Ligand Pro

# QSAR анализ ингибиторов KRAS

Тиньков Олег Викторович,
к.х.н., e-mail: oleg.tinkov.chem@mail.ru

16 июля, Москва

**QSAR (Quantitative Structure-Activity Relationship)** — это вычислительный метод, который связывает химическую структуру соединений с их биологической активностью или физико-химическими свойствами с помощью математических моделей. Его цель — **предсказать активность новых соединений без дорогостоящих лабораторных испытаний**.

🎯 **Основные задачи QSAR**:

**1.Оптимизация лекарств**: увеличение эффективности, снижение токсичности.

**2.Предсказание ADME-свойств** (всасывание, распределение, метаболизм, выведение).

**3.Оценка токсичности** (экотоксикология, канцерогенность)

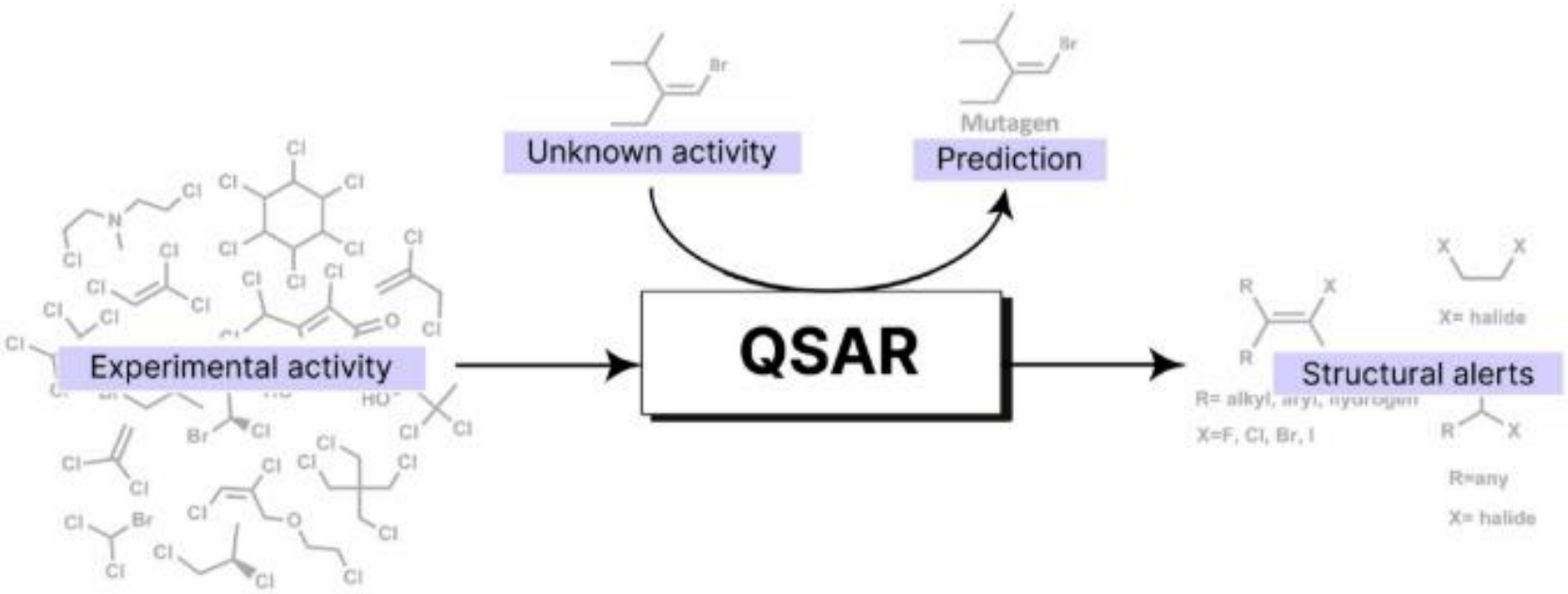
**4.Изучение механизмов** действия химических веществ.

**Примеры успешного применения**:

1. Разработка противогриппозного препарата Zanamivir (Relenza®)

2. Антигипертензивный препарат Captopril (Capoten®)

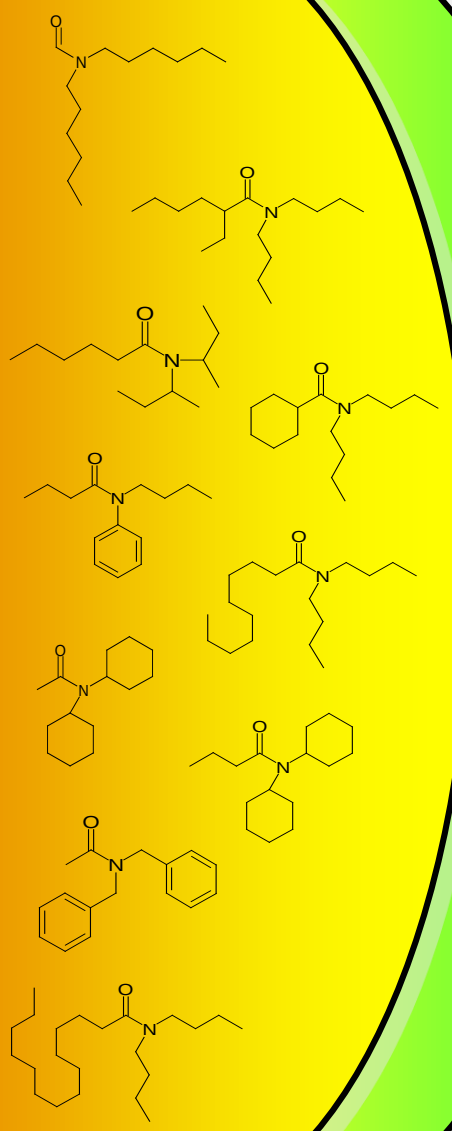
3. Предсказание токсичности фторхинолонов



# Принципы QSAR моделирования

QSAR - Quantitative Structure-Activity Relationship

СОЕДИНЕНИЯ



ДЕСКРИПТОРЫ

Quantitative
Structure
Property
Relationships

**Построение моделей** с использованием методов машинного обучения (PLS, RF, NN, SVM, и др.);

**Валидация моделей,** определение их областей применения (domains applicability), интерпретация.

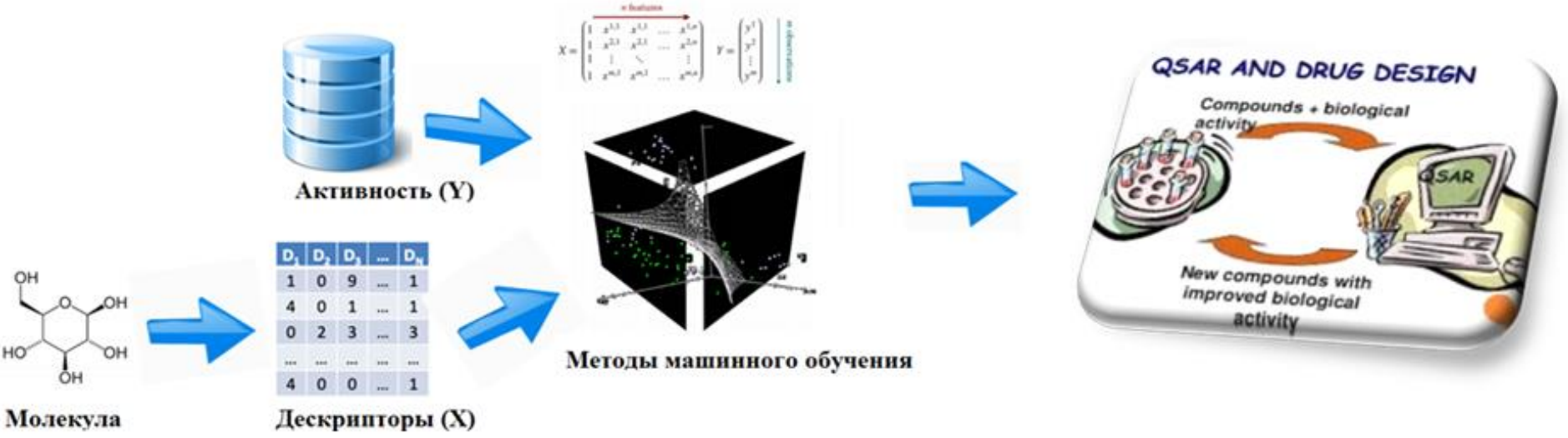
0.613
0.380
-0.222
0.708
1.146
0.491
0.301
0.141
0.956
0.256
0.799
1.195
1.005

СВОЙСТВА

# Цель и задачи этапа исследования:

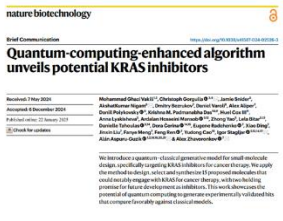
**Целью** данного этапа исследования явилась разработка регрессионных QSAR-моделей для предсказания ингибирующей активности KRAS. Для достижения указанной цели нам необходимо было решить нижеследующие **задачи**:

1. Собрать, проанализировать и проверить выборки соединений по ингибиторам KRAS;
2. Провести разведывательный анализ данных;
3. Разделить общую выборку на обучающей и тестовый наборы;
4. Рассчитать молекулярные дескрипторы;
5. Валидировать QSAR модели.



# Основные этапы исследования:

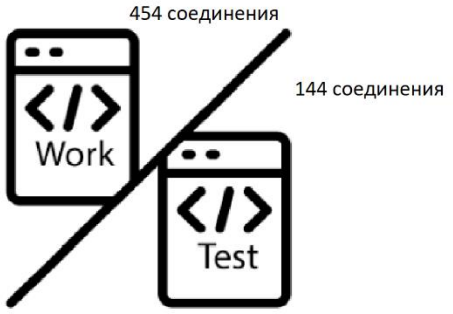
**Источники данных**



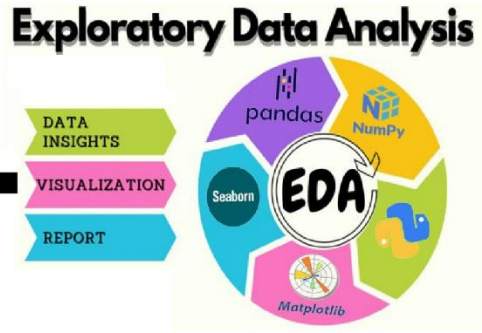

645 соединений

**Сбор данных**

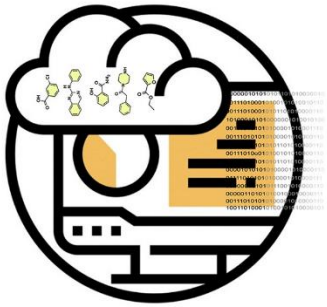

**Проверка и обработка данных**



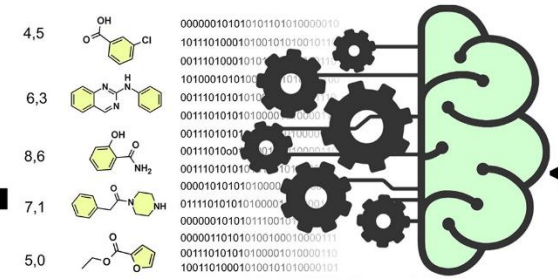

**Разделение на обучающий и тестовый наборы**




**Разведывательный анализ данных**



**Расчёт молекулярных дескрипторов**




$$f(\text{molecule}) = \left[ pIC_{50}^{KRAS} \right]$$

**QSAR моделирование**

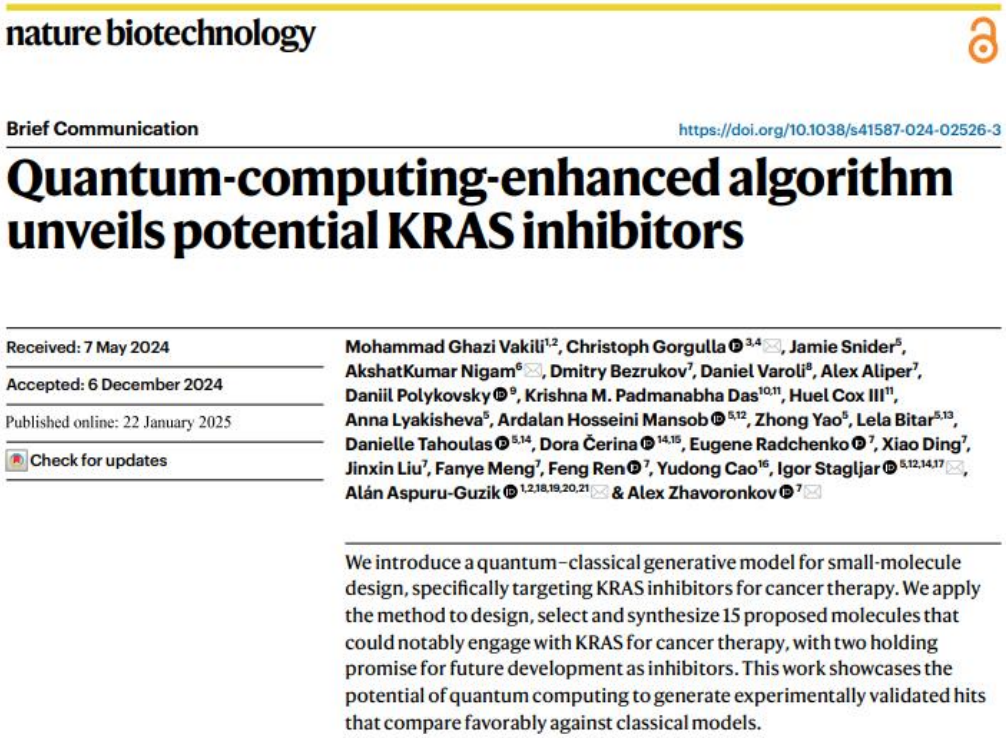
**Задачи на перспективу**

https://github.com/LigandPro/QSAR_KRAS_inhibitors

# Материалы и методы:

## Выборка для QSAR моделирования

Источник данных по ингибиторам KRAS - публикация «Quantum-computing-enhanced algorithm unveils potential KRAS inhibitors»



nature biotechnology

Brief Communication

https://doi.org/10.1038/s41587-024-02526-3

**Quantum-computing-enhanced algorithm unveils potential KRAS inhibitors**

Received: 7 May 2024

Accepted: 6 December 2024

Published online: 22 January 2025

Check for updates

Mohammad Ghazi Vakili[1,2], Christoph Gorgulla[3,4], Jamie Snider[5], AkshatKumar Nigam[6], Dmitry Bezrukov[7], Daniel Varoli[8], Alex Aliper[7], Daniil Polykovsky[9], Krishna M. Padmanabha Das[10,11], Huel Cox III[11], Anna Lyakisheva[5], Ardalan Hosseini Mansob[5,12], Zhong Yao[5], Lela Bitar[5,13], Danielle Tahoulas[5,14], Dora Čerina[14,15], Eugene Radchenko[7], Xiao Ding[7], Jinxin Liu[7], Fanye Meng[7], Feng Ren[7], Yudong Cao[16], Igor Stagljar[5,12,14,17], Alán Aspuru-Guzik[1,2,18,19,20,21] & Alex Zhavoronkov[7]

We introduce a quantum–classical generative model for small-molecule design, specifically targeting KRAS inhibitors for cancer therapy. We apply the method to design, select and synthesize 15 proposed molecules that could notably engage with KRAS for cancer therapy, with two holding promise for future development as inhibitors. This work showcases the potential of quantum computing to generate experimentally validated hits that compare favorably against classical models.

В дополнительных материалах к статье (https://zenodo.org/records/11137638) представлена выборка из 645 соединений.

# Материалы и методы:

## Проверка Выборки для QSAR моделирования

В соответствии с общепринятыми рекомендациями (https://doi.org/10.1021/ci100176x) перед построением QSAR моделей необходима проверка исходных данных по определенному алгоритму

*J. Chem. Inf. Model.* **2010,** *50,* 1189–1204

**Trust, But Verify: On the Importance of Chemical Structure Curation in Cheminformatics and QSAR Modeling Research**

Denis Fourches,[†] Eugene Muratov,[†,‡] and Alexander Tropsha*[,†]

Laboratory for Molecular Modeling, Eshelman School of Pharmacy, University of North Carolina, Chapel Hill, North Carolina 27599, and Laboratory of Theoretical Chemistry, Department of Molecular Structure, A.V. Bogatsky Physical-Chemical Institute NAS of Ukraine, Odessa, 65080, Ukraine

Received May 5, 2010

Проверка выборки проходила в два этапа.

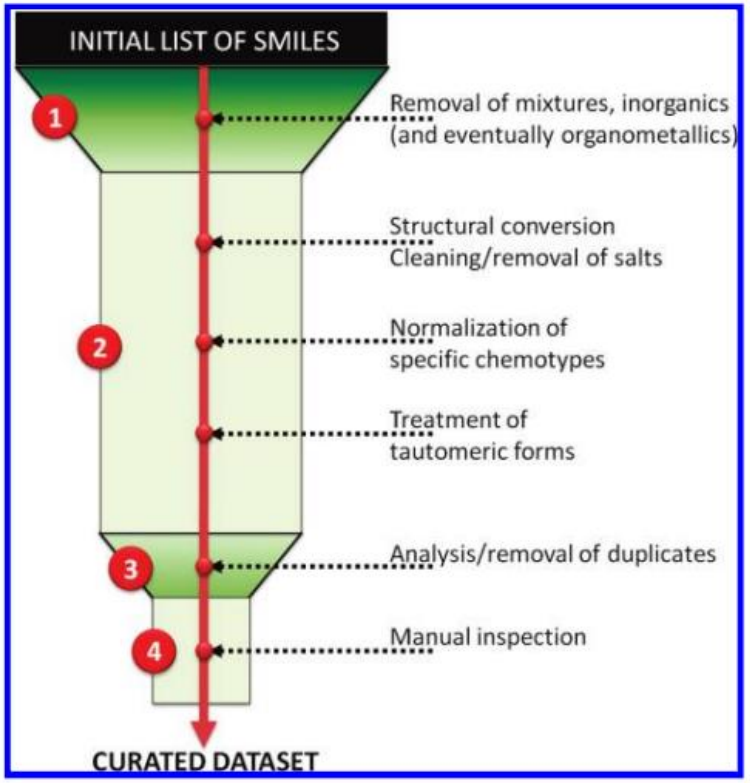


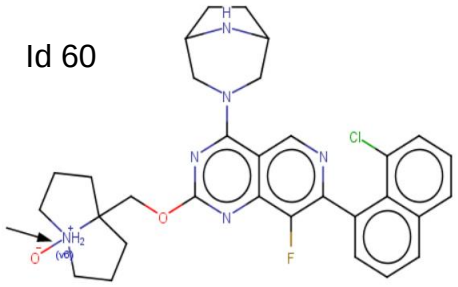

**Figure 1.** General data set curation workflow.

# Материалы и методы:

## I этап проверки выборки для QSAR моделирования

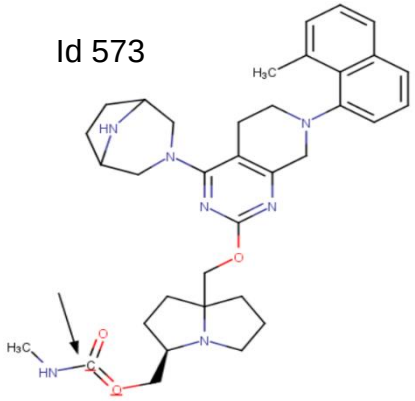
Была проведена корректности SMILES 645 представленных структур. Выявлено 7 ошибочных структур, которые были исключены, так как не представлены общеизвестные идентификаторы (CAS, PubChem CID, ChEMBL ID и др.) .

Id 60



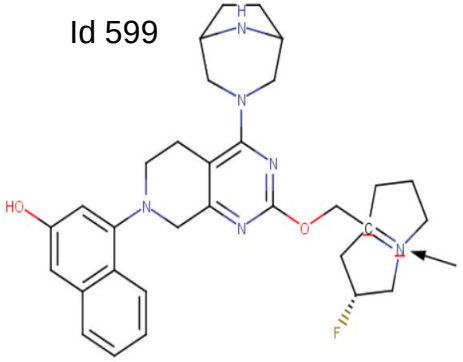
Explicit valence for ato m # 1 N, 6, is greater t han permitted

Id 573



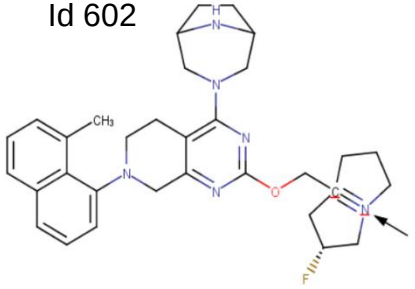
Explicit valence for atom # 2 C, 5, is greater than permitted

Id 599



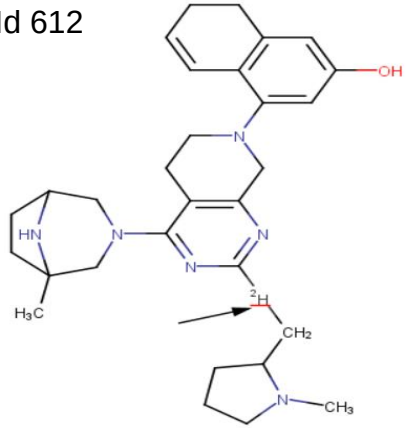
Explicit valence for atom # 27 C, 5, is greater than permitted

Id 602



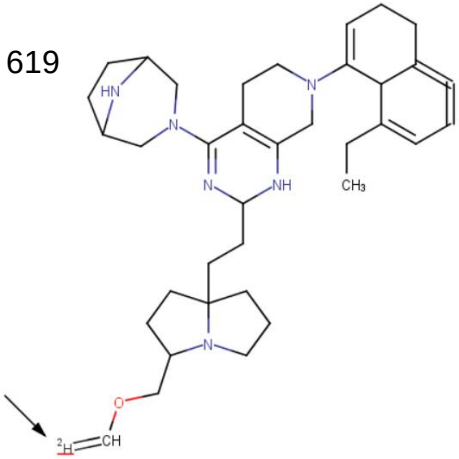
Explicit valence for atom # 25 C, 6, is greater than permitted

Id 612



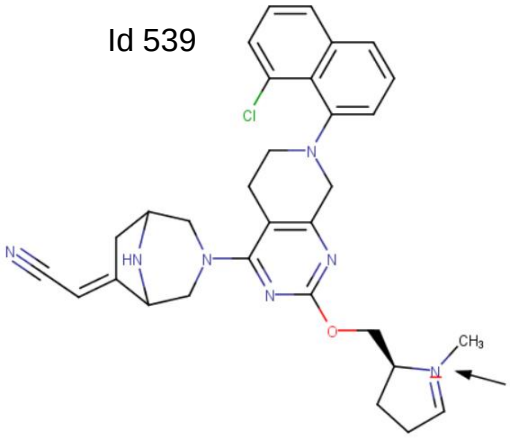
Explicit valence for atom # 7 H, 2, is greater than permitted

Id 619



Explicit valence for atom # 0 H, 2, is greater than permitted

Id 539



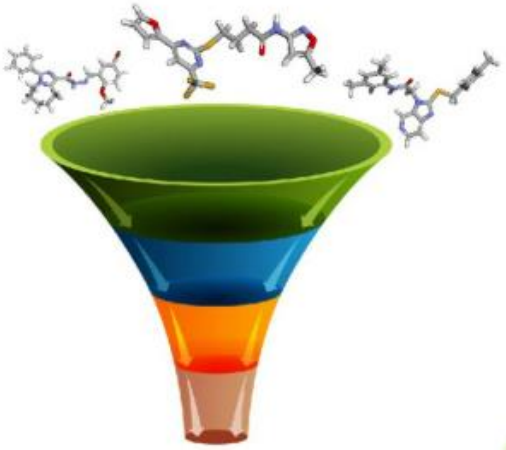
Explicit valence for atom # 1 N, 4, is greater than permitted

# Материалы и методы:

## II этап проверки и предобработки выборки для QSAR моделирования

Оставшиеся после I этапа 638 структур были дополнительно проверены:

1. Поскольку было принято решение разрабатывать регрессионные QSAR модели, позволяющие прогнозировать $IC_{50}$ в отношение KRAS, при формировании выборки в нее включались только те записи, для которых поле «KRAS G12D binding IC50 (nM)» имело фактические значения, при этом исключались записи, содержащие знаки «>» или «<» - размер выборки сократился до 603 соединений;
2. Экспериментальные значения активности, выраженные с помощью концентрации полумаксимального ингибирования ($IC_{50}$, nM), были сконвертированы в отрицательный десятичный логарифм данной величины ($pIC_{50}$)
3. Смеси, соединения в солевой форме не включались в общую выборку – исключено 4 соединения (id 118, 119, 121, 182)



# Материалы и методы:

II этап проверки и предобработки выборки для QSAR моделирования

4. Поскольку планировалось использовать 2D дескрипторы, представленные в оригинальной выборке isomeric SMILES были конвертированы в канонические SMILES;
5. Для соединений, имеющих два и более экспериментальных значения $pIC_{50}$, вычислялось среднее и стандартное отклонение. В выборку с рассчитанным средним значением активности включали только те соединения, для которых стандартное отклонение $pIC_{50}$ в логарифмических величинах не превышало 0.5, в соответствии с предложенным ранее подходом обработки дубликатов химических соединений [https://doi.org/10.1021/acs.jcim.9b00526] - 568 соединений;
6. С целью оценки предсказательной способности QSAR моделей, первоначальную общую выборку разделяли на обучающий (ws) и тестовый (ts) наборы. Для этого общий набор соединений упорядочивали по возрастанию $pIC_{50}$ и каждое пятое соединение помещали в тестовый набор, оставшиеся соединения представляли собой обучающий набор. Исходя из этого, обучающий набор представлен 454 соединениями, тестовый набор – 144 соединением.

# Материалы и методы:

## EDA обучающей и тестовой выборки перед QSAR моделированием

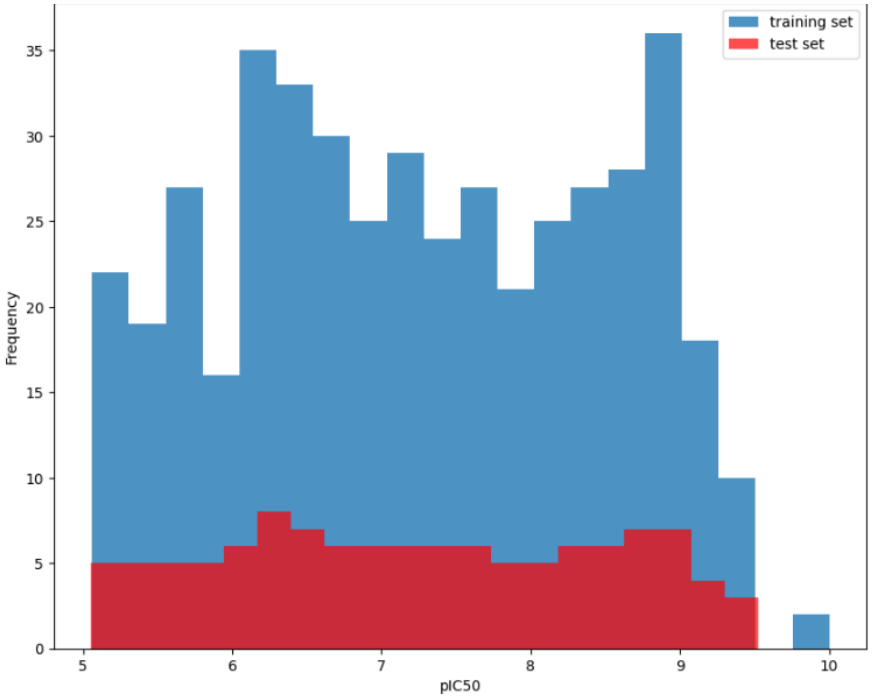



Рис. 2. Распределение значений ингибирующей активности в обучающей и тестовой выборках

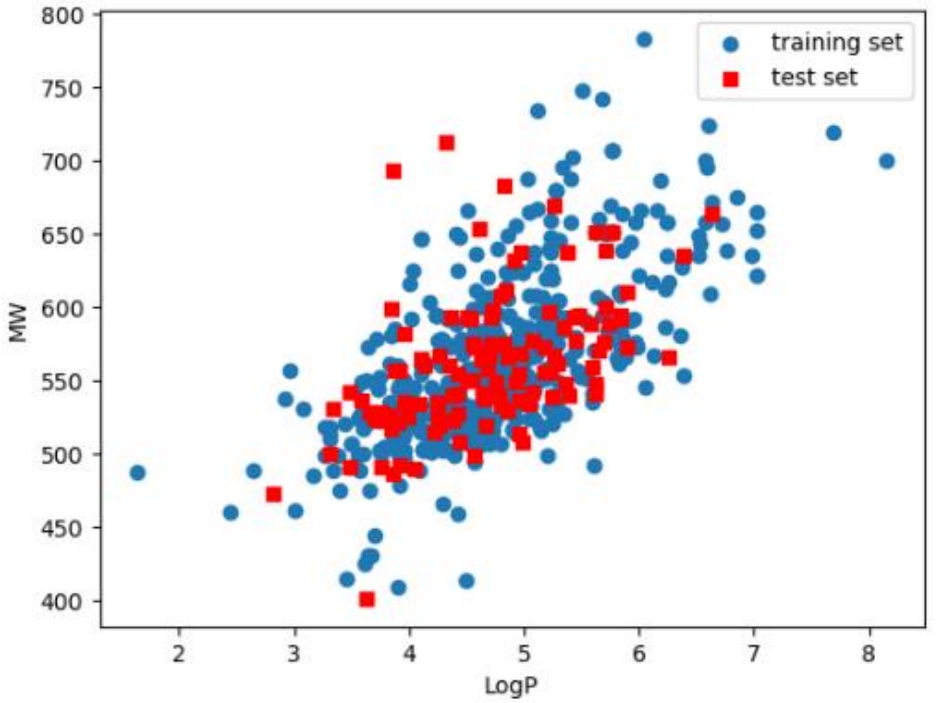



Рис. 3. Визуализация химического пространства обучающей и тестовой выборок в системе координат молекулярный вес – коэффициент липофильности

Анализируя на рисунке 2 распределение соединений по экспериментальным значениям ингибирующей активности можно отметить схожесть интервалов и частот величины $pIC_{50}$ для обучающей и тестовой выборок. Как показано на рисунке 2 для исследуемых соединений характерен достаточно широкий интервал $pIC_{50}$ - более пяти логарифмических единиц, что позитивно влияет на описательную и предсказательную способности разрабатываемых QSAR моделей. Ранее [https://pubs.acs.org/doi/10.1021/ci050413p] было показано, что для получения адекватной QSAR модели необходим интервал исследуемой активности, по крайне мере, в одну логарифмическую единицу.

Анализируя рисунок 3 можно отметить достаточно высокий уровень химического разнообразия обучающей и тестовой выборок соответственно, о чём могут свидетельствовать широкие диапазоны молекулярной массы и липофильности изучаемых веществ.

# Материалы и методы:

## EDA обучающей и тестовой выборки перед QSAR моделированием

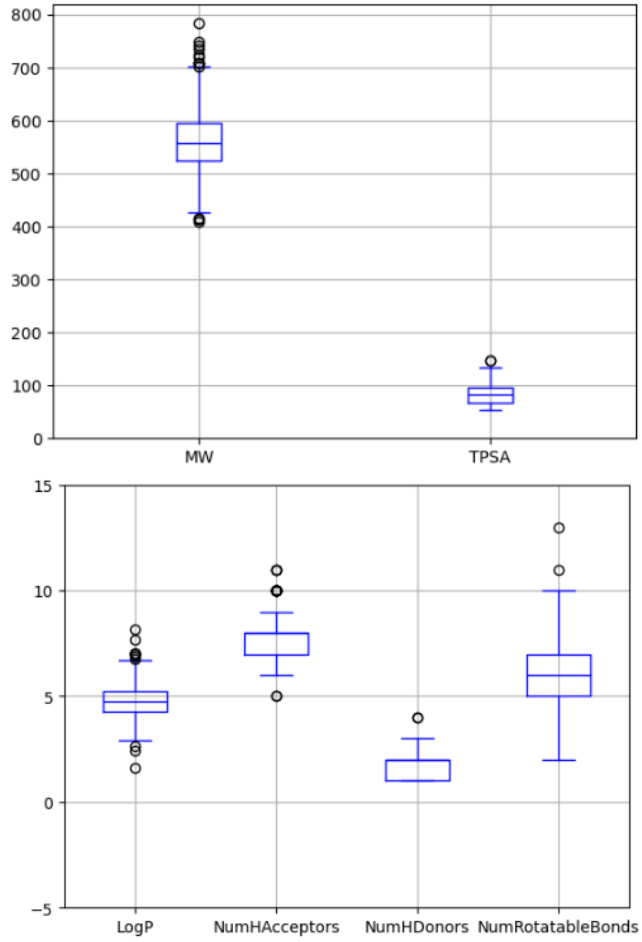



Рис. 4. Визуализация химического пространства обучающей выборки

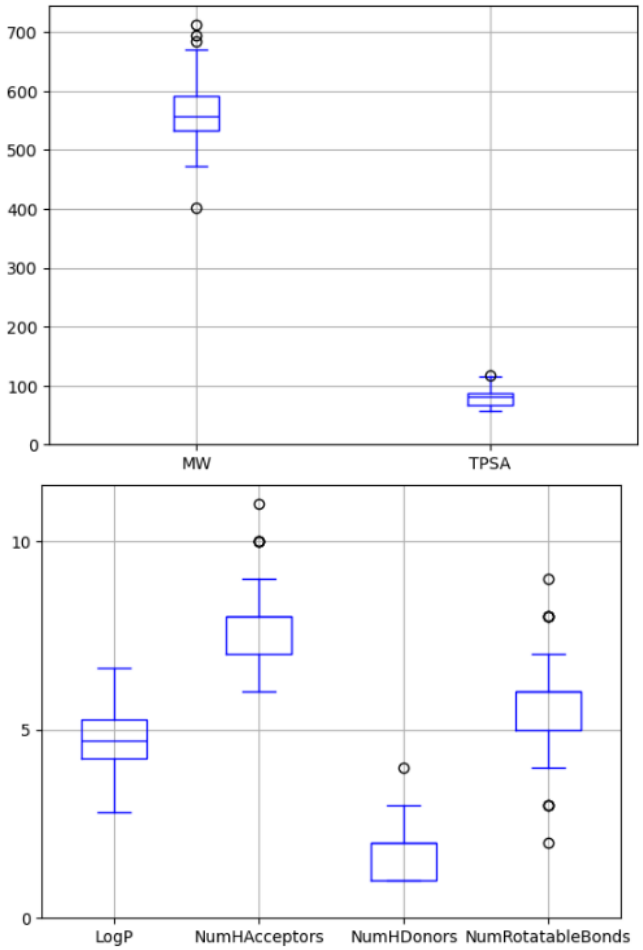



Рис. 5. Визуализация химического пространства тестовой выборки с помощью количества доноров, акцепторов водорода, вращающихся связей

При этом данные рисунков 4, 5 (средние величины total polar surface area (TPSA) ≤ 140 Å2, hydrogen bond donor < 5, hydrogen bond acceptor and rotatable bonds < 10) указывают на то, что большинство соединений в обучающей и тестовой выборках не в полной мере соответствуют общеизвестным требованиям биодоступности, в частности, Lipinski's rule, Veber filter. Таким образом, использование данных выборок при разработке QSAR моделей снижает вероятность выявления новых соединений среди биодоступных веществ, что, в конечном счете, понижает результативность высокопроизводительного виртуального скрининга. Вывод: в последующем нужен поиск иных выборок (chemble)

# Материалы и методы:

## Средства разработки

Python:

- **RDKit** - библиотека для решения различных задач в хемоинформатике https://github.com/rdkit/rdkit-tutorials;
- **PaDELPy** – библиотека для расчета молекулярных дескрипторов https://github.com/ecrl/padelpy;
- **Scikit-lear** – библиотека, в которой реализованы различные методы машинного обучения https://github.com/scikit-learn/scikit-learn
- **CatBoost** – библиотека для построения моделей градиентным бустингом https://catboost.ai/

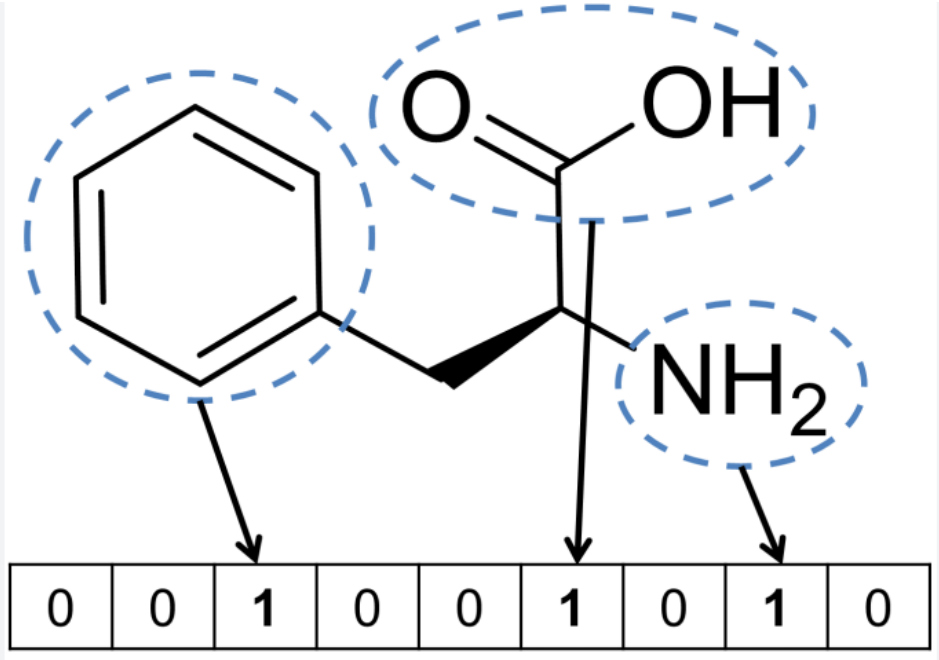
# Материалы и методы:

## Методы построения QSAR моделей

1. Метод градиентного бустинга (Gradient Boosting Method, CatBoost)
2. Метод опорных векторов (Support Vector Machine, SVM)
3. Multi-layer Perceptron (MLP) Regressor - нейронная сеть прямого распространения

## Молекулярные дескрипторы

1) Morgan fingerprints (MF);
2) MACCS-166.
3) PubChem fingerprints;
4) 2D дескрипторы RDKit;



# Материалы и методы:

## Валидация и применимость QSAR моделей

Для оценки устойчивости и косвенно предсказательной способности моделей, была использована пятикратная кросс-валидация (5-fold cross validation, CV).

Вхождение соединений тестовой выборки в область применимости (Applicability Domain, AD) рассчитывали с помощью «расстояния сходства» (similarity distance). Считается, что соединение тестовой выборки принадлежит области применимости QSAR модели, если его расстояние сходства не превышает порогового значения Dc, вычисляемого по формуле (1):

$$D_c = Z\sigma + \bar{y} \qquad (1)$$

где: $\bar{y}$ и σ – это, соответственно, среднее значение и стандартное отклонение величин евклидового расстояния в химическом пространстве дескрипторов между всеми объектами из обучающей выборки и их ближайшими соседями в ней; Z – константа, которая, как правило, принимается равной 0.5. Охват данных (data coverage, Cov) в области применимости рассчитывался как отношение числа соединений из тестовой выборки, вошедших в область применимости, к общему числу соединений тестовой выборки.

# Критерии оценки качества QSAR моделей

$$Q^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - y_{mean})^2}$$

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

где m-количество молекул в выборке; $y_i$ -заданное значение активности, $\hat{y}_i$ -вычисленное (предсказанное) значение активности для i-го соединения, $\hat{y}_i^{cv}$ - значение активности, вычисленное в условиях кросс-валидации с выбросом по одному для i-той молекулы, $y_{i\,mean}$ – среднее значение активности для всех соединений обучающей выборки.

# Результаты QSAR моделирование

## Статистические характеристики QSAR ингибиторов KRAS

| Descriptors | Method | Training set, 5-fold CV | | Test set | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | All compounds | | Cov | Compounds in AD | |
| | | $Q^2_{cv}$ | RMSE | $Q^2_{ts}$ | RMSE | | $Q^2_{ts}$ | RMSE |
| **MF** | **CATBOOST** | **0.61** | **0.76** | **0.7** | **0.67** | | **0.73** | **0.61** |
| MF | SVM | 0.64 | 0.73 | 0.76 | 0.61 | **0.8** | 0.76 | 0.57 |
| MF | MLP | 0.59 | 0.78 | 0.68 | 0.69 | | 0.69 | 0.65 |
| | | | | | | | | |
| MACCS | CATBOOST | 0.45 | 0.9 | 0.51 | 0.86 | | 0.4 | 0.9 |
| MACCS | SVM | 0.45 | 0.91 | 0.48 | 0.88 | 0.74 | 0.37 | 0.91 |
| MACCS | MLP | 0.42 | 0.93 | 0.48 | 0.88 | | 0.38 | 0.91 |
| | | | | | | | | |
| PubChem | CATBOOST | 0.56 | 0.81 | 0.69 | 0.68 | | 0.66 | 0.67 |
| PubChem | SVM | 0.58 | 0.79 | 0.67 | 0.7 | 0.76 | 0.61 | 0.72 |
| PubChem | MLP | 0.55 | 0.82 | 0.6 | 0.78 | | 0.51 | 0.81 |
| | | | | | | | | |
| RDKit | CATBOOST | 0.59 | 0.78 | 0.66 | 0.72 | | 0.65 | 0.7 |
| RDKit | SVM | 0.58 | 0.8 | 0.57 | 0.81 | 0.86 | 0.6 | 0.75 |
| RDKit | MLP | 0.46 | 0.9 | 0.49 | 0.88 | | 0.51 | 0.83 |

Выделенная модель, разработанные с использованием Morgan fingerprints и градиентного бустинга, по показателям коэффициента детерминации для обучающей выборки в условиях пятикратной кросс-валидации, а также для тестовой выборки, соответствуют общепризнанным требованиям к адекватным QSAR моделям, пригодным для целей регулирования ($Q^2_{cv} > 0.5$; $Q^2_{test} > 0.6$) [https://onlinelibrary.wiley.com/doi/10.1002/minf.201000061 ]

## Заключение:

- ✓ Проведена валидация и препроцессинг выборки соединений с экспериментальными значениями $IC_{50}$ ,
- ✓ Проведен разведывательный анализ данных
- ✓ Разработанные QSAR-модели показали высокую устойчивость, предсказательную способность, что подтверждено кросс-валидацией, внешним независимым тестированием

**Планы:**

- ✓ Провести структурную интерпретацию разработанных QSAR моделей + SAR анализ с целью выявления наиболее значимых фрагментов, повышающих и понижающих ингибирующую активность;
- ✓ Интегрировать предложенные модели в веб-приложение KRAS_VS_assistant, которое позволит проводить виртуальный скрининг с одновременной оценкой острой токсичности потенциальных ингибиторов KRAS G12D

# СПАСИБО ЗА ВНИМАНИЕ!