



Квадрицепс

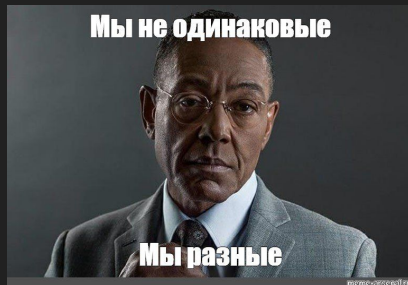
datacon2025

in silico drug discovery
Болезнь Альцгеймера

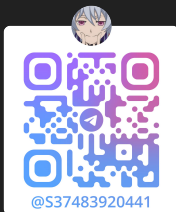
Проблематика

Основные сложности:

- Труднопрогнозируемая **проходимость** через ВВВ и активный транспорт
- Мультифакторная патогенез/политаргетность (амилоид, тау, нейровоспаление, митохондрии) **усложняют** **single-target** стратегии
- **Глубокая гетерогенность** пациентов → слабая переносимость моделей, «one-size-fits-all» не работает
- **Исторические провалы** (соланезумаб, и др.) показывают неуверенность в выборе мишени и gap между in silico / in vivo / клиникой
- Скудные CNS-AD **датасеты** → переобучение, bias в ML-моделях
- Гибкие и неструктурированные мишени (tau, α -Syn) делают docking менее надёжным
- **Необходимость** моделировать **полифармакологию** (MTDL) и сетевые эффекты, а не одну IC₅₀

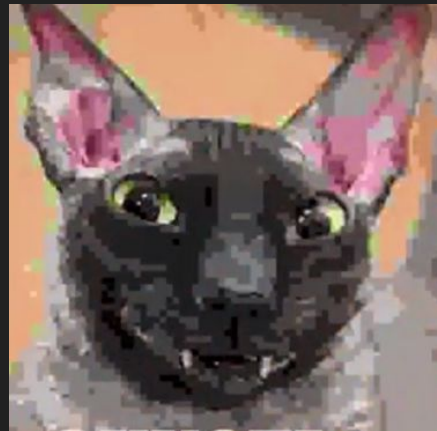


Команда



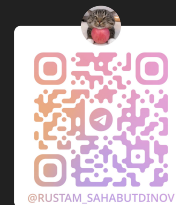
Клещенок Максим

mle, mlops, devops



Сахабутдинов Рустам

ds, mle








Плайлайн

- Анализ и выбор мишени
- Сбор данных и QSAR моделирование
- Генерация молекул
- Подготовка лигандов
- Докинг
- Отбор хитов

Выбор мишени

DYRK1A - оптимальный выбор

DYRK1A выбран как оптимальная мишень потому что:

-  Научная актуальность: Участвует в патогенезе Тау и Аβ
-  Данные по лигандам: Десятки известных ингибиторов с IC₅₀ в нМ
-  Структура: Решены кристаллографические структуры с лигандами
-  In silico дизайн: Классический АТР-карман, хорошо изучен
-  Лекарственная перспективность: Селективные ингибиторы уже показали эффективность в доклинике

Преимущества DYRK1A:

- Более специфичная экспрессия vs GSK-3β (меньше побочных эффектов)
- Воздействует на две ключевые патологии одновременно (Тау + Аβ)
- Успешные доклинические результаты (SM07883)
- Проходит через ГЭБ, обратимое действие



rnd

Сбор данных и QSAR моделирование

- Сбор данных: Извлечение данных по активности соединений из ChEMBL
- Расчет дескрипторов: Мордред, PaDEL, RDKit дескрипторы
- Feature selection: Отбор наиболее информативных признаков
- Scaffold split для химически корректной валидации
- XGBoost

Test RMSE 0.650 | R^2 0.696

Генерация молекул



В процессе разработки были опробованы различные методы генерации молекул:

- VAE модели: `SELFIES VAE`, `Transformer VAE` - показали околонулевую валидность сгенерированных структур
- Fine-tuning: DPO (Direct Preference Optimization) и RLHF (Reinforcement Learning from Human Feedback) - сложность в настройке и нестабильность обучения
- Docking-guided generation: Попытки генерации с учетом docking scores - технические сложности интеграции (не успели)



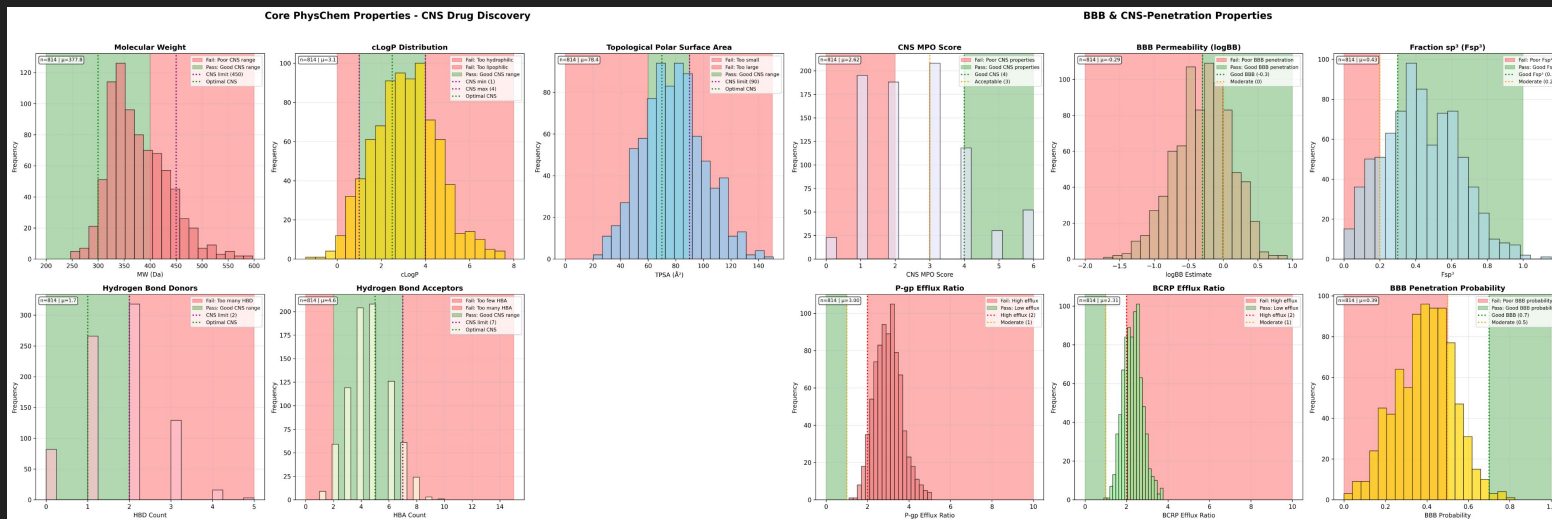
Финальный подход: Остановились на `fine-tuning` предобученной модели (`entropy/gpt2_zinc_87m`) на 3 эпохи, так как:

- Loss стабильно падал с $2.71 \rightarrow 0.66 \rightarrow 0.32$
- Дальнейшее обучение приводило к переобучению и падению метрик
- Достигнута валидность 981 молекулы из 1000 сгенерированных

Виртуальный скрининг

- Подготовка белка: Очистка и подготовка структуры DYRK1A
- Подготовка лигандов: Конвертация в PDBQT формат
- GPU-ускоренный докинг: AutoDock Vina с CUDA
- Ранжирование: Композитный скор с учетом активности, липидофильности, токсичности

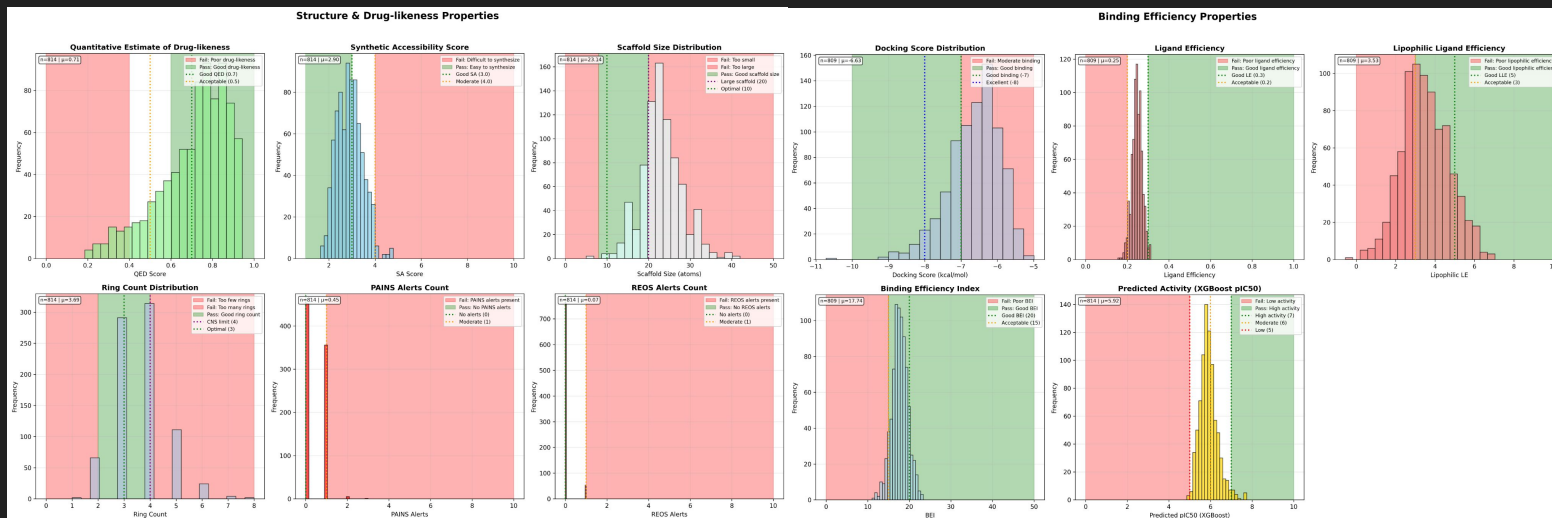
Метрики сгенерированных молекул(1)



Core Physicochemical Properties

BBB & CNS Penetration

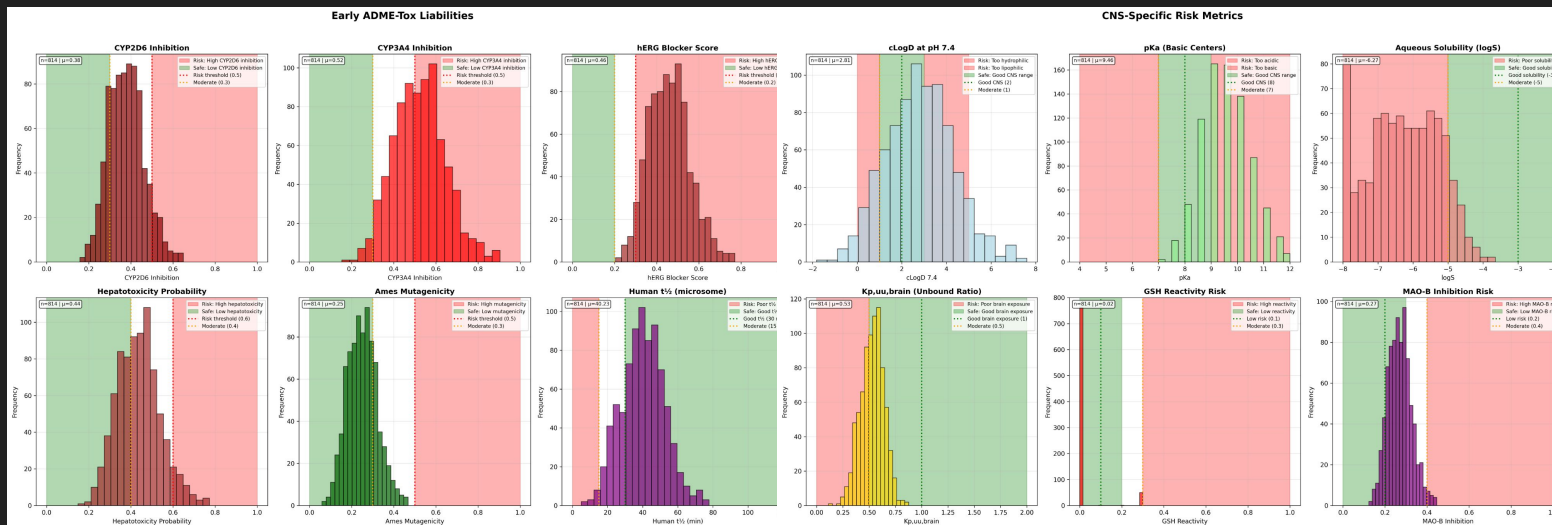
Метрики сгенерированных молекул(2)



Structure & Drug-likeness

Binding Efficiency

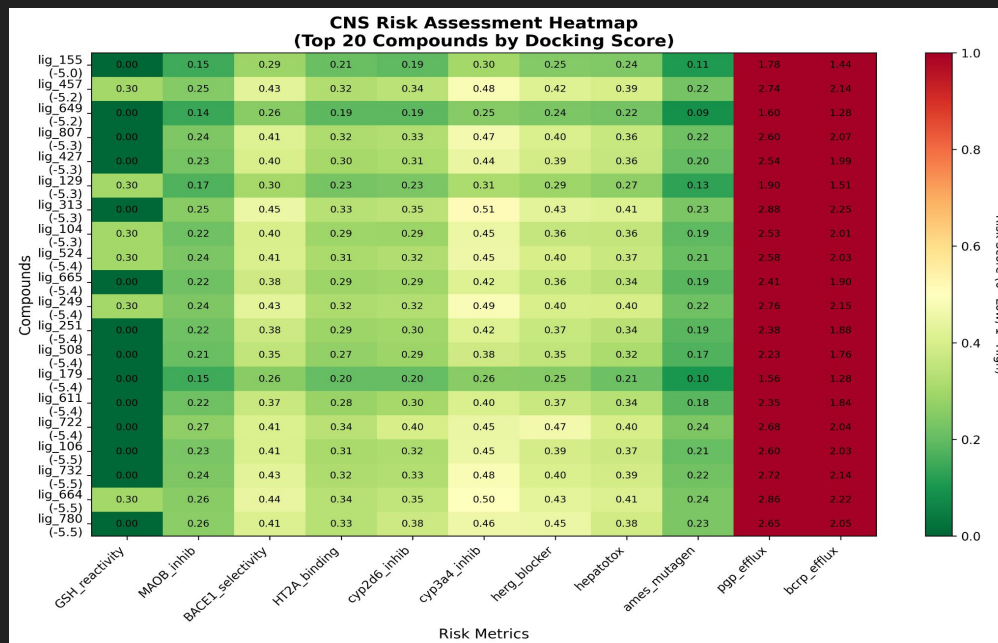
Метрики сгенерированных молекул(3)



ADME & Toxicity

CNS-Specific Parameters

Метрики сгенерированных молекул



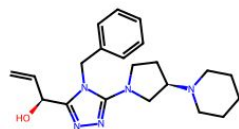
CNS Risk Assessment Heatmap

Финальные критерии отбора молекул

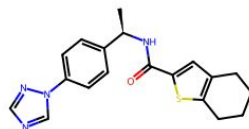
Базовый минимум

- > - P-gp efflux: должно быть <2 (чем меньше, тем лучше проникновение в мозг)
- > - Docking: должно быть < -8 (чем меньше, тем выше аффинность)
- > - hERG: должно быть <0.3 (чем меньше, тем ниже риск кардиотоксичности)
- > - QED: 0.4-0.8 (оптимальный диапазон лекарственной привлекательности)
- > - CNS_MPO: ≥ 4 (оптимально для CNS)
- > - TPSA: <70 (оптимально для проникновения через ГЭБ)
- > - logBB: > -0.3 (лучше проникновение через ГЭБ)

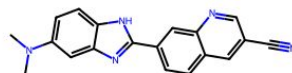
Топ5 молекул (Общие признаки)



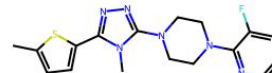
lig_107
Criteria: 10/13



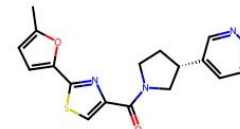
lig_109
Criteria: 10/13



lig_23
Criteria: 10/13



lig_77
Criteria: 10/13



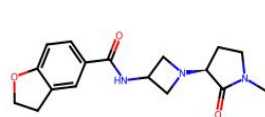
lig_90
Criteria: 10/13

| | MW | LogP | TPSA | HBD | HBA | RotB | RingCount | CNS_MPO | QED | LE | LLE | docking_score | logBB_est | bbb_prob |
|---------|-------|------|------|-----|-----|------|-----------|---------|------|------|------|---------------|-----------|----------|
| lig_107 | 367.2 | 2.61 | 57.4 | 1 | 6 | 6 | 4 | 5 | 0.80 | 0.25 | 4.02 | -6.63 | -0.15 | 0.39 |
| lig_109 | 352.1 | 3.70 | 59.8 | 1 | 5 | 4 | 4 | 4 | 0.78 | 0.25 | 2.68 | -6.37 | -0.02 | 0.48 |
| lig_23 | 313.1 | 3.72 | 68.6 | 1 | 4 | 2 | 4 | 4 | 0.61 | 0.26 | 2.55 | -6.27 | -0.15 | 0.46 |
| lig_77 | 358.1 | 2.71 | 50.1 | 0 | 7 | 3 | 4 | 5 | 0.72 | 0.27 | 4.01 | -6.72 | -0.17 | 0.42 |
| lig_90 | 339.1 | 3.74 | 59.2 | 0 | 5 | 3 | 4 | 4 | 0.73 | 0.26 | 2.58 | -6.32 | -0.15 | 0.48 |

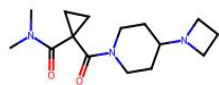
ОДНАКО:

Ни одна молекула не проходит все строгие критерии, но топ-5 молекул проходят 10 из 13 критериев.
Основные барьеры – P-gp efflux, docking score и hERG.

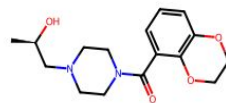
Top5 молекул (composite score)



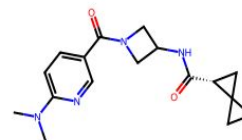
lig_674
Score: 0.207



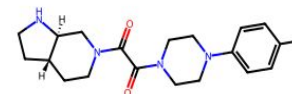
lig_155
Score: 0.093



lig_232
Score: 0.081



lig_794
Score: 0.066



lig_601
Score: 0.056

| | MW | LogP | TPSA | HBD | HBA | RotB | RingCount | CNS_MPO | QED | LE | LLE | docking_score | logBB_est | bbb_prob |
|---------|-------|------|------|-----|-----|------|-----------|---------|------|------|------|---------------|-----------|----------|
| lig_674 | 315.2 | 0.27 | 61.9 | 1 | 4 | 3 | 4 | 4 | 0.87 | 0.26 | 5.79 | -6.06 | -0.58 | 0.20 |
| lig_155 | 279.2 | 0.55 | 43.9 | 0 | 3 | 3 | 3 | 4 | 0.71 | 0.25 | 4.44 | -4.99 | -0.41 | 0.26 |
| lig_232 | 306.2 | 0.60 | 62.2 | 1 | 5 | 3 | 3 | 4 | 0.89 | 0.25 | 4.94 | -5.54 | -0.53 | 0.22 |
| lig_794 | 314.2 | 0.89 | 65.5 | 1 | 4 | 4 | 4 | 4 | 0.90 | 0.25 | 4.75 | -5.64 | -0.54 | 0.24 |
| lig_601 | 356.2 | 0.85 | 55.9 | 1 | 4 | 1 | 4 | 4 | 0.76 | 0.27 | 6.17 | -7.02 | -0.40 | 0.26 |

Composite Score - наш композитный скор, учитывающий:

Docking score (связывание с мишенью), ADMET свойства (токсичность, метаболизм), CNS-специфичные параметры (проникновение в мозг), Лигандную эффективность (LE, LLE), Химическую привлекательность (QED, SA score)

ЧТО МЫ НЕ ПОПРОБОВАЛИ А СТОИЛО

Базовый минимум

> - модель на CNS-активных молекул

Роскошный максимум

> - REINVENT 4, MolDQN, GENTRL, GraphAF (+ RL-fine-tune

> - ChemCrow, DrugGen, InstructMol / MolReGPT, ChemGPT-подобные

> - DiffDock-guided RL, «Docking ENRICH» workflow, sample-efficient RL-AL

> - DiffDock (лиганд-белок), DiffDock-PP (белок-белок)

> - AlphaFold 3, ESMFold, Oxford ODDI AD-KG

СПАСИБО ЗА ХАКАТОН

мы хотим еще хакатонов, но уже с лидербордом



project repo

