БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ МЕДИЦИНСКИЙ УНИВЕРСИТЕТ
КАФЕДРА ФАРМАКОЛОГИИ

ПОСТРОЕНИЕ QSAR МОДЕЛИ ДЛЯ ПРЕДСКАЗАНИЯ АКТИВНОСТИ ИНГИБИТОРОВ РЕНИНА

Автор: Евстафьева Виктория Алексеевна, 3 курс, лечебный факультет
Научный руководитель: ассистент Кашкур Юрий Витальевич

МИНСК, 2021

ЦЕЛЬ И ЗАДАЧИ

Цель: построить модель машинного обучения на основе алгоритма “случайных лесов” (random forest), которая позволит предсказывать активность ингибиторов ренина, основываясь на структуре молекул.

Задачи:

1. Собрать данные об уже изученных лигандах ренина;
2. Провести обработку данных, отобрав только подходящие для построения модели лиганды;
3. Построить модель машинного обучения на основе алгоритма “случайных лесов”, предсказывающую активность ингибиторов ренина на основе их молекулярной структуры;
4. Проверить эффективность модели на тестовых данных;
5. Использовать данную модель для предсказания активности найденных in-silico потенциальных ингибиторов ренина.

АКТУАЛЬНОСТЬ

Сердечно-сосудистые заболевания являются основной причиной смерти и инвалидизации во всем мире. Поиск новых способов лечения данных патологий является важнейшей проблемой современного медицинского научного сообщества. Одной из перспективных групп лекарственных средств являются **ингибиторы ренина. Ренин** – это протеолитический фермент, осуществляющий гидролиз ангиотензиногена до ангиотензина I. Он является первым звеном ренин-ангиотензин-альдостероновой системы, которая участвует в регуляции артериального давления. Таким образом, препараты данной группы могут использоваться для лечения артериальной гипертензии. На данный момент единственным препаратом на рынке из этой группы является алискирен.

Для предсказания активности ингибиторов ренина можно использовать **методы машинного обучения**, что значительно облегчит поиск новых потенциальных лекарственных соединений.

МАТЕРИАЛЫ И МЕТОДЫ

Машинное обучение (machine learning, ML) – совокупность методов искусственного интеллекта, позволяющих строить алгоритмы (модели), которые способны обучаться на каких-либо данных.

QSAR (Quantitative Structure–Activity Relationship) – частный случай применения машинного обучения для построения моделей, способных по химическому строению молекул предсказывать их различные свойства.

QSAR для предсказания **активности** соединений можно использовать, как для задачи классификации (то есть отнесения молекулы к классу активных, либо неактивных соединений), так и для задачи регрессии (прогнозирования числовых показателей активности соединения).

В данной работе методы машинного обучения использовались **для классификации** молекул на активные и неактивные.

МАТЕРИАЛЫ И МЕТОДЫ

Скрипты для обработки данных и построения модели были написаны на языке программирования **Python**. Для 1D представления структуры молекул использовался генератор фингерпринтов (FingerprintGenerator) из библиотеки **RDKit**. Построение модели машинного обучения осуществлялось с помощью алгоритма “случайных лесов” (random forest) из программной библиотеки **scikit-learn**.

МАТЕРИАЛЫ И МЕТОДЫ

Для отбора ингибиторов ренина использовалась база данных **ChEMBL**. Всего было найдено 5154 лиганда.

Для дальнейшей работы были отобраны лиганды со следующими свойствами:

- Тип измеренной активности (standard_type): **IC50**;
- Единица активности (standard_units): **nM**;
- Тип анализа (assay_type): **B** (binding);
- Анализируемый организм (assay_organism): **Homo sapiens**;
- Целевой организм (target_organism): **Homo sapiens**.

Данным условиям соответствовало **2190** соединений.

```
data_last = data_ed.query("standard_type ='IC50' &
standard_units ='nM' & assay_type='B' & assay_organism ='Homo
sapiens'& target_organism ='Homo sapiens'")
```

МАТЕРИАЛЫ И МЕТОДЫ

Фингерпринты – представление молекул в виде битовой строки, где каждый бит соответствует наличию (1) либо отсутствию (0) в молекуле какой-то определенной структуры. В данной работе каждый лиганд был закодирован в 2048-битную строку.

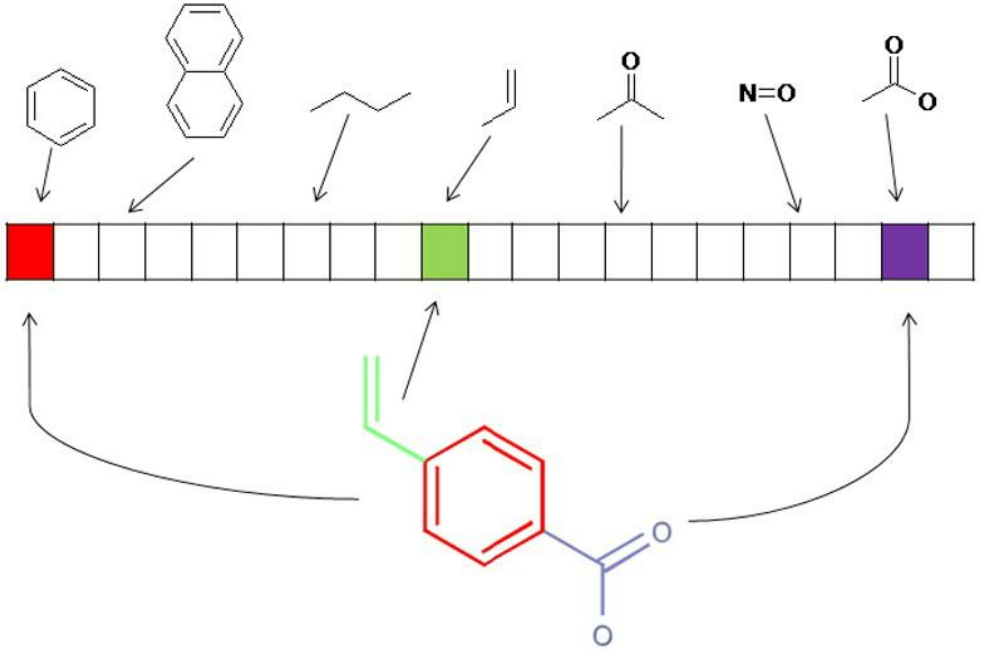

Рис. 1 – Схематичное представление принципа генерации фингерпринтов.

МАТЕРИАЛЫ И МЕТОДЫ

Решающее дерево (дерево принятия решений, decision tree) – алгоритм машинного обучения, структура которого представляет собой “узлы” и “листья”, где каждый узел – это какое-либо условие, а “лист” – это результат, получаемый при соблюдении либо несоблюдении данного условия.

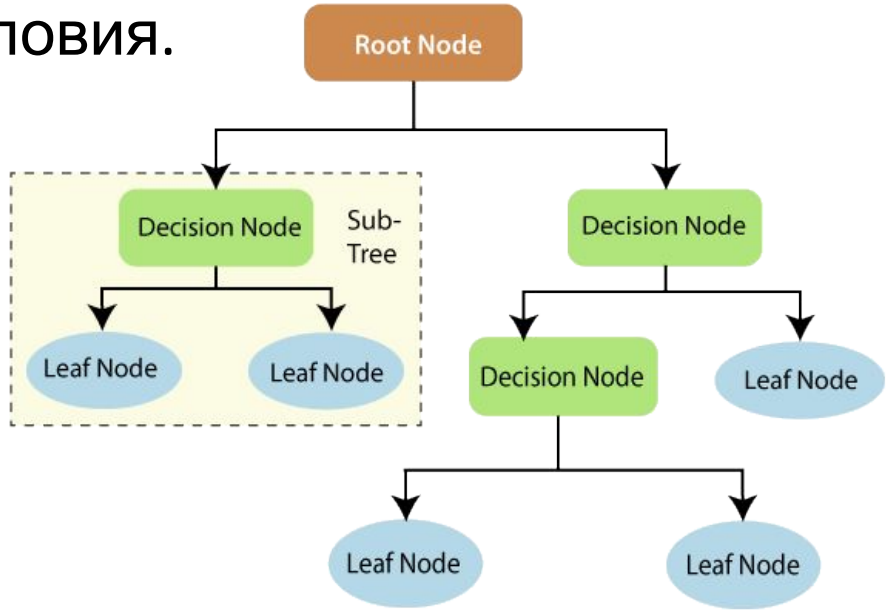

Рис. 2 – Обобщенная схема дерева решений.

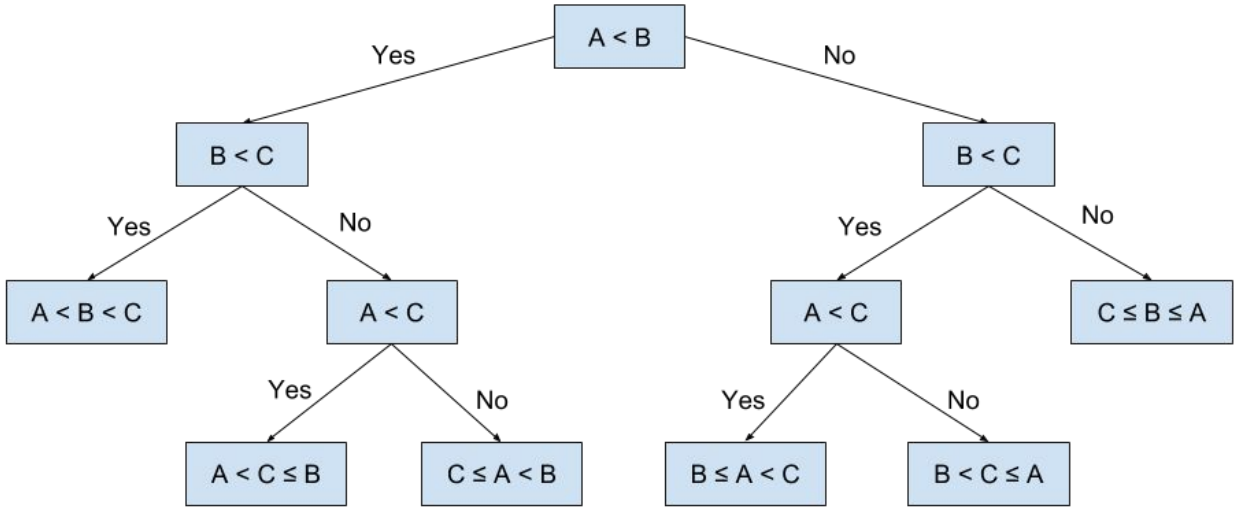

Рис. 3 – Схема дерева решений, сравнивающего три числа друг с другом.

МАТЕРИАЛЫ И МЕТОДЫ

Случайный лес (random forest) – алгоритм машинного обучения, основанный на использовании множества решающих деревьев. Данный метод позволяет получить более точные результаты, так как невысокое качество прогноза каждого дерева корректируется предсказаниями других деревь

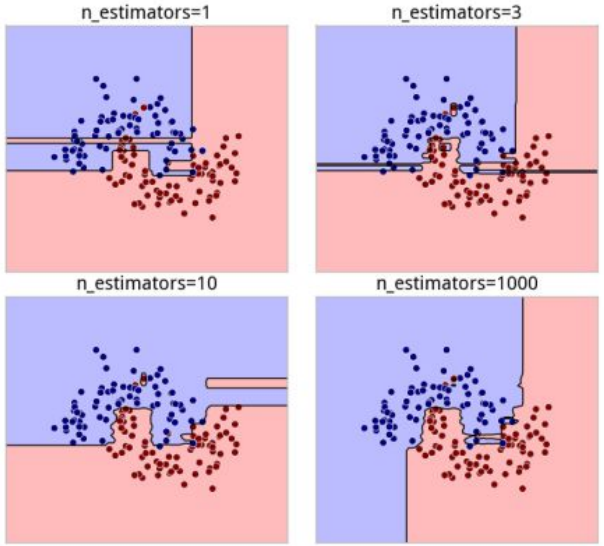

Рис. 4 – Увеличение точности предсказания с увеличением числа деревьев.

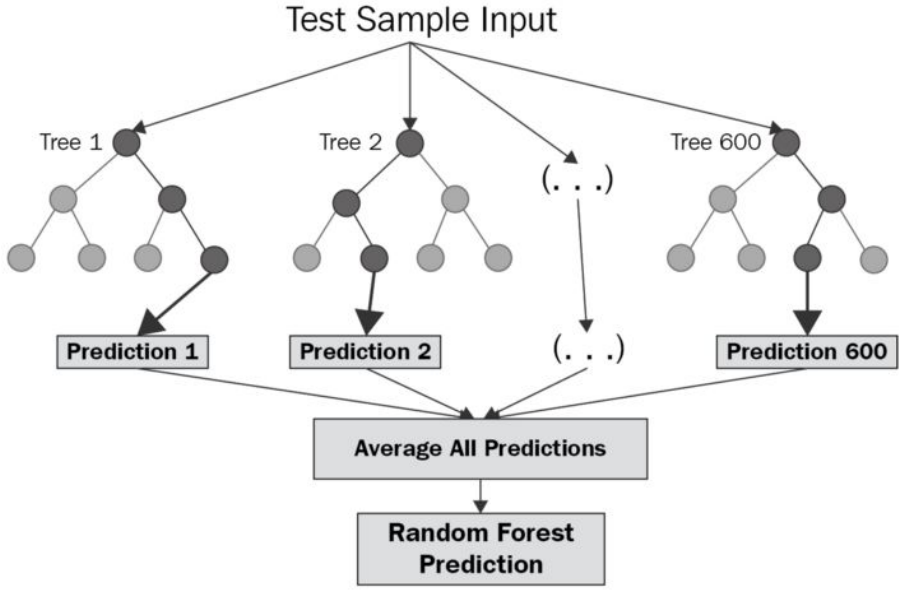

Рис. 5 – Схематичное представление случайного леса

МАТЕРИАЛЫ И МЕТОДЫ

В качестве меры активности молекул использовалась IC50. Для удобства анализа данной меры активности был получен ее логарифмический показатель - **pIC50**. Активными считались соединения, у которых **pIC50 ≥ 6**, а неактивными - соединения, у которых pIC50 < 6.

```
data['bioactivity_class'] = np.where(data['pIC50'] >= 6.0, 1, 0)
#1 - active, 0 - inactive
```

Для проверки правильности работы модели перед ее построением набор данных был разделен случайным образом на **тестовую** (30 % данных) и **тренировочную** (70 %) части (train_test_split, библиотека scikit-learn).

```
from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size = 0.3)
```

РЕЗУЛЬТАТЫ И ИХ ОБСУЖДЕНИЕ

С помощью метода RandomForestClassifier была построена QSAR модель “случайных лесов”. Обучение модели проводилось на тренировочных данных. В качестве **независимых** переменных передавались значения фингерпринтов, в качестве **зависимой** – класс соединения: активное или неактивное.

Для поиска оптимальных значений гиперпараметров модели использовался метод **RandomizedSearchCV**, позволяющий задать

```
randomized_search_cv_clf_rf = RandomizedSearchCV(clf_rf, parameters, cv=5)
randomized_search_cv_clf_rf.fit(x_train, y_train)
randomized_search_cv_clf_rf.best_estimator_
```

Лучшей оказалась модель с числом “деревьев” (**n_estimators**), равным 14, глубиной “деревьев” (**max_depth**), равной 19, и минимальным количеством соединений для разделения узла “дерева” (**min_samples_split**), равным 17.

РЕЗУЛЬТАТЫ И ИХ ОБСУЖДЕНИЕ

Метрики качества модели составили:

- **Точность** предсказания на **тренировочных данных**: 0,9472;
- **Точность** предсказания на **тестовых данных**: 0,9482;
- **Precision** на **тестовых данных** (отношение true positives к сумме true positives и false positives): 0,949;
- **Recall** на **тестовых данных** (отношение true positives к сумме true positives и false negatives): 0,997;
- **F1-score** (двойное произведение precision и recall, деленное на их сумму): 0,972.

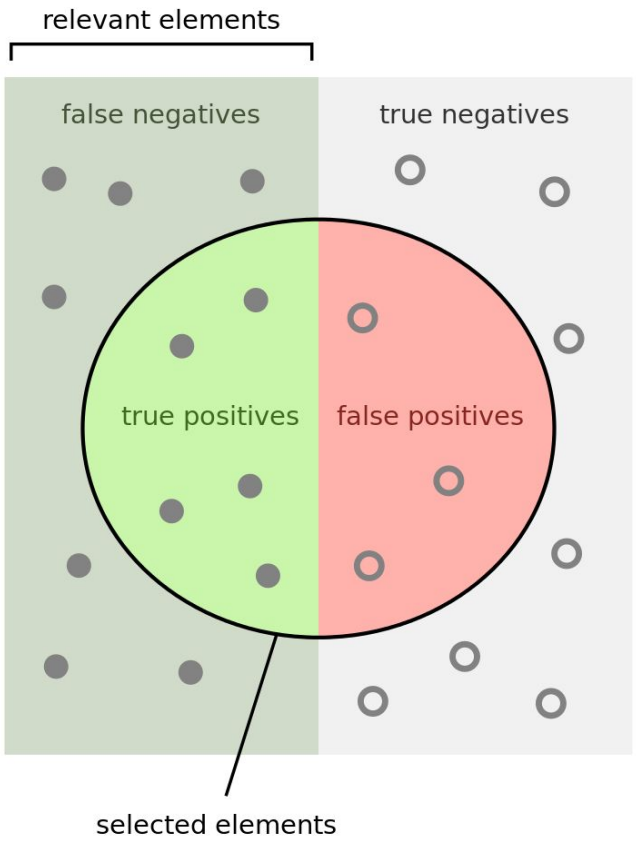

Рис. 6 – Разделение объектов на группы в зависимости от реальных и предсказанных значений.

РЕЗУЛЬТАТЫ И ИХ ОБСУЖДЕНИЕ

Конфузионная матрица (confusion matrix) – таблица, представляющая результаты предсказания в сравнении с реальными классами данных.

		Настоящий класс	
		Active	Inactive
Предсказанный класс	Active	598 (TP)	32 (FP)
	Inactive	2 (FN)	25 (TN)

Табл. 1 – Конфузионная матрица для тестовых данных (TP - true positives, FP - false positives, FN - false negatives, TN – true negatives).

РЕЗУЛЬТАТЫ И ИХ ОБСУЖДЕНИЕ

По метрикам качества модели можно сделать вывод, что данная модель **находит практически все активные соединения** из имеющегося набора молекул (recall = 0,997, что очень близко к 1,0) , однако иногда ошибочно причисляет неактивные соединения к активным (precision = 0,949, что тоже достаточно близко к 1,0).

На следующем этапе работы полученная модель была использована для проверки активности соединений, найденных с помощью построения **фармакофора**.

Для построения фармакофора использовались только активные молекулы.

```
data_only_active = data.query("pIC50 >= 6.0")
```

РЕЗУЛЬТАТЫ И ИХ ОБСУЖДЕНИЕ

Была проведена **кластеризация** активных лигандов с помощью **Butina Clustering**. Всего было получено 77 кластеров. Для построения фармакофора использовались центроиды четырех кластеров.

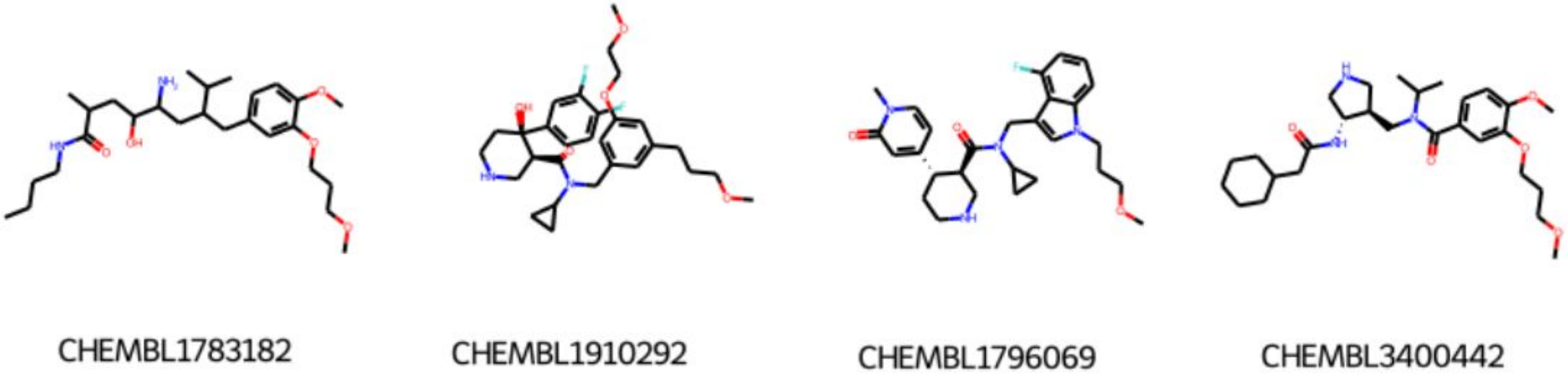

Рис. 7 – Центроиды кластеров, использованных для построения фармакофора.

РЕЗУЛЬТАТЫ И ИХ ОБСУЖДЕНИЕ

Построение фармакофора проводилось с помощью **алгоритма MAPex**. Полученный фармакофор имеет 4 фармакофорных центра: 3 акцептора водорода (показаны красным цветом) и 1 гидрофобный центр (желтый цвет).

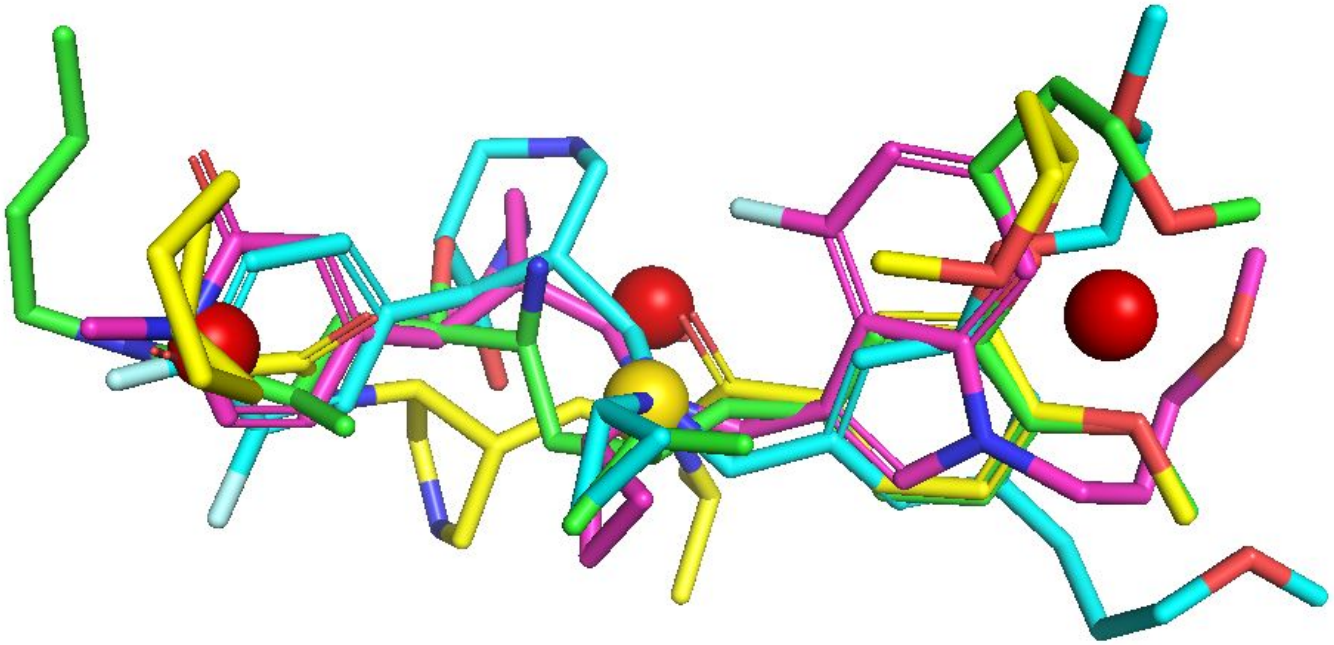

Рис. 8 – Визуальное представление выровненных молекул и фармакофорных центров.

РЕЗУЛЬТАТЫ И ИХ ОБСУЖДЕНИЕ

По данному фармакофору был проведен поиск в базе данных **ZINCPharmer**, были сохранены только те соединения, **RMSD** (root-mean-square deviation, мера среднего расстояния между центрами фармакофора и атомами молекул) которых составило меньше 0,3. Такому критерию удовлетворяло **2602 соединения**.

Далее был проведен анализ данных соединений на предмет совпадения с лигандами, которые уже анализировались. Оставшиеся **2002 соединения** были проверены на соответствие правилам Липински.

```
a = LipinskiCalc(smiles)
data = a.lipinski_table
data_fulfilled = data.query('Fulfill == True')
```

РЕЗУЛЬТАТЫ И ИХ ОБСУЖДЕНИЕ

1811 соединений, которые удовлетворяли критериям Липински были проверены по **фильтру PAINS** (Pan-assay interference compounds).

```
p = PAINS(list(data_fulfilled['Smile’]))
no_pains = p.exclude()
```

Через фильтр прошло **1785 соединений**. Далее они были проверены на предмет наличия нежелательных структур (длинных алифатических цепей, нитрогрупп и т.п.). “Чистых” структур осталось **1245**.

```
u = UnwantedSubs(no_pains)
unwanted, clean = u.get_unwanted()
```

РЕЗУЛЬТАТЫ И ИХ ОБСУЖДЕНИЕ

Для оставшихся **1245 соединений** были получены их фингерпринты. С использованием QSAR модели, которая была построена ранее на тренировочных данных об известных ингибиторах ренина, были предсказаны классы найденных молекул.

```
x_search = np.array(fingerprints_2)
y_search_predicted = clf_rf.predict(x_search)
```

Среди данных молекул **868 оказались активными**. Данные соединения можно использовать в дальнейшем для молекулярного докинга, так как есть большая вероятность, что среди них будут активные ингибиторы ренина.

ВЫВОДЫ

1) QSAR моделирование – это современный способ анализа свойств потенциальных лекарственных веществ, основанный на методах машинного обучения.
2) В ходе данной работы была построена QSAR модель, которая позволяет достаточно эффективно предсказывать активность ингибиторов ренина на основе их химического строения (метрики качества модели составили: precision - 0,949; recall - 0,997; F1-score - 0,972).
3) Данная модель была применена для предсказания активности найденных с помощью фармакофора молекул. Среди этих молекул потенциально активными и удовлетворяющими критериям лекарственных веществ являются 868 соединений.
4) В дальнейшем эти соединения можно использовать для молекулярного докинга.
5) Код и построенная модель доступны на GitHub: https://github.com/walking-chaos/QSAR_random_forest.git