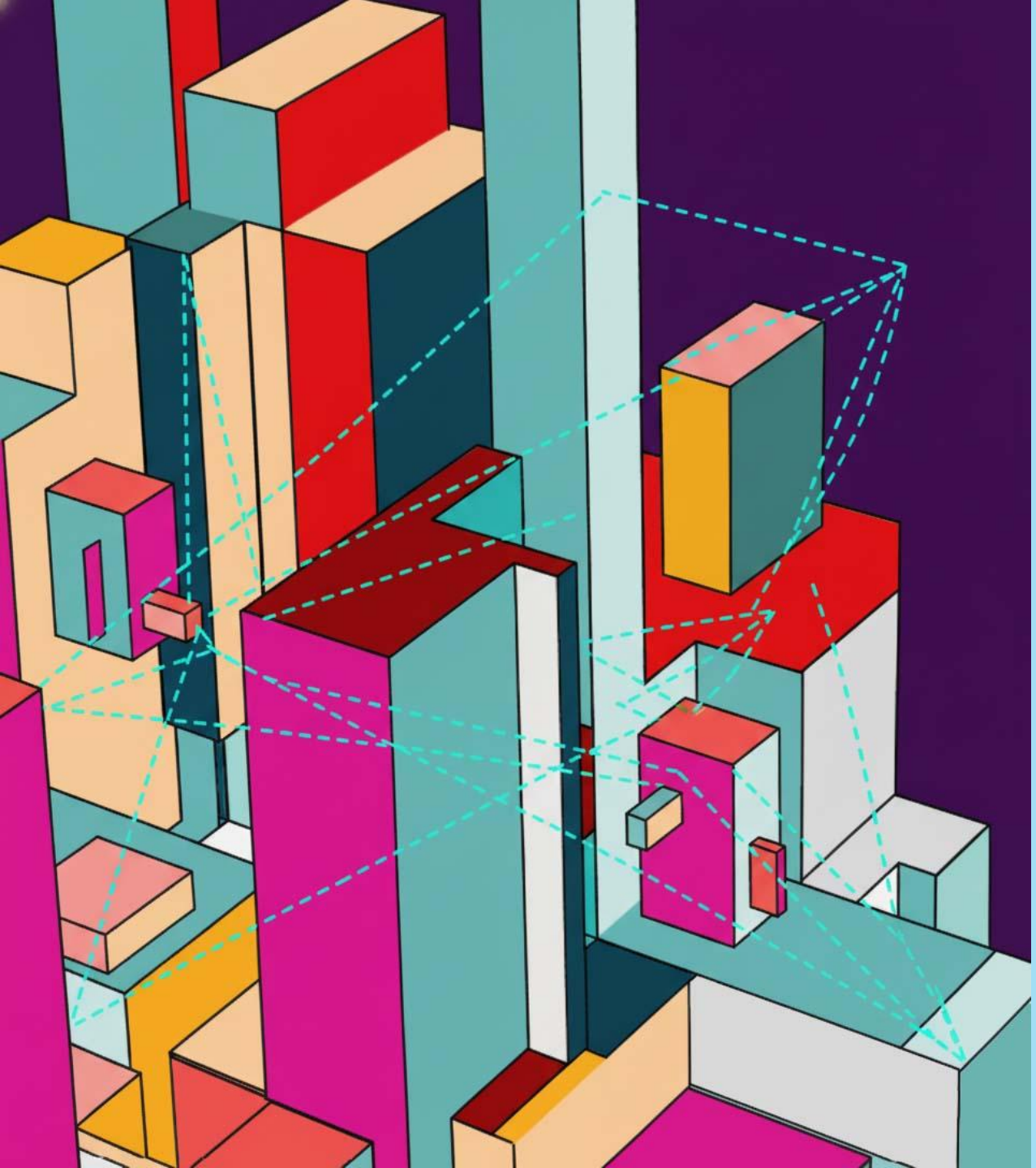


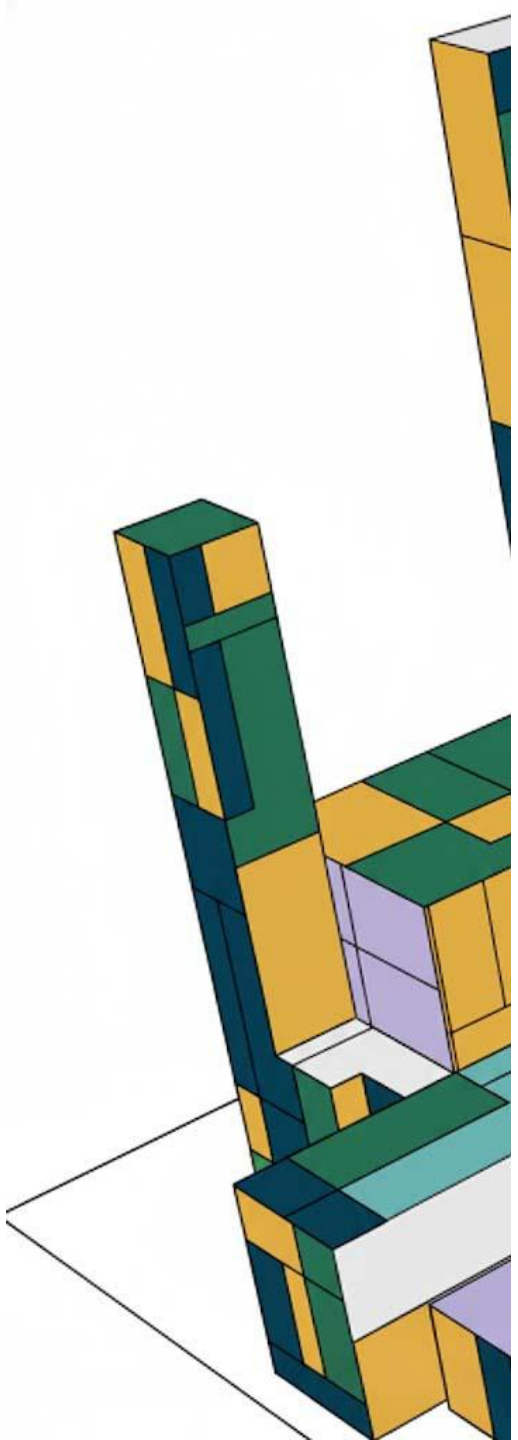
# ОПРЕДЕЛЕНИЕ РЕЗИСТЕНТНОСТИ И КЛАССИФИКАЦИЯ ДОМИНИРУЮЩИХ В РФ СУБТИПОВ ВИЧ-1 С ПРИМЕНЕНИЕМ МЕТОДОВ ГЛУБОКОГО ОБУЧЕНИЯ НА ПРОГРАММНО-АППАРАТНОМ КОМПЛЕКСЕ ONSITESEQ

МФТИ

OnSiteSeq.io

# СПИКЕР

*Роман Горбенко. Автор внутривузовского (МФТИ) medtech-стартапа OnSiteSeq по предсказанию лекарственной устойчивости бактериальных и вирусных инфекций с помощью нанопорового секвенирования ML и GPGPU-вычислений.*
*Ведущий специалист в АО “Валента Фарм”*
*Научный руководитель: Радченко Никита Сергеевич*



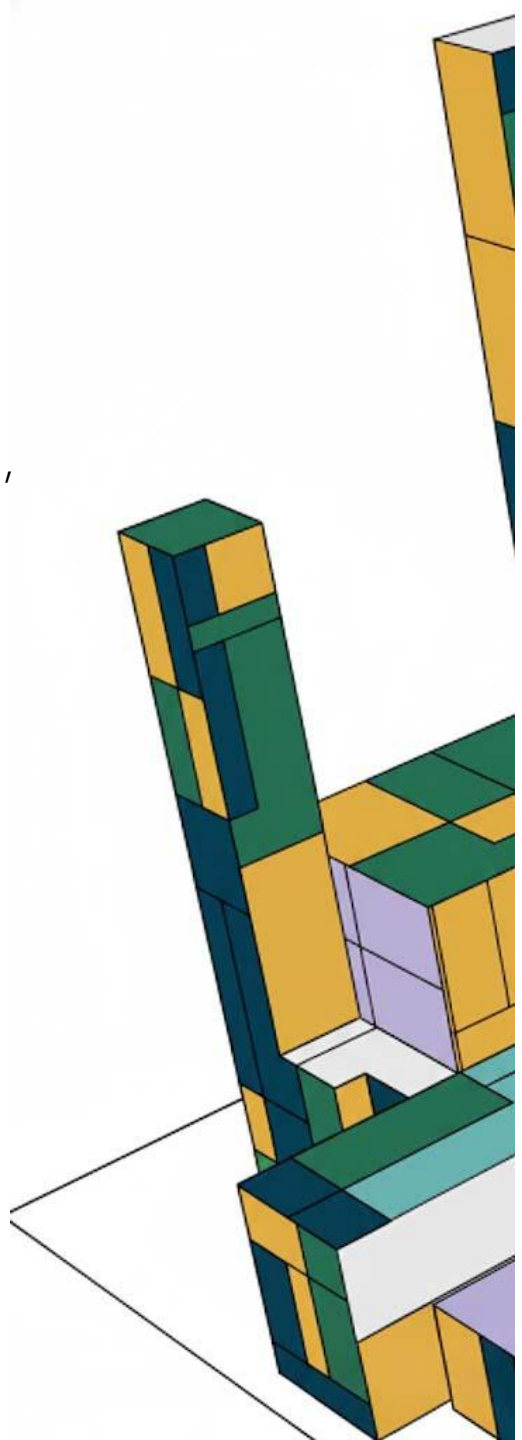
# КОНТЕКСТ ПРОБЛЕМЫ

В Российской Федерации в настоящее время синдемия*, образованная двумя наиболее социально значимыми патогенами (**ВИЧ**-1 **и Туберкулёз**)

***Синдемия** - термин, предложенный антропологом Мерриллом Сингером, описывает ситуацию, когда две и более болезни не просто сосуществуют, а взаимно усиливают друг друга, нанося организму урон, больший, чем простая сумма заболеваний.

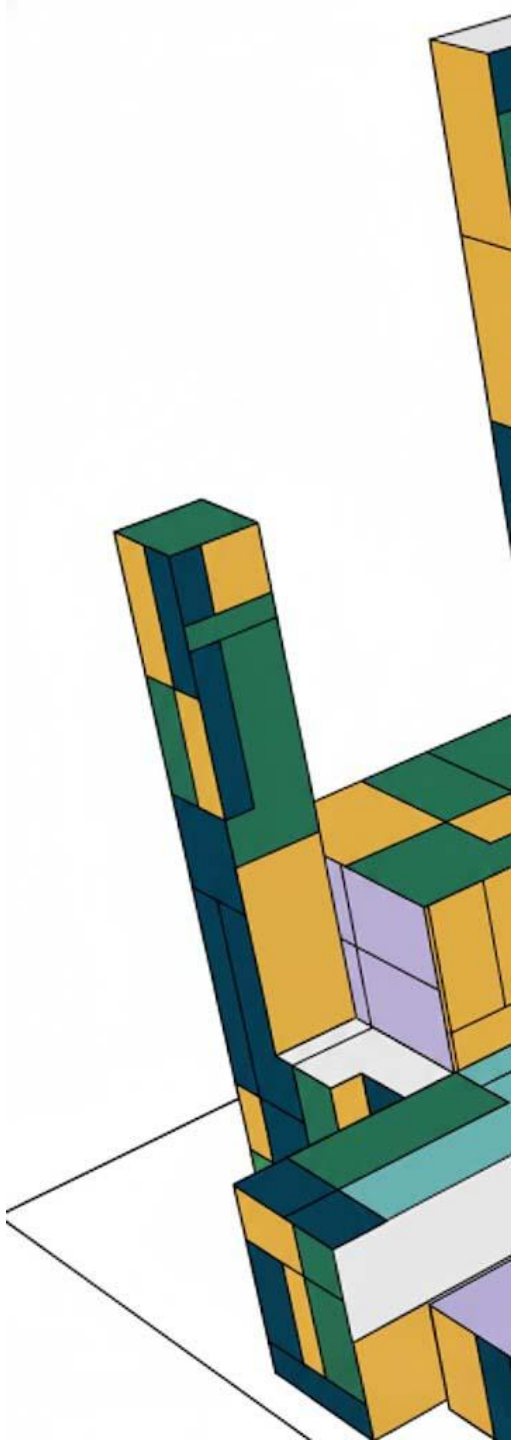
По данным Роспотребнадзора, каждый год от ВИЧ-ассоциированных заболеваний умирает не менее 30 тысяч человек – больше 80 в день. В 2024 году умерли 33 269 человек с ВИЧ, в 2023 году – 34 254. Средний возраст умершего – 42 года.
Главная причина летальных исходов среди ВИЧ-положительных – туберкулез, на него приходится около 39%.

В качестве ключевого драйвера синдемии исследователи называют аномально высокое по обще мировым меркам количество лекарственно-устойчивых субтипов вируса ВИЧ-1 и M. Tuberculosis.



# ПАК ONSITESEQ. АРХИТЕКТУРА HARDWARE


Программно-аппаратный комплекс Onsiteseq. Слева направо: (A) Кассета Flongle (B) Нанопоровый NGS сенвенатор “Нанопорус” (C) Сенсорный экран (D) Шприц-дозатор (E) Коммуникационный блок Wi-Fi/4G (F) CPU/GPU процессор на базе платформы Nvidia Jetson (G) Амплификатор (H) Термопринтер.



# ПАК ONSITESEQ. ПРИМЕРЫ ОТЧЁТОВ (ВИЧ-1)

## Результат исследования: Резистентность и субтип ВИЧ-1

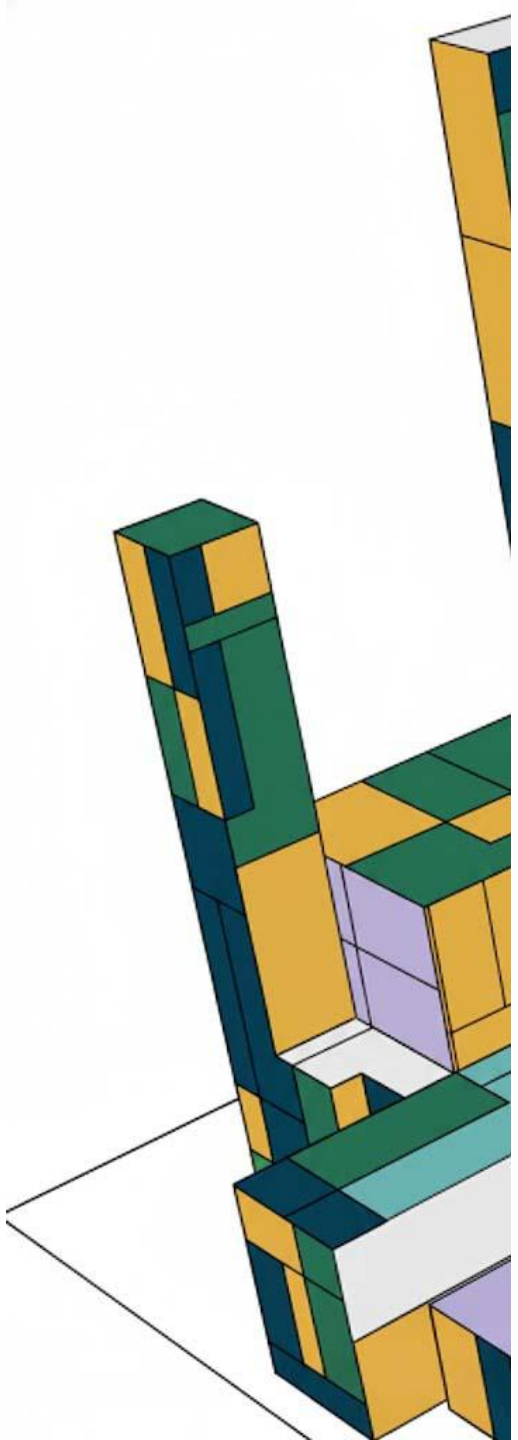
**ID образца:** DRR537804_1_full.fasta
**Дата формирования:** 2026-02-24 16:59:47
**Версия пейплайна onsiteseq_hiv:** 1.0.0
**Версия модуля HIV-1-M-Env-Rus:** 1.0.0
**Версия модуля HIV-1-Resist-Rus:** 1.0.0

### 1. Определение субтипа

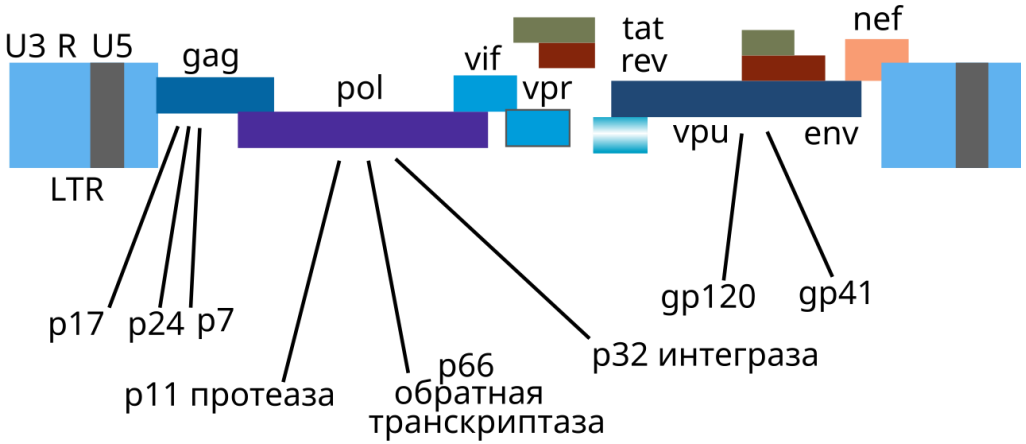
| | |
|---|---|
| **Определенный субтип** | **B** |
| **Уверенность нейросети** | 99.98% |

### 2. Профиль резистентности к АРВТ

| Препарат (Код) | Чувствительность |
|---|---|
| Ламивудин (3TC) | Восприимчив (95.6%) |
| Абакавир (ABC) | Восприимчив (97.1%) |
| Зидовудин (AZT) | Восприимчив (98.7%) |
| Тенофовир (TDF) | Восприимчив (99.4%) |



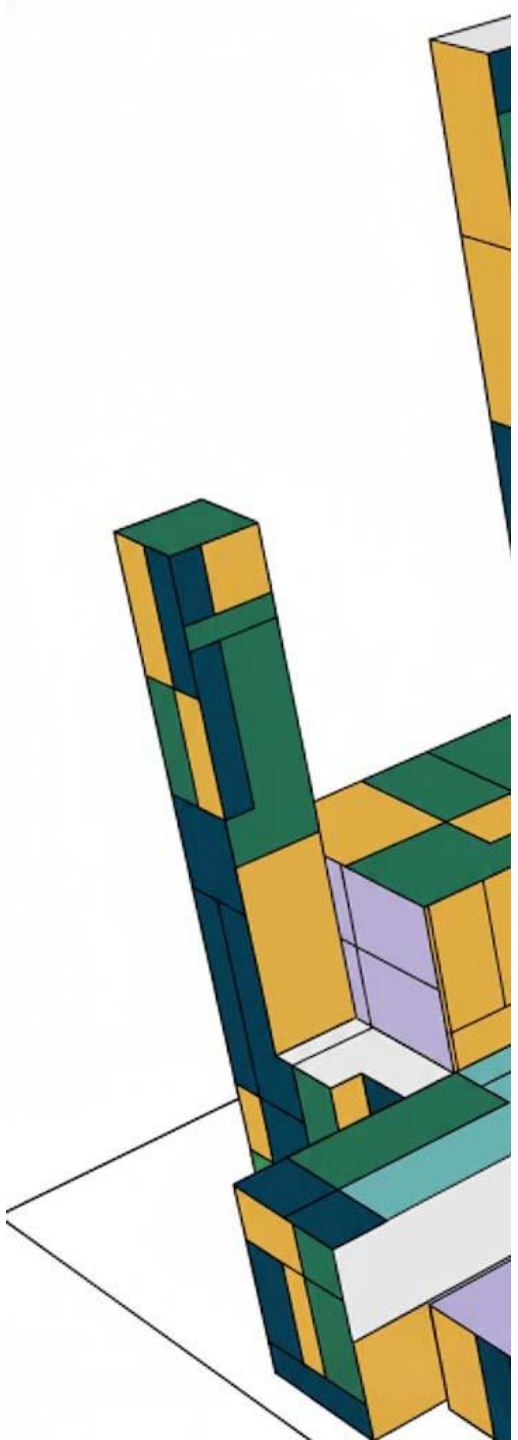
# ЗАДАЧА 1. КЛАССИФИКАЦИЯ СУБТИПОВ ВИЧ-1




Для определения субтипа используется ген оболочки env, характеризующийся наивысшей степенью генетической изменчивости

Основная проблема машинного обучения в данной задаче – неравномерное распределение данных в мировых базах (например, преобладание субтипа B и дефицит данных по CRF02_AG)

Значительный дисбаланс классов может привести к переобучению модели на более многочисленные классы.

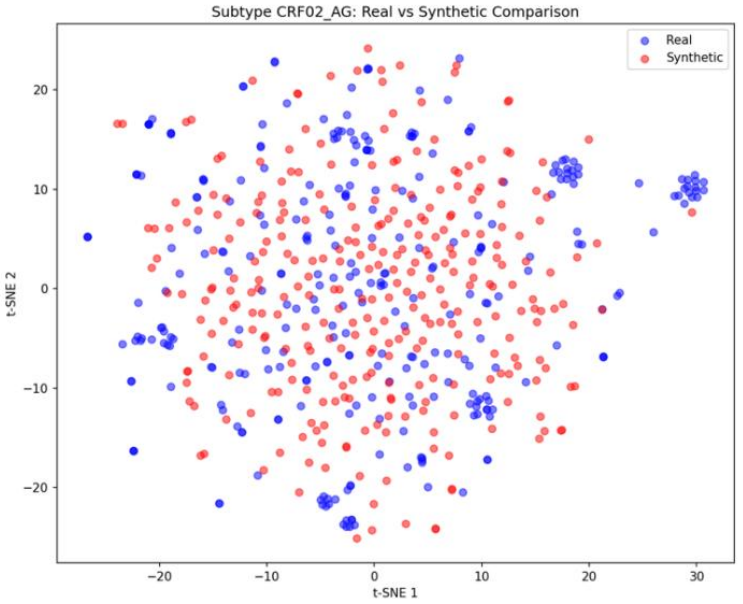


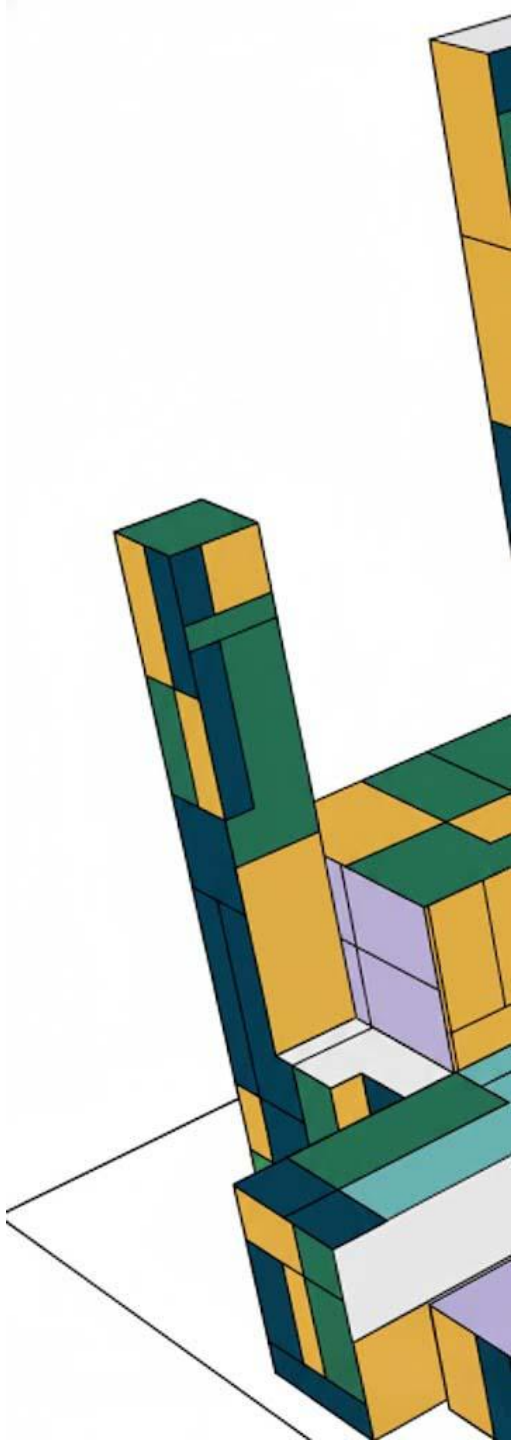
# ИСКУССТВЕННАЯ МОЛЕКУЛЯРНАЯ ЭВОЛЮЦИЯ (AME)

Для аугментации обучающей выборки и устранения дисбаланса классов применен метод искусственной молекулярной эволюции

Генерация синтетических данных включает синонимичные и несинонимичные мутации, вставки, делеции и рекомбинацию

Валидация методом t-SNE подтвердила, что синтетические образцы успешно воспроизводят топологию многообразия реальных данных без артефактов.



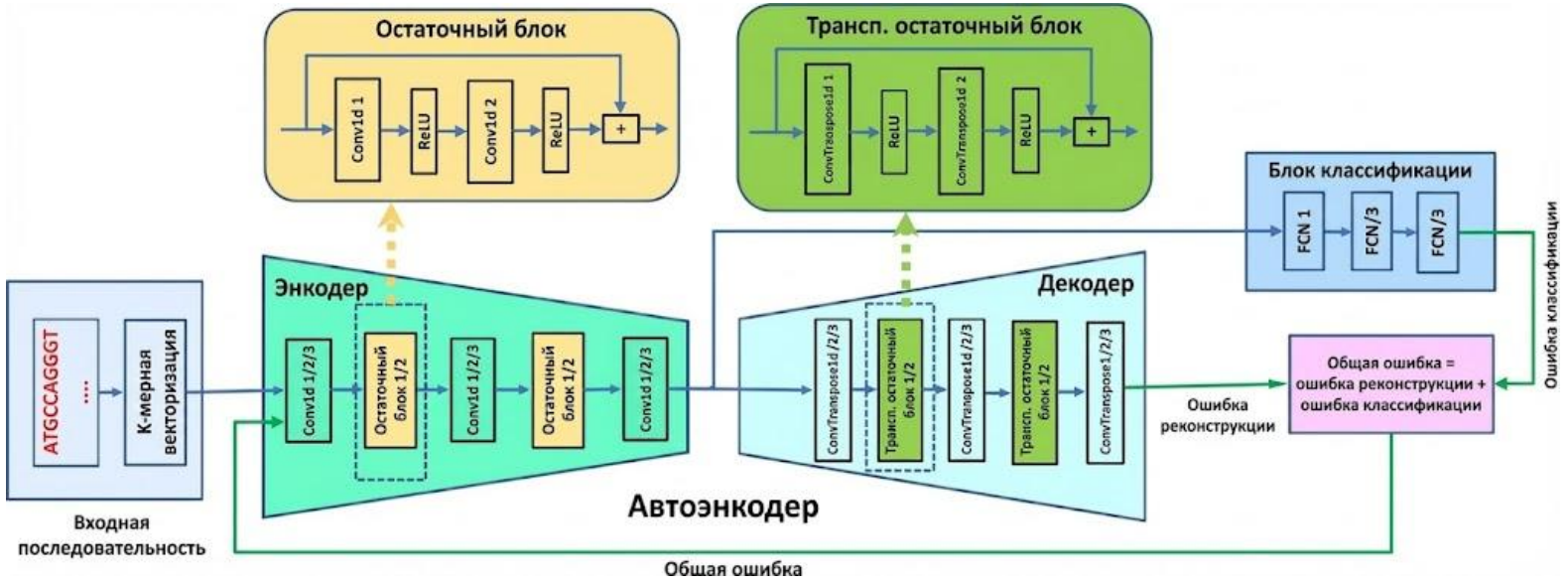



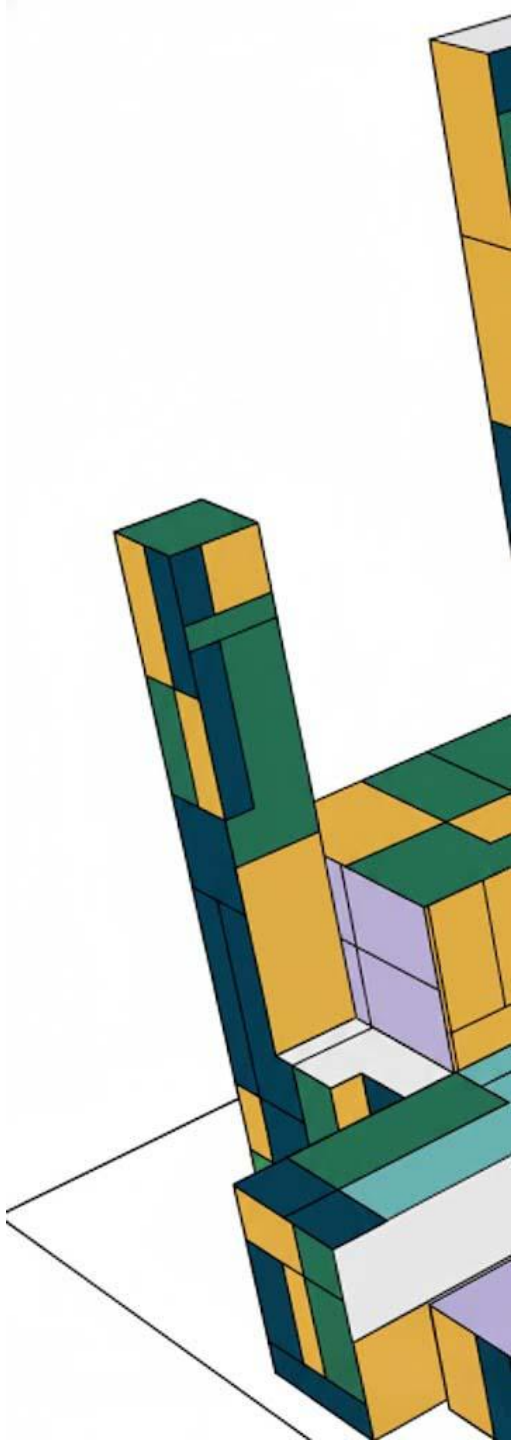
# АРХИТЕКТУРА МОДЕЛИ HIV-1-M-ENV-RUS

Входные данные преобразуются в численные векторы признаков на основе частот k-меров.

Используется архитектура 1D-свёрточного автоэнкодера с остаточными связями (ResNet) и транспонированными свёртками.

Обучение классификатора проводится с комбинированной функцией потерь, состоящей из ошибки реконструкции и ошибки классификации.



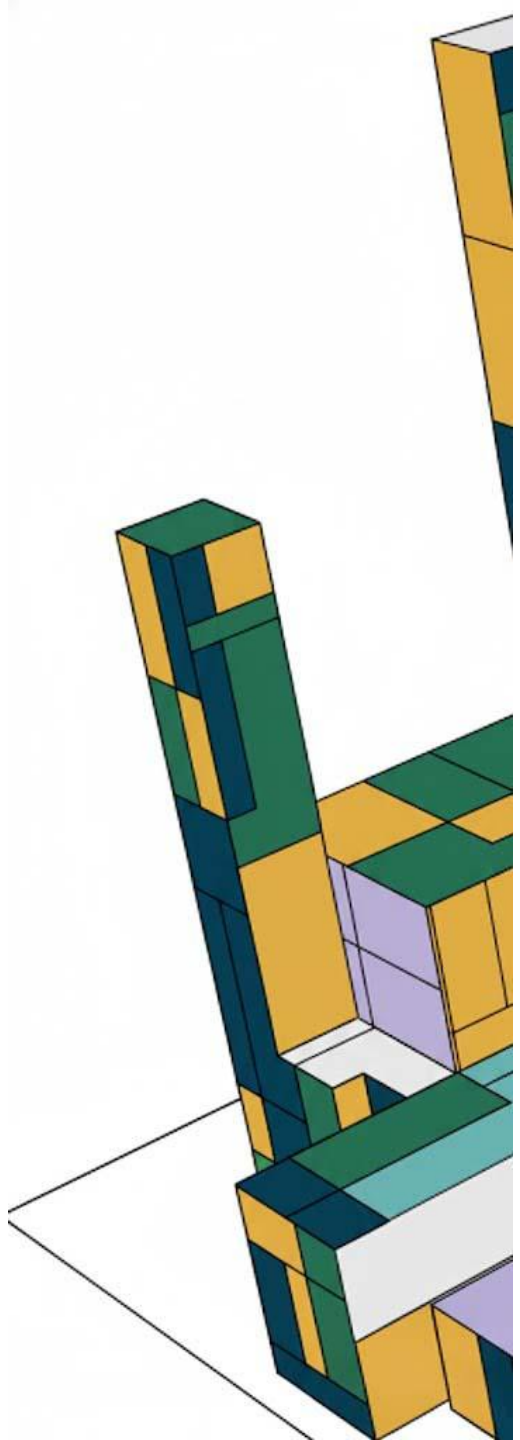
# РЕЗУЛЬТАТЫ КЛАССИФИКАЦИИ СУБТИПОВ

Итоговая точность (Accuracy) модели HIV-1-M-Env-Rus на независимой валидационной выборке составила 99,64%

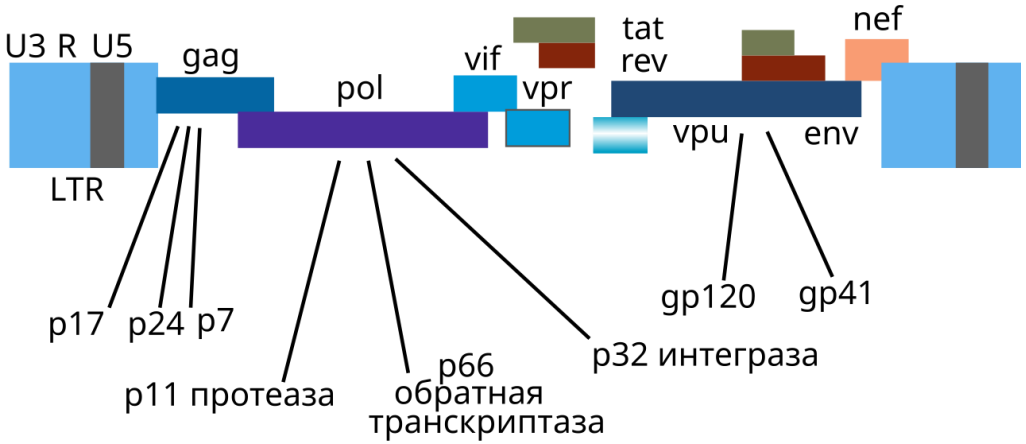
Для доминирующих в РФ субтипов A6 и CRF63_02A6 достигнуты абсолютные показатели Precision и Recall (1.00)

Модель успешно дифференцирует 12 различных субтипов вируса, демонстрируя низкий уровень межклассовой интерференции.4

Точность модели на представленном ФБУН ЦНИИ Эпидемиологии составила 100%

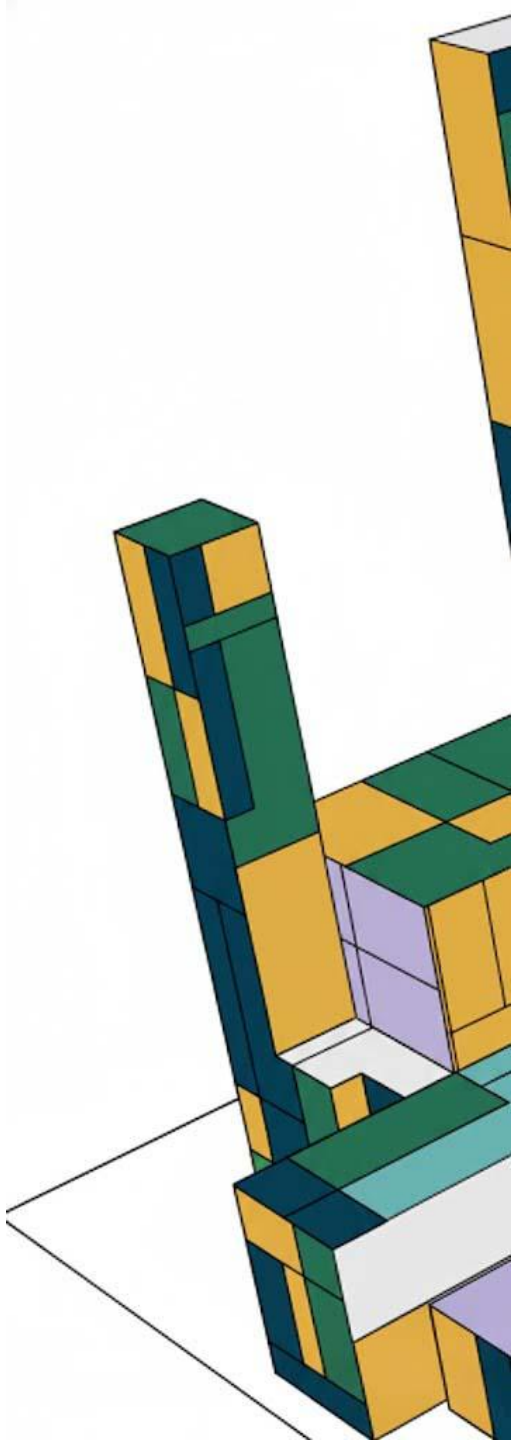
# ЗАДАЧА 2. ПРОГНОЗИРОВАНИЕ ЛЕКАРСТВЕННОЙ УСТОЙЧИВОСТИ




Прогнозирование фенотипа резистентности осуществляется на основе последовательностей генов pol (протеаза, обратная транскриптаза, интеграза)

Базовые архитектуры одномерных свёрточных сетей (1D-CNN) ограничены локальным рецептивным полем.

Для точного предсказания необходимо моделирование эпистаза – нелинейного взаимодействия между пространственно удаленными аминокислотными остатками



# ИНЖЕНЕРНЫЕ ВЫЗОВЫ И ВНЕДРЕНИЕ BIOMLOPS

Биоинформатические системы подвержены «дрейфу данных» (data drift) из-за постоянного пополнения клинических баз генотипов.

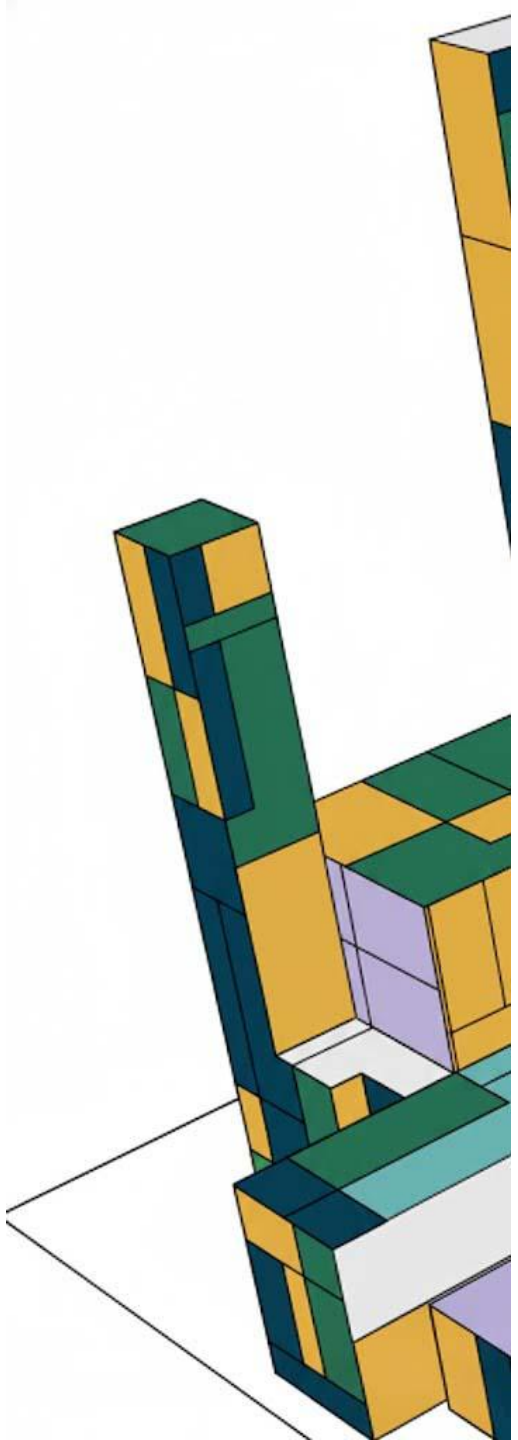
Отсутствие стандартов версионирования приводит к кризису воспроизводимости экспериментов.

Впервые предложена и внедрена методология BioMLOps (с использованием DVC) для версионирования данных и автоматического отслеживания метрик.

```yaml
stages:
  # ==========================================
  # 1. КОНВЕРТАЦИЯ СЫРЫХ ДАННЫХ
  # ==========================================
  convert_nrti:
    cmd: python src/convert.py --input data/raw/NRTI_DataSet.Full.txt --output data/interim/
    deps:
      - data/raw/NRTI_DataSet.Full.txt
      - src/convert.py
    outs:
      - data/interim/NRTI_stanford.csv

  convert_nnrti:
    cmd: python src/convert.py --input data/raw/NNRTI_DataSet.Full.txt --output data/interim
    deps:
      - data/raw/NNRTI_DataSet.Full.txt
      - src/convert.py
    outs:
      - data/interim/NNRTI_stanford.csv

  convert_pi:
    cmd: python src/convert.py --input data/raw/PI_DataSet.Full.txt --output data/interim/PI
    deps:
      - data/raw/PI_DataSet.Full.txt
      - src/convert.py
    outs:
      - data/interim/PI_stanford.csv
```

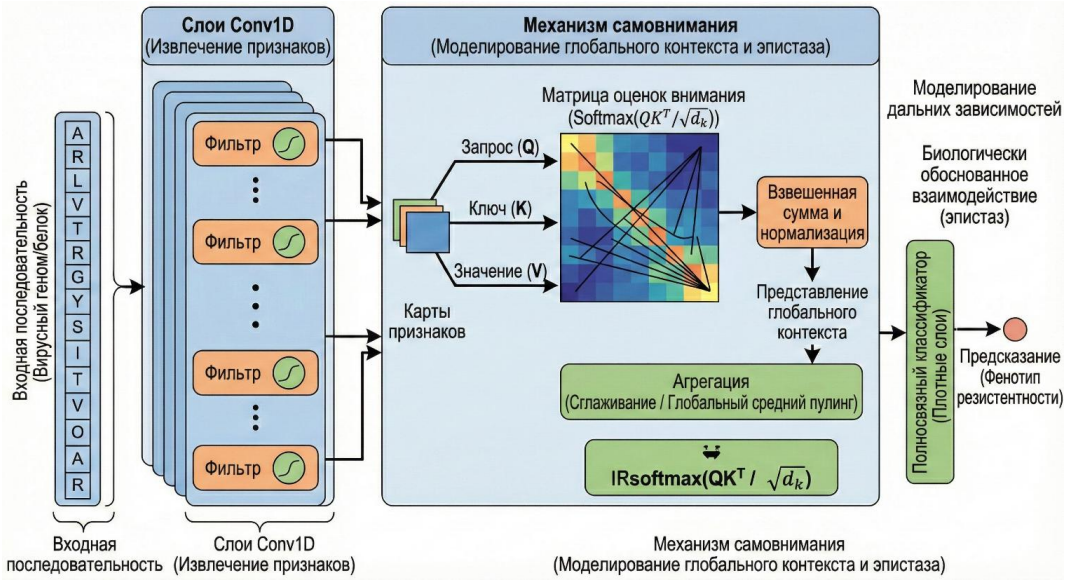


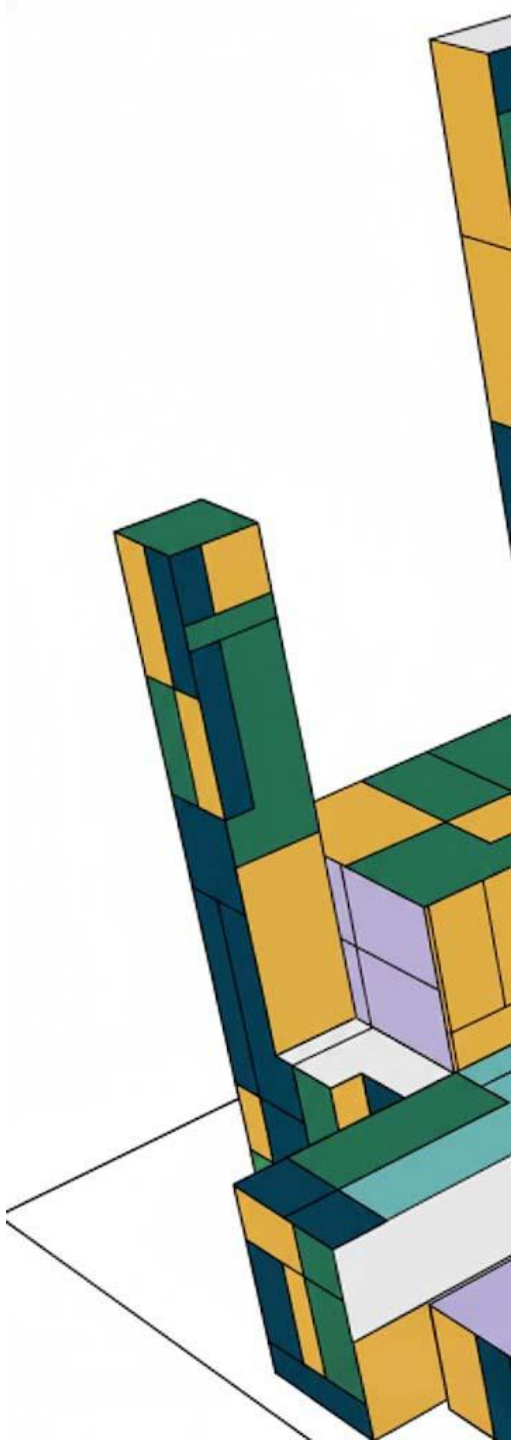
# АРХИТЕКТУРА МОДЕЛИ HIV-1-RESIST-RUS

Базовая CNN-архитектура модифицирована слоем механизма самовнимания (Self-Attention)

Механизм вычисляет матрицу внимания, позволяя модели динамически фокусироваться на значимых комбинациях мутаций независимо от расстояния между ними.

Гибридная модель формирует глобальное представление вирусной последовательности, улавливая дальние зависимости.



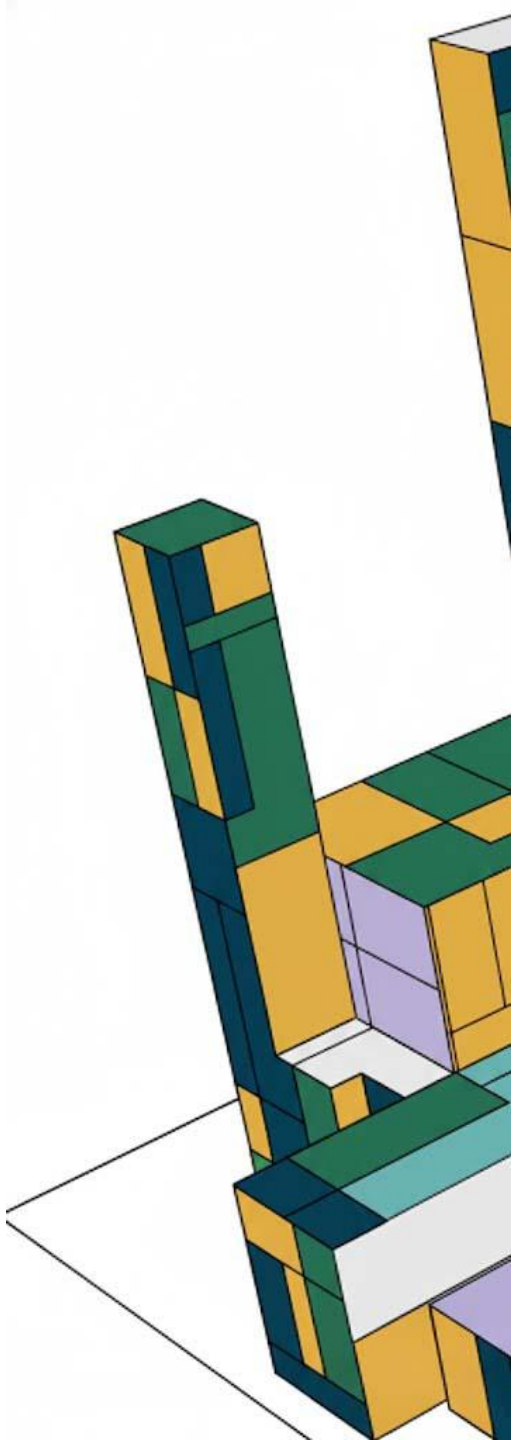



# ВАЛИДАЦИЯ ДЕТЕКЦИИ ЭПИСТАЗА (IN SILICO STRESS TEST)

Разработана стратегия стресс-тестирования на синтетических данных, моделирующих редкие генотипы субтипа A6.

В качестве маркера использовался полиморфизм L74I, который сам по себе не вызывает резистентности, но компенсирует дефекты репликации от других мутаций.

Гибридная модель формирует глобальное представление вирусной последовательности, улавливая дальние зависимости.



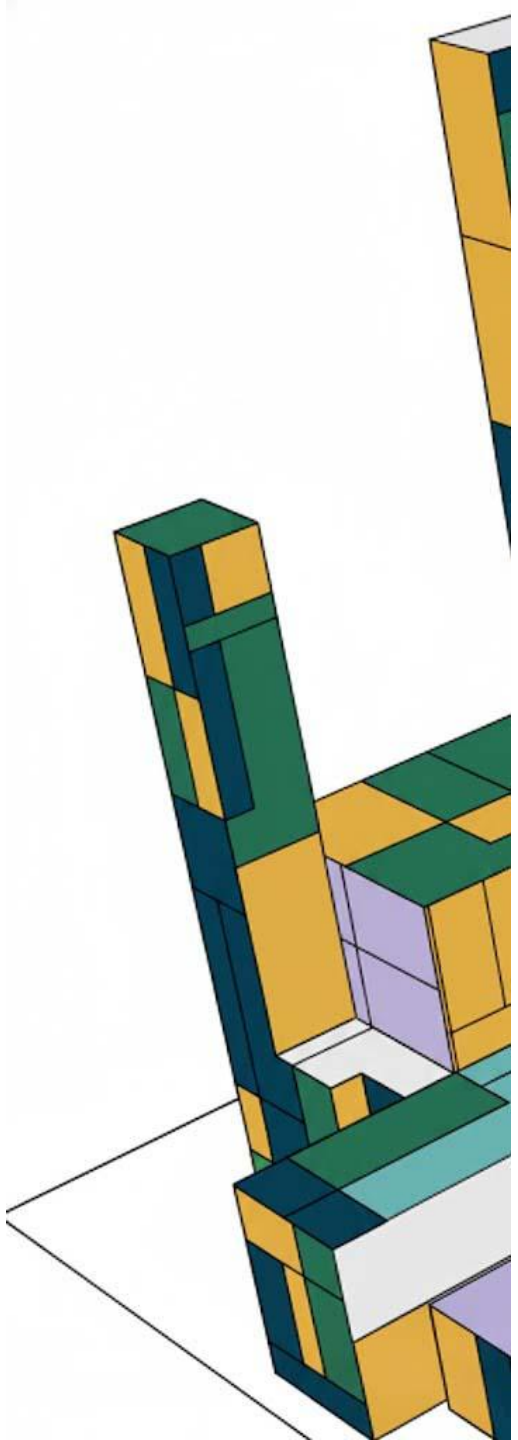
# РЕЗУЛЬТАТЫ МОДЕЛИРОВАНИЯ ДИСТАЛЬНЫХ ВЗАИМОДЕЙСТВИЙ

**Ограничения базовой модели** (CNN): Классические сверточные сети отлично находят локальные мутации (G118R), но ограничены размером рецептивного поля. При добавлении вторичной мутации (L74I) CNN практически не меняет прогноз (реакция на уровне шума, Δ +0.047

**Биологическая адекватность гибридной модели**: В реальности L74I компенсирует дефекты репликации вируса, усиливая профиль резистентности. Модель CNN+Attention успешно «выучила» эту нелинейную зависимость.

**Метрика чувствительности**: Механизм Self-Attention отреагировал на синергию мутаций **более чем в 2.1 раза острее**, чем базовая CNN (прирост уверенности +0.102 против +0.047)

**Вывод**: Использование механизма Attention – это не просто формальный рост метрик, а достижение соответствия модели реальным компенсаторным механизмам ВИЧ-1 (субтипа A6).



# ПОЛЬЗОВАТЕЛЬСКИЙ ИНТЕРФЕЙС (GUI) И ГЕНЕРАЦИЯ ОТЧЕТОВ

Для удобного взаимодействия с врачом-клиницистом разработано графическое приложение на базе фреймворка PyQt

Генерируются два отчета: технический (QC) для биоинформатика и финальный медицинский бланк с профилем резистентности и определенным субтипом.

## [QC Report] Pipeline Execution Metrics

> TIMESTAMP: 2026-02-23 22:15:31

> PIPELINE: onsiteseq_hiv **v1.0.0**

> ML MODELS: HIV-1-M-Env-Rus **v1.0.0** | HIV-1-Resist-Rus **v1.0.0**

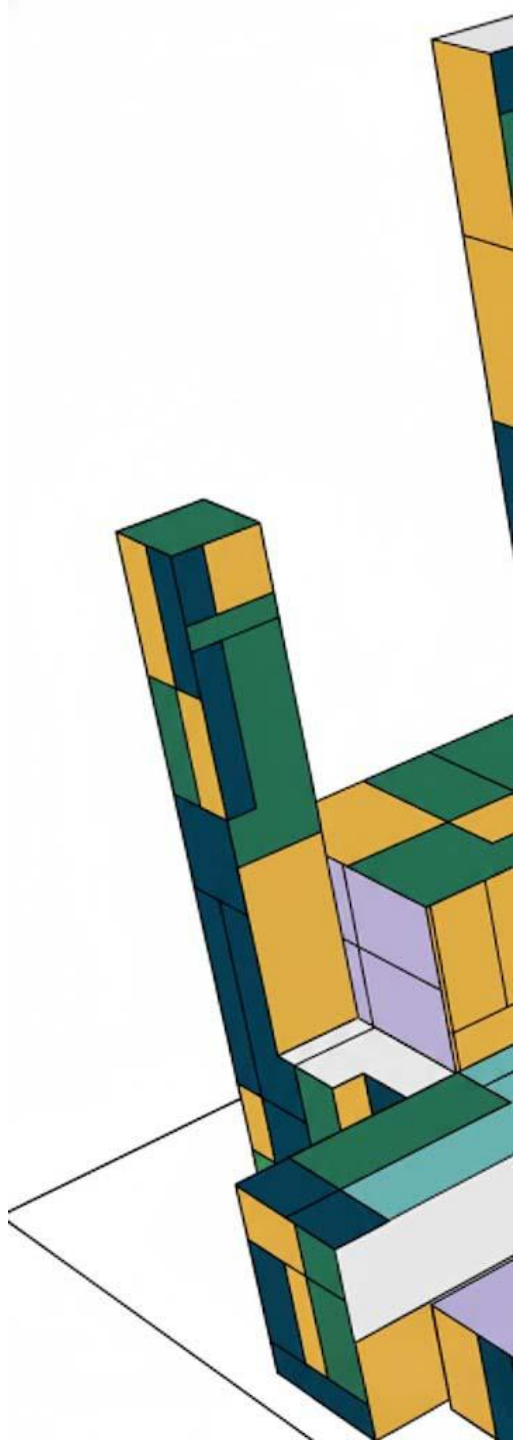
### > Настройки окружения (Tools & Parameters)

| Инструмент | Аргументы / Конфигурация |
|---|---|
| **porechop_abi** | ab-initio trimming (-v 0) |
| **NanoFilt** | min_len=500, min_qual=9 (-l 500 -q 9) |
| **minimap2** | Nanopore preset (-ax map-ont) |
| **medaka** | Model r1041_e82_400bps_sup_v4.2.0 |

### > Метрики консенсуса и покрытия

| ID образца | Референс | Длина консенсуса | Слепые зоны (N) | Покрытие (%) | Ср. Глубина | PR | RT | IN |
|---|---|---|---|---|---|---|---|---|
| DRR537804_1 | Ref.B.FR.83.HXB2_LAI_IIIB_BRU.K03455 | 9769 bp | 8 | 94.3774% | 4298.4x | ✅ | ✅ | ✅ |

* Примечание: ❌ означает, что ген не найден или отбракован из-за низкого качества.
* Если глубина (Ср. Глубина) ниже 100x, качество сборки может быть снижено.



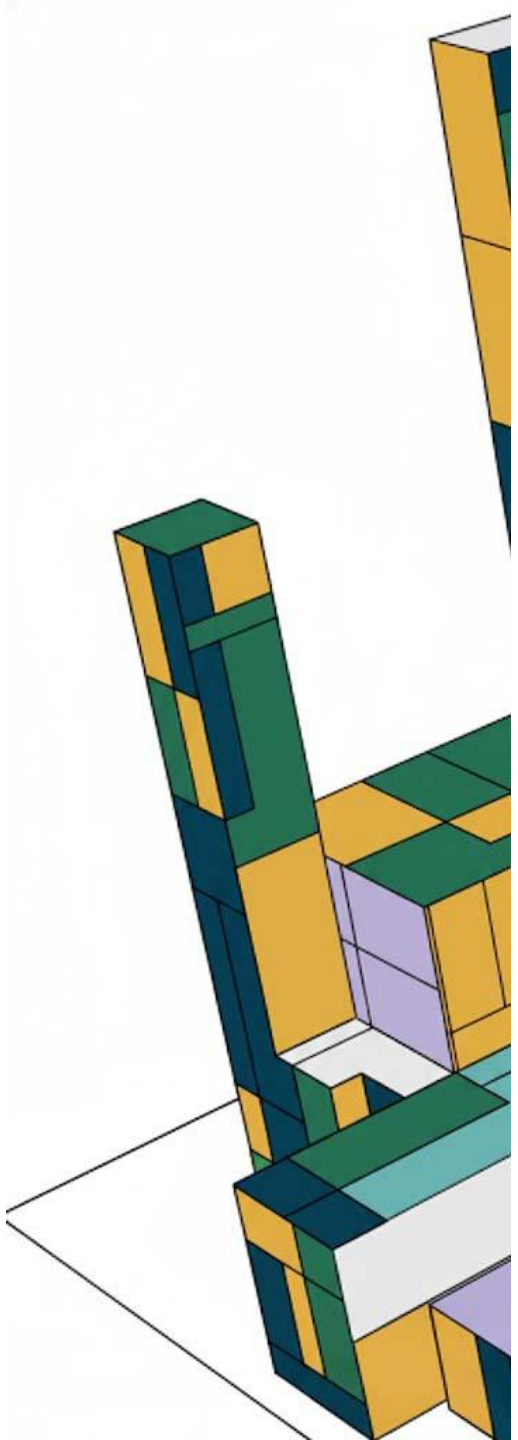
# ТЕХНОЛОГИЧЕСКИЙ СУВЕРЕНИТЕТ И ДОСТУПНОСТЬ

Исходный код и веса обученных моделей опубликованы на отечественной платформе GitVerse.

Обеспечивается возможность локального анализа геномных данных без их передачи на зарубежные веб-серверы (такие как Stanford HIVDB).

Инструмент полностью гарантирует соблюдение требований к суверенитету медицинских данных РФ.

Создание устройств, методов, алгоритмов, которые направлены на определение лекарственной устойчивости с использованием генетических технологий напрямую соответсвтует определенным в Указе Президента N145 приоритетам научно-технологического развития.



# Спасибо за внимание!

Роман Горбенко
gorbenko.ra@phystech.edu
+79162691517, tg: gorbenkoteh
Web: onsiteseq.io


МФТИ