

## Почему биологические данные важнее в ИИ-открытии лекарств

**Источник:** AI News Healthcare

**Дата публикации:** 2025-12

**Оригинал:** <https://www.artificialintelligence-news.com/news/gsk-relation-therapeutics-ai-drug-discovery-biological-data/>

GSK

target identification

биологические данные

машинное обучение

одноклеточные данные

открытие лекарств

фармацевтика

### GSK заключила исследовательское сотрудничество с Relation Therapeutics на сумму до \$110 миллионов

GSK заключила исследовательское сотрудничество с британской биотехнологической компанией Relation Therapeutics на сумму до \$110 миллионов, расширяя существующую работу компаний в области открытия лекарств с помощью ИИ (искусственного интеллекта).

Согласно договору, Relation будет генерировать крупномасштабные наборы данных, измеряющие, как человеческие клетки реагируют на генетические изменения и лекарственные вмешательства. Данные будут использоваться для обучения моделей ИИ, предназначенных для идентификации потенциальных мишеней для лекарств, включая модели в рамках платформы MORGAN компании Relation.

Соглашение объединяет генерацию биологических данных с разработкой моделей ИИ. Исследовательский подход Relation связывает вычислительный анализ с экспериментами, которые генерируют новую информацию о человеческих клетках.

Сотрудничество основывается на более ранних соглашениях между GSK и Relation, сосредоточенных на фиброзных заболеваниях и остеоартрите. Эти проекты включали наблюдательные исследования, предназначенные для создания двух функциональных наборов данных о заболеваниях для анализа с использованием платформы Lab-in-the-Loop компании Relation.

Предыдущая работа объединила человеческую генетику, одноклеточную мультиомику, полученную из человеческой ткани, функциональные анализы и машинное обучение для идентификации и валидации потенциальных мишеней заболеваний.

## **Как Relation генерирует биологические данные**

Relation описывает свой подход Lab-in-the-Loop как комбинацию лабораторных экспериментов и вычислительного анализа. Их работа включает профилирование тканей, одноклеточную и пространственную транскриптомику, секвенирование и валидацию мишеней, в то время как машинное обучение используется для идентификации мишеней, приоритизации, валидации и дизайна экспериментов.

Компания также проводит эксперименты с возмущениями, которые измеряют, как генетические изменения влияют на клеточные характеристики, связанные с заболеванием. Эти результаты затем могут быть проанализированы вместе с генетическими и полученными от пациентов биологическими данными.

Публичные репозитории остаются важным источником обучающего материала для фундаментальных биологических моделей, хотя комбинирование информации, полученной в разных исследованиях, может создавать технические проблемы.

Обзор 2025 года в журнале *Experimental & Molecular Medicine* отметил, что репозитории, включая CZ CELLxGENE, Human Cell Atlas и NCBI Gene Expression Omnibus, предоставляют исследователям доступ к большим объемам одноклеточных данных. Только CZ CELLxGENE предоставляет доступ к более чем 100 миллионам стандартизированных клеток, согласно обзору.

Методы выборки, протоколы секвенирования, экспериментальные процедуры и конвейеры обработки могут различаться между исследованиями. Одноклеточные данные также могут содержать технический шум и другие

артефакты, требующие тщательного отбора набора данных, фильтрации, балансировки композиции и контроля качества во время обучения фундаментальной модели.

Перекрытие наборов данных представляет ещё одну проблему. Обзор отметил, что одни и те же или похожие клетки могут появляться в нескольких публичных ресурсах, потенциально придавая им непропорциональное влияние во время обучения и создавая риски утечки данных, когда обучающие и тестовые наборы данных перекрываются.

Обзор обнаружил, что сбор высококачественного, не избыточного набора данных так же важен, как и архитектура модели, при создании надёжных фундаментальных одноклеточных моделей.

## **Большие биологические наборы данных не гарантируют лучшие модели**

Исследование, опубликованное в Nature Methods в июне этого года, изучало, как размер и разнообразие данных предварительного обучения влияли на фундаментальные одноклеточные модели с использованием корпуса из 22,2 миллиона клеток. Исследователи обучили 400 моделей и оценили их в 6 400 экспериментах.

Исследование обнаружило, что текущие фундаментальные одноклеточные модели, как правило, достигали плато производительности после обучения только на части доступного корпуса. В отличие от больших языковых моделей, оценённые системы не демонстрировали чётких законов масштабирования данных, при которых постоянное увеличение обучающих данных последовательно давало лучшие результаты.

Исследователи обнаружили, что ёмкость модели, размер набора данных и вычислительные ресурсы необходимо балансировать, а не просто увеличивать вместе. Исследование не установило, что меньшие или проприетарные наборы данных по своей сути лучше, но обнаружило, что добавление большего количества биологических обучающих данных не последовательно приводило к дальнейшему повышению производительности.

Отдельное исследование, опубликованное в Genome Biology в 2025 году, оценило две фундаментальные одноклеточные модели, Geneformer и scGPT, по нескольким задачам оценки без предварительного обучения. Модели не последовательно превосходили более простые подходы, в то время как

исследователи также выявили проблемы, связанные с эффектами партий, и предостерегли от предположения, что более крупные предварительно обученные модели автоматически создают лучшие биологические представления.

## **Фармацевтические компании pursue специализированные наборы данных**

Relation уже применила свой подход к генерации данных к Osteomics, который компания описывает как проприетарный функциональный одноклеточный атлас костей. Проект использует полученные от пациентов образцы и сочетает одноклеточную и пространственную омику с визуализацией, геномикой, протеомикой и данными клинических фенотипов.

По данным компании, Osteomics используется для исследования биологии заболеваний, терапевтических мишеней, биомаркеров и подгрупп пациентов при остеопорозе. Больницы и исследовательские партнёры в Великобритании и Австралии участвуют в обсервационном исследовании.

Исследование, опубликованное в Nature Genetics в прошлом месяце, также изучало клеточные и генетические детерминанты скелетных заболеваний с использованием одноклеточного анализа, генетических данных и функциональной валидации. Несколько исследователей Relation были среди авторов исследования.

Анализ Nature Biotechnology 2025 года сделок в области биофармацевтики, ориентированных на ИИ, идентифицировал провайдеров специализированных наборов данных как одну из нескольких тенденций, возникающих из недавних партнёрств. Другие тенденции включали более крупные авансовые платежи, новые терапевтические модальности и большее участие крупных биотехнологических компаний.

Анализ показал, что высококачественные, специфичные для заболеваний наборы данных становятся важным входным материалом для причинных и генеративных моделей машинного обучения. В нём приводилось отдельное соглашение GSK с Ochre Bio на сумму \$37,5 миллиона за лицензирование данных, включающее одноклеточные данные человеческой печени и данные перфузированных органов.

Другой пример включал соглашение AstraZeneca и Pathos AI с Tempus на сумму \$200 миллионов в 2025 году. Согласно договорённости, Pathos должна была разработать фундаментальные модели онкологии с использованием обезличенных клинических, геномных данных и данных визуализации, охватывающих более 150 000 пациентов.

Доступ к достаточным высококачественным данным остаётся ограничением в открытии лекарств с помощью ИИ. Исследовательский обзор Nature о федеративном обучении в фармацевтических исследованиях идентифицировал ограниченный доступ к подходящим обучающим данным как основное узкое место для приложений ИИ, отметив при этом, что компании также могут сталкиваться с ограничениями на обмен проприетарной информацией.

Соглашения между ИИ и биофармацевтическими компаниями поэтому различаются по тому, как компании получают данные и вычислительные возможности. Некоторые сосредоточены на доступе к платформам ИИ, в то время как другие охватывают совместную разработку, лицензирование данных или создание новых биологических наборов данных.

Соглашение GSK-Relation включает как генерацию данных, так и разработку моделей. Relation будет производить наборы данных человеческих клеток в рамках сотрудничества и использовать их для обучения моделей ИИ для идентификации потенциальных мишеней для лекарств.

(Фото CDC)