МЕТОДОЛОГИЯ УПРАВЛЕНИЯ ЖИЗНЕННЫМ ЦИКЛОМ МАШИННО-ОБУЧАЕМЫХ МОДЕЛЕЙ В БИОМЕДИЦИНСКИХ ПРИЛОЖЕНИЯХ: ИССЛЕДОВАНИЕ НА ДАННЫХ О РЕЗИСТЕНТНОСТИ ВИЧ-1

Горбенко Роман Анатольевич
e-mail: gorbenko.ra@phystech.edu

Аннотация

В работе впервые предложена концепция BioMLOps представляющая собой адаптацию методологий MLOps для задач вычислительной биологии. Эффективность подхода продемонстрирована на примере создания HIV-1-Resist-Rus, системы прогнозирования лекарственной устойчивости (ЛУ) ВИЧ-1. Внедрение BioMLOps обеспечило управляемость жизненного цикла моделей: от выбора архитектуры нейросети до непрерывного мониторинга метрик (Validation Loss, Accuracy) на эталонных данных.

Разработанный инструмент реализует автоматическое версионирование экспериментов с сохранением метрик (включая F1-score) в JSON-формате, что позволяет проводить регрессионное тестирование. Реализован пайплайн дообучения при обновлении внешних баз данных (Stanford HIVDB) с автоматической детекцией деградации качества (model degradation) или фиксацией значимых улучшений на отложенной выборке (Hold-out Set).

Исходный код и веса моделей опубликованы на платформе GitVerse. Это обеспечивает возможность интеграции в научную и клиническую практику для локального анализа геномных данных, гарантируя соблюдение требований к суверенитету данных без их передачи на зарубежные серверы.

.

Ключевые слова: Машинное обучение, ML, MLOPS, BioMLOps, сверточные нейронные сети, Self-Attention, ВИЧ-1, лекарственная устойчивость.

Введение

Вирус иммунодефицита человека (ВИЧ-1) остается глобальной угрозой общественному здравоохранению. По данным ВОЗ, в 2025 году число людей, живущих с ВИЧ, достигло 40,8 млн, из которых 1,25 млн проживают в Российской Федерации. Ключевым препятствием для эффективной антиретровирусной терапии (АРТ) является лекарственная устойчивость (ЛУ) вируса. В России проблема приобрела критический характер: локальные исследования фиксируют резистентность у более чем 50% пациентов с опытом приема терапии [1,2].

Современные подходы к прогнозированию ЛУ все чаще опираются на методы глубокого обучения. Исследования [3] подтверждают эффективность сверточных нейронных сетей (CNN) для анализа нуклеотидных последовательностей из баз данных Stanford HIVDB [4] и RuHIV [5]. В настоящей работе базовая архитектура CNN была модифицирована добавлением слоя Self-Attention [6] для учета дистальных взаимодействий (эпистаза) между аминокислотными остатками, что соответствует передовым практикам в области белковых языковых моделей.

Однако внедрение таких моделей в клиническую практику сопряжено с рядом инженерных вызовов. В отличие от классической разработки ПО, биоинформатические системы подвержены «дрейфу данных» (data drift) из-за постоянного пополнения баз генотипов. Отсутствие строгих стандартов версионирования связки «Датасет-Модель-Гиперпараметры» приводит к кризису воспроизводимости экспериментов и невозможности объективного сравнения архитектур (например, CNN против CNN+Attention).

Для преодоления этих ограничений в работе впервые предложена методология BioMLOps представляющую собой адаптацию хорошо зарекомендовавших себя практик MLOps [7] для специфики вычислительной биологии. На примере разработанного инструмента HIV-1-Resist-Rus [8] продемонстрирована реализация ключевых концепций BioMLOps:

1. Версионирование данных: строгий контроль версий наборов «генотип-фенотип».
2. Отслеживание экспериментов: автоматическая фиксация метрик и параметров обучения.
3. CI/CD for Bio: автоматизированные пайплайны переобучения и валидации при поступлении новых клинических данных.

Исходные данные для исследования

В качестве основы для обучения и валидации моделей использовался курируемый набор данных **Stanford HIV Drug Resistance Database (HIVDB)** [4], содержащий сопоставленные пары «генотип–фенотип». Выборка включает нуклеотидные последовательности генов *pol* ВИЧ-1, кодирующих ферменты-мишени антиретровирусной терапии (протеазу, обратную транскриптазу и интегразу), а также соответствующие им количественные показатели лекарственной устойчивости (Fold Change), полученные фенотипическими методами *in vitro*.

Анализ проводился для четырех основных классов антиретровирусных препаратов, составляющих основу современных схем АРТ:

1. **Нуклеозидные ингибиторы обратной транскриптазы (НИОТ / NRTI).**

В обучающую выборку были включены данные по резистентности к ключевым препаратам первой линии и нуклеозидным основам схем: *абакавир (ABC), зидовудин (AZT), ламивудин (3TC), тенофовир (TDF)*.

2. **Ненуклеозидные ингибиторы обратной транскриптазы (ННИОТ / NNRTI).**

Анализируемый набор данных включал профили устойчивости к препаратам первого и второго поколений: э*фавиренз (EFV), невирапин (NVP), этравирин (ETR), рилпивирин (RPV), доравирин (DOR)*.

3. **Ингибиторы протеазы (ИП / PI).**

Анализируемый набор данных включал профили устойчивости к препаратам: *атазанавир (ATV), дарунавир (DRV), лопинавир (LPV)*.

4. **Ингибиторы интегразы (ИИ / INI или INSTI).**

В выборку вошли данные по устойчивости к препаратам: *биктегравир (BIC), долутегравир (DTG), элвитегравир (EVG), ралтегравир (RAL)*.

Общий объем выборки после препроцессинга и фильтрации дубликатов составил **30 921** пару «генотип–фенотип», охватывающую **8 275** уникальных вариантов вирусного генома. Каждая запись содержала аннотацию уровня устойчивости, классифицированную согласно алгоритму Stanford HIVDB (Susceptible, Low, Intermediate, High).

Для оценки обобщающей способности моделей на данных, специфичных для эпидемиологической ситуации в РФ, был дополнительно сформирован синтетический валидационный набор данных. Основой для генерации послужили результаты исследования Hu Z et al. (2023) "Effect of the L74I Polymorphism on Fitness of Cabotegravir-Resistant Variants of Human Immunodeficiency Virus 1 Subtype A6" [9].

Данное исследование демонстрирует клиническую значимость полиморфизма **L74I** в гене интегразы, который является маркерным для субтипа **A6** (доминирующего в Российской Федерации [1, 2, 5]). Установлено, что L74I сам по себе не вызывает резистентности, но способен компенсировать снижение фитнеса вируса при возникновении мутаций устойчивости к ингибиторам интегразы второго поколения (в частности, к **долутегравиру (DTG) и каботегравиру**). Поскольку эти препараты обладают схожим профилем резистентности, а долутегравир является основой терапии первой линии в РФ и имеет наиболее полную обучающую выборку, валидация модели проводилась именно на DTG.

Генерация данных производилась методом программного сайт-направленного мутагенеза (*in silico site-directed mutagenesis*). В консенсусную последовательность интегразы были внесены следующие комбинации мутаций для моделирования различных фенотипических сценариев:

1. **Варианты с двойными мутациями (A6):** Комбинации L74I + G118R и L74I + R263K - ожидаемый статус *лекарственно устойчивый* с высоким риском вирусологической неудачи.
2. **Варианты с одиночными мутациями (B):** Изолированные мутации резистентности (например, G118R) без фона L74I.

Использование данного синтетического набора позволило проверить способность нейросетевых архитектур учитывать эпистатические взаимодействия между удаленными позициями генома (кодоны 74 и 118/263) и корректно интерпретировать профили лекарственной устойчивости для субтипа A6, недостаточно представленного в обучающей выборке Stanford HIVDB.

Архитектура моделей нейронных сетей

1. Базовая модель: Сверточная нейронная сеть (1D-CNN)

В качестве базового решения (baseline) была выбрана архитектура одномерной сверточной нейронной сети (1D-Convolutional Neural Network), доказавшая свою эффективность в задачах анализа геномных последовательностей [3] .

2. Модифицированная архитектура: CNN с механизмом самовнимания (CNN + Self-Attention)

Для преодоления ограничений базовой модели и учета глобального контекста была разработана гибридная архитектура, дополненная механизмом самовнимания Self-Attention, впервые предложенный в работе [6] . Биологическое обоснование данной модификации заключается в необходимости моделирования эпистаза являющегося нелинейным взаимодействием между пространственно удаленными аминокислотными остатками, которые могут сближаться в трехмерной структуре вирусного белка.

В предложенной архитектуре карты признаков, сформированные сверточными слоями (Conv1D), не подаются сразу на полносвязные слои, а служат входом для блока Self-Attention. Данный механизм вычисляет матрицу внимания (attention scores), определяющую степень взаимосвязи каждого элемента последовательности со всеми остальными элементами. Это позволяет модели динамически фокусироваться на значимых комбинациях мутаций, независимо от расстояния между ними в первичной структуре генома.

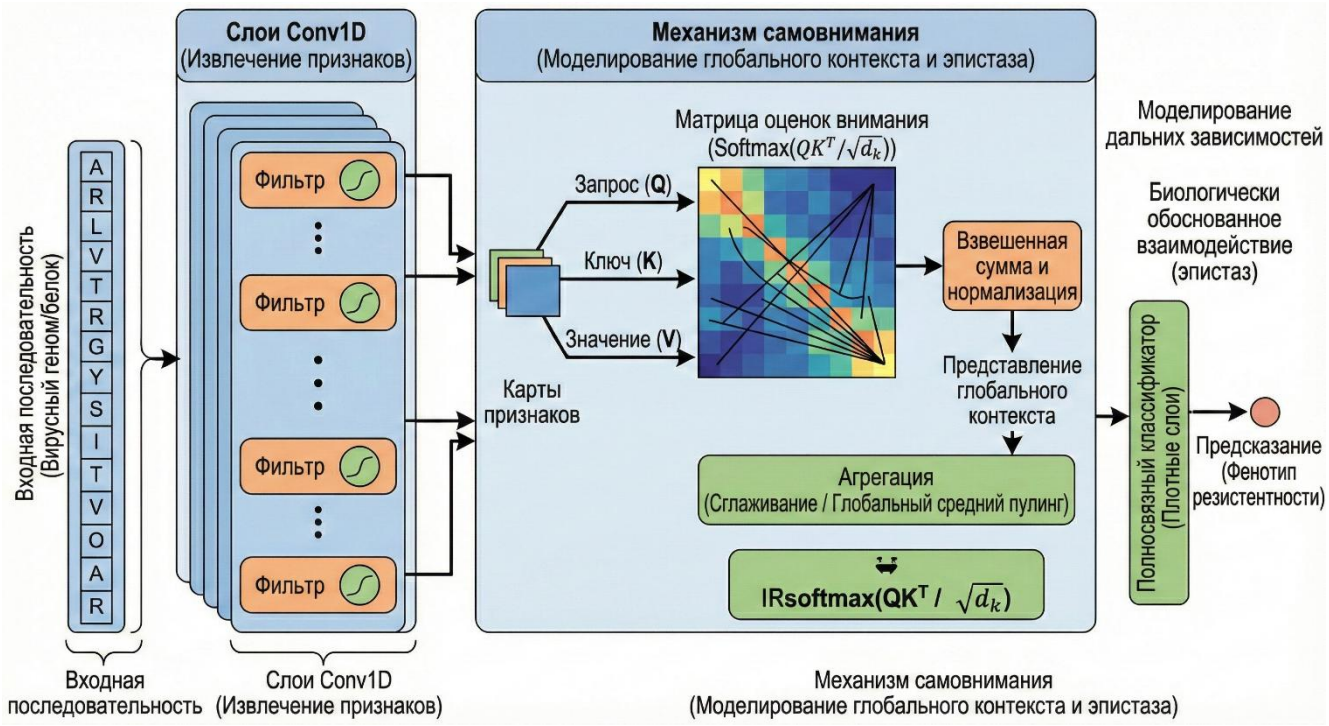

Рис. 1. Модифицированная архитектура: CNN с механизмом самовнимания (CNN + Self-Attention)

Интеграция слоя Self-Attention позволяет модели формировать глобальное представление (global representation) вирусной последовательности, эффективно улавливая дальние зависимости (long-range dependencies), критически важные для точного прогнозирования фенотипа резистентности в сложных мутационных профилях. Выход блока внимания агрегируется (Global Average Pooling или Flatten) и передается в полносвязный классификатор.

Обучение и валидация нейронных сетей

Для обеспечения воспроизводимости экспериментов и строгого контроля версий данных и моделей была реализована методология BioMLOps с использованием инструмента DVC (Data Version Control) [10]. Весь жизненный цикл моделирования (от препроцессинга сырых данных до обучения и валидации) был организован в виде автоматизированного пайплайна, зафиксированного в конфигурационном файле dvc.yaml.

1. Сравнительный анализ архитектур (A/B Testing)
В ходе исследования проводилось сравнение двух различных архитектур глубокого обучения для оценки их способности выявлять паттерны лекарственной устойчивости, включая сложные эпистатические взаимодействия мутаций. Эксперимент проводился параллельно для препарата Долутегравир (DTG), что отражено в соответствующих стадиях пайплайна (train_dtg_cnn и train_dtg_attn).

2. Протокол обучения
Обучение моделей производилось с использованием фреймворка PyTorch [11]. Для всех экспериментов использовались унифицированные гиперпараметры, зафиксированные в параметрах DVC

3. Валидация на синтетических данных (In Silico Stress Test)
Помимо стандартной валидации на отложенной выборке Stanford HIVDB, была разработана стратегия стресс-тестирования на синтетических данных (in silico), моделирующих редкие и сложные для детекции генотипы, характерные для субтипа A6 (распространенного в РФ).
Генерация тестового набора основывалась на данных исследования [9], доказавшего влияние полиморфизма L74I на восстановление репликативной способности вируса при наличии мутаций резистентности к ингибиторам интегразы. С помощью скрипта create_test.py были созданы последовательности, содержащие комбинации L74I с ключевыми мутациями резистентности (G118R, R263K).

Результаты и их обсуждение
Оценка эффективности разработанных моделей проводилась в два этапа: (1) анализ метрик обучения на валидационной выборке с использованием BioMLOps-пайплайна и (2) тестирование на синтетических данных для проверки биологической интерпретируемости (Biological Sanity Check).
1. Динамика обучения и метрики (Валидация)
Из визуального анализа результатов (рис. 2), следует, что внедрение механизма Self-Attention стабилизирует процесс обучения. Предложенная гибридная архитектура достигает плато функции потерь (Loss convergence) к 22-й эпохе, в то время как базовой CNN требуется 32 эпохи.

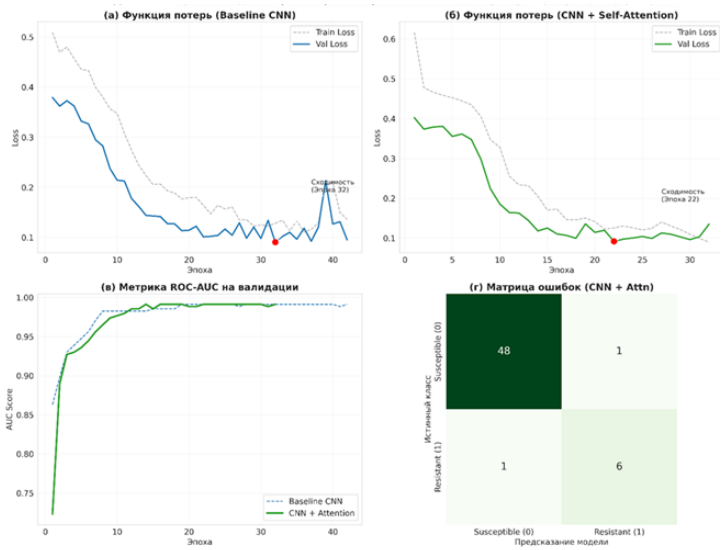

Рис. 2. Динамика обучения моделей для препарата Долутегравир (DTG).

2. Тестирование на синтетических данных

Критически важным этапом валидации, реализованным через отдельную стадию BioMLOps -пайплайна (compare_synthetic), стала проверка способности моделей детектировать эпистаз.
Для этого использовались синтетические профили: "дикий тип" (WT), профиль с одиночной первичной мутацией (G118R), профиль с одиночной вторичной мутацией (L74I) и их комбинация. Результаты инференса представлены в Таблице 1.

Таблица 1.

Сравнительный анализ предсказаний вероятности резистентности (DTG) на синтетических данных

ID / Мутации	*CNN (Prob)*	*Attention (Prob)*	*Интерпретация*
Wild_Type_B (WT)	*0.003*	*0.011*	*Обе модели верно определяют чувствительность (Prob ≈ 0).*
A6_Consensus (L74I)	*0.006*	*0.022*	*Одиночная вторичная мутация не вызывает резистентности.*
B_Mutant_1 (G118R)	*0.878*	*0.801*	*Ключевая мутация вызывает резистентность в обеих моделях.*
A6_Mutant_1 (L74I + G118R)	*0.925*	*0.903*	*Эффект синергии мутаций.*
Прирост уверенности (Δ)	*+0.047*	*+0.102*	*Вклад L74I в контексте G118R.*

Из Таблицы 1 видно, что обе модели успешно справляются с классификацией очевидных случаев (WT и G118R). Однако анализ дельты вероятностей выявляет фундаментальное различие в архитектурах:

1. **Моделирование эпистаза:** При добавлении вторичной мутации L74I к первичной G118R, базовая модель CNN увеличила вероятность резистентности лишь незначительно (на **0.047**). В то же время, модель с механизмом Self-Attention отреагировала на эту комбинацию приростом уверенности на **0.102** (более чем в 2 раза сильнее).
2. **Биологическое обоснование:** Известно, что мутация L74I сама по себе мало влияет на резистентность к Долутегравиру, но компенсирует дефекты репликации, вызванные мутацией G118R, усиливая общий профиль устойчивости. Модель **CNN + Attention** успешно "уловила" эту дистантную зависимость между 74-й и 118-й аминокислотами, чего не смогла сделать обычная сверточная сеть, ограниченная размером ядра свертки.

Заключение

В рамках настоящего исследования разработан биоинформатический инструмент HIV-1-Resist-Rus для прогнозирования лекарственной устойчивости ВИЧ-1, основанный на передовых методах глубокого обучения и адаптированных практиках MLOps.
Основные итоги работы можно сформулировать следующим образом:

1. **Методология BioMLOps:** Впервые формализована и успешно апробирована концепция BioMLOps. Внедрение инструментов версионирования данных (DVC) и автоматизации экспериментов позволило решить критическую для биоинформатики проблему «дрейфа данных» и кризиса воспроизводимости. Разработанный пайплайн обеспечивает прозрачное отслеживание эволюции моделей при обновлении клинических баз данных (Stanford HIVDB).
2. **Архитектурное преимущество:** Сравнительный анализ продемонстрировал превосходство гибридной архитектуры **CNN + Self-Attention** над базовыми сверточными сетями. Модифицированная модель показала более высокую скорость сходимости (достижение плато функции потерь на 22-й эпохе против 32-й у базовой CNN) и стабильность обучения.
3. **Биологическая интерпретируемость:** С помощью стресс-тестирования на синтетических данных (in silico) доказано, что механизм самовнимания (Self-Attention) позволяет модели

учитывать глобальный контекст и **эпистатические взаимодействия** между удаленными аминокислотными остатками. Модель с вниманием более чем в 2 раза чувствительнее реагировала на синергию мутаций L74I и G118R (дельта уверенности +0.102 против +0.047 у CNN), что подтверждает её способность моделировать сложные компенсаторные механизмы, характерные для субтипа A6, доминирующего в РФ.

4. **Практическая значимость:** Исходный код и предобученные модели опубликованы на отечественной платформе GitVerse, что обеспечивает технологический суверенитет решения. Система готова к локальному развертыванию в медицинских и исследовательских центрах РФ, позволяя проводить анализ геномных данных без их передачи на зарубежные сервера и обеспечивая высокую точность прогнозирования для актуальных схем антиретровирусной терапии.

Список использованных источников

[1] 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.

[2] 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:10.36233/0372-9311-56

[3] Steiner M. C., Gibson K. M., Crandall K. A. Drug Resistance Prediction Using Deep Learning Techniques on HIV-1 Sequence Data // Viruses. 2020. Vol. 12, № 5. P. 560. DOI: 10.3390/v12050560.

[4] Tang, M. W. The HIVdb System for HIV-1 Genotypic Resistance Interpretation / M. W. Tang, T. F. Liu, R. W. Shafer // Intervirology. 2012. Vol. 55, № 1. P. 47–53. DOI: 10.1159/000331998.

[5] Киреев Д.Е., Кириченко А.А., Лопатухин А.Э., Шлыкова А.В., Галкин Н.Ю., Савельев 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

[6] Vaswani A., Shazeer N., Parmar N. et al. Attention Is All You Need // arXiv preprint arXiv:1706.03762. 2017. DOI: 10.48550/arXiv.1706.03762.

[7] Eken B., Pallewatta S., Tran N. K., Tosun A., Ali Babar M. A Multivocal Review of MLOps Practices, Challenges and Open Issues // ACM Computing Surveys. 2026. Vol. 58, № 2. DOI: 10.1145/3747346.

[8] Исходный код и веса модели HIV-1-Resist-Rus, URL: https://gitverse.ru/onsiteseq/HIV-1-Resist-Rus, (дата обращения: 03.02.2026).

[9] Hu Z., Cordwell T., Nguyen H., Li J., Jeffrey J.L., Kuritzkes D.R. Effect of the L74I Polymorphism on Fitness of Cabotegravir-Resistant Variants of Human Immunodeficiency Virus 1 Subtype A6 // The Journal of Infectious Diseases. 2023. Vol. 228, no. 10. P. 1352–1356. DOI: 10.1093/infdis/jiad291.

[10] Petrov, D. DVC: Data Version Control for Data Science Projects [Электронный ресурс] / D. Petrov [et al.]. Электрон. дан. Zenodo, 2020. Режим доступа: https://doi.org/10.5281/zenodo.3677553 (дата обращения: 15.02.2026).

[11] 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