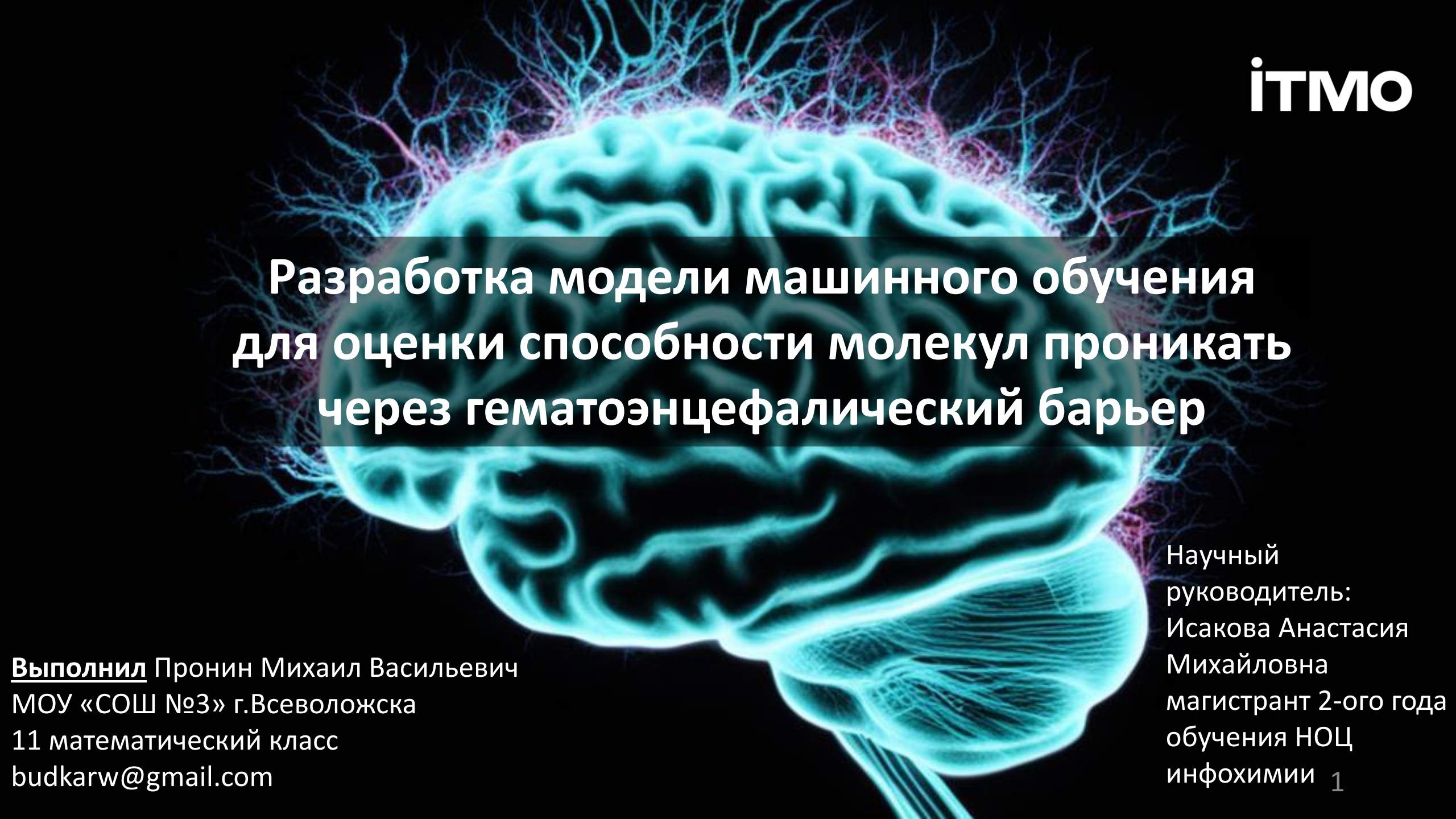


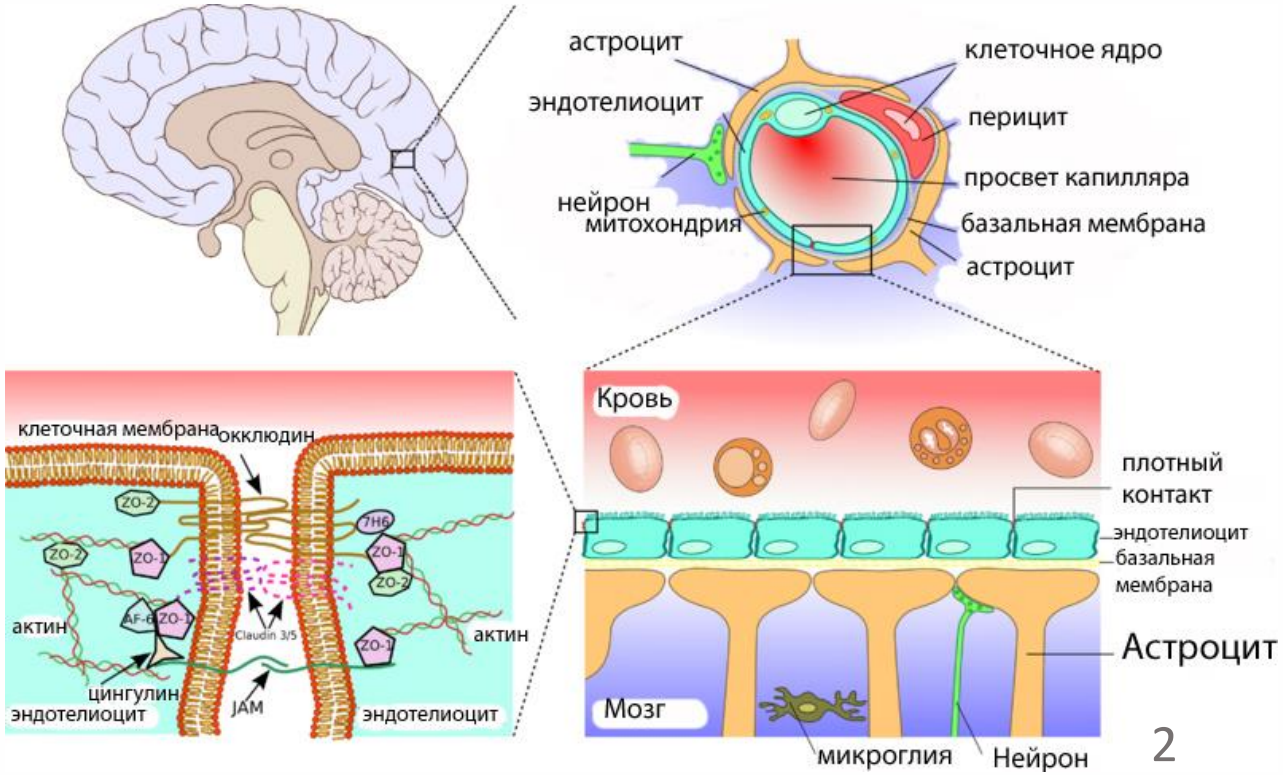
Разработка модели машинного обучения для оценки способности молекул проникать через гематоэнцефалический барьер

Выполнил Пронин Михаил Васильевич
МОУ «СОШ №3» г.Всеволожска
11 математический класс
budkarw@gmail.com

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

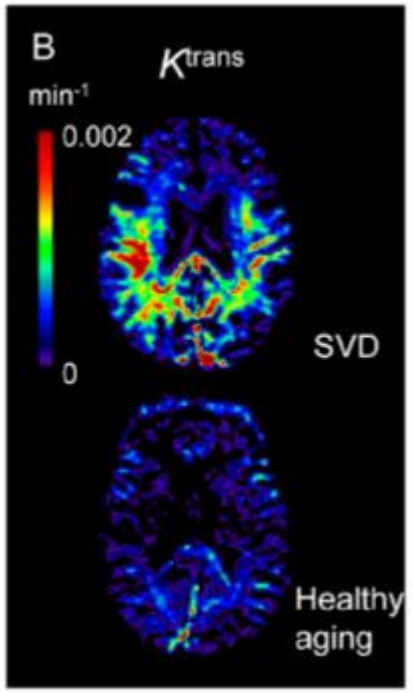
Проблематика

- Гематоэнцефалический барьер (ГЭБ) - это полупроницаемый барьер между кровью и нервной тканью.
- Способность проникать через ГЭБ является одной из важнейших характеристик молекул в фармакологии.

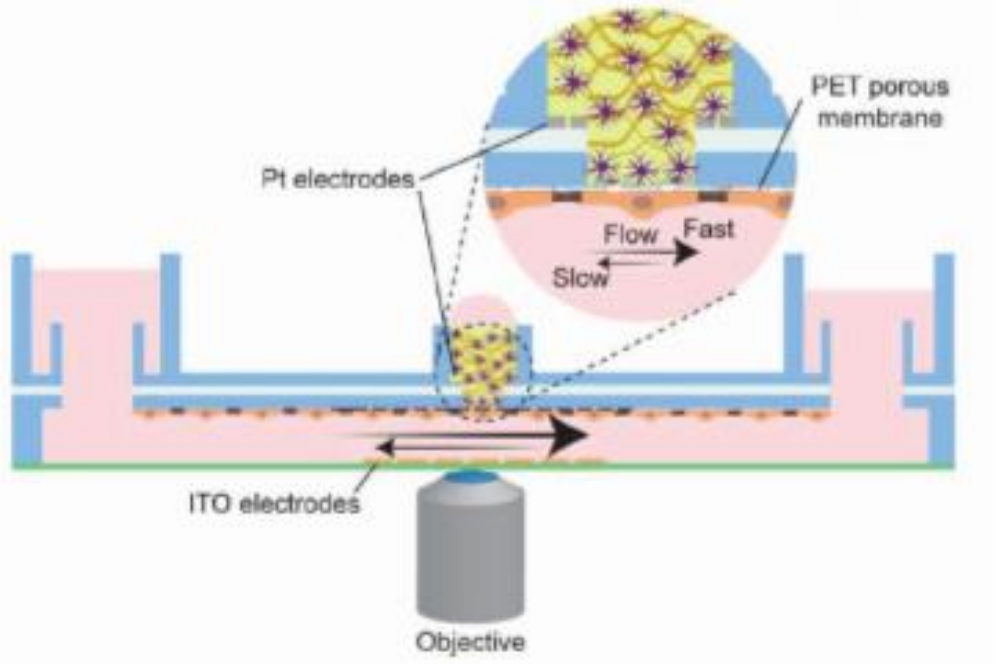


Методы анализа

Экспериментально *in vivo*



Экспериментально *in vitro*



Модель *in silico*

```
            nn.Linear(in_features=n_input_layer, out_features=n_inner_layer),
            #nn.Dropout(0.5),
            nn.LeakyReLU(),

            nn.Linear(in_features=n_inner_layer, out_features=1)
        )

    def forward(self, x):
      return self.regressor(x)
```

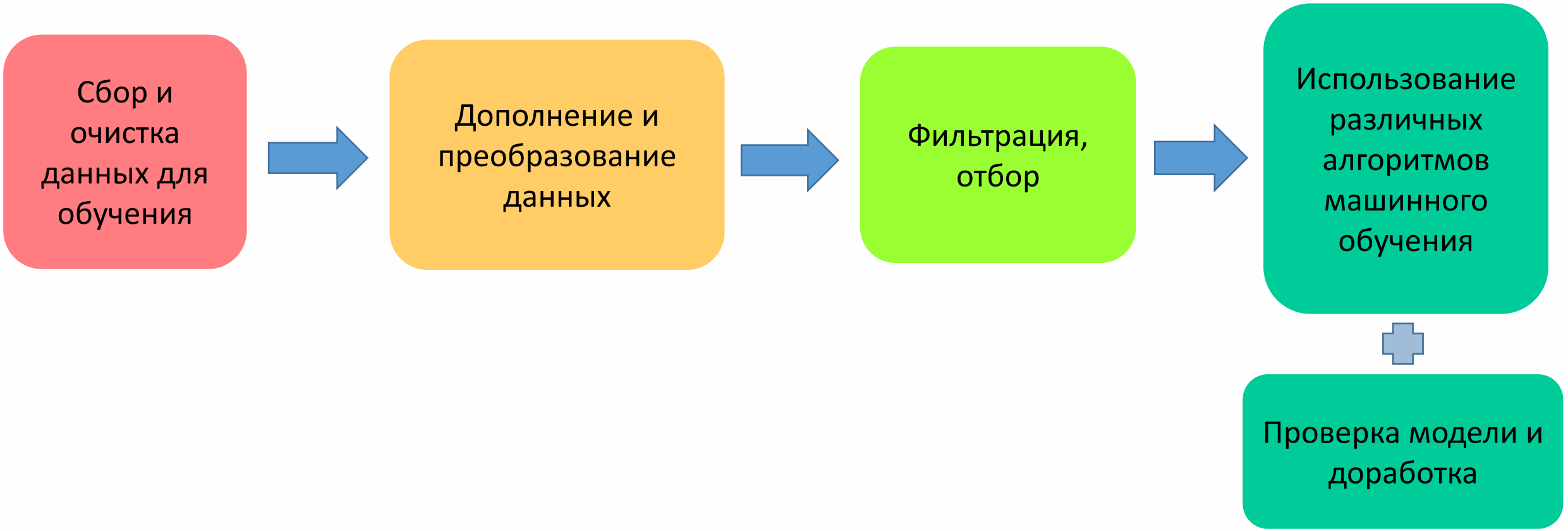
```
# initiate model
def initiate_model():
  logBB_net = logBB_Net(n_input_features=397, n_input_layer=350, n_inner_layer=350)

  # define some other parameters
  n_epochs = 100
  criterion = nn.MSELoss() # criterion for training
  optimizer = torch.optim.Adam(logBB_net.parameters(), lr=3e-4) # define optimizer with lr
  return logBB_net, n_epochs, criterion, optimizer
```

Лабораторное оборудование, стерильность, ткани/животные **VS** Интернет, компьютер

Вывод: использовать модели выгодно

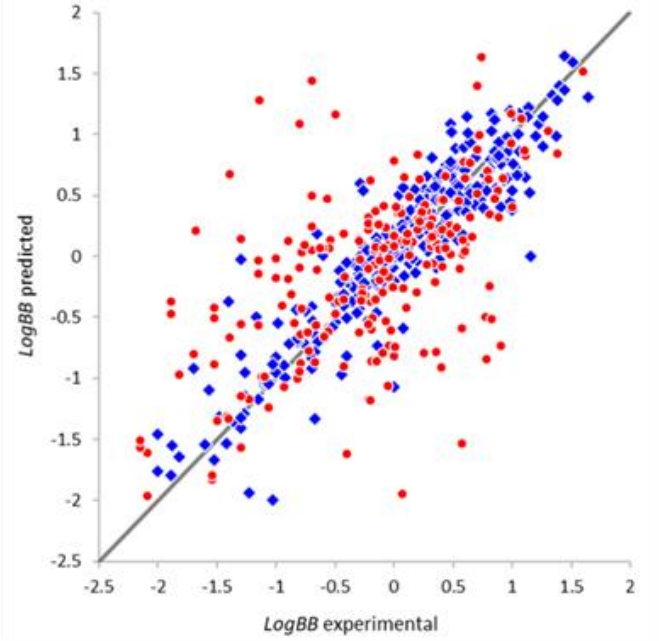
Этапы создания модели



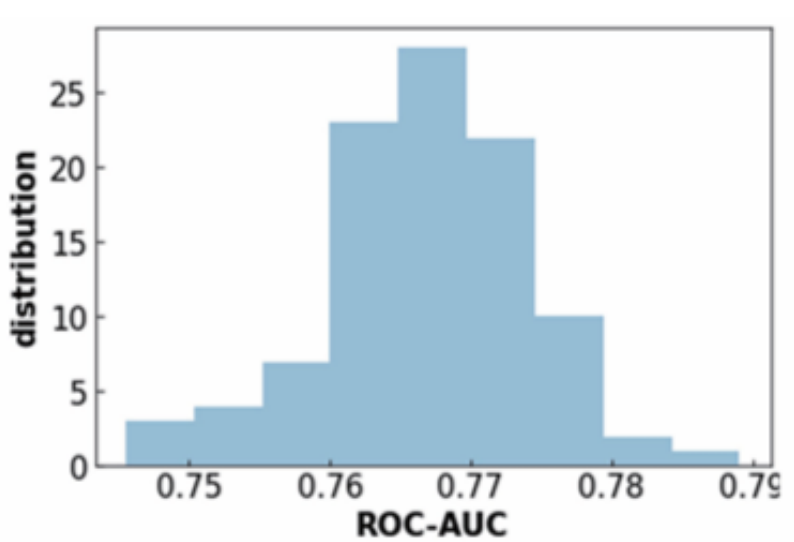
Предсказание значения logBB

Проблема 1: модели обучены на небольших наборах данных/искусственных данных -> результаты недостаточно надежны.

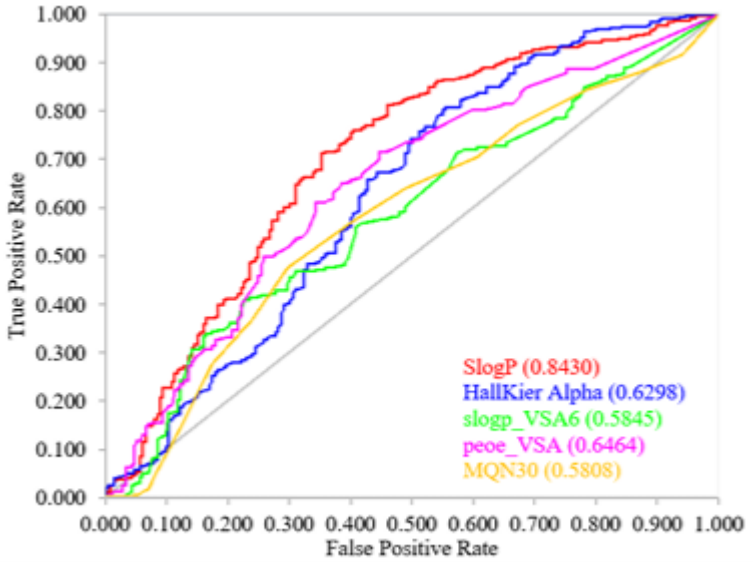
Проблема 2: модели для предсказания logBB изменяются и совершенствуются, не все перспективные варианты исследованы



Radchenko et. al (2020)
529 молекул



Sakiyama, Motoki, Okuno et al. (2023)
1957 молекул



Feng, Edros, Ghani, Mokhtar, Dong et. al(2024)
7236 молекул

Цели и задачи

Создать ML-модель для предсказания logBB, обученную на большом количестве достоверных данных. Проверить эффективность различных алгоритмов машинного обучения и разных параметров молекул в качестве входных данных.

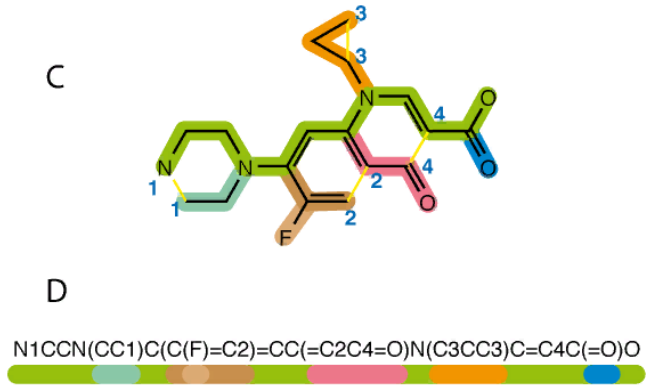
Поиск базы данных, проверка и нормирование

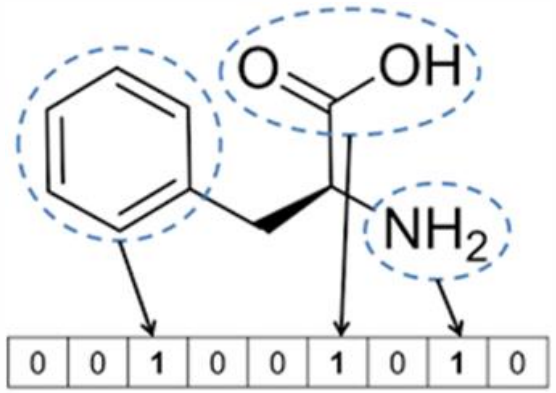
Использование различных библиотек для извлечения характеристик

Подбор перцентиля важности параметров

Проверка различных алгоритмов

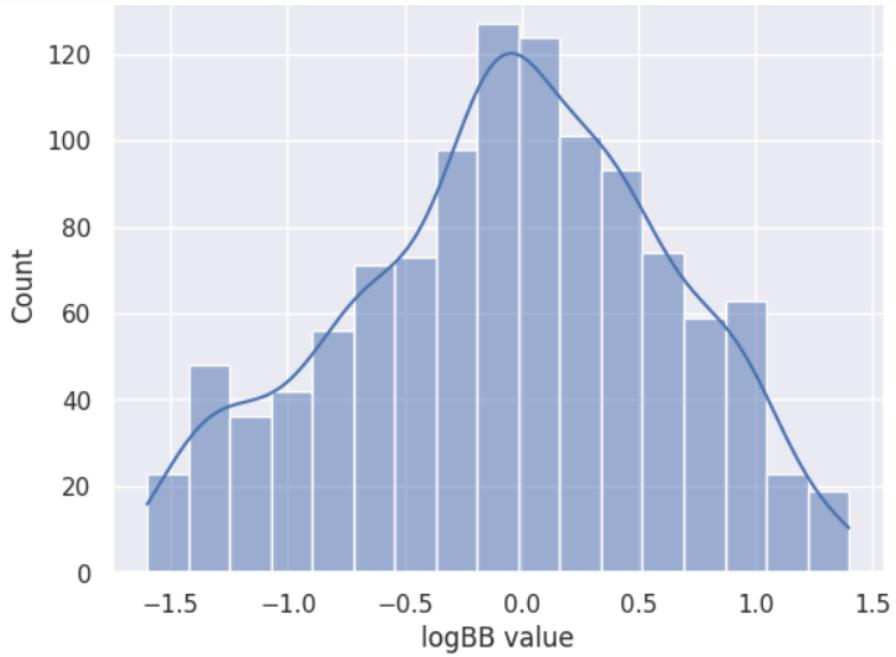
Использование меток класса





Сбор и предобработка данных

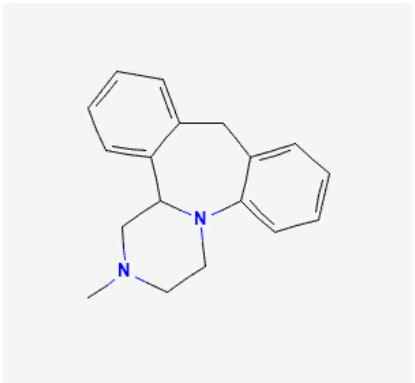
- Источник данных: база данных **B3DB**, статья ***Tevosyan et. al***
- Все SMILES переведены в канонический вид
- Удалены дубликаты и отсутствующие значения
- Удалены неорганические молекулы
- Удалены выбросы



B3DB

Tevosyan et al.

После обработки осталось 1130 молекул



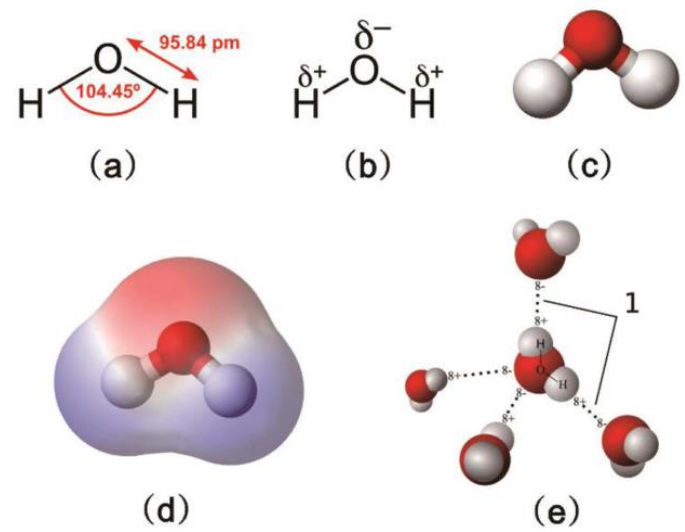
CN1CCN2C3=CC=CC=C3CC3=CC=CC=C3C2C1

mianserin logBB=0,99

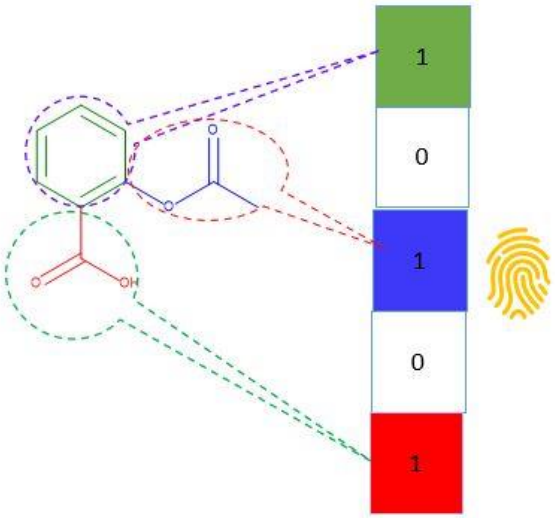
Создание параметров молекул

По данным молекул были сгенерированы дескрипторы RDKit, фингерпринты Avalon и PubChem. Сравнивались результаты для различных комбинаций признаков.

RDKit
Open-Source Cheminformatics and Machine Learning

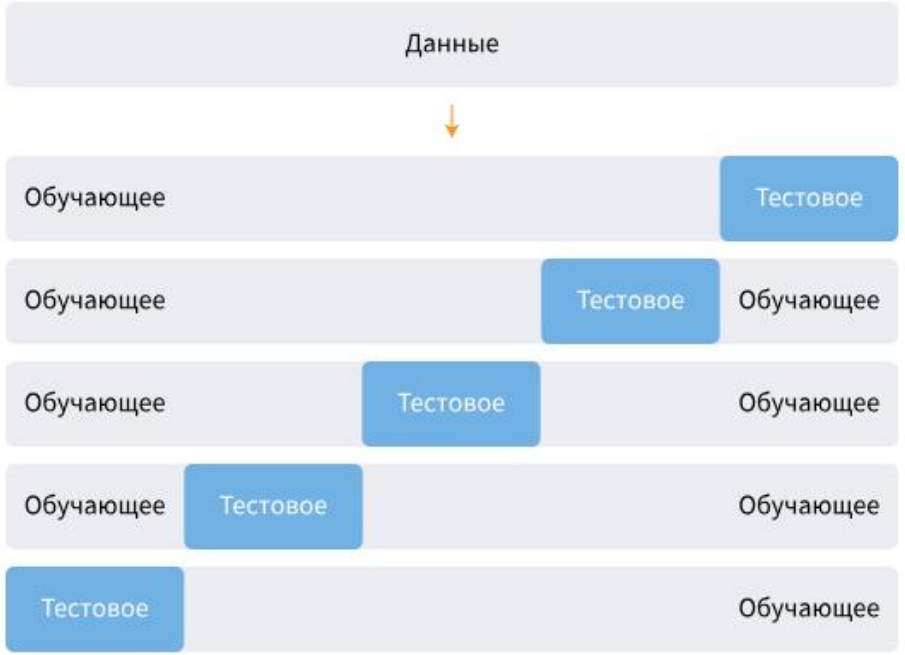


PubChem

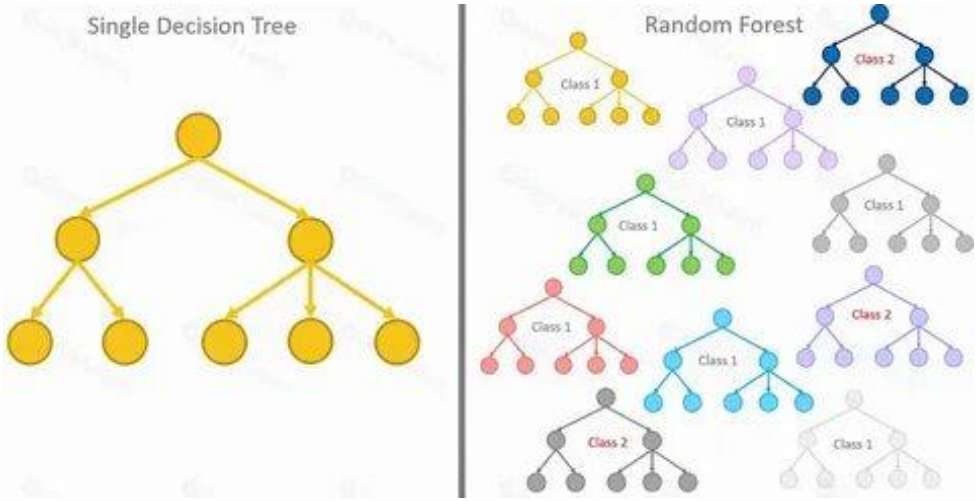


Использование разного количества данных

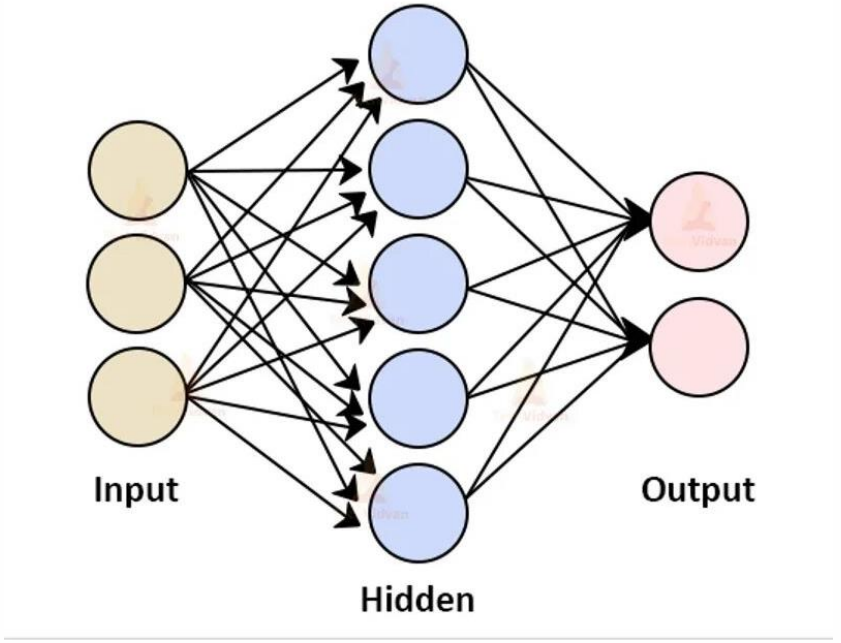
- Для обучения отбирались 25%, 50%, 75%, 100% лучших признаков. Отбор происходил по критерию взаимной информации
- Проведена кросс-валидация. На 80% данных модели обучались, 20% были оставлены для тестирования



Алгоритмы машинного обучения



Решающие деревья



Нейронная сеть

Метка класса

Добавление информации о классе, предсказанном с помощью модели классификации

Результаты

- Лучшие показатели без использования метки класса у нейронной сети.
- Использование метки класса улучшает показания метрик Q^2 и $\mathrm{RMSE}_{\mathrm{cv}}$ на 8% и 14% соответственно.
- Использование параметров PubChem даёт результаты значительно хуже, чем параметры Avalon.
- Лучшие показатели получаются при использовании 75% признаков.

Модель	Значение метрики		
	RMSE	Q2	R2
k-nearest neighbours Avalon 50%	0,49	0,44	0,78
decision tree Avalon 25%	0,50	0,40	0,93
random forest Avalon 75%	0,39	0,64	0,95
light gradient boosting Avalon 75%	0,38	0,65	0,98
neural network Avalon 75%	0,33	0,70	0,98
light gradient boosting PubChem 25%	0,64	0,01	0,94
neural network PubChem 75%	0,35	0,04	0,05

Заключение

- Созданы надежные модели для предсказания logBB (см.Github)
- Лучшие результаты были получены с использованием модели neural network regression
- Использование pubchem не эффективно для этой задачи по сравнению с Avalon
- При отборе признаков по критерию взаимной информации оптимально использовать 75% лучших параметров

Github-репозиторий

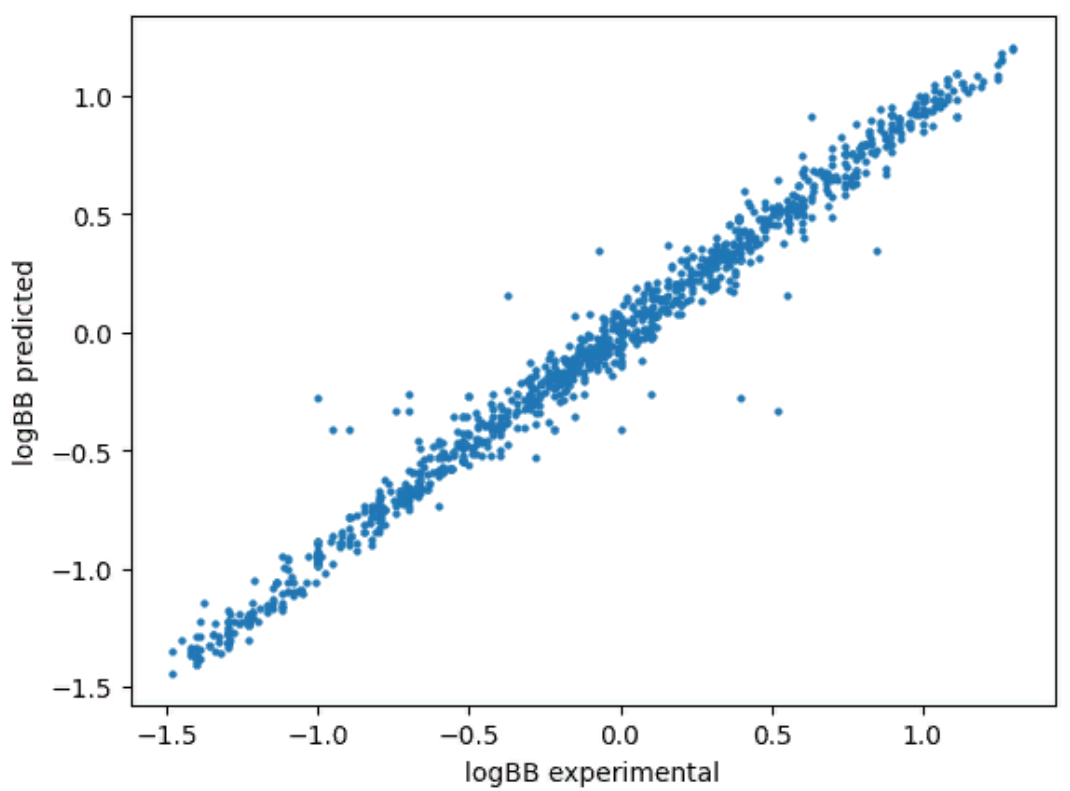


График предсказанных и реальных значений logBB молекул с использованием оптимальной модели

Спасибо за внимание

Метрики качества модели

Среднеквадратичная ошибка (RMSE)

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

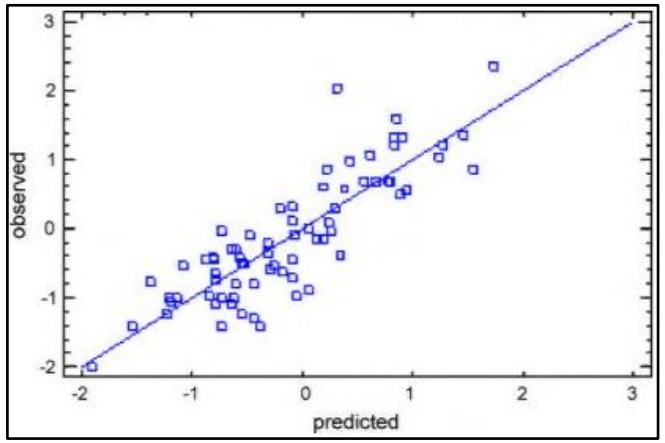
Разница между предсказанными значениями и экспериментально проверенными. Идеальное значение среднеквадратичной ошибки – 0

Коэффициент детерминации (R^2)

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

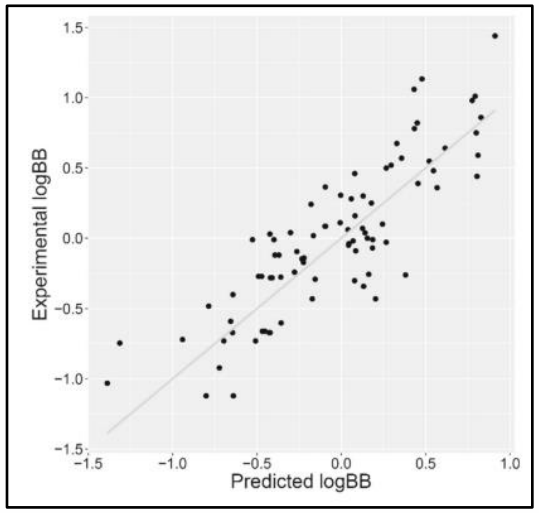
Насколько хорошо модель работает на разных наборах данных. Лучшее значение метрики – 1. Q^2 – R^2, но при обучении и проверке на полном датасете

Задача регрессии: предсказание значения logBB



Mensch et. al (2010)

- ✅ $R^2 = 0.774$, $Q^2 = 0.635$
- ❌ В закрытом доступе
- ❌ Очень маленький набор данных(88 молекул)



Zhu et. al (2018)

- ✅ R^2=0.938, Q^2=0.788, $RMSE_{cv}$=0,324*
- ❌ В закрытом доступе
- ❌ Маленький набор данных (287 молекул)

Radchenko et. al (2020)

- ✅ $Q^2 = 0.816$, $RMSE_{cv} = 0.318$
- ✅ Публично доступно
- ❌ Маленький набор данных (529 молекул)

**идеальные значения R^2, Q^2 и $RMSE_{cv}$: 1, 1 и 0 соответственно*

Модель	Значение метрики		
	RMSE	Q2	R2
k-nearest neighbours Avalon 50%	0,49	0,44	0,78
k-nearest neighbours Pubchem 25%	0,65	0,10	0,19
decision tree Avalon 25%	0,50	0,40	0,93
decision tree Pubchem 50%	0,81	0,11	0,24
random forest Avalon 75%	0,39	0,64	0,95
random forest Pubchem 75%	0,61	0,08	0,87
light gradient boosting Avalon 75%	0,35	0,71	0,98
light gradient boosting Pubchem 25%	0,64	0,01	0,94
neural network Pubchem 75%	0,35	0,04	0,05
neural network Avalon 75%	0,33	0,70	0,98