Классификация доминирующих в РФ субтипов ВИЧ-1 с применением методов глубокого обучения.

Горбенко Роман. E-mail: gorbenko.ra@phystech.edu
Московский физико-технический институт (МФТИ),
141701, Московская область, г. Долгопрудный, Институтский переулок, д.9
Научный руководитель. Радченко Никита Сергеевич
E-mail: nikita@radchenko.tech

Научный руководитель. к.б.н Киреев Дмитрий Евгеньевич
E-mail: dmitry.kireev@pcr.ms

Аннотация

В 2021 г. во время пандемии COVID-19 Всемирная организация здравоохранения (ВОЗ) призвала [1] страны усилить роль геномного эпидемиологического надзора для более глубокого изучения особенностей передачи возбудителей инфекционных болезней. Это требование совпало с начавшимся широким применением нейронных сетей, машинного обучения (ML) и глубокого обучения (DL) в биомедицинской области. Полученные результаты продемонстрировали огромный потенциал этих технологий, особенно в вирусологических исследованиях [2]. В контексте изучения вируса ВИЧ-1 в университете штата Джорджия были разработаны инструменты для определения субтипа вируса на основе машинного обучения. Однако разработанные в США методы [3] не учитывают специфику Российской Федерации, где эпидемия ВИЧ-1 остается монофилетической. Так, согласно пилотному исследованию в Орловской области [4,5], а также статистике базы RuHIV [6], основным субтипом в Российской Федерации остаётся субтип A6 и его рекомбинантная форма CRF63_02A6, которые крайне редко встречаются в США [7].

Данное исследование направлено на создание инструмента HIV-1-M-Env-Rus для определения субтипа вируса на основе машинного обучения, учитывающего специфику эпидемии ВИЧ в Российской Федерации. Предложенный в рамках данного исследования подход использует архитектуру свёрточного автоэнкодера, где энкодер содержит два резидуальных блока, а декодер -два резидуальных блока с транспонированными свёртками.

Выходные признаки в этой архитектуре подаются в полносвязный нейросетевой блок. При этом модель, лежащая в основе инструмента HIV-1-M-Env-Rus, преобразует сложные многомерные данные о последовательностях ДНК вируса ВИЧ-1 в лаконичные, информативные и низкоразмерные представления, достигая исключительной точности классификации. В результате независимой проверки данными, полученными из российской базы данных устойчивости ВИЧ к антиретровирусным препаратам [6], точность, достоверность, полнота и показатель F1 модели HIV-1-M-Env-Rus продемонстрировали значения выше 99%, что подтверждает её способность точно идентифицировать наиболее распространённые в Российской Федерации субтипы. Код и веса модели выложены в российском хранилище исходных кодов GitVerse [8], что даёт возможность встроить HIV-1-M-Env-Rus в биоинформатические пайплайны и использовать его в рутинной научной и клинической практике.

Введение

Вирус иммунодефицита человека, также известный как ВИЧ, - это ретровирус, вызывающий ослабление иммунной системы человека. Этот вирус атакует и постепенно разрушает иммунную систему, оставляя организм беззащитным перед оппортунистическими инфекциями, что в конечном итоге без применения терапии приводит к переходу заболевания в стадию СПИДа.

ВИЧ-1 возник в районе бассейна реки Конго в Африке и является наиболее распространённым штаммом в мире, ответственным за

глобальную эпидемию (на него приходится 95% всех случаев заражения). ВИЧ-2 в основном встречается в Западной Африке, в Российской Федерации случаи заражения единичны. ВИЧ-1 принято разделять на 4 группы: M, N, O и P, при этом группа M наиболее широко распространена в мире и делится еще на ряд различных субтипов. Согласно исследованию [9] 2012 года, субтип A6 (в старой классификации, актуальной в 2010-х годах, обозначен как A1) занимает 89-90%. Несмотря на то, что в 2020-х годах доля A6 незначительно снизилась за счёт увеличения доли субтипа CRF02_AG (связанного с миграционным потоком из Средней Азии) и субтипа CRF63_02A6, имеющего более высокую скорость распространения [10,11,12], в целом эта тенденция сохраняется и в настоящее время. Это было показано на примере пилотного исследования в Орловской области [4,5] и отражено в статистике российской базы данных устойчивости ВИЧ к антиретровирусным препаратам RuHIV [6]. Доли субтипов ВИЧ-1 в Российской Федерации представлены на Рис. 1 и сведены в Таблицу 1 в Приложении 1.

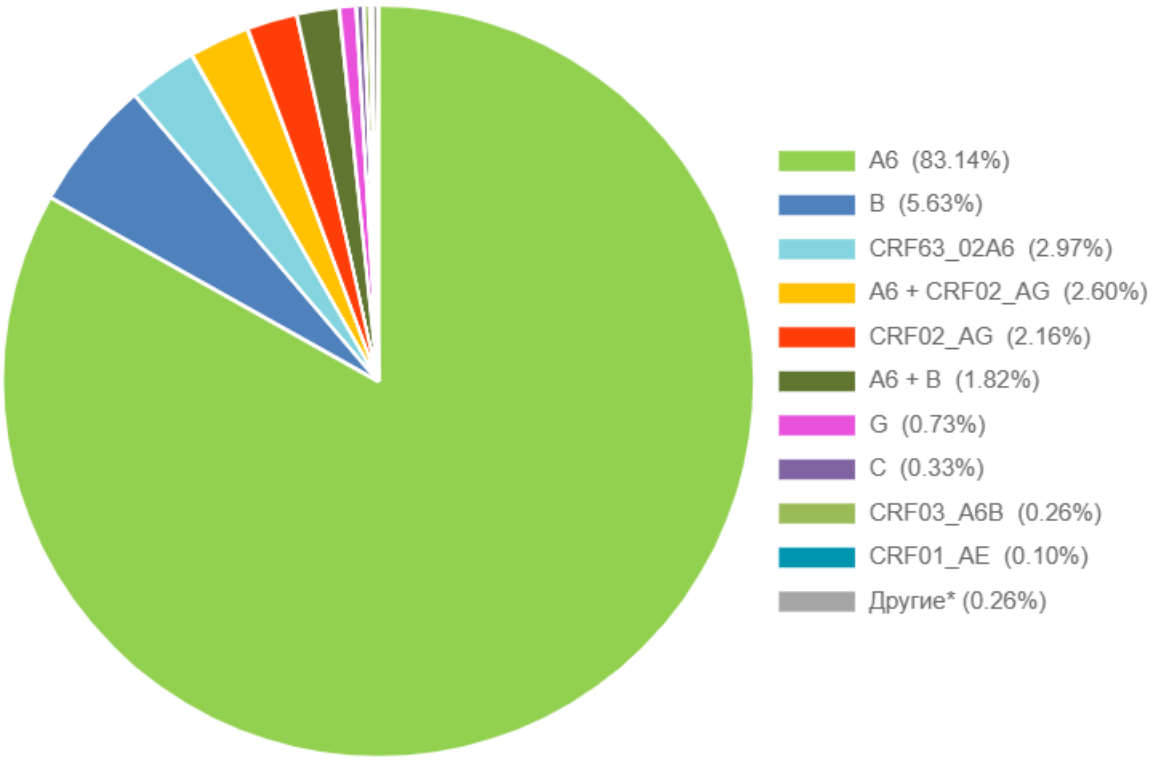

Рис. 1. Доли субтипов ВИЧ-1 в Российской Федерации

Представленность субтипов в США значительно отличается от Российской Федерации. Так, согласно исследованию Diversity and characterization of HIV-1 subtypes in the United States, 2008–2016 [7], распространённость субтипов в этой стране выглядит иным образом (см. Рис. 2 и Таблицу 2 в Приложении 1).

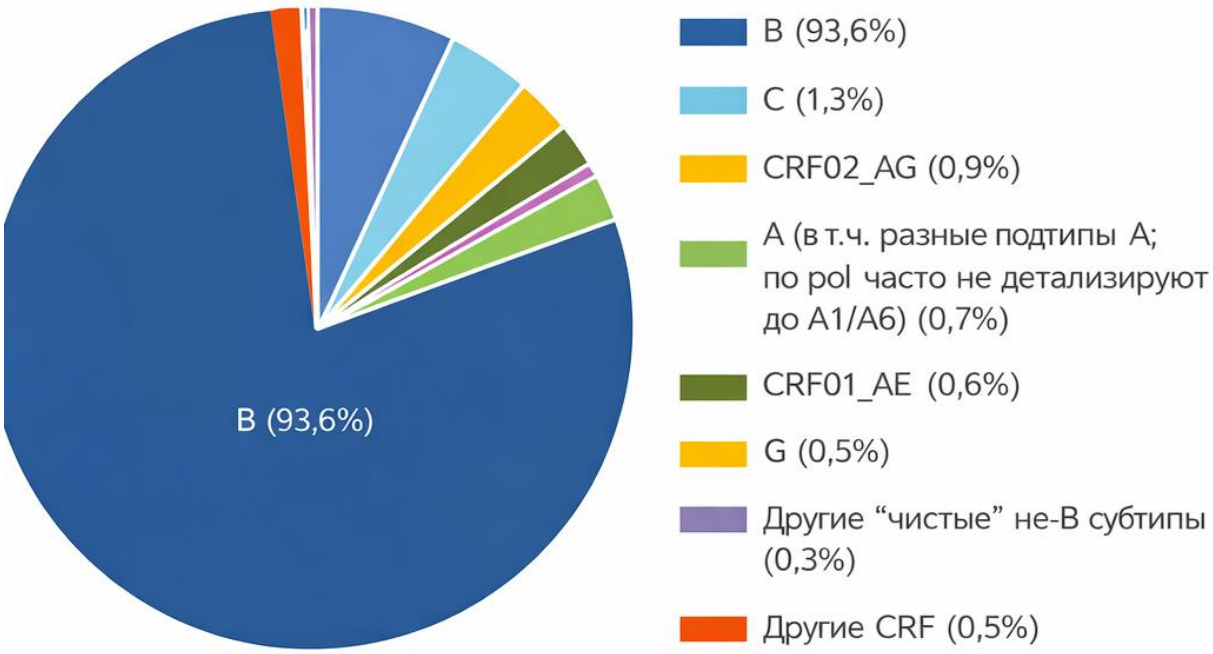

Рис.2. Доли субтипов ВИЧ-1 в США.

Сравнительный анализ эпидемий в Российской Федерации и США представлен в Таблице 3.

Таблица 3. Сравнительная характеристика эпидемии ВИЧ-1 в РФ и США

Параметр	Российская Федерация (РФ)	США
Доминирующий субтип	A6 (~83%)	B (~94%)
Ключевой рекомбинант	CRF63_02A6	Редко
Характер эпидемии	Монофилетический (сдвиг в сторону A6)	Полифилетический (но с доминантой B)

Такое значительное отличие приводит к тому, что субтипы A6 и CRF63_02A6 не находятся в фокусе интересов исследователей из ведущих университетов США, что, в свою очередь, даёт пространство для появления оригинальных работ в Российской Федерации. Важно отметить, что формирование различных субтипов ВИЧ-1 группы M с неоднородностью представления в разных популяциях и появление CRF-рекомбинантов является результатом непрерывной молекулярной эволюции вируса. Правильная классификация субтипов важна для разработки вакцин, лекарственных препаратов, эпидемиологического надзора и назначения терапии [13,14].

Машинное обучение (ML) в биомедицинской сфере

Применение машинного обучения и глубокого обучения продемонстрировало огромный потенциал в биомедицинской сфере, особенно в вирусологических исследованиях [2]. Эти технологии стали важнейшими инструментами для понимания поведения вирусов и ускорения разработки вакцин и лекарств. Немаловажную роль в ускорении

внедрения этих методов сыграла пандемия COVID-19. Согласно публикациям, множество биоинформатических подходов с использованием ML, разработанных для SARS-CoV-2, ученые планируют адаптировать и для других патогенов, в том числе для ВИЧ-1.
Так, в статье "Analysis of 6.4 million SARS-CoV-2 genomes identifies mutations associated with fitness" [15] показана возможность применения машинного обучения для геномного надзора. Исследователи разработали инструмент PyR$_0$, использующий анализ больших массивов данных секвенирования, чтобы быстро оценивать фитнес (темп экспоненциального роста) вирусных линий и выделять мутации, статистически связанные с этим ростом. Модель выдаёт апостериорные оценки фитнеса линий и вкладов отдельных мутаций. Например, она оценила Omicron BA.2 как наиболее "фитную" линию (примерно 8.9× относительно исходной Wuhan/A-линии) и отметила его повышенный потенциал уже к середине декабря 2021 года на основе всего 76 геномных последовательностей, что впоследствии полностью подтвердилось в клинической практике.
Еще один пример представлен в статье "Classification and specific primer design for accurate detection of SARS-CoV-2 using deep learning" [16], где показано, как глубокое обучение можно использовать не только для классификации геномов, но и как генератор биологически интерпретируемых признаков — вплоть до кандидатов в ПЦР-праймеры. Предсказанные моделью праймеры были синтезированы и проверены в лабораторной практике, показав лучшую специфичность по сравнению с наборами праймеров референс-лабораторий ВОЗ.

Материалы и методы

Архитектура приложения

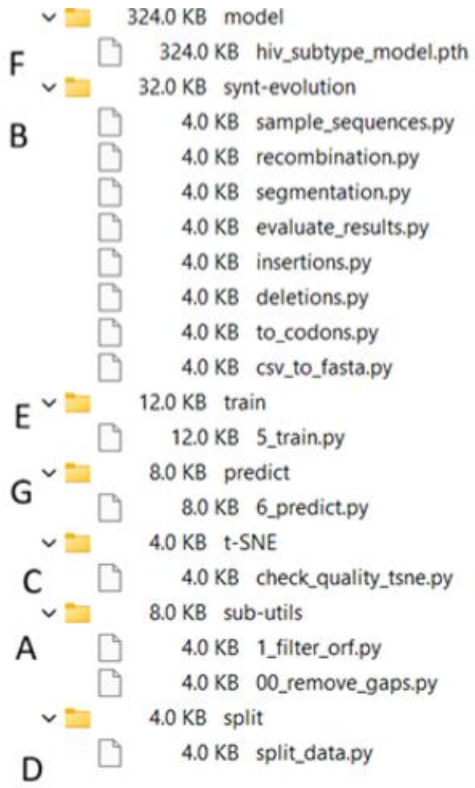

Рис.3. Репозиторий приложения HIV-1-M-Env-Rus [8]

A (sub-utils). Утилиты для извлечения и контроля качества исходных биологических данных.
Функционал: Извлечение гена env из полных геномов в формате GenBank[17] и фильтрация на валидную открытую рамку считывания (ORF), наличие старт-кодонов (ATG, GTG, TTG) и стоп-кодонов, отсутствие внутренних стоп-кодонов.
B (synt-evolution). Утилиты для моделирования синтетической эволюции.
Функционал: Переосмысленный и доработанный метод, предложенный в статье [3], включает 5 типов операций синтетической эволюции:

1. Синонимичные мутации.
2. Несинонимичные мутации.
3. Вставки (Insertions).
4. Делеции (Deletions).
5. Рекомбинация (Recombination).

C(t-SNE). Проверка генерации синтетических данных выполнена путём визуализации распределения методом t-SNE [18,19]
D (split). Утилиты для подготовки и разделения наборов данных.

Функционал: Объединение синтетических и реальных последовательностей различных субтипов в единый репрезентативный датасет. Случайное перемешивание (shuffling) данных для исключения систематических ошибок и разделение выборки на обучающую (Train) и валидационную (Validation) в заданном соотношении (80/20) с сохранением меток классов в заголовках FASTA-файлов.
E (train). Модуль обучения глубокой нейронной сети.
Функционал: Преобразование биологических последовательностей в численные векторы признаков на основе частот k-меров. Реализация архитектуры глубокого обучения на базе сверточных нейронных сетей (CNN) с использованием остаточных блоков (Residual Blocks). Оптимизация гиперпараметров, расчет функции потерь (Cross-Entropy Loss) и построение метрик качества (графики обучения, матрица ошибок/Confusion Matrix) для оценки точности классификации субтипов.
F (model). Хранилище обученных состояний моделей.
Функционал: Сохранение весов обученной нейронной сети в формате .pth (PyTorch). Обеспечивает версионность моделей и возможность их повторного использования без необходимости повторного обучения, фиксируя структуру и параметры, показавшие наилучшую точность на валидационном наборе данных.
G (predict). Модуль классификации и вероятностной оценки. Функционал: Программный интерфейс для выполнения инференса (предсказания) на новых, ранее не размеченным данных. Архитектура модуля включает автоматический пайплайн предобработки (извлечение и нормализацию частот k-меров), загрузку весов предобученной модели и процедуру классификации. Ключевой особенностью модуля является расчет метрики уверенности (Confidence Score) для каждого предсказания, вычисляемой через функцию активации Softmax. Итоговый отчет содержит не только прогнозируемый субтип (из перечня в Таблице 5), но и вероятность правильного прогноза в процентах. Это позволяет фильтровать результаты с низкой достоверностью (например, для новых, ранее не описанных рекомбинантов), что критически важно для клинического применения.

Таблица 4. Возможные варианты вывода и интерпретации итогового отчета работы модуля G (predict) на тестовых данных.

Patient_Sample_042	A6	99.98%	Высокая достоверность, типичный вариант для РФ

Patient_Sample_115	CRF63_02A6	98.45%	Надежная идентификация рекомбинанта
Patient_Sample_009	B	99.91%	Классический вариант для США
Unknown_Sample_X	Unclassified	0.00%	Низкое качество прочтения (Low Quality)
Patient_Sample_303	CRF02_AG	52.10%*	*Низкая уверенность: требует экспертной проверки

Датасет

Вирус ВИЧ-1 имеет всего 9 генов, которые кодируют 15 различных белков.

Ген env (Envelope — оболочка) кодирует поверхностные гликопротеины (gp120 и gp41), с помощью которых вирус прикрепляется к клетке [20]. Иммунная система атакует именно эти белки, поэтому вирус вынужден эволюционно постоянно менять структуру *env*, чтобы уходить от иммунного ответа.

Высокая изменчивость делает ген *env* особенно информативным для задач классификации, так как он накапливает достаточное количество различий между субтипами, что важно для обучения модели глубокого обучения (DL).

Формирование датасета было осуществлено в следующие два этапа.

Этап 1.

Полные геномы и последовательности генов env были загружены из базы данных геномов ВИЧ [21], поддерживаемой Лос-Аламосской национальной лабораторией. Загруженные последовательности прошли через извлечение и фильтрацию, которая была выполнена инструментами из раздела "A" архитектуры приложения.

Итоговое количество загруженных и извлеченных последовательностей гена env для каждого субтипа представлено в Таблице 5.

Таблица 5. Субтипы, загруженных из базы данных LANL [21]

Субтип	Количество
A1	3159

A2	10
B	89961
C	47021
D	1532
F1	4393
F2	209
G	600
H	10
A6	1452
CRF63_02A6	47
CRF02_AG	748

Этап 2.
Из таблицы 4 видно, что распределение имеющихся данных неравномерно. Например, для субтипа CRF02_AG доступно всего 748 последовательностей, тогда как для субтипа A1 — порядка 3159. Данное неравномерное распределение может существенно повлиять на обучение модели и привести к переобучению на более многочисленные классы, т.е при значительном дисбалансе классов модель получит высокую точность (accuracy) ~99% определяя самый представленный субтипов и будет бесполезной для более редких суб-топов.
Для преодоления этой проблемы на данном этапе был применён метод синтетической генерации данных. Используя доработанную технику искусственной молекулярной эволюции описанную в работе [3] были созданы дополнительные образцы данных для менее представленных субтипов. Генерация синтетических данных была выполнена инструментами из раздела "B" архитектуры приложения и обеспечила более сбалансированное распределение классов, что, в свою очередь,

позволило улучшить обучение модели, повысив её обобщающую способность.

Для преодоления получившего распространённость в вычислительной биологии "кризиса воспроизводимости"[22, 23] последовательность запуска утилит оформлена в Snakemake-пейплайн, что является "золотым" стандартом в биоинформатике [24]

Проверка генерации синтетических данных выполнена путём визуализации распределения методом t-SNE [18] в соответствии с рекомендациями по настройке гиперпараметров для биологических данных [19]. На графиках (Рис 4 и Рис.5,6 в Приложении 1) наблюдается значительное перекрытие (overlap) реальных и синтетических образцов, что указывает на отсутствие фундаментальных различий в их статистических свойствах. Синтетическая модель успешно воспроизводит топологию многообразия реальных данных, о чем свидетельствует равномерное смешивание классов и отсутствие выраженных кластеров-артефактов, свойственных исключительно синтетической выборке.

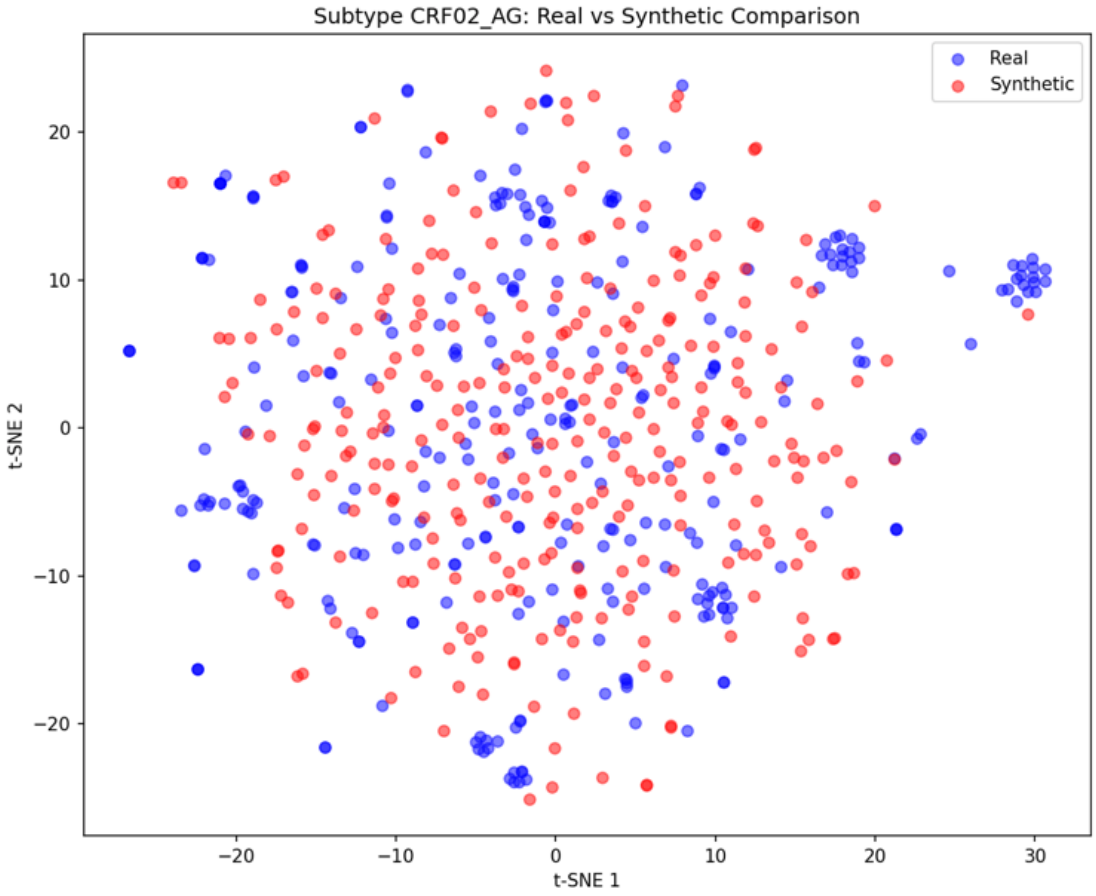

Рис.4. Проверка генерации синтетических данных для субтипа CRF02_AG методом t-SNE

Итоговое количество загруженных последовательностей совместно с синтетическими данными для каждого субтипа представлено в Таблице 6

Таблица 6. Итоговое количество последовательностей для обучения.

Субтип	Количество
A1	500(3159)
A2	500(10)
B	500(89961)
C	500(47021)
D	500(1532)
F1	500(4393)
F2	500(209)
G	500(600)
H	500(10)
A6	500(1452)
CRF63_02A6	500(47)
CRF02_AG	500(748)

Объединённые синтетические и реальные последовательности различных субтипов помещаются в единый репрезентативный датасет путём случайной выборки 500 образцов каждого субтипа. Объединённый датасет проходит процедуру случайного перемешивания (shuffling) данных для

исключения систематических ошибок и разделение выборки на обучающую (Train) и валидационную (Validation) в заданном соотношении (80/20) и подаётся на вход глубокой сверточной нейронной сети (CNN) в виде векторов частот k-меров для обучения модели классификации и минимизации функции потерь.

Архитектура нейронной сети

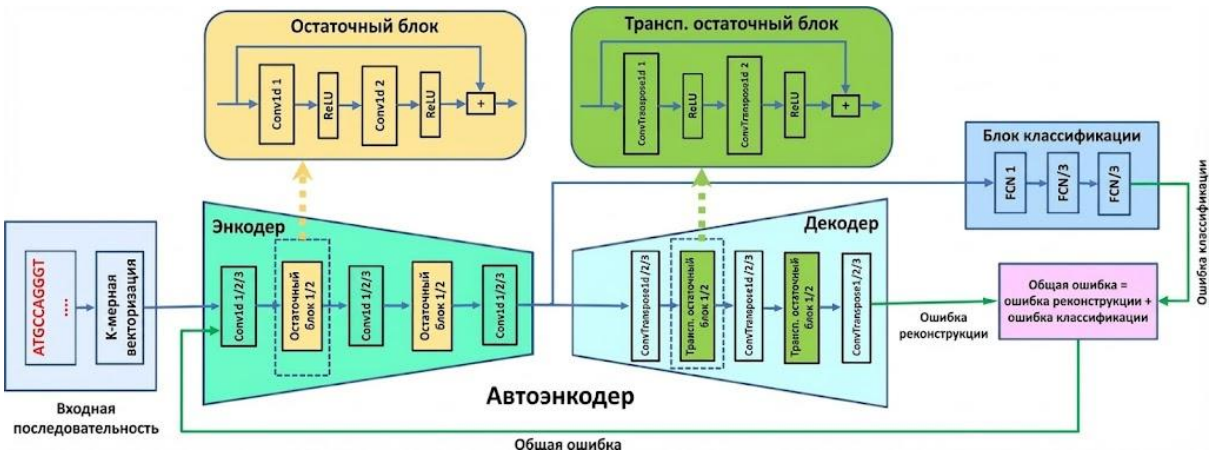

Рис 7. Архитектура нейронной сети, лежащая в основе HIV-1-M-Env-Rus

Блок 1.1. Входная последовательность

В качестве входных данных модель принимает нуклеотидные последовательности гена env вируса иммунодефицита человека 1-го типа

(ВИЧ-1). Выбор гена env обусловлен тем, что он кодирует оболочечные гликопротеины вируса и характеризуется наивысшей степенью генетической изменчивости среди всех областей генома ВИЧ. Эта вариабельность делает его наиболее информативным маркером для филогенетического анализа и точного определения субтипов.

Формально входная последовательность S представляет собой упорядоченную строку символов произвольной длины L заданную над алфавитом Σ = {A, C, G, T}, где символы соответствуют азотистым основаниям ДНК:
A — Аденин (Adenine)
C — Цитозин (Cytosine)
G — Гуанин (Guanine)
T — Тимин (Thymine)
Данные поступают в формате FASTA или в виде «сырых» (raw) текстовых строк. Перед этапом векторизации производится предварительная обработка (препроцессинг):
Приведение всех символов к верхнему регистру.
Фильтрация или исключение последовательностей, содержащих недопустимые символы (например, символы вырожденности кода IUPAC, такие как 'N', 'R', 'Y'), если они превышают установленный пороговый уровень шума, чтобы гарантировать чистоту обучающей выборки.

Полученная последовательность S передается в блок k-мерной векторизации для преобразования в числовой формат. Процесс векторизации включает сопоставление каждого уникального k-мера w индексу idx(w) в пространстве признаков размерностью 4^k. Для этого используется отображение азотистых оснований на целые числа: μ(A)=0, μ(C)=1, μ(G)=2, μ(T)=3

Индекс для произвольного k-мера $w = s_1 s_2 \dots s_k$
вычисляется как значение полинома в четверичной системе счисления:

Блок 1.2. Векторизация последовательности ДНК на основе k-меров

Пусть нуклеотидная последовательность S имеет длину L и задана над алфавитом Σ = {A, C, G, T}. k-мером (k-mer) называют непрерывную

подстроку длины k, представляющую собой фрагмент последовательности S:
w = S[i … i+k−1],
где 1 ≤ i ≤ L−k+1.

Совокупность всех k-меров, получаемых при проходе скользящим окном (sliding window) длины k с шагом 1 по последовательности, называют k-мерным спектром. В зависимости от задачи, спектр рассматривают либо как множество уникальных k-меров, либо как мультимножество с учётом их частотности (кратностей).
Пример: Рассмотрим последовательность S = AGCGAT. При k=2 спектр составляет: {AG, GC, CG, GA, AT}. При k=3 спектр составляет: {AGC, GCG, CGA, GAT}.
Количественные характеристики: Число k-меров, извлекаемых из последовательности длины L, рассчитывается по формуле:
N_kmers = L − k + 1 (при k ≤ L).
Количество теоретически возможных уникальных k-меров над алфавитом мощности n (где n = |Σ|) определяется как n^k. Для последовательностей ДНК, где n=4, размерность пространства признаков составляет 4^k [25].

Шаг 1. Формирование вектора признаков. Для каждой анализируемой нуклеотидной последовательности строится признаковый вектор размерности 4^k. Такая размерность обусловлена тем, что алфавит ДНК содержит четыре основания (A, C, G, T), поэтому множество всех возможных k-меров включает 4^k уникальных комбинаций.
Шаг 2. Сопоставление оснований с числами: Сначала каждому основанию сопоставляется число. Например, A = 0, C = 1, G = 2, а T = 3.
Шаг 3. Вычисление индекса k-мера. Для заданного k-мера числовые коды, сопоставленные каждому нуклеотиду, конкатенируются и используются для определения позиции этого k-мера в признаковом векторе. Удобно трактовать k-мер как число в четверичной системе счисления (основание 4) и затем перевести его в десятичную систему — полученное значение и задаёт индекс.
Шаг 4. Подсчёт частот k-меров. Для каждой последовательности последовательно извлекают все k-меры (обычно скользящим окном), для каждого k-мера определяют его индекс в признаковом векторе и увеличивают значение соответствующей компоненты. В результате

элементы вектора отражают число наблюдений каждого k-мера в данной последовательности.

Блок 2. Автоэнкодер, состоящий из энкодера и декодера.

В 1987 году Янн ЛеКун построил нейронную сеть с энкодером и декодером на основе многослойного перцептрона (MLP) и применил её для подавления шума в данных, что делает эту работу одной из самых ранних реализаций автокодировщиков [26]. В 1988 году Бурлар и Камп использовали автокодировщик на базе MLP для исследования методов снижения размерности данных [27]. Позднее, в 1994 году, Хинтон и Ричард С. Земел предложили первую генеративную модель, основанную на автокодировщиках [28].

Автоэнкодер — это нейронная сеть, в которой входные данные совпадают с целевыми значениями при обучении. Архитектура обычно включает две части: энкодер, преобразующий вход в компактное скрытое представление, и декодер, восстанавливающий исходный сигнал из этого представления. Пусть входное пространство $X \in \mathcal{X}$, а пространство признаков (скрытых представлений) $h \in \mathcal{F}$. Тогда автокодировщик ищет два отображения f и g так, чтобы минимизировать ошибку реконструкции входа.

После решения задачи функция скрытого слоя h (выходные данные кодировщика), известная как «кодированные признаки», может рассматриваться как представление входных данных X. Автокодировщики можно разделить на разреженные, шумоподавляющие, вариационные, свёрточные и другие.

Блок 2.1. Одномерная свёртка (Conv1d).

На схеме в блоке энкодера и остаточных блоках явно указаны слои Conv1d. В отличие от двумерных свёрток (Conv2D), одномерная свёртка применяется для анализа последовательных данных. Слой Conv1d перемещает ядро (фильтр) вдоль одного измерения, извлекая локальные паттерны из k-мерных представлений ДНК. Это позволяет модели находить характерные мотивы и взаимосвязи между нуклеотидами независимо от их абсолютной позиции.

Блок 3. Резидуальная нейронная сеть.

Резидуальная нейронная сеть (ResNet) — это архитектура, использующая концепцию «остаточного» соединения (skip connections). Данная архитектура была представлена в работе He K. et al. «Deep Residual Learning for Image Recognition» [29]. Принцип работы ResNet заключается во введении «пропускного соединения» для прямой передачи данных на последующие слои.
Предположим, что на вход слоя подаётся x. После прохождения через нелинейное преобразование на выходе получается F(x, {W_i}), где F — функция отображения, а {W_i} — весовые параметры. В остаточной сети на выходе этого слоя получается:
y = F(x, {W_i}) + x
Где x — это входные данные, переданные через «пропускной канал», а F(x, {W_i}) — остаточная функция (разница между входом и результатом).

Блок 3.1. Функция активации ReLU.
Между слоями свёртки присутствует блок ReLU (Rectified Linear Unit) — функция активации, определяемая формулой: f(x) = max(0, x). Она вносит нелинейность и помогает решить проблему затухающего градиента.

Блок 4. Транспонированная остаточная сеть.
Транспонированная остаточная сеть (Transposed Residual Network) сочетает принципы остаточных связей и транспонированных свёрток. Данная архитектура была представлена в работе “Transpose convolution based model for super-resolution image reconstruction” [30]. Математическое описание выглядит следующим образом:
y = T(F(x, {W_i}) + x)
где:
x и y — входной и выходной векторы блока;
F(x, {W_i}) — функция остаточного отображения;
T(·) — операция транспонированной свертки (upsampling), применяемая к сумме входа и остатка.

Блок 5. Блок классификации (FCN)

В правой части схемы изображен «Блок классификации», который принимает на вход сжатое представление данных (латентный вектор) из выхода энкодера (синяя стрелка от центральной части автоэнкодера на Рис.7). Блок состоит из слоев FCN 1, FCN 2, FCN 3. Аббревиатура FCN (Fully Connected Network) обозначает полносвязные слои (аналог многослойного перцептрона). В данной архитектуре этот блок выполняет задачу обучения с учителем: он преобразует абстрактные признаки, выделенные энкодером, в вероятности принадлежности входного образца к определённому субтипу ВИЧ.

Блок 6. Комбинированная функция потерь
Нижняя и правая части схемы демонстрируют логику обучения модели через блок **«Общая ошибка»** (розовый блок). Схема показывает, что обучение управляется двумя потоками ошибок:

1. Ошибка реконструкции: Вычисляется как разница между исходной последовательностью и выходом декодера. Она заставляет автоэнкодер сохранять максимум информации в латентном пространстве.
2. Ошибка классификации: Вычисляется на выходе блока FCN. Она направляет энкодер на выделение именно тех признаков, которые важны для различения субтипов вируса. Итоговая целевая функция модели представляет собой сумму этих двух компонентов:

$$Loss_{total} = Loss_{reconstruction} + Loss_{classification}$$

Такой подход (Multi-task learning) позволяет сформировать признаковое пространство, которое является одновременно информативным (восстанавливаемым) и дискриминативным (разделимым по классам).

Реализация архитектуры и анализ результатов

Разработка и обучение глубокой нейронной сети проводились с использованием программной библиотеки с открытым исходным кодом PyTorch [31], обеспечивающей эффективную работу с тензорами и автоматическое дифференцирование. Использованная архитектура 1D-CNN с остаточными блоками (Residual Blocks) позволила эффективно извлекать иерархические признаки из векторов частот k-меров, минимизируя риск исчезновения градиентов при глубоком обучении.

Результаты численного эксперимента и оценка качества классификации
Для оценки прогностической способности модели HIV-1-M-Env-Rus было проведено тестирование на независимой валидационной выборке, включающей 7200 последовательностей (сбалансировано по ~500 образцов для каждого из 12 субтипов). Использование сбалансированного набора данных позволило исключить смещение модели в сторону доминирующих классов.
По итогам 100 эпох обучения модель достигла исключительных показателей эффективности, что подтверждается динамикой функции потерь, представленной на Рис.8. Кривая обучения демонстрирует стабильную сходимость алгоритма (convergence) без признаков переобучения. Как видно из графика, основное снижение ошибки происходит в течение первых 20 эпох, после чего модель выходит на плато, достигая минимума функции потерь (Cross-Entropy Loss) к 100-й эпохе. Это свидетельствует о том, что нейросеть успешно выделила устойчивые паттерны, характерные для каждого из 12 субтипов, и оптимизировала веса для их безошибочной классификации.

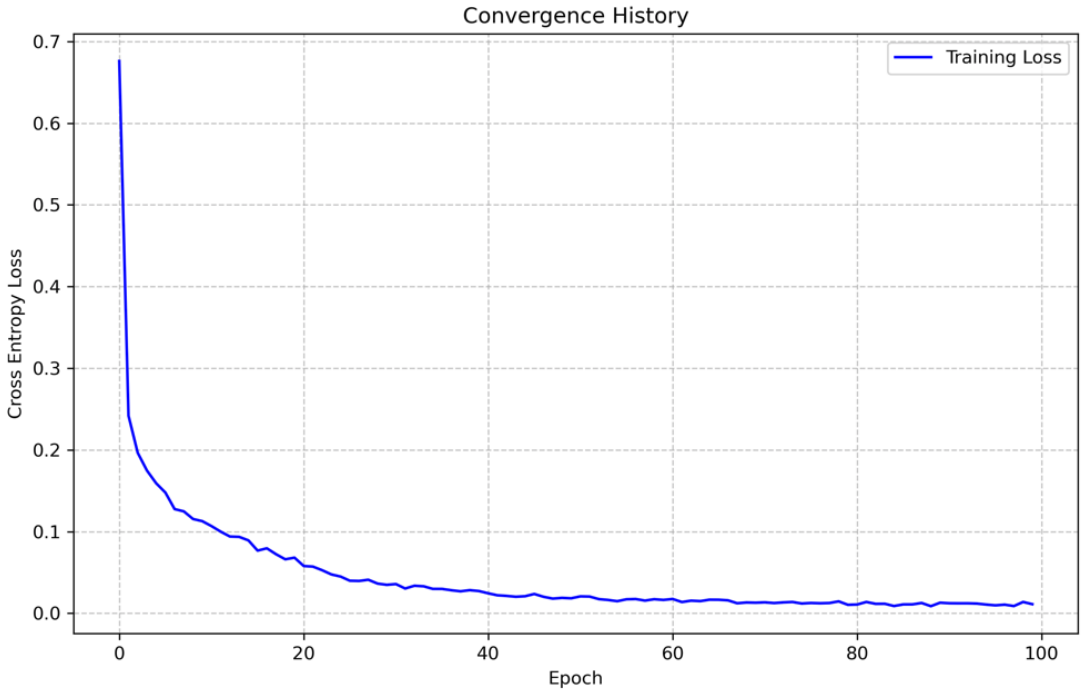

Рис. 8. Кривая сходимости.

Итоговая точность (Accuracy) составила 0.9964 (99,64%), что свидетельствует о практически безошибочной идентификации субтипов.

Детальная структура предсказаний визуализирована с помощью Матрицы ошибок (Confusion Matrix), представленной на Рис. 9.

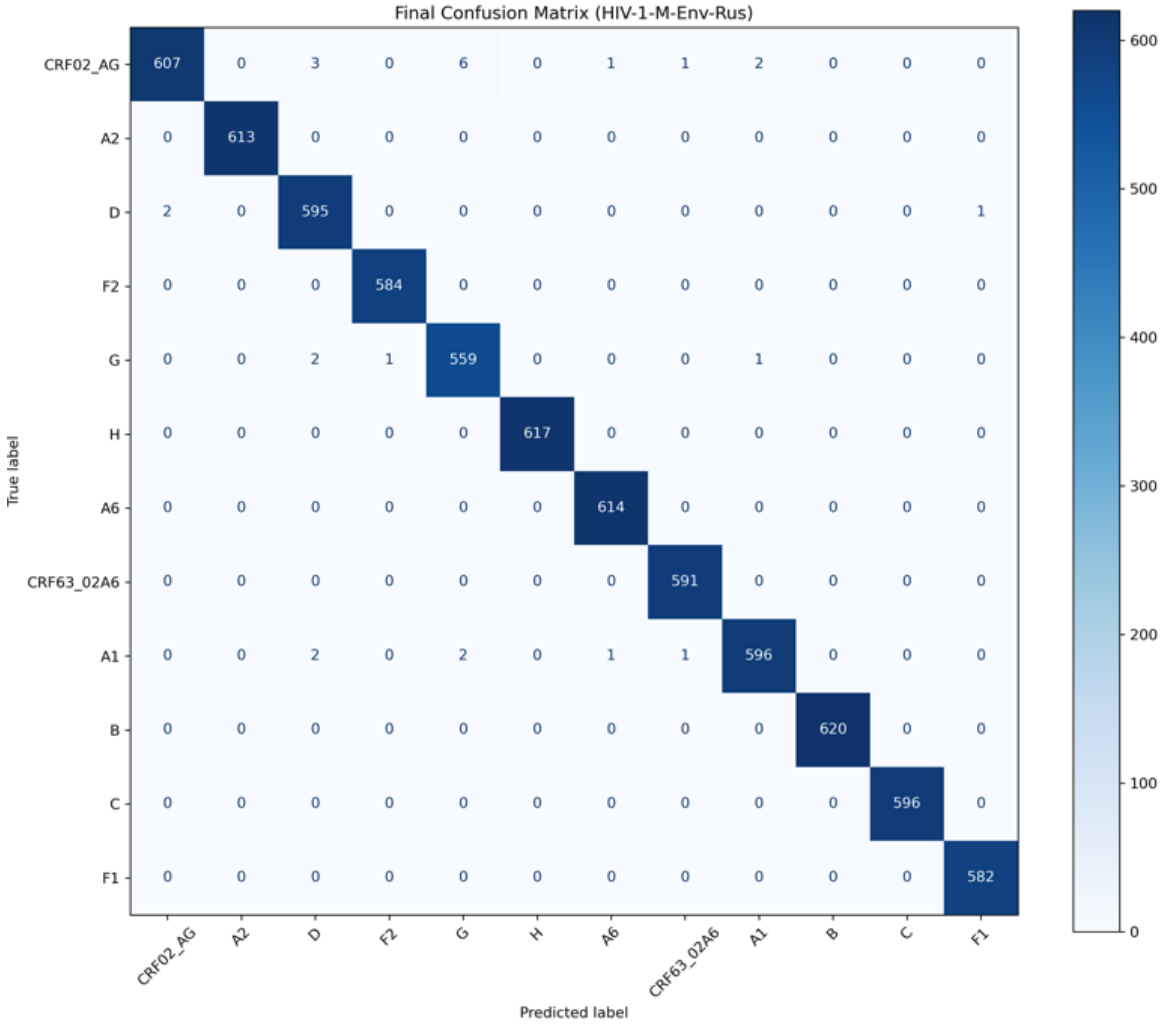

Рис. 9. Матрица ошибок (Confusion Matrix)

На диаграмме наблюдается ярко выраженная диагональная структура, где высокая концентрация значений вдоль главной оси соответствует верным предсказаниям (True Positives) для каждого из 12 классов. Отсутствие значимых скоплений вне главной диагонали подтверждает низкий уровень межклассовой интерференции: модель уверенно дифференцирует даже филогенетически близкие варианты вируса. Единичные случаи дисперсии (ошибочной классификации) носят стохастический характер и не образуют устойчивых ложных кластеров, что подтверждает высокую специфичность разработанной архитектуры нейронной сети

Таблица 7. Анализ ключевых метрик.

Субтип	Precision	Recall	F1-score	Особенности классификации
A6	1	1	1	Абсолютная точность для доминирующего типа в РФ
CRF63_02A6	1	1	1	Безошибочное определение ключевого рекомбинанта
H, A2, F2	1	1	1	Успешная работа с редкими типами после аугментации (AME)
CRF02_AG	1	0.98	0.99	Незначительная погрешность из-за высокой изменчивости

Специфичность для Российской Федерации: Особо следует отметить, что для субтипов A6 и CRF63_02A6, являющихся основными драйверами эпидемии в Российской Федерации, модель продемонстрировала абсолютные показатели Precision и Recall (1.00). Это подтверждает, что инструмент HIV-1-M-Env-Rus полностью адаптирован к региональной специфике.

Эффективность синтетической эволюции: Для субтипов с критически малым объемом исходных данных (H, A2), где количество реальных последовательностей не превышало десяти, модель достигла показателя F1-score = 1.00. Это является прямым доказательством эффективности примененного метода искусственной молекулярной эволюции: сгенерированные синтетические данные успешно имитируют естественную вариабельность вируса.

Стабильность классификации: Значения F1-score для всех классов находятся в диапазоне 0.99–1.00. Минимальное снижение полноты (Recall) до 0.98 для субтипа CRF02_AG объясняется филогенетической сложностью данной рекомбинантной формы, однако и этот результат

значительно превосходит точность существующих зарубежных аналогов при работе с короткими фрагментами гена env.

Валидация модели на данных базы RuHIV

Для оценки эффективности инструмента HIV-1-M-Env-Rus в условиях реальной эпидемиологической ситуации в Российской Федерации была проведена дополнительная валидация на независимых образцах из российской базы данных устойчивости ВИЧ к антиретровирусным препаратам RuHIV. В выборку были включены как наиболее распространенный в РФ субтип A6, так и ключевой рекомбинант CRF63_02A6 и субтип B, занимающий второе место по представленности в стране. Итоговые результаты валидации сведены в Таблицу 8.

Таблица 8. Результаты валидации модели на данных базы RuHIV

Идентификатор образца	**Прогнозируемый субтип**	**Уверенность (Confidence)**	**Примечание**
A6_1env	A6	99.22%	Высокая достоверность
A6_2env	A6	99.49%	Высокая достоверность
A6_3env	A6	54.52%	Требуется экспертная проверка
B_1env	B	99.56%	Надежная идентификация
CRF63_02A6env	CRF63_02A6	99.71%	Безошибочное определение рекомбинанта

Обсуждение.

Разработанный инструмент HIV-1-M-Env-Rus может быть интегрирован в качестве аналитического модуля в перспективные программно-аппаратные комплексы для оперативной диагностики. Это позволит клиницисту непосредственно во время приема получать данные о субтипе вируса, что имеет определяющее значение для прогнозирования естественной резистентности.

В частности, оперативная идентификация доминирующего в РФ субтипа A6 позволяет с высокой долей вероятности предположить наличие специфических природных полиморфизмов, таких как L74I в гене интегразы. Данная мутация характерна для восточноевропейской линии субтипа A и способна снижать генетический барьер резистентности к ингибиторам интегразы второго поколения (например, каботегравиру).

Таким образом, использование HIV-1-M-Env-Rus способствует реализации принципов персонализированной медицины, позволяя оптимизировать схемы антиретровирусной терапии на самых ранних этапах обследования

Выводы

В данной работе предложен биоинформатический инструмент HIV-1-M-Env-Rus — новый метод глубокого обучения для точной классификации субтипов группы M вируса ВИЧ-1 по последовательностям гена *env*. Ключевая особенность подхода — использование методов искусственной молекулярной эволюции для синтеза дополнительных образцов ДНК, что позволяет компенсировать одну из основных проблем ML/DL в биоинформатике: дефицита обучающих данных. Еще одна из особенностей HIV-1-M-Env-Rus - это фокус на получивших распространение в Российской Федерации субтипах A6, CRF02_AG и CRF63_02A6

Свёрточный автокодировщик с встраиванием ResNet извлекает информативные признаки из последовательностей с высокой точностью тем самым обеспечивает последующему классификатору очень высокое качество распознавания. Точность определения субтипа инструментом HIV-1-M-Env-Rus была валидирована на данных полученных из курируемой ФБУН ЦНИИ эпидемиологии Роспотребнадзора базы данных RuHIV.

Библиографический список

[1] Global genomic surveillance strategy for pathogens with pandemic and epidemic potential 2022–2032. URL: https://www.who.int/initiatives/genomic-surveillance-strategy (дата обращения: 03.01.2026).
[2] Bowyer S., Allen D.J., Furnham N. Unveiling the ghost: machine learning's impact on the landscape of virology // Journal of General Virology. 2025. Vol. 106, Issue 1. Article 002067. doi:10.1099/jgv.0.002067.
[3] Peng S.HIV-1 M group subtype classification using deep learning approach. Computers in Biology and Medicine. 2024 Dec;183:109218. Epub 2024 Oct 5. doi: 10.1016/j.compbiomed.2024.109218.
[4] Safina K.R., Sidorina Y., Efendieva N., Belonosova E., Saleeva D., Kirichenko A., Kireev D., Pokrovsky V., Bazykin G.A. Molecular epidemiology of HIV-1 in Oryol Oblast, Russia // Virus Evolution. 2022. Т. 8. № 1. veac044. doi:10.1093/ve/veac044.
[5] Кириченко А.А., Киреев Д.Е., Сидорина Ю.Н., Абашина Н.Д., Брусенцева Е.Е., Акимкин В.Г. Пилотное исследование по изучению особенностей распространения резистентных вариантов ВИЧ-1 с помощью молекулярных кластеров // Журнал микробиологии, эпидемиологии и иммунобиологии. 2024. Т. 101. № 5. С. 581–593. doi:10.36233/0372-9311-565
[6] Киреев Д.Е., Кириченко А.А., Лопатухин А.Э., Шлыкова А.В., Галкин Н.Ю., Савельер Е.В., Глазов М.Б., Покровский В.В., Акимкин В.Г. Российская база данных устойчивости ВИЧ к антиретровирусным препаратам. Журнал микробиологии, эпидемиологии и иммунобиологии. 2023;100(2):219–227. doi: https://doi.org/10.36233/0372-9311-345
[7] Kline R.L., Saduvala N., Zhang T., Oster A.M. Diversity and characterization of HIV-1 subtypes in the United States, 2008–2016 // Annals of Epidemiology. 2019. Vol. 33. P. 84–88. doi:10.1016/j.annepidem.2019.02.010.
[8] Исходный код и веса модели HIV-1-M-Env-Rus, URL: https://gitverse.ru/onsiteseq/HIV-1-M-Env-Rus (дата обращения: 03.01.2026).
[9] Дементьева Н.Е., Сизова Н.В., Лисицина З.Н., Беляков Н.А.Молекулярно-эпидемиологическая характеристика ВИЧ-инфекции в

Санкт-Петербурге // Медицинский академический журнал. 2012. Т. 12. № 2. С. 97–104. doi:10.17816/MAJ12297-104.
[10] Sivay M.V. et al.Spatiotemporal dynamics of HIV-1 CRF63_02A6 sub-epidemic.Frontiers in Microbiology (2022). DOI: 10.3389/fmicb.2022.946787
[11] Baryshev P.B., Bogachev V.V., Gashnikova N.M.HIV-1 Genetic Diversity in Russia: CRF63_02A1, a New HIV Type 1 Genetic Variant Spreading in Siberia.AIDS Research and Human Retroviruses. 2014;30(6):592–597. DOI: 10.1089/aid.2013.0196
[12] Ульянова Я.С. и соавт.Клинико-лабораторная характеристика острой ВИЧ-инфекции у взрослых в Новосибирской области. Журнал инфектологии. 2019;11(2):40–44. DOI: 10.22625/2072-6732-2019-11-2-40-44.
[13] Kirichenko A. et al. Genetic Features of HIV-1 Integrase Sub-Subtype A6 Predominant in Russia and Predicted Susceptibility to INSTIs. Viruses. 2020;12(8):838. DOI: 10.3390/v12080838.
[14] [Hu Z. et al. Effect of the L74I Polymorphism on Fitness of Cabotegravir-Resistant Variants of HIV-1 Subtype A6. J Infect Dis. 2023;228(10):1352–1356
[15] Obermeyer F., Abecasis A.B., Smit M., et al. Analysis of 6.4 million SARS-CoV-2 genomes identifies mutations associated with fitness // Science. 2022. Vol. 376. P. eabm1208. doi:10.1126/science.abm1208.
[16] Lopez-Rincon A., Tonda A., Mendoza-Maldonado L. et al. Classification and specific primer design for accurate detection of SARS-CoV-2 using deep learning // Scientific Reports. — 2021. — Vol. 11. — Art. 947. — DOI: 10.1038/s41598-020-80363-5
[17] GenBank 2024 Update / E. W. Sayers, M. Cavanaugh, K. Clark [et al.] // Nucleic Acids Research. – 2024. – Vol. 52, no. D1. – P. D134–D137. – DOI 10.1093/nar/gkad903
[18] Van der Maaten L., Hinton G. Visualizing data using t-SNE // Journal of machine learning research. – 2008. – Vol. 9. – No. 11. – P. 2579-2605.
[19] Kobak D., Berens P. The art of using t-SNE for single-cell transcriptomics // Nature Communications. – 2019. – Vol. 10. – No. 1. – P. 5416. – DOI 10.1038/s41467-019-13056-x.
[20] Дмитрюкова М.Ю., Киреев Д.Е., Лопатухин А.Э., Лаповок И.А., Шипулин Г.А. ТОЧНОСТЬ ОПРЕДЕЛЕНИЯ СУБТИПА ВИЧ-1 НА ОСНОВАНИИ АНАЛИЗА НУКЛЕОТИДНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ V3 ПЕТЛИ ГЕНА gp120. ВИЧ-инфекция и иммуносупрессии. 2015;7(1):40-44. https://doi.org/10.22328/2077-9828-2015-7-1-40-44

[21] Kuiken C., Korber B., Shafer R.W. HIV Sequence Databases [Электронный ресурс]. — 2003. — DOI: 10.1097/01.aids.0000174691.16088.98. — Режим доступа: https://www.hiv.lanl.gov/. — Дата обращения: 14.12.2025.
[22] Peng RD (2011) Reproducible research in computational science (Science). DOI: 10.1126/science.1213847.
[23] Sandve GK et al. (2013)Ten Simple Rules for Reproducible Computational Research (PLOS Computational Biology). DOI: 10.1371/journal.pcbi.1003285
[24] Köster J., Rahmann S. Snakemake—a scalable bioinformatics workflow engine // Bioinformatics. 2012. Т. 28, № 19. С. 2520–2522. DOI: 10.1093/bioinformatics/bts480.
[25] Киселев С.С., Озолинь О.Н., Панюков В.В. Использование k-меров для внутривидового типирования бактерий // Доклады Международной конференции «Математическая биология и биоинформатика» / под ред. В.Д. Лахно. Пущино: ИМПБ РАН, 2018. Т. 7. Статья № e59. DOI: 10.17537/icmbb18.101.-
[26] Le Cun Y., Fogelman-Soulié F. Modèles connexionnistes de l’apprentissage // Intellectica. 1987. No. 2–3. P. 114–143. doi:10.3406/intel.1987.1804.
[27] Bourlard H., Kamp Y. Auto-association by multilayer perceptrons and singular value decomposition // Biological Cybernetics. 1988. Vol. 59. P. 291–294. doi:10.1007/BF00332918.
[28] Hinton G.E., Zemel R.S. Autoencoders, Minimum Description Length and Helmholtz Free Energy // Advances in Neural Information Processing Systems. 1994. Vol. 6. P. 3–10. doi:10.5555/2987189.2987190.
[29] He K., Zhang X., Ren S., Sun J. Deep Residual Learning for Image Recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016. P. 770-778. https://doi.org/10.1109/CVPR.2016.90.
[30] Sahito F., Zhiwen P., Sahito F., Ahmed J. Transpose convolution based model for super-resolution image reconstruction. Applied Intelligence. 2023. Vol. 53. P. 10574–10584. https://doi.org/10.1007/s10462-023-10368-9
[31] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... & Chintala, S. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. Advances in Neural Information Processing Systems, 32.DOI:10.48550/arXiv.1912.01703

Приложение 1.

Таблица 1. Доля Субтипов ВИЧ-1 в Российской Федерации в %

Вариант ВИЧ-1 (РФ)	Доля (%)
A6	83.14%
B	5.63%
CRF63_02A6	2.97%
A6 + CRF02_AG	2.60%
CRF02_AG	2.16%
A6 + B	1.82%
G	0.73%
C	0.33%
CRF03_A6B	0.26%
Другие*	0.26%
CRF01_AE	0.10%

Таблица 2. Доля Субтипов ВИЧ-1 в США в %

Вариант ВИЧ-1 (США)	Доля, %
B	93.6
C	1.3
CRF02_AG	0.9

A	0.7
CRF01_AE	0.6
G	0.5
Другие “чистые” не-B субтипы	0.3
Другие CRF	0.5

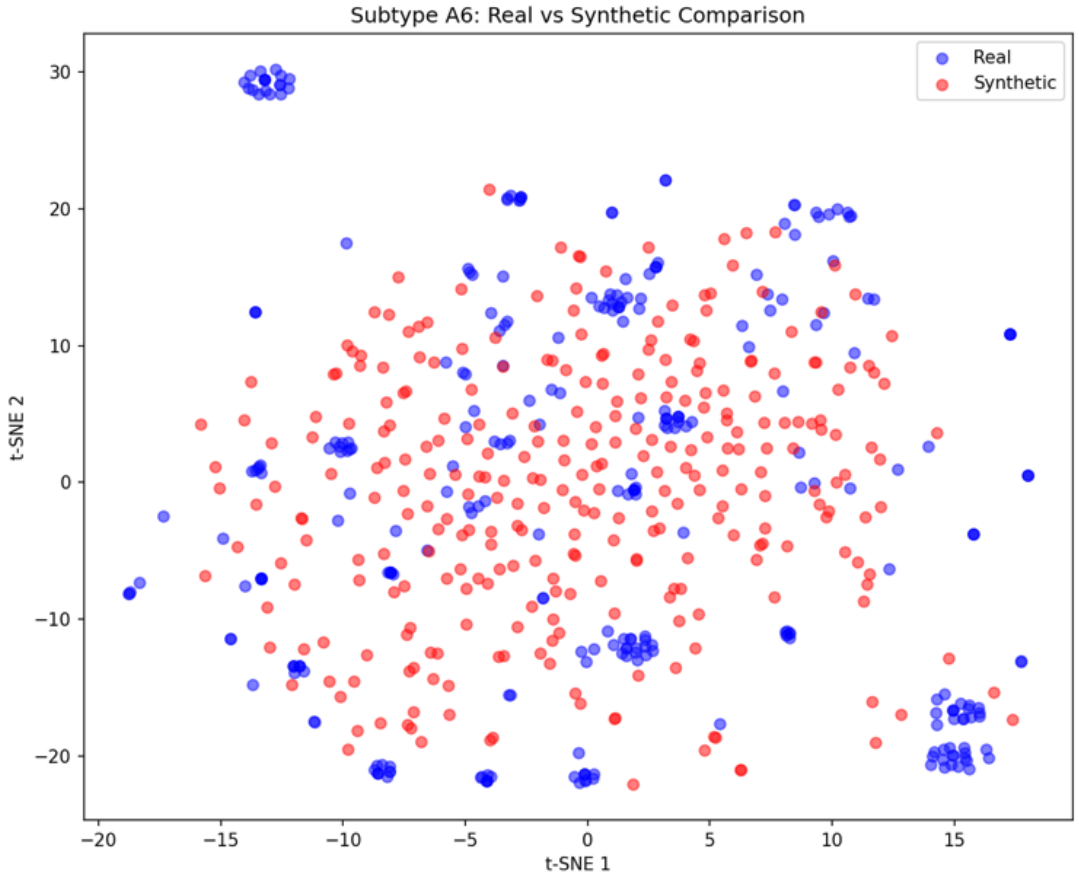

Рис.4. Проверка генерации синтетических данных для субтипа A6 методом t-SNE

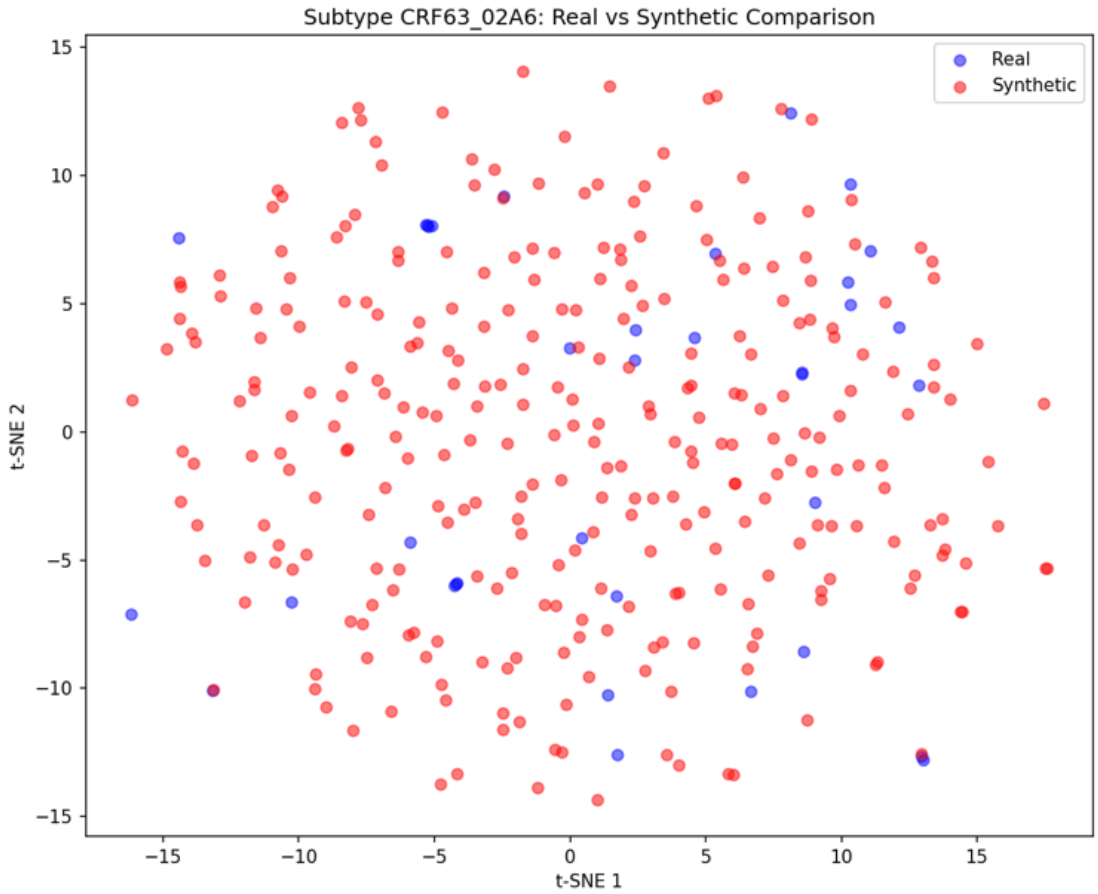

Рис.5. Проверка генерации синтетических данных для субтипа CRF63_02A6 методом t-SNE