УДК 616.12-07-08:577.21:004.8

ИИ-МОДЕЛИ ОДНОКЛЕТОЧНОЙ МУЛЬТИОМИКИ В ДИАГНОСТИКЕ И ЛЕЧЕНИИ СЕРДЕЧНО-СОСУДИСТЫХ ЗАБОЛЕВАНИЙ

Адылова Ф.Т., Давронов Р.Р.
Институт математики им. В.И.Романовского АН РУз

XULOSA

Kardiologiyada bir hujayrali texnologiyalardan foydalanish fundamental tadqiqotlardan tortib klinik diagnostika va shaxsiylashtirilgan terapiyalarni ishlab chiqishgacha bo'lgan keng ko'lamli qo'llanmalarni qamrab oladi. Bir hujayrali RNK ketma-ketligi (scRNA-seq) yurak va qon tomir to'qimalarida an'anaviy usullar bilan aniqlab bo'lmaydigan noyob hujayra turlarini va yangi subpopulyatsiyalarni aniqlash imkonini beradi. Ushbu texnologiyalar ateroskleroz, miokard infarkti, yurak yetishmovchiligi va gipertrofik kardiyomiyopatiyada molekulyar yo'llarni aniqlash uchun ishlatiladi. Hujayra rivojlanish traektoriyalarini tahlil qilish regenerativ tibbiyot uchun juda muhim bo'lgan o'zak hujayralarining shakllanishi va differentsiatsiyasi jarayonlarini tushunishga yordam beradi. Hujayra o'zaro ta'sir modellari bizga sog'lom va kasal holatlarda turli hujayra turlari (masalan, kardiomiotsitlar, fibroblastlar va immun hujayralari) o'rtasidagi aloqani simulyatsiya qilish imkonini beradi. Ushbu maqolada zamonaviy GPT (generativ oldindan o'qitilgan transformer) arxitekturasidan foydalangan holda keng ko'lamli, bir hujayrali, ko'p o'zgaruvchan ma'lumotlarni tahlil qilish uchun ishlab chiqilgan yangi scGPT bazaviy modeliga e'tibor qaratib, bir hujayrali modellarni qo'llash yondashuvlariga umumiy nuqtai nazar keltirilgan. Model genlar va hujayralar haqidagi tegishli biologik ma'lumotlarni samarali ravishda to'playdi va hujayra turini annotatsiya qilish, genetik buzilishlarni bashorat qilish va gen tarmog'ini chiqarish kabi turli vazifalarda yuqori samaradorlikka erishish uchun osongina sozlanishi mumkin. Muhimi, scGPT modeli https://github.com/bowang-lab/scGPT manzilida ochiqchasiga mavjud.

Kalit so'zlar: *kardiologiya, bitta hujayrali multiomika, aniq diagnostika, fazoviy transkriptomika, maqsadli terapiya, generativ model, sun'iy intellekt.*

SUMMARY

The application of single-cell models (single-cell technologies) in cardiology covers a wide range of tasks, from basic research to clinical diagnostics and the development of personalized therapy. Single-cell RNA sequencing (scRNA-seq) makes it possible to identify rare cell types and new subpopulations in heart and vascular tissues that cannot be detected by traditional methods. The technologies are used for decoding molecular pathways in atherosclerosis, myocardial infarction, heart failure, and hypertrophic cardiomyopathy. The analysis of cell development trajectories helps to understand the processes of heart formation and stem cell differentiation, which is critical for regenerative medicine. Cellular interaction models allow us to model communication between different types of cells (for example, cardiomyocytes, fibroblasts, and immune cells) in a healthy and pathological state. The article provides an overview of approaches to the application of single-cell models, while focusing on the new basic scGPT model, developed for the analysis of large-scale single-cell multidimensional data using the modern GPT architecture (generative pre-trained transformer). The model effectively captures meaningful biological information about genes and cells and can be easily customized to achieve high performance in other tasks, including cell type annotation, prediction of genetic perturbations, and gene network inference. Importantly, the scGPT model is publicly available at https://github.com/bowang-lab/scGPT.

Keywords: *cardiology, single-cell multiomics, accurate diagnosis, spatial transcriptomics, targeted treatment, generative model, artificial intelligence.*

За последние 10 лет достижения в технологиях секвенирования РНК отдельных клеток (scRNA-seq) оказали преобразующее воздействие на биомедицинские исследования, позволив проводить профилирование и анализ транскриптомов отдельных клеток. В частности, scRNA-seq способствовало идентификации новых или редких типов клеток, анализу построения траекторий отдельных клеток и дифференцировке стволовых или прогениторных клеток, а также сравнению здоровых и пораженных тканей на

уровне отдельных клеток. Эти приложения сыграли решающую роль в развитии сердечно-сосудистых исследований за последнее десятилетие, о чем свидетельствует создание клеточных атласов сердца, выяснение механизмов, участвующих в развитии сердечно-сосудистой системы и дифференцировке стволовых или прогениторных клеток.

scRNA-seq кардинально меняет уровень понимания кардиологии, позволяя анализировать транскриптом каждой клетки в отдельности, а не усредненные показатели ткани. scRNA-seq помогает проследить «траектории» развития клеток от предшественников до зрелых форм, а при патологиях (инфаркт миокарда, сердечная недостаточность, аневризма) метод позволяет увидеть, какие именно группы клеток подвергаются стрессу, как меняется их коммуникация и какие сигнальные пути активируют фиброз или воспаление.

Детальная дешифровка ответов конкретных клеток на повреждение, позволяет находить новые мишени для лекарств, не затрагивая здоровые типы клеток. scRNA-seq выявил высокую гетерогенность сердечных клеток, позволив описать новые подтипы кардиомиоцитов, функциональные подгруппы фибробластов и специфические макрофаги, отвечающие за воспаление или регенерацию. Технология позволяет идентифицировать специфические клеточные мишени для лечения инфаркта и сердечной недостаточности, несмотря на технологические сложности и высокую стоимость. [25,14,7].

Что даёт применение одноклеточной мультиомики в кардиологии?

1. Точная диагностика: определение новых биомаркеров на уровне отдельных клеток для раннего выявления сосудистых патологий, таких как аневризма и расслоение аорты;

2. Персонализированная терапия: использование моделей на основе индуцированных плюрипотентных стволовых клеток (iPSCs) пациента в сочетании с одноклеточным анализом для тестирования лекарств;

3. Пространственная транскриптомика:ожидается внедрение методов, позволяющих не просто изучать отдельные клетки, но и картировать их точное расположение в ткани сердца, что раскрывает влияние микроокружения на функции органа;

4. Таргетное лечение: идентификация конкретных патологических субпопуляций макрофагов или Т-клеток в атеросклеротических бляшках для точечного иммунотерапевтического воздействия.

Применение мультиомики единичных клеток (single-cell multi-omics) в сочетании с искусственным интеллектом (ИИ) является современным «золотым стандартом» в кардиологических исследованиях. Если scRNA-seq изучает только РНК, то мультиомика рассматривает клетку комплексно: ДНК (эпигеномика), РНК (транскриптомика) и белки (протеомика) одновременно. Мультиомные данные имеют большие объёмы, но алгоритмы машинного обучения (например, scVI или MOFA+) позволяют объединять данные разных слоев (например, состояние хроматина и экспрессию генов) в единую модель. ИИ анализирует мультиомные профили тысяч клеток из биопсии или крови пациента и находит «скрытые» признаки болезни на самых ранних стадиях. Модели ИИ (например, CellRank) позволяют моделировать будущее состояние клетки. В диагностике это помогает понять пути восстановления сердца после инфаркта: регенерация или образование рубца (фиброза). ИИ использует мультиомные данные для моделирования ответа конкретных типов клеток на препарат. ИИ-мультиомика определяет мишени для CRISPR -терапии, т.к. находит сравнение их эффективности в конкретных задачах. Это может быть конкретный участок ДНК, открытый для транскрипции в клетках проводящей системы сердца при аритмии, который подвергается воздействию. Следовательно, эти модели на базе ИИ дают врачу прогностические инструменты, которые позволяют лечить пациента на основе его уникального «клеточного атласа» сердца и сосудов [17,30,35].

Сегодня исследования одиночных клеток распространены, но есть необходимость в использовании базовой модели, предварительно подготовленной на крупномасштабных данных [27]. Поэтому в данной статье рассматривается новая базовая модель scGPT, разработанная для анализа крупномасштабных одноклеточных многомерных данных с использованием современной архитектуры GPT (generative pre-trained transformer). Модель эффективно улавливает значимую биологическую информацию о генах и клетках и может быть легко настроена для достижения высокой производительности в различных задачах, включая аннотацию типа клетки, предсказание генетических возмущений и вывод генной сети. Важно, что модель scGPT находится в открытом доступе по адресу https://github.com/bowang-lab/scGPT.

Релевантные исследования

Следует отметить, что генеративные предварительно обученные модели в последнее время успешно применяются во многих областях [22,12,23,24]. Например, модели DALL-E2 (диффузионная модель с 3,5 миллиардами параметров) и GPT-4, следующие парадигме предварительной подготовки трансформеров на крупномасштабных разнообразных наборах данных [23,24] могут быть легко адаптированы к различным задачам и сценариям из кардиологии. Модели демонстрируют более высокую производительность при выполнении различных задач [31,8,34]. С точки зрения данных, широкомасштабные атласы секвенирования одноклеточной РНК (scRNA-seq), такие, как Human Cell Atlas, уже охватывают десятки миллионов клеток, но объем доступных данных omic продолжает расти в геометрической прогрессии [12,3,33]. И это открывает широкие возможности для использования новых методов ИИ, позволяющих изучать различные типы клеток, тканей и интегрировать их с различными органами.

В работе [11] представлена фундаментальная базовая модель одной клетки, - scGPT, полученная путем предварительного генеративного обучения на более чем 10 миллионах клеток. Предварительно обученная модель отражает значимую биологическую информацию как на генном, так и на клеточном уровнях. Карты изученных генов расшифровывают известные пути, группируя вместе гены, которые функционально значимы. Благодаря обучению в режиме zero-shot модель правильно классифицирует объекты, и поэтому может выявлять значимые кластеры клеток в новых наборах данных. Благодаря точной настройке (fine-tuning) модель обеспечивает хорошую производительность при решении разных задач, включая пакетную коррекцию, многомерную интеграцию, аннотацию типа клетки, предсказание генетических возмущений, генерацию псевдоклеток и вывод генной сети.

Реализация модели scGPT направлена на продвижение будущих исследований, поскольку внедрение предварительно подготовленных базовых моделей углубит понимание клеточной биологии и заложит основу для диагностики и лечения заболеваний в области кардиологических, нейродегенеративных и других заболеваний.

За последние два года появилось семейство фундаментальных моделей для биологии (Single-Cell Foundation Models), конкурирующих со scGPT. Как и scGPT, они обучаются по принципу больших языковых моделей (LLM).

Geneformer, главный конкурент scGPT,- фундаментальная трансформерная модель, которая, в отличие от scGPT, обучается на ранжированных списках экспрессии генов (в порядке их активности), а не на абсолютных значениях. Это делает её особенно устойчивой к «шуму» в данных и позволяет успешно применять в кардиологии. Geneformer использовался для идентификации потенциальных терапевтических мишеней при различных формах кардиомиопатий. Модель обучалась на атласах здорового и больного сердца, что позволило ей с высокой точностью предсказывать, какие гены «выходят из строя» при патологии. Geneformer может имитировать удаление или активацию гена в цифровой клетке, что помогает понять, как генетический сбой приведет к сердечной недостаточности или нарушению ритма, не проводя тысячи реальных экспериментов на мышах Geneformer лучше сохраняет сигналы от редких типов клеток, что важно для изучения постинфарктного восстановления, когда малые группы клеток запускают процесс регенерации или фиброза. Хотя Geneformer пока остается инструментом исследователей, его результаты уже используются для выбора генов-кандидатов в генной терапии. Например, данные модели помогают точнее настраивать системы CRISPR/Cas9 для коррекции мутаций, вызывающих наследственные аритмии или гипертрофическую кардиомиопатию. https://huggingface.co/ctheodoris/Geneformer.

SCimilarity,- модель глубокого обучения (метрический трансформер), разработанная исследователями из Genentech. Её главная задача: находить функционально похожие клетки в разных исследованиях, органах, независимо от того, как были собраны данные. SCimilarity,-лучший инструмент для сравнения и поиска специфических клеточных состояний между тысячами пациентов. Например, если у части пациентов с сердечной недостаточностью лекарство не работает, SCimilarity может помочь найти общие черты в их клетках, которые отличают их от «ответчиков» (responders). Модель работает в десятки раз быстрее, чем классические методы интеграции данных, что позволяет анализировать гигантские биобанки данных пациентов.https://genentech.github.io/scimilarity/index.html.

scBERT (single-cell Bidirectional Encoder Representations from Transformers),- модель, основанная на архитектуре BERT от Google, адаптированной специально для биологии отдельных клеток. Если Geneformer фокусируется на регуляции генов, а SCimilarity – на поиске похожих клеток, то scBERT - метод автоматической классификации и аннотации типов клеток. В кардиологии scBERT решает несколько задач. Поскольку после инфаркта миокарда состав клеток сердца стремительно меняется, scBERT помогает точно и быстро разметить тысячи клеток из биопсий, выделяя специфические субпопуляции, ответственные за воспаление или заживление. Модель способна уловить тонкие переходы между здоровым кардиомиоцитом и клеткой, находящейся в состоянии стресса или дегенерации, что важно для понимания прогрессирования сердечной недостаточности. В атеросклеротических бляшках или при миокардитах scBERT помогает идентифицировать редкие типы Т-клеток, которые могут быть виновниками агрессивного течения болезни. https://github.com/TencentAILabHealthcare/scBERT.

xTrimoGene xTrimoGene, - одна из самых масштабных фундаментальных моделей для анализа одноклеточных данных (Single-Cell Foundation Model), разработанная компанией BioMap. Если модели (scBERT, Geneformer) можно сравнить с небольшими библиотеками, то xTrimoGene,-огромный архив, обученный на колоссальном объеме данных. Благодаря обучению на 100 млн клеток, xTrimoGene лучше других предсказывает «траекторию болезни». Например, прогнозирует скорость перехода здорового кардиомиоцита в патологический при хронической гипертензии. Модель способна выявлять скрытые признаки клеточного стресса в образцах тканей пациентов с терминальной стадией сердечной недостаточности, которые не определяются обычными методами. За счет огромной базы знаний xTrimoGene используется для виртуального тестирования влияния новых молекул на экспрессию генов в клетках эндотелия сосудов или миокарда. https://www.biorxiv.org/content/10.1101/2023.03.24.534055v1.full. В таблице 1 представлены

характеристики этих моделей.

Таблица 1

Характеристики фундаментальных трансформерных моделей

Модель	Главное преимущество	Роль в кардиологии
scBERT	Точная аннотация	Определение типа клетки в биопсии.
Geneformer	Регуляторные сети	Предсказание эффекта от выключения гена.
xTrimoGene	Масштаб и точность	Глубокий анализ сложных состояний и поиск новых лекарств.

В данной работе мы покажем основные характеристики и результаты работы новой фундаментальной модели scGPT [17].

МЕТОДОЛОГИЯ

Базовая модель одной клетки

Секвенирование отдельных клеток позволяет получить генетические профили на уровне отдельных клеток. Новейшие клеточные эталонные карты, такие как «Атлас клеток человека», содержат миллионы отдельных клеток из различных органов и тканей, предлагая беспрецедентное представление о клеточной гетерогенности [12,9]. scGPT (single-cell Generative Pre-trained Transformer),- первая фундаментальная модель для биологии единичных клеток, построенная на архитектуре Transformer. Модель позволяет виртуально «выключить» ген или «ввести» лекарство и предсказать, как изменится состояние клетки. В лечении ССЗ это означает имитацию эффекта препарата без проведения реального опыта на клетках пациента. scGPT может объединять данные разных модальностей (РНК и доступность хроматина ATAC-seq), заполняя «пробелы» в данных, если один из анализов был неполным или зашумленным. Но главная ценность scGPT с точки зрения ИИ состоит в переносе обучения. Исследователям можно не обучать модель с нуля на данных конкретного исследования сердца; достаточно взять предобученную scGPT и «дообучить» (fine-tune) её на небольшом наборе данных по конкретной патологии с гарантией получения точного результата. scGPT помогает выявить не простые корреляционные, а сложные причинно-следственные связи в генных сетях, тем самым определяя транскрипционные факторы развития патологического процесса в сердце.

На рисунке 1A показан двухэтапный рабочий процесс, включающий предварительное обучение и тонкую настройку scGPT. На этапе предварительного обучения собрали более 10,3 миллионов данных scRNA-seq клеток крови и костного мозга с портала CellXGene [6] для обучения. В процессе обучения модель постепенно учится генерировать экспрессию генов в клетках на основе простых сигналов о клеточной или генной экспрессии. На этапе точной настройки можно применить предварительно обученную модель к новым наборам данных и конкретным задачам.

Модель scGPT изучает представления клеток и генов на основе разнообразных данных об отдельных клетках с помощью моделирования экспрессии генов. В представлении генов использовали прогнозирование экспрессии генов (Gene Expression Prediction, GEP), а для представлений клеток разработали программу прогнозирования экспрессии генов для клеточного моделирования (Gene Expression Prediction for Cell Modelling, GEPC), в которой модель предсказывает значения экспрессии генов на основе представлений клеток. Так создается прямая связь между профилем экспрессии генов и клеточной гетерогенностью, что позволяет проводить совместную оптимизацию в рамках модели scGPT.

Представляя надежную и унифицированную структуру, модель scGPT позволяет исследователям одной клетки легко использовать предварительно подготовленную базовую модель в конкретных исследованиях.

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Интеграция разнообразных данных моделью scRNA-seq

Кластеризация и визуализация данных секвенирования отдельных клеток сопряжены со значительными трудностями из-за наличия пакетных эффектов, возникающих при использовании разных наборов данных в качестве входных данных. Используя процесс точной настройки, модель эффективно решает эту задачу, сохраняя при этом истинные биологические сигналы, присущие данным. Модель обеспечивает современную производительность, сохраняя биологическую изменчивость интегрированных наборов данных. Для доказательства этого утверждения авторы провели сравнение scGPT с тремя популярными методами интеграции: scVI [29], Seurat [28] и Harmony [15] на двух наборах данных Immune Human [20] и PBMC 10K [1]. Как показано на рисунке 2A, в наборе данных Immune Human модель scGPT успешно объединила все партии CD4+ T-клеток, CD8+ T-клеток и CD14+ моноцитов в отдельные кластеры, в то время как Seurat создал несколько подкластеров, соответствующих секвенированным партиям для каждого из этих типов клеток. Модели также удалось отделить дендритные клетки, полученные из моноцитов, от моноцитов CD16+, но scVI и Harmony обнаружили значительное совпадение двух кластеров. Эффективность кластеризации scGPT отражена в показателе биологической сохранности, где scGPT достигает среднего балла по шкале AvgBIO, равного 0,812, что на 5% выше, чем у Seurat и Harmony, и на 10% выше, чем у метода глубокого обучения scVI. На рисунке 2C показаны оценки по всем показателям кластеризации типов клеток,где scGPT заняла первое место по общему показателю.

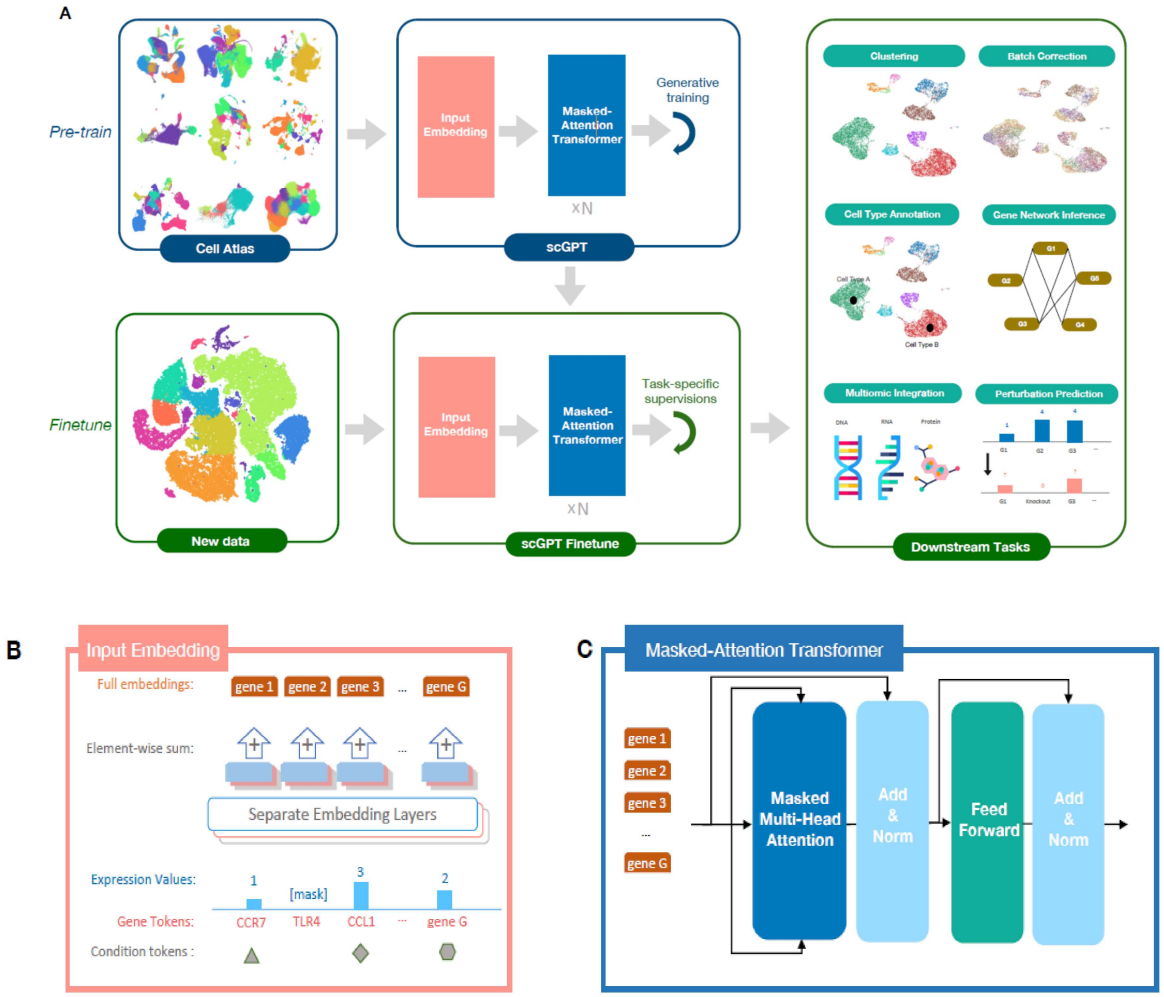

Рис. 1. Схема модели scGPT. (A).Сначала модель генерируется на основе крупномасштабных данных scRNA-seq из Cell Atlas. Основной компонент scGPT содержит блоки генеративного обучения. (B) Входные данные, которые содержат три уровня информации: маркер гена, значение экспрессии и маркеры условий. (C) Слой scGPT-преобразователя для проведения генеративной предварительной подготовки по данным секвенирования отдельных клеток.

Отметим преимущества предобучения, которое значительно повышает производительность в модели с точной настройкой по сравнению с моделью в режиме zero-shot на наборе данных PBMC 10K (рис. 2B). В отличие от традиционных генетических карт, которые показывают физическое расположение на хромосоме, генные карты визуализируют пространства генов, где гены со сходными признаками или картинками совместной экспрессии сгруппированы вместе.

Аннотация типов клеток

Аннотация типов клеток является важным этапом анализа отдельных клеток после кластеризации, поскольку позволяет выявить гетерогенность в секвенированных тканях и закладывает основу дальнейшего изучения функций клеток и генов для получения биологической и патологической информации. Было предложено несколько методов для аннотаций, - cellAssign [2], singleR [10] и Chetah [19], которые обычно требуют уменьшения размерности, что может привести к потере информации. В отличие от этого, модель scGPT может непосредственно и объективно учитывать экспрессию генов, используя в качестве входных данных полный набор генов с высокой вариабельностью. Для задачи аннотирования типов клеток авторы применили тонкую настройку предварительно обученной модели scGPT, используя перекрестную потерю энтропии в сравнении с метками достоверности из нового набора справочных данных.

Чтобы подтвердить эффективность модели scGPT в аннотации типов клеток, её сравнили с моделью TOSICA (Transformer for One-Stop Interpretable Cell type Annotation) [18], которая сегодня является новейшим методом аннотации клеток. scGPT превосходит TOSICA по всем показателям,- точности, чувствительности, специфичности и MacroF1.

Прогнозирование генетических возмущений

Методы секвенирования и редактирования генов недавно позволили проводить эксперименты, позволяющие исследовать реакцию клеток на множественные генетические возмущения. Подход имеет огром-

ные перспективы для раскрытия новых взаимодействий генов и развития регенеративной медицины. Но обширное комбинаторное пространство потенциальных нарушений в генах выходит за пределы экспериментальной осуществимости. Чтобы преодолеть это ограничение, модель scGPT может использовать знания, полученные о клеточных реакциях в известных экспериментах, и экстраполировать их для прогнозирования реакций в неизвестных сценариях.

Использование механизмов саморегуляции в генах позволяет кодировать сложные взаимодействия между возмущенными генами и реакциями других генов. Используя эту возможность, scGPT может эффективно извлекать уроки из существующих экспериментальных данных и точно прогнозировать экспрессию генов после возмущения.

Для решения задачи прогнозирования возмущений в [37] использовали два предварительно обработанных набора данных о возмущениях: 1. набор данных Pertub-seq для клеточной линии лейкемии K562 [4], который содержит 87 одногеновых возмущений, приблизительно по 100 клеток на возмущение и минимум 7000 невозмущенных клеток и 2. набор данных Norman Perturb-Seq [32], состоящий из 131 двухгенового возмущения и 105 одногеновых возмущений. Прогноз возмущения оценили вычислением корреляции Пирсона (corr) между предсказанными и истинными значениями экспрессии гена после возмущения.

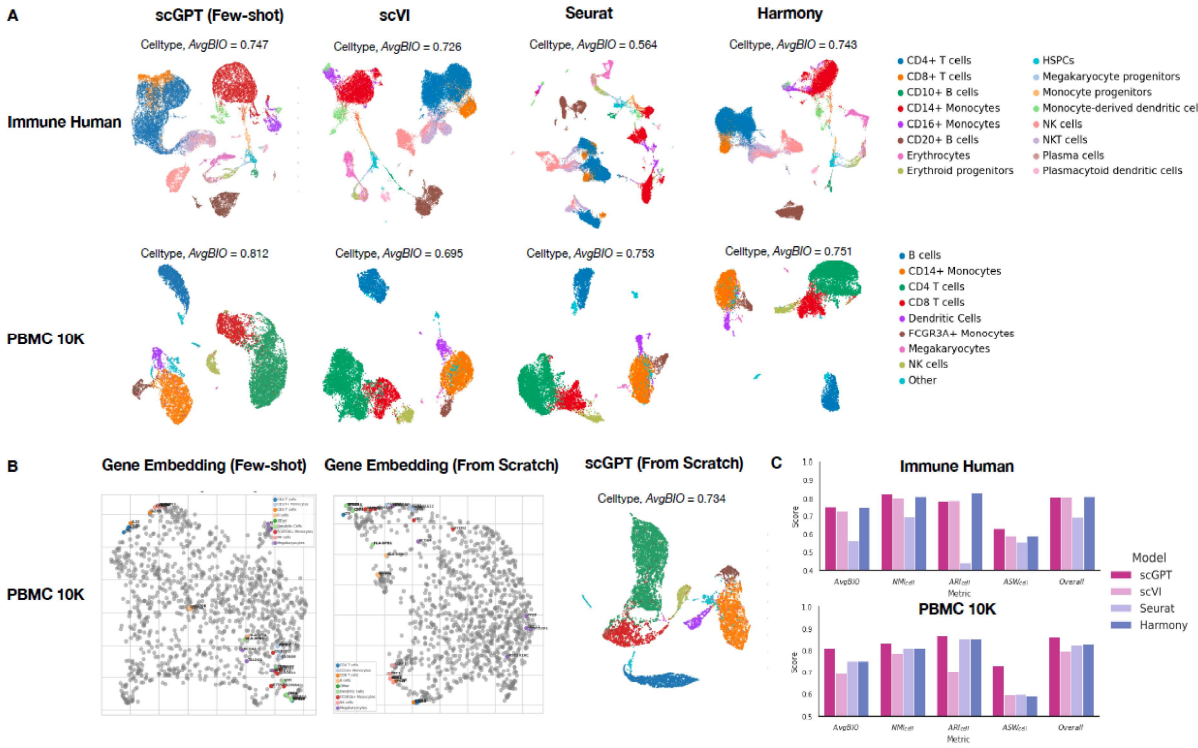

Рис. 2. (A) Сравнение модели scGPT с scVI [29], Seurat Seurat [28] и Harmony [15] на наборах Immune Human и PBMC 10K для кластеризации типов клеток после интеграции пакетов. (B) Сравнение карт генов в моделях scGPT, полученных с помощью нескольких кадров, и в моделях *zero-shot*. Выделены гены с высокой вариабельностью в каждом типе клеток. (C) Сравнение модели scGPT с другими тестируемыми методами на AvgBIO, с показателями биологического сохранения (NMIcell, ARIcell, ASWcell) и общей оценке.

Также ввели вариант показателя Пирсона, обозначаемый как corr(Δ), который измеряет корреляцию на основе величины изменения экспрессии после возмущения по сравнению с контролем. Показатели Пирсона представили для различных наборов генов, а именно, для всех генов (ALL) и 20 лучших дифференциально экспрессируемых генов (DE). Провели сравнение производительности между scGPT, новейшим методом GEARS и базовым уровнем (многослойный перцептрон, MLP). Результаты показывают, что scGPT достигает наивысшей корреляции по семи из восьми оценочных показателей.

Таким образом, оценка дифференциально экспрессируемых генов, в частности, столбцов DE в таблице 2, дает убедительные доказательства. scGPT демонстрирует значительные улучшения в корреляции изменения (Δ) в 20 ведущих дифференциально экспрессируемых генах, что, возможно, является ключевым показателем.

Мультимодальная интеграция модели

Одноклеточные многомерные (scMultiomic) данные представляют сразу несколько аспектов генетической регуляции, включая эпигенетическую, транскриптомную и трансляционную активность [16,38]. Это расширяет возможности для улучшения изучения признаков и представления клеток, помимо экс-

прессии генов. Однако проблема заключается в том, как надежно объединить представления клеток из нескольких видов, сохраняя при этом биологические сигналы.

Таблица 2

Результаты прогнозирования возмущений

Model	Norman et al. [32]				Adamson et al. [4]			
		DE	ALL			DE	ALL	
	corr	corr (Δ)	corr	corr (Δ)	corr	corr(Δ)	corr	corr (Δ)
MLP	0.909	0.428	0.987	0.408	0.948	0.729	0.991	0.656
GEARS	0.917	0.508	0.98G	0.387	0.9G1	0.726	0.991	0.652
scGPT	0.923	0.546	0.988	0.459	0.971	0.775	0.992	0.647

Каждый тип omic в scMulti-omic данных (например, экспрессия генов, доступность хроматина и содержание белка) можно рассматривать как отдельный язык. Модель scGPT поддерживает совместную оптимизацию с использованием различных методов секвенирования и также позволяет легко добавлять новые методы секвенирования в существующую предварительно обученную сеть. В тестовых экспериментах scGPT демонстрирует наилучшую производительность в изучении представлений клеток и задачах многоатомной пакетной интеграции по сравнению с существующими современными методами. scGPT эффективно извлекает интегрированные представления клеток из парных многомерных данных SCM (Structural Causal Models), которые используются для анализа сложных связей между множеством переменных (генотипом, экспрессией генов, метаболитами и фенотипами), выявляя причинно-следственные связи.

Был проведен сравнительный анализ scGPT с двумя самыми современными методами scGLUE[39] и Seurat v4[36] по эффективности кластеризации типов клеток. scGPT оказался единственным методом, который позволил получить четкий отдельный кластер для наивных клеток CD8, в то время как два других метода потерпели неудачу. scGPT также отличал B-клетки памяти от кластеров наивных B-клеток и промежуточных B-клеток, в то время как scGLUE создавал объединенный кластер из всех трех типов B-клеток. scGPT разделил группы клеток CD4 и CD8 на две отдельные группы кластеров, превзойдя результаты Seurat v4. Таким образом, scGPT демонстрирует высокую эффективность кластеризации типов клеток в целом (среднее значение BIO=0,767) и надежность по различным показателям биологического сохранения.

Встраивание генов для построения генной регуляторной сети

Взаимодействие между факторами транскрипции, кофакторами и генами-мишенями, лежащее в основе генной регуляторной сети (Gene Regulatory Network, GRN), опосредует важные биологические процессы. Модель scGPT способна группировать функционально родственные гены и дифференцировать функционально отличающиеся гены с помощью своей сети представлений генов. На рисунке 3A показана сеть сходства антигенов человеческих лейкоцитарных антигенов (HLA) на основе предварительно обученных генов.

В условиях zero-shot модель scGPT выделяет два кластера, соответствующие двум хорошо охарактеризованным классам HLA, которые вызывают различные иммунные реакции, а именно HLA класса I и HLA класса II [26]. Точно настроенная модель scGPT на основе набора данных Immune Human исследовала антигенную сеть CD(Cluster of Differentiation antigens), -набор маркеров, специфичных для типов иммунных клеток, присутствующих в этом наборе данных (см. рис. 3C). scGPT реконструирует значимые генные программы: на рисунке 3D визуализированы генные программы, извлеченные с помощью точно настроенной модели scGPT из набора данных Immune Human, и их дифференциальную экспрессию по типам клеток. Эти генные программы выбираются неконтролируемым образом путем предварительной кластеризации генов, а затем установления порогового значения для кластеров, состоящих из 5 или более генов, в соответствии с процедурами, предложенными Ceglia и соавторами [21].

Предполагаемые генные программы scGPT соответствуют функциональным группам, которые являются биологически значимыми. А ближайшие соседи гена CD8A скорее всего являются частью пути R-HSA-168256 иммунной системы чем гены, расположенные дальше (рис. 3B). Точно настроенная модель выявила один дополнительный ген GZMM, который обычно обогащен NK, NKT клетками: NK-клетки (натуральные киллеры) и NKT-клетки (натуральные киллеры Т-клетки и подкластерами Т-клеток [5], которые являются основными типами клеток, обнаруженных в наборе данных иммунной системы человека.

Среди пар генов существует положительная корреляция между показателем сходства генов и количеством общих путей, разделяемых этими генами, при этом показатель корреляции Пирсона равен 0,316. Эти результаты показывают, что scGPT усвоил значимые биологические закономерности, что говорит о способности модели выполнять неконтролируемое обнаружение генных программ на новых наборах данных наряду с другими задачами анализа на клеточном уровне, используя предварительно обучен-

ную модель.

ЗАКЛЮЧЕНИЕ

Сегодня (апрель 2026 г.) scGPT применяется в фундаментальных и трансляционных исследованиях сердечно-сосудистых заболеваний (ССЗ), а не в рутинной клинической практике. Основная роль модели заключается в расшифровке сложных клеточных механизмов при инфаркте миокарда (ИМ) и сердечной недостаточности (СН) на уровне отдельных клеток. В исследованиях сердечно-сосудистой системы scGPT используется для анализа данных секвенирования единичных клеток (scRNA-seq), что позволяет получить следующие результаты: идентификация новых клеточных подтипов, которые, например, активируются при инфаркте миокарда и способствуют фиброзу или регенерации;с помощью scGPT исследователи моделируют, как клетки сердца (кардиомиоциты) будут реагировать на ишемию или генетические изменения, что критично для понимания патогенеза сердечной недостаточности; scGPT эффективно объединяет данные разных исследований (разные пациенты, лаборатории и технологии), создавая единый «атлас» сердца для поиска новых терапевтических мишеней.

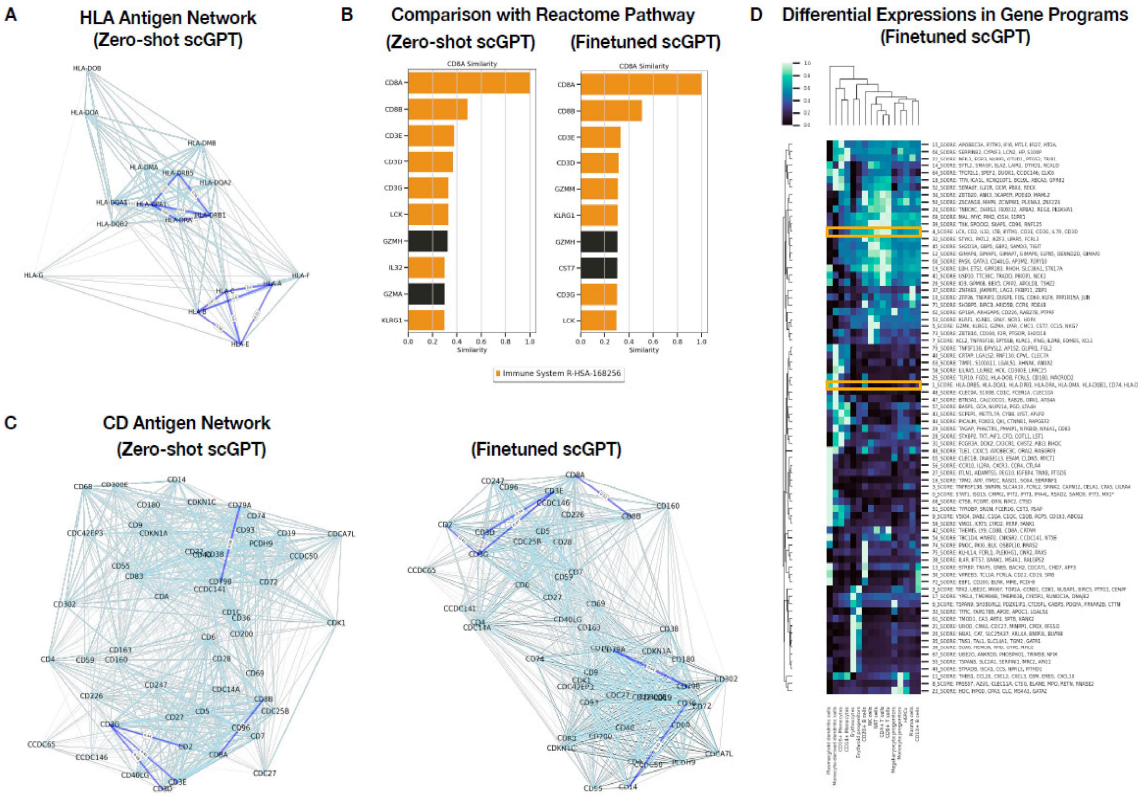

Рис. 3. (А) Антигенная сеть HLA, полученная из scGPT в режиме zero-shot. (В) Соседи гена CD8A из моделей scGPT zero-shot и точной настройки, ранжированные по сходству «основа-истина» из Reactome. (С) Сеть антигенов CD из набора данных Immune Human, обработанная н zero-shot и точно настроенного scGPT. (Д) Дифференциальные экспрессии генных программ, выделенных с помощью scGPT, по типам клеток в наборе данных Immune Human.

Таким образом, scGPT представляет собой первую базовую модель, в которой используются предварительно обученные трансформеры, изучающие более 10 миллионов данных из одной клетки [22,23]. Очевидны преимущества предварительного обучения, поскольку такая модель сама по себе является универсальным средством извлечения признаков и имеет возможности экстраполяции на невидимые наборы данных, представляя значимую кластеризацию клеток в экспериментах без предобучения. Изученные генные сети также отражают известные генные программы и их функциональные роли, что даёт уверенность в том, что предварительно обученная модель не только запомнила, но и синтезировала паттерны на основе крупномасштабных одноклеточных данных.

Однако, у scGPT есть технические проблемы при анализе с использованием одиночных клеток. Во-первых, несмотря на обширную предварительную подготовку, эти модели требуют значительной настройки для достижения приемлемой точности на основе новых данных. Во-вторых, представления генов вычисляются взвешиванием информации от всех других генов во входной последовательности.

В-третьих, современный вариант работает изолированно от применений, что создает разрыв между анализом отдельных клеток и практическими терапевтическими разработками. Чтобы решить эти проблемы, разработали scKAN, интерпретируемую платформу глубокого обучения для анализа транскриптомных данных отдельных клеток [24,13]. Цель scKAN - выполнить точную аннотацию типа клетки, одновременно идентифицируя маркерные гены, специфичные для этого типа клетки; эта модель используется в качестве базовой модель scGPT.

Ещё раз подчеркнем, что прямое использование scGPT «у постели больного» пока ограничено, но модель закладывает базу для персонализированной медицины. Каким образом? Модель выявляет генные сигнатуры, которые в будущем могут лечь в основу диагностических тестов для раннего обнаружения сердечной дисфункции. scGPT позволяет также предсказать, как определенные препараты повлияют на экспрессию генов в клетках сердца, ускоряя разработку таргетной терапии для лечения сердечной недостаточности. Во многих исследовательских работах результаты scGPT сопоставляются с данными визуализации (например, эхокардиографией) для верификации клеточных изменений, наблюдаемых при структурных заболеваниях сердца.

ЛИТЕРАТУРА

1. Adam Gayoso et al. “A Python library for probabilistic analysis of single-cell omics data”.In: Nature Biotechnology 40.2 (2022), pp. 163–166.
2. Allen W Zhang et al. “Probabilistic cell-type assignment of single-cell RNA-seq for tumor microenvironment profiling”. In: Nature methods 16.10 (2019), pp. 1007–1015.
3. Aviv Regev et al. “Science forum: the human cell atlas”. In: elife 6 (2017), e27041.
4. Britt Adamson et al. “A multiplexed single-cell CRISPR screening platform enables systematic dissection of the unfolded protein response”. In: Cell 167.7 (2016), pp. 1867–1882.
5. Britta Bade et al. “Differential expression of the granzymes A, K and M and perforin in human peripheral blood lymphocytes”. In: International immunology 17.11 (2005), pp. 1419– 1428.
6. Chanzuckerberg Initiative. CZ CELLxGENE Discover. https://cellxgene.cziscience.com/. Online; accessed 26 December 2022. 2022.
7. Chen Z, Wei L, Duru F, Chen L. Single-cell RNA Sequencing: In-depth Decoding of Heart Biology and Cardiovascular Diseases. Curr Genomics. 2020 Dec;21(8):585-601. doi:10.2174/1389202921999200604123914
8. Chi Sun et al. “How to fine-tune bert for text classification?” In: Chinese Computational Linguistics: 18th China National Conference, CCL 2019, Kunming, China, October 18–20, 2019, Proceedings 18. Springer. 2019, pp. 194–206.
9. Dennis Thompson Healthday Reporter. ’human cell atlas’ maps 1 million cell types in 33 organs. 2022. url: https://medicalxpress.com/news/2022-05-human-cell-atlas-million.html.doi: 10.1038/s41569-020-0359-y.
10. Dvir Aran et al. “Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage”. In: Nature immunology 20.2 (2019), pp. 163–172.
11. Haotian Cui1,Chloe Wang, Hassaan Maan, Bo Wang scGPT: Towards Building a Foundation Model for Single-Cell Multi-omics Using Generative AI doi: https://doi.org/10.1101/2023.04.30.538439
12. HCA. HCA DCP. https://data.humancellatlas.org/. Online; accessed 12 April 2023.
13. He, H., Tang, Z., Chen, G. et al. scKAN: interpretable single-cell analysis for cell-type-specific gene discovery and drug repurposing via Kolmogorov-Arnold networks. Genome Biol 26, 300 (2025). https://doi.org/10.1186/s13059-025-03779-0
14. Hu Y, Zhang Y, Liu Y, Gao Y, San T, Li X, Song S, Yan B and Zhao Z (2022)Advances in application of single-cellRNA sequencing in cardiovascular research.Front. Cardiovasc. Med. 9:905151. doi:10.3389/fcvm.2022.905151
15. Ilya Korsunsky et al. “Fast, sensitive and accurate integration of single-cell data with Harmony”. In: Nature methods 16.12 (2019), pp. 1289–1296.
16. Indhupriya Subramanian et al. “Multi-omics data integration, interpretation, and its application”. In: Bioinformatics and biology insights 14 (2020), p. 1177932219899051
17. Jiajia Liu et al. “Machine intelligence in single-cell data analysis: advances and new chal855 lenges”. In: Frontiers in Genetics 12 (2021), p. 655536.
18. Jiawei Chen et al. “Transformer for one stop interpretable cell type annotation”. In: Nature Communications 14.1 (2023), p. 223.
19. Jurrian K De Kanter et al. “CHETAH: a selective, hierarchical cell type identification method for single-cell RNA sequencing”. In: Nucleic acids research 47.16 (2019), e95–e95.
20. Malte D Luecken et al. “Benchmarking atlas-level data integration in single-cell genomics”.In: Nature methods 19.1 (2022), pp. 41–50.
21. Nicholas Ceglia et al. “GeneVector: Identification of transcriptional programs using densevector representations defined by mutual information”. In: bioRxiv (2022), pp. 2022–04.
22. OpenAI. CZ CELLxGENE Discover. https://openai.com/blog/chatgpt. Online; accessed 10 April 2023. 2023.
23. OpenAI. GPT-4 Technical Report. 2023. arXiv: 2303.08774 [cs.CL].
24. OpenAI.CZ CELLxGENE Discover. https://openai.com/product/dall-e-2. Online; accessed 10 April 2023.
25. Paik DT, Cho S, Tian L, Chang HY, Wu JC. Single-

cell RNA sequencing in cardiovascular development, disease and medicine. Nat Rev Cardiol. 2020 Aug;17(8):457-473.
26. Paola Cruz-Tapias, John Castiblanco, and Juan-Manuel Anaya. “Major histocompatibility complex: antigen processing and presentation”. In: Autoimmunity: From Bench to Bedside [Internet]. El Rosario University Press, 2013.
27. Philipp Angerer et al. “Single cells make big data: New challenges and opportunities in transcriptomics”. In: Current opinion in systems biology 4 (2017), pp. 85–91.
28. Rahul Satija et al. “Spatial reconstruction of single-cell gene expression data”. In: Nature biotechnology 33.5 (2015), pp. 495–502.
29. Romain Lopez et al. “Deep generative modeling for single-cell transcriptomics”. In: Nature methods 15.12 (2018), pp. 1053–1058.
30. Sergio Oller-Moreno et al. “Algorithmic advances in machine learning for single-cell expres876 sion analysis”. In: Current Opinion in Systems Biology 25 (2021), pp. 27–33.
31. Suchin Gururangan et al. “Don’t stop pretraining: Adapt language models to domains and tasks”. In: arXiv preprint arXiv:2004.10964 (2020).
32. Thomas M Norman et al. “Exploring genetic interaction manifolds constructed from rich single-cell phenotypes”. In: Science 365.6455 (2019), pp. 786–793.
33. Xiaoping Han et al. “Mapping the mouse cell atlas by microwell-seq”. In: Cell 172.5 (2018),pp. 1091–1107.
34. Xipeng Qiu et al. “Pre-trained models for natural language processing: A survey”. In: Science China Technological Sciences 63.10 (2020), pp. 1872–1897.
35. Yuge Ji et al. “Machine learning for perturbational single-cell omics”. In: Cell Systems 12.6 (2021), pp. 522–537.
36. Yuhan Hao et al. “Integrated analysis of multimodal single-cell data”. In: Cell 184.13 (2021), pp. 3573–3587.
37. Yusuf Roohani, Kexin Huang, and Jure Leskovec. “GEARS: Predicting transcriptional outcomes of novel multi-gene perturbations”. In: bioRxiv (2022).
38. Zhen Miao et al. “Multi-omics integration in the age of million single-cell data”. In: Nature Reviews Nephrology 17.11 (2021), pp. 710–724.
39. Zhi-Jie Cao and Ge Gao. “Multi-omics single-cell data integration and regulatory inference with graph-linked embedding”. In: Nature Biotechnology 40.10 (2022), pp. 1458–1466.
