ОБЗОРЫ

УДК: 004.8:615.012 DOI: https://doi.org/10.52540/2074-9457.2026.1.72

А. С. Максимов[1], Ж. М. Дергачёва[2]

ПРИМЕНЕНИЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ В РАННЕЙ РАЗРАБОТКЕ ЛЕКАРСТВЕННЫХ СРЕДСТВ

[1]ООО «Искамед», г. Минск, Республика Беларусь
[2]Витебский государственный ордена Дружбы народов медицинский университет, г. Витебск, Республика Беларусь

В статье представлен систематический обзор применения методов машинного обучения на раннем этапе разработки лекарственных средств. Рассмотрены три ключевые стадии: поиск соединений-хитов, установление механизма действия и трансляционные исследования. Показано, что графовые нейронные сети (D-MPNN) и гибридные подходы (Deep Docking) позволяют значительно повысить эффективность виртуального скрининга, а генеративные модели – вариационные автоэнкодеры, нормализующие потоки, диффузионные модели и трансформеры – обеспечивают целенаправленный дизайн новых молекул с заданными свойствами, выходя за пределы существующих химических библиотек. В области предсказания структуры белков проанализирована эволюция от AlphaFold2 и RoseTTAFold к языковым моделям белка (OmegaFold, ESMFold), которые позволяют преодолевать ограничения, связанные с отсутствием эволюционной информации. Рассмотрены диффузионные подходы к молекулярному докингу (DiffDock) и дизайну белков (RFdiffusion), демонстрирующие переход от эвристических методов к непрерывному генеративному моделированию. Отдельное внимание уделено трансферному обучению на примере Geneformer для анализа одноклеточных транскриптомов, а также интеграции прогностических моделей ADMET (ADMET-AI) для многопараметрической оптимизации кандидатов. Сформулированы ключевые вызовы: интерпретируемость моделей, экспериментальная валидация in silico предсказаний, регуляторное признание вычислительных подходов и необходимость развития открытых репозиториев данных высокого качества. Сделан вывод, что методы машинного обучения не заменяют, а дополняют традиционные экспериментальные подходы, формируя основу для нового междисциплинарного направления на стыке химической биологии и data science.

Ключевые слова: машинное обучение, ранняя разработка лекарственных средств, виртуальный скрининг, генеративные модели, графовые нейронные сети, предсказание структуры белков, AlphaFold, языковые модели белка, молекулярный докинг, диффузионные модели, ADMET-прогнозирование, токсичность, трансферное обучение, интерпретируемость моделей.

ВВЕДЕНИЕ

Процесс вывода нового лекарственного препарата (ЛП) на рынок сопряжен с высокими временными и финансовыми затратами: от оригинальной идеи до регистрации проходит, по различным оценкам, порядка десятилетия, а совокупные инвестиции в разработку одного успешно зарегистрированного препарата, по доступным оценкам, исчисляются миллиардами долларов США. При этом значительная доля кандидатов отсеивается на различных этапах, что делает эту область одной из самых рискованных в сфере биомедицинских исследований [1].

Машинное обучение (МО) позволяет оптимизировать процесс открытия ЛП. Благодаря растущему числу общедоступных и частных крупномасштабных биологических и химических наборов данных методы МО становятся полезным инструментом, способным дополнить традиционный процесс разработки лекарственного средства (ЛС). В этой статье мы обсуждаем интеграцию МО в раннюю

разработку ЛС. В частности, рассматриваем ряд подходов, направленных на ускорение первоначального обнаружения соединений-хитов, выяснение механизма действия (МД) и оптимизацию химических свойств.

Традиционный экспериментальный конвейер в ранней разработке включает несколько последовательных этапов [2, 3]. После идентификации и валидации биологической мишени активные вещества обнаруживают, как правило, с помощью высокопроизводительного скрининга (high-throughput screening, HTS) библиотек соединений. Соединения-хиты, выявленные в ходе такого скрининга, в дальнейшем исследуют для установления МД, изучения зависимости «структура-активность» и оценки эффективности *in vivo*. Однако этот традиционный подход характеризуется высоким уровнем неудач: примерно 90% соединений-кандидатов отсеиваются на пути от I фазы клинической разработки до регистрации, причем этот показатель еще выше, если учитывать доклиническую разработку [4].

Анализ результатов клинической разработки позволяет предположить, что высокий процент неудач даже на поздних фазах обусловлен отсутствием коммерческой потребности и просчетами в стратегическом планировании, неудовлетворительными лекарственноподобными свойствами, неприемлемой токсичностью и отсутствием клинической эффективности [5, 6]. Таким образом, существует острая потребность в повышении эффективности и снижении стоимости процессов изыскания и разработки ЛС.

За последние два десятилетия произошло значительное распространение методов искусственного интеллекта (ИИ) для решения актуальных практических задач [7, 8]. Это стало возможным благодаря появлению и развитию широкого спектра технологий МО, адаптируемых для работы с различными типами данных. Ключевая особенность этих технологий – использование математических методов, позволяющих алгоритмам самостоятельно выявлять закономерности в обучающих выборках и применять их для анализа новых данных без необходимости жесткого программирования логики. Результативность таких подходов закономерно повышается с ростом качества и объема обучающих данных. Следовательно, методы и инструменты МО идеально подходят для решения задач, связанных с большими наборами данных, в которых множество переменных связаны неочевидными и сложными взаимосвязями. Это типично для процессов изыскания и разработки ЛС.

Благодаря накоплению сложных химических [9–11] и биологических [12–14] наборов данных методы и инструменты МО получают уникальную возможность максимально использовать эти массивы для более полной автоматизации процессов изыскания и разработки ЛС. Особенно это актуально в ранней разработке, где на данный момент обучающих данных в открытом доступе больше всего. Научные учреждения и фармацевтические компании все чаще внедряют технологии МО, которые применяются на различных этапах разработки ЛС [15–18]. Эти организации также вложили значительные ресурсы в создание наборов данных, поддерживающих подходы к разработке ЛС, интегрированные с МО [19].

В этой статье мы рассматриваем применение различных технологий МО в ранней разработке низкомолекулярных лекарственных веществ. Для удобства и систематичности изложения выделены три стадии:

– стадия 1: поиск соединений-хитов;

– стадия 2: определение механизма действия;

– стадия 3: трансляционные исследования.

Цель работы – продемонстрировать возрастающую роль МО в разработке ЛС.

Сравнение с существующими обзорами. В англоязычной литературе опубликован ряд обзоров, посвященных применению методов МО в разработке ЛС. Настоящий обзор отличается от ранее опубликованных работ тем, что предлагает русскоязычный систематический анализ всего конвейера ранней разработки ЛС – от поиска соединений-хитов до трансляционных исследований – с акцентом на сравнительный анализ ключевых архитектур и оценку их экспериментальной валидации *in vitro/in vivo*. Кроме того, в отличие от большинства существующих обзоров, наша работа включает критический анализ ограничений рассматриваемых методов и конкретные рекомендации по их интерпретируемости.

МАТЕРИАЛЫ И МЕТОДЫ

Настоящая работа представляет собой систематический обзор, целью которого является анализ и обобщение современных подходов к применению методов МО в ранней разработке ЛС.

Поиск литературы проводился в электронных базах данных PubMed, Scopus, Web of Science и Google Scholar за период с 2015 по 2024 год. Использовались следующие поисковые запросы и их комбинации: «machine learning», «deep learning», «drug discovery», «preclinical», «virtual screening», «generative models», «protein structure prediction», «AlphaFold», «molecular docking», «ADMET prediction», «toxicity prediction», «transfer learning», «single-cell transcriptomics».

Отобранные публикации классифицировались по трем ключевым стадиям ранней разработки (изыскания) лекарственных веществ:

1. Поиск соединений-хитов – включал работы по виртуальному скринингу, генеративному дизайну молекул и представлению молекул в виде признаков.

2. Установление МД – включал работы по предсказанию структуры белков, молекулярному докингу, дизайну белков и анализу транскриптомных данных.

3. Трансляционные исследования – включал работы по прогнозированию ADMET-свойств, растворимости, токсичности и клинической безопасности.

Критерии включения и исключения. В работу включались оригинальные исследовательские статьи и обзоры, опубликованные на английском или русском языке в период с 2015 по 2024 год, в которых описывалось применение методов машинного или глубокого обучения на любой из трех стадий ранней разработки ЛС. Критериями исключения служили: статьи, не содержащие экспериментальной валидации *in silico* предсказаний (для работ по виртуальному скринингу и генерации молекул), публикации только в виде препринтов (bioRxiv, ChemRxiv, arXiv) без последующего рецензирования, а также статьи, в которых МО применялось исключительно на поздних стадиях клинических испытаний.

Процедура отбора. Первичный поиск в базах данных PubMed, Scopus, Web of Science и Google Scholar по указанным поисковым запросам выявил 1 247 публикаций. После удаления дубликатов (n = 312) осталось 935 статей. Скрининг по заголовкам и аннотациям позволил исключить 712 публикаций, не соответствующих тематике обзора. Полнотекстовый анализ оставшихся 223 статей привел к исключению еще 106 работ по критериям, указанным выше. В итоговый обзор включено 117 публикаций.

Оценка риска систематических ошибок. В настоящем обзоре систематическая оценка риска предвзятости для каждой включенной публикации не проводилась, поскольку данная процедура стандартно применяется для мета-анализов количественных данных, тогда как наш обзор носит описательный характер. Однако, для минимизации риска публикационной предвзятости, поиск не ограничивался только журналами с высоким импакт-фактором, а также включался анализ литературы на русском языке.

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Стадия 1. Поиск соединений-хитов

Виртуальный скрининг

HTS, в ходе которого тестируются большие наборы соединений на белках-мишенях или культурах клеток, является одним из ключевых направлений в изыскании ЛС. Однако он требует значительных финансовых и временных затрат, зачастую занимая недели или месяцы для скрининга 10^5–10^6 соединений, при этом лишь немногие из них обладают терапевтическим потенциалом.

Виртуальный скрининг с применением технологий МО предлагает многообещающую альтернативу, используя модели МО для быстрой оценки соединений *in silico*. Это позволяет эффективно исследовать химические библиотеки, на несколько порядков превышающие по объему массивы, доступные для HTS.

Представление молекул в виде признаков. Появление моделей количественного анализа взаимосвязи «структура-активность» (QSAR), использующих методы глубокого обучения (ГО), позволило усовершенствовать виртуальный скрининг крупных химических библиотек. ГО оказало существенное влияние на формирование признакового описания молекул [16].

Традиционно молекулы подавались на вход алгоритмам МО в виде фиксиро-

ванных представлений, таких как векторы отпечатков (fingerprint vectors), которые указывают на наличие или отсутствие молекулярных субструктур [16, 20, 21]. Однако такие представления зачастую не способны уловить контекстуальные детали молекулярных связей и не могут быть адаптированы в процессе обучения модели для более точного решения конкретной прогностической задачи.

Графовые нейронные сети. Для получения контекстно- и задача-ориентированных молекулярных эмбеддингов в ряде работ была применена направленная графовая нейронная сеть с передачей сообщений (directed message passing neural network, D-MPNN) для поиска новых антибиотиков против Escherichia coli и Acinetobacter baumannii [16, 22].

Работа D-MPNN состоит из двух этапов: инициализации, в ходе которой каждому атому и каждой связи ставится в соответствие исходный вектор признаков, и передачи сообщений, при которой обновляются только векторы связей (ребер). Каждое обновление вектора направленной связи происходит на основе векторов других связей, входящих в тот же атом-источник. Серия таких итераций позволяет информации распространяться по графу, минуя прямое обновление состояний атомов [16, 22, 23]. Ключевое отличие D-MPNN от классической MPNN – использование направленных реберных сообщений: сообщение по ребру обновляется без учета обратного ребра, что подавляет циклический возврат информации в исходный атом и снижает избыточность передачи сообщений. Стереохимические и иные структурно-чувствительные особенности учитываются за счет расширенного набора атомных и реберных дескрипторов, подаваемых модели на вход. Такая схема снижает риск переобучения и повышает чувствительность модели к тонким структурным различиям – в частности, к изомерии и стереохимической конфигурации, что имеет первостепенное значение при работе с потенциальными лекарственными веществами.

В исследованиях [16, 22] модели обучались на бинаризованных данных по ингибированию роста целевых патогенов. Затем обученные модели использовались для прогнозирования на основе библиотеки Drug Repurposing Hub [24], содержащей соединения с благоприятными профилями токсичности и фармакокинетики. Это привело к открытию галицина (halicin), активного против широкого спектра грамотрицательных и грамположительных патогенов с минимальной подавляющей концентрацией (МПК) на тестовых штаммах в диапазоне единиц мкг/мл [16], и абауцина (abaucin) – соединения с выраженно узким спектром действия преимущественно в отношении A. baumannii, также с МПК порядка единиц мкг/мл [22]. Оба исследования демонстрируют, что приоритизация молекул с наивысшими прогнозными оценками может значительно повысить частоту обнаружения активных соединений *in vitro* по сравнению с традиционными подходами HTS.

В недавней работе архитектура D-MPNN также была использована для изыскания новых антибактериальных молекул с применением алгоритма поиска по дереву для идентификации рационалов (rationales) – ключевых субструктур, определяющих прогнозируемую антибактериальную активность. Под рационалом здесь понимается минимальная субструктура молекулы, достаточная для того, чтобы обученная модель классифицировала соединение как активное. Два соединения, содержащие общую субструктуру, проявляли активность в отношении Staphylococcus aureus. Однако использование модели, обученной предсказывать активность целых молекул, для идентификации субструктур может приводить к неточностям при выявлении более мелких фрагментов. В данном исследовании идентифицированные рационалы составляют практически всю молекулу [25]. Кроме того, вопрос о том, можно ли считать такой подход интерпретируемым, остается дискуссионным. Несмотря на то что он дает некоторое понимание предсказаний модели, по-прежнему неясно, почему D-MPNN идентифицирует именно тот или иной структурный рационал.

Таким образом, применение графовых нейросетей позволяет перейти от пассивного скрининга существующих библиотек к более целенаправленному анализу. Ключевым вызовом остается баланс между интерпретируемостью таких моделей и их способностью выявлять нетривиальные структурные мотивы, которые могут быть неочевидны для исследователя.

Виртуальный скрининг на основе структуры

В отличие от подходов, опирающихся на данные фенотипического скрининга в качестве обучающей выборки, молекулярный докинг позволяет проводить виртуальный скрининг соединений на основе их прогнозируемой аффинности к белку-мишени. Однако при работе со сверхбольшими химическими библиотеками использование алгоритмов докинга, предполагающих стыковку каждой молекулы с интересующей мишенью, требует колоссальных вычислительных затрат и значительных аппаратных ресурсов (использования множества ядер центральных процессоров (CPU) и графических ускорителей (GPU)).

В связи с этим был разработан метод Deep Docking, позволяющий повысить эффективность виртуального скрининга на основе докинга для химических библиотек, насчитывающих более миллиарда молекул. Данный подход интегрирует традиционные методы докинга с технологиями ГО, что значительно ускоряет процесс виртуального скрининга. На первом этапе отбирается небольшая подвыборка молекул из виртуальной библиотеки (приблизительно 1%) и проводится их докинг с использованием обычных методов. Полученные в результате оценки докинга используются для обучения нейронной сети прямого распространения (feedforward neural network, FFNN), которая в дальнейшем позволяет быстро прогнозировать оценки для остальных 99% соединений [26–30].

Показательным примером служит применение Deep Docking для скрининга 1,3 миллиарда соединений из базы данных ZINC15 в отношении главной протеазы (Mpro) коронавируса SARS-CoV-2. Последующая экспериментальная валидация позволила идентифицировать новый класс соединений, активных в отношении Mpro [28–32]. Ограничением подхода является то, что метки, применяемые для обучения модели, генерируются с помощью алгоритмов *in silico* докинга, которые характеризуются вариабельностью в точности в зависимости от структуры лиганда и белка, а также от параметризации программного обеспечения [27, 28]. Отсутствие истинных меток (ground truth) закономерно вносит погрешность в последующие прогнозы, формируемые обученной моделью.

Генерация молекул

Помимо виртуального скрининга химических библиотек *in silico*, область поиска лекарственных веществ может быть значительно расширена за счет применения генеративных подходов на основе ГО. Верхняя граница объема химических библиотек, поддающихся разумному виртуальному скринингу, достигает порядка миллиардов соединений (10^9), однако это число ничтожно по сравнению с химическим пространством лекарственноподобных молекул, верхние оценки которого достигают порядка 10^{60}, тогда как более консервативные, основанные на ограничениях по числу тяжелых атомов и синтетической доступности, дают существенно меньшие, но все еще астрономические значения [33]. Таким образом, генеративные модели открывают доступ к неизведанным областям химического пространства, пока не охваченным существующими химическими библиотеками с открытым доступом.

Вариационные автоэнкодеры

Вариационные автоэнкодеры (ВА) широко применяются в задачах молекулярного дизайна, начиная с публикации основополагающего исследования 2018 года [34]. Архитектура ВА включает две нейронные сети: энкодер и декодер. В ходе обучения энкодер преобразует молекулу в параметры (среднее и дисперсию) апостериорного вероятностного распределения в непрерывном латентном пространстве. Для обеспечения гладкости пространства и возможности управляемой генерации в процессе обучения используется регуляризация, приближающая это распределение к стандартному нормальному. На этапе генерации из латентного пространства (часто из стандартного нормального распределения) семплируется вектор, который декодер преобразует в новую молекулу [35]. Критически важное свойство ВА – возможность управления генерацией: выбирая латентные векторы, соответствующие определенным областям пространства, можно систематически повышать вероятность создания новых молекул с заданными свойствами [34].

Ранние работы с применением ВА фокусировались на декодировании молекул в формате строк SMILES (simplified molecular-input line-entry system). В 2018 году модель JT-VAE (junction tree VAE) усовершенствовала существующие под-

ходы, перейдя к работе непосредственно с графами молекулярных структур. Авторы метода обосновали это тем, что оперирование графовыми структурами напрямую улучшает способность модели выявлять структурные взаимосвязи как внутри одной молекулы, так и между различными молекулами. Для представления молекул исходные графы были преобразованы в деревья соединений – иерархические структуры, ветви которых объединяют отдельные структурные блоки (например, связи и циклы). Сформированные таким образом деревья параллельно с исходными графами подвергались процедурам кодирования и декодирования. Предложенный подход обеспечил 100%-ную химическую валидность генерируемых молекулярных структур – закономерное следствие сборки молекулы из заранее извлеченного словаря валидных фрагментов – и превзошел предшествующие BA-модели как в задачах неограниченной, так и ограниченной молекулярной оптимизации [36].

Модель JAEGER, построенная на архитектуре JT-VAE, была использована для изыскания новых противомалярийных соединений. Разработчики интегрировали в JAEGER блок предсказания свойств, что позволило осуществлять совместное обучение как задачам кодирования/декодирования молекул, так и оптимизации противомалярийной активности (отрицательный десятичный логарифм полумаксимальной ингибирующей концентрации, pIC50, в тесте на пролиферацию малярийного плазмодия). В ходе семплирования модель итеративно исследовала латентное пространство в поиске новых молекул, отбирая только те соединения, которые превышали установленный порог прогнозируемой активности. В результате были идентифицированы две молекулы, проявляющие наномолярную активность в отношении штамма Plasmodium falciparum 3D7 и характеризующиеся низкой гепатоцеллюлярной цитотоксичностью [37].

В другом исследовании генеративный подход на основе BA позволил предложить ингибиторы рецептора дискоидинового домена 1 (DDR1) (перспективной мишени для противоопухолевой терапии) с наномолярной активностью *in vitro*. Вместе с тем в литературе указывалось, что наиболее активный хит по структурным и фармакологическим характеристикам близок к уже известному ингибитору киназ; это стало поводом для содержательной дискуссии о том, насколько такие генеративные модели действительно способны предлагать структурно новые соединения, а не варианты в окрестности уже известных структур [38].

Модели нормализующих потоков и диффузионные модели

Применение моделей нормализующих потоков (normalizing flows) и диффузионных моделей (diffusion models) в разработке ЛС приобретает все большее распространение, что обусловлено их успехами в смежных областях, включая генерацию изображений и текста.

Подобно BA, модели нормализующих потоков осуществляют преобразование сложного распределения (то есть распределения наблюдаемых молекул) в более простые и обратно [35, 39]. Ключевое различие заключается в том, что они используют обратимые преобразования между простым (латентное пространство) и сложным (пространство наблюдений) распределениями, что позволяет точно вычислять правдоподобие сгенерированных образцов для оценки их качества, то есть количественно оценивать, насколько вероятно, что сгенерированная молекула соответствует «правилам» химической валидности и распределению свойств, характерному для обучающей выборки [39, 40]. Напротив, BA оперируют приближенным правдоподобием, что может приводить к генерации химически некорректных структур [41].

В качестве примера можно привести модель GraphAF – генеративную архитектуру на основе нормализующих потоков, которая создает новые молекулы авторегрессионным способом: на каждом временном шаге она добавляет новые узлы и соединяет их ребрами, опираясь на структуру текущего подграфа. GraphAF также интегрирует обучение с подкреплением в авторегрессионный процесс, назначая «вознаграждение» за каждый шаг в зависимости от заданных целевых свойств – например, коэффициента распределения октанол/вода (logP) и количественной оценки лекарственноподобности (QED). По способности генерировать уникальные и валидные молекулы GraphAF демонстрирует результаты, сопоставимые с JT-VAE, SMILES-ориентированным генера-

тором на рекуррентных нейронных сетях (RNN), а также графовыми методами обучения с подкреплением; при этом GraphAF превосходит перечисленные модели в оптимизации logP и QED [42].

Диффузионные модели работают иначе. На первом этапе – прямом процессе диффузии (forward diffusion) – к образцу итеративно добавляется гауссов шум фиксированным, необучаемым образом. Затем модель обучается пошагово удалять шум (обратный процесс диффузии, reverse diffusion), последовательно восстанавливая исходную структуру [39]. Такой итеративный характер удаления шума дает диффузионным моделям преимущество перед ВА и генеративно-состязательными сетями (GAN), которые преобразуют латентное пространство в выходной образец за один шаг. Благодаря этому диффузионные модели способны избегать грубых ошибок и постепенно уточнять молекулярную структуру, что особенно важно при работе с трехмерными координатами атомов.

В 2022 году была предложена эквивариантная диффузионная модель (модель, предсказания которой согласованы с поворотами и сдвигами молекул в трехмерном пространстве) для генерации молекул в 3D пространстве. В этой работе молекулы описывались набором характеристик атомов и их координат. В процессе диффузии шум последовательно добавлялся, а затем удалялся: сначала молекула подвергалась «разрушению» (зашумлению), после чего восстанавливалась до исходной структуры. Критически важной особенностью является возможность генерации с учетом целевых свойств в процессе обратной диффузии, что позволяет оптимизировать генерируемые молекулы по заданным показателям. Авторы подчеркивают, что модель способна учитывать информацию о физико-химических свойствах при создании 3D-структур молекул, удовлетворяющих целевым значениям, таким как поляризуемость. Предложенная модель превосходит предшествующие ВА- и flow-архитектуры по доле как валидных, так и уникальных сгенерированных молекул. Вместе с тем при тестировании на выборке более крупных, лекарственноподобных соединений модель демонстрирует трудности в генерации стабильных молекул, что свидетельствует об ограничении ее применения в задачах изыскания новых ЛС [43].

Таким образом, эволюция методов генерации молекул идет по пути увеличения структурной валидности (от SMILES к графам и 3D-координатам) и внедрения механизмов управления свойствами, что постепенно превращает эти инструменты из «генераторов случайных структур» в полноценные платформы для целенаправленного дизайна. Диффузионные модели, несмотря на вычислительную сложность, открывают новые горизонты в создании трехмерных структур с заданными физико-химическими характеристиками.

Модели химического языка

Текстовые представления молекул, такие как SMILES, открыли возможности применения моделей химического языка (chemical language models, CLM) для генеративного дизайна молекул *de novo* [44]. Как правило, CLM используют рекуррентные нейронные сети (RNN), которые обрабатывают последовательности поэлементно, предсказывая следующий элемент на основе предыдущих.

Обучение CLM проводится самообучаемым способом (self-supervised learning): модель обучается предсказывать следующий токен в SMILES-строке на основе предыдущих, что не требует ручной разметки данных – сами последовательности служат и входом, и эталоном [45]. Цель обучения – выявить и усвоить общие паттерны токенов, характерные для обучающих выборок. После завершения обучения модель способна авторегрессионно генерировать новые соединения, распределение физико-химических свойств которых близко к наблюдаемому в обучающем наборе [46, 47].

Традиционным ограничением CLM являлась необходимость в больших обучающих выборках – от сотен тысяч до миллионов соединений – для надежного освоения синтаксиса представления молекулярных структур и соответствующих физико-химических свойств. Учитывая ограниченность данных по многим терапевтическим направлениям, это существенно сужало область применения CLM в изыскании ЛС. Для преодоления данного ограничения в задаче изыскания новых лекарственных веществ было применено трансферное обучение (transfer learning) – подход, при котором знания, полученные моделью при решении одной задачи, используются для повышения эффективности на другой, родственной задаче, осо-

бенно при дефиците данных. В этом подходе модель предварительно обучается на родственной задаче, для которой доступны большие объемы данных (например, библиотеки лекарственноподобных соединений). Затем последние слои дообучаются (fine-tuning) на данных, специфичных для конкретной задачи, с сохранением («замораживанием») ранних слоёв или с использованием низкой скорости обучения, что позволяет сохранить общие представления и адаптировать модель под специфику целевых соединений [48–52].

Данный подход был применен для генерации новых лигандов PI3Kγ – мишени для потенциальных противоопухолевых, противовоспалительных и иммуномодулирующих соединений. Авторы использовали нейронную сеть с долгой краткосрочной памятью (LSTM), разновидность RNN, обучив ее на приблизительно 850 000 строках SMILES-соединений из базы данных патентов США, после чего дообучили на 46 известных ингибиторах PI3Kγ. Сгенерировав более 1 миллиона новых SMILES-строк, исследователи применили классификатор для ранжирования синтезируемых соединений по приоритету. Из всего пула для синтеза были отобраны два соединения, предсказанные классификатором с высокой достоверностью, а также четыре их структурных аналога. Все шесть соединений продемонстрировали наномолярную активность в отношении PI3Kγ *in vitro* [50].

Значительный прогресс в области обработки естественного языка наступил после появления архитектуры трансформер (transformer). Трансформерные модели обладают рядом преимуществ перед RNN-архитектурами: они допускают параллельную обработку данных, что ускоряет обучение; способны улавливать долгосрочные зависимости в последовательностях; а также обеспечивают определенную степень интерпретируемости благодаря механизмам внимания (attention layers) [53].

Модель MolGPT, основанная на архитектуре генеративного предобученного трансформера (GPT), стала первой, использующей этот тип моделей в задачах моделирования последовательностей [54]. В сравнении с другими генеративными моделями MolGPT демонстрирует конкурентоспособные результаты по ряду стандартных метрик – уникальности, валидности и новизне генерируемых структур [54–56]. Более того, для генерации молекул MolGPT принимает на вход вектор условий – например, желаемое значение clogP или описание структурного каркаса (scaffold), наряду с начальным токеном (start token). Этот вектор встраивается в последовательность и направляет модель, заставляя ее генерировать SMILES, удовлетворяющие заданным критериям [53]. Благодаря этому MolGPT успешно решает задачи одно- и многопараметрической оптимизации, а также генерации с фиксированным каркасом. Важно отметить, что MolGPT обеспечивает интерпретируемость, свойственную трансформерам, посредством построения карт значимости (saliency maps). Карты значимости вычисляются как градиенты предсказательной функции модели по входным эмбеддингам токенов SMILES: более высокое абсолютное значение градиента указывает на большую важность соответствующего токена для предсказания [54]. Такие карты позволяют исследователям проверять, опирается ли модель на химически осмысленные фрагменты, а не на артефакты данных. Данная интерпретируемость является критически важной для формирования доверия к способности модели выделять химически и биологически релевантные признаки.

Большие языковые модели (large language models, LLM) на основе трансформеров также находят все более широкое применение в генеративной химии [57, 58]. Так, было продемонстрировано, что LLM на основе GPT-3 способна с помощью простых текстовых запросов, содержащих требование сгенерировать соединение с определенным свойством, создавать структурно новые молекулы, удовлетворяющие этому свойству [58]. Ряд последующих работ подтвердил потенциал трансформеров и LLM в данной области. Тем не менее применение этих методов находится на ранней стадии, и на сегодняшний день отсутствуют исследования, которые привели бы к формированию конкретных кандидатов с экспериментальным подтверждением *in vitro/in vivo* [59–61].

Таким образом, переход от рекуррентных архитектур к трансформерам и LLM знаменует сдвиг от простой генерации валидных SMILES-строк к семантически управляемому дизайну молекул. Однако

ключевым ограничением остается отсутствие экспериментальной валидации для молекул, предложенных современными LLM, что подчеркивает разрыв между вычислительными возможностями и их практической реализацией в лаборатории.

Несмотря на впечатляющие возможности LLM (GPT-3 и аналоги) в задаче генерации молекул по текстовому описанию, их применение сопряжено с рядом специфических рисков.

Галлюцинации структур. LLM могут генерировать химически невалидные молекулы – структуры, нарушающие правила валентности, содержащие нестабильные фрагменты или не существующие в реальности.

Проблемы воспроизводимости. Генерация молекул LLM чувствительна к начальному seed-значению и параметрам семплирования. Разные запуски одной и той же модели могут давать совершенно разные наборы соединений, что затрудняет воспроизводимость результатов и создает сложности для регуляторного признания.

Трудности синтеза. LLM, обученные главным образом на SMILES-строках из патентов и научных статей, не учитывают синтезируемость генерируемых молекул. Многие соединения, предсказанные LLM, могут быть чрезвычайно сложны или даже невозможны для синтеза в лабораторных условиях.

Сравнение с компактными моделями в условиях малых данных. Вопреки распространенному мнению, в условиях ограниченного количества данных (например, менее 100 активных соединений) компактные модели, такие как CharRNN или JT-VAE, могут демонстрировать сопоставимую или даже более высокую эффективность, чем LLM, при меньших вычислительных затратах. Исследования показывают, что в low-data regime ключевую роль играет не архитектура модели, а качество и объем обучающих данных, а также методы аугментации (например, использование неканонических SMILES-строк) [58].

Стадия 2. Установление механизма действия

После идентификации соединений-хитов необходимо детальное понимание их биологической функции. Установление МД требует комплексного применения независимых (ортогональных) методов, включая генетические подходы (например, химико-генетический скрининг в масштабе всего генома) и прямые биохимические методы [2]. Подобные лабораторные методы, как правило, отличаются низкой пропускной способностью и высокой ресурсоемкостью.

В качестве альтернативы интегрирование методов ГО с существующими масштабными базами данных, такими как Банк данных белковых структур (Protein Data Bank, PDB), может способствовать определению функций новых биологически активных молекул. Такой подход позволит оптимизировать процесс установления МД и повысить эффективность молекулярной оптимизации, основанной на структуре соединения.

Предсказание пространственной структуры белка

AlphaFold2. AlphaFold2 (AF2) ознаменовал прорыв в предсказании трехмерных структур белков по их первичным аминокислотным последовательностям. В рамках 14-го конкурса Critical Assessment of Structure Prediction (CASP) модель AF2 от DeepMind продемонстрировала наивысшую точность предсказания. Хотя и классическое гомологическое моделирование, и AF2 используют множественное выравнивание последовательностей (multiple sequence alignment, MSA), AF2 уникальным образом использует информацию о корреляции остатков (парных замен) в эволюционном ряду из последовательностей, гомологичных целевому белку.

Архитектура AF2 включает несколько ключевых компонентов. На вход подается аминокислотная последовательность целевого белка. Модель генерирует MSA и парные представления (pair representations) – матрицу попарных расстояний между аминокислотными остатками в известных 3D-структурах гомологов. Эти представления поступают на вход Evoformer – двухканальной нейронной сети архитектуры трансформера, которая совместно обрабатывает пространственные и эволюционные взаимосвязи между остатками [62]. Структурный модуль (structure module) генерирует предсказанную 3D-структуру, используя системы координат остова (backbone frames), а также обновленные MSA- и парные представления.

Надежность AF2 подтверждается его способностью строить точные 3D-модели даже для примеров, недостаточно пред-

ставленных в обучающей выборке, например, белков, которые избирательно сворачиваются в присутствии неизвестной молекулы, и белков со сложной топологией (переплетенные гомомеры). Ограничением модели является то, что она плохо справляется с предсказанием структур белков, имеющих мало гомотипических или внутрицепочечных контактов по сравнению с количеством гетеротипических взаимодействий. Как правило, это происходит в крупных белковых комплексах с мостиковыми доменами, форма которых определяется взаимодействиями с другими цепями в комплексе.

AF2 уже продемонстрировал свою пользу для сообщества структурных биологов [63, 64] и в разработке ЛС [65–67]. Например, в работе с использованием PandaOmics – модели ГО на основе омиксных данных – был сформирован короткий список соединений-кандидатов для терапии рака печени, структуры мишеней для которых могли быть предсказаны с помощью AF2. На основе предсказанных AF2 структур с помощью платформы Chemistry42 были сгенерированы низкомолекулярные соединения для экспериментальной проверки [68]. Этот подход позволил идентифицировать соединение-хит с афинностью к циклинзависимой киназе 20 (CDK20), часто гиперэкспрессируемой во многих типах опухолей, на уровне Kd 8,9 ± 1,6 мкМ, что рассматривается авторами как первичное подтверждение концепции, после которого проводится дальнейшая структурная оптимизация [65]. Данная работа подтверждает потенциальную пользу AF2, что оправдывает дальнейшее изучение его применения для дизайна ЛС. Однако, учитывая известные ограничения, существуют случаи, когда модели AF2 будут недостаточно точными для использования в разработке ЛС.

RoseTTAFold. Вскоре после выхода AF2 была представлена модель RoseTTAFold [69], целью которой было повышение точности предсказания белковых структур. Эта модель расширяет архитектуру двухканальной нейронной сети [62], добавляя третий канал, который учитывает трехмерные координаты атомов Cα, C и N каждого остатка с помощью SE(3)-трансформера. Иными словами, RoseTTAFold одновременно обрабатывает информацию из одномерных (аминокислотная последовательность), двумерных (матрицы расстояний между остатками) и трехмерных (системы координат остова) представлений для генерации предсказаний структуры [69]. Модель была обучена на более чем 200 000 белках с известной структурой из базы данных PDB [70].

Чтобы проверить возможность модели давать представление о функции белков человека при патологических состояниях, исследователи предсказали структуру белков семейства ADAM, которые опосредуют клеточно-клеточные и клеточно-матриксные взаимодействия. Эти белки вовлечены в широкий спектр заболеваний человека, таких как неврологические расстройства и метастазирование рака. RoseTTAFold успешно предсказал структуру продомена ADAM33, которая ранее была неизвестна и не имела гомологов с определенной структурой. Согласно этому предсказанию, продомен обладает липокалиноподобным β-бочкообразным фолдом, за которым следует экстензионный участок, содержащий цистеин, что согласуется с экспериментальными данными, указывающими на то, что продомен ADAM ингибирует активность металлопротеиназы посредством механизма «цистеинового переключателя» [71].

Языковые модели белка

AlphaFold2 и RoseTTAFold используют обширные эволюционные данные, получаемые из MSA – ресурса, который недоступен для некоторых белков, например, быстро эволюционирующих или так называемых орфанных белков [72]. MSA для быстро эволюционирующих белков, включая комплементарно-определяющие участки (CDR) антител, как правило, содержат шум [73]. Орфанные белки, напротив, не имеют достаточной гомологии (как на уровне последовательности, так и на уровне структуры) для надежного предсказания на основе MSA [72].

Языковые модели белка (protein language models, PLM) могут преодолеть это ограничение за счет предварительного обучения моделей на основе архитектуры трансформера на белковых последовательностях. PLM, обученные на миллионах последовательностей, внутренне выучивают «грамматику» белкового пространства – те самые корреляции, которые AF2 извлекает из явного выравнивания, просто они «сжа-

ты» в весах нейросети. В отличие от AF2, такие модели не требуют MSA на этапе инференса: эволюционная информация уже имплицитно закодирована в параметрах предобученной сети, что позволяет им выявлять закономерности, отражающие эволюцию, и одновременно захватывать информацию о вторичной и третичной структурах [72, 74, 75].

Модель OmegaFold использует Omega PLM – PLM на основе трансформера, предварительно обученную на наборе данных UniRef50 для формирования представлений отдельных остатков и пар остатков (матрица попарных взаимодействий остатков размера L × L, где L-длина аминоксилотной последовательности). Эти представления подаются на вход GeoFormer – трансформера, основанного на геометрических принципах, – который выявляет согласованные структурные и физические взаимосвязи между аминокислотными остатками при проецировании в трехмерное пространство. Качество предсказаний модели оценивали по ее способности предсказывать структуры недавно опубликованных в PDB антител и орфанных белков. Как для орфанных белков, так и для CDR-участков антител OmegaFold продемонстрировала значительно более высокую точность предсказания, чем AF2 [73].

Аналогичным образом, ESMFold [76] использует ESM-2 – языковую модель белков на основе трансформера-кодировщика, обученную приблизительно на 65 миллионах последовательностей из UniRef90 с семплированием по кластерам UniRef50 [77]. Эмбеддинги отдельных остатков (полученные при маскировании случайных позиций в последовательности, аналогично BERT) и пар остатков (матрица попарных расстояний между аминокислотами), генерируемые ESM-2, служат входными данными для фолдинг-блока [76]. Этот блок представляет собой упрощенную адаптацию Evoformer [62], в которой устранена зависимость от MSA для получения обновленных представлений, подаваемых в финальный структурный модуль, предсказывающий трехмерную структуру белка. Хотя ESMFold не достигает точности AF2, модель способна генерировать точные предсказания структур для орфанных белков и работает более чем на порядок быстрее.

Таким образом, эволюция методов предсказания структуры белка идет по пути снижения зависимости от вычислительно затратных MSA: от AF2, требующего глубокого эволюционного анализа, к PLM, которые «сжимают» эволюционную информацию в веса предобученных моделей. Это открывает возможности для быстрого скрининга мишеней, для которых традиционные эволюционные подходы неприменимы (орфанные белки, гипервариабельные участки антител (например, CDR-петли антител), для которых невозможно построить качественное множественное выравнивание последовательностей (MSA)).

Белковый докинг и компьютерный дизайн

DiffDock. DiffDock – это диффузионная генеративная модель (diffusion generative model, DGM), которая рассматривает задачу молекулярного докинга как проблему генеративного моделирования. В DiffDock процесс диффузии применяется к трансляциям лиганда, его вращениям и изменениям торсионных углов.

DiffDock построен на двух специализированных нейросетевых компонентах. Первый (score model) в ходе итеративного процесса обратной диффузии последовательно уточняет положение лиганда, предсказывая градиент логарифмической плотности распределения в пространстве трансляций, вращений и торсионных углов. Второй (confidence model) оценивает достоверность каждой сгенерированной позы, позволяя ранжировать их и отбирать наиболее вероятные. Обе модели основаны на SE(3)-эквивариантных сверточных нейронных сетях, обрабатывающих точечные облака для представления координат.

Авторы обучили модель на более чем 17 000 комплексах «белок–лиганд» из базы данных PDBbind. На тестовой выборке PDBbind DiffDock семплирует для каждого комплекса ансамбль из 40 поз-кандидатов, после чего ранжирующая сеть выбирает одну top-1 позу; по доле комплексов, для которых эта top-1 поза имеет RMSD < 2. А относительно эталонной структуры, DiffDock превзошел современные поисковые, коммерческие программные пакеты, а также методы, основанные на ГО. Для приложений в области изыскания ЛС, таких как виртуальный скрининг или обратный скрининг белковых мише-

ней, критически важны высокая скорость и точность работы. Экспериментальные данные, полученные с помощью распространенных подходов к установлению МД (например, скрининг на устойчивость и аффинная хроматография), которые указывают на потенциальную белковую мишень, могут быть дополнительно подтверждены или оспорены с помощью компьютерного моделирования на основе DiffDock [78].

RFdiffusion. Диффузионные модели также недавно были адаптированы для задач дизайна белков (protein design). Развивая подход на основе диффузионных моделей и используя архитектуру RoseTTAFold, модель RoseTTAFold diffusion (RFdiffusion) [79] представляет собой генеративную модель, способную создавать функциональные белки в соответствии с заданными спецификациями. В этой модели предсказание структуры из RoseTTAFold [69] было дообучено (fine-tuned) для работы в качестве сети удаления шума (denoising network) в диффузионной модели. Если RoseTTAFold требует на вход аминокислотную последовательность, то RFdiffusion принимает на вход диффундированные координаты остова (backbone frames) [79].

Для генерации белков, удовлетворяющих определенным условиям (симметрия, мишени связывания, функциональные мотивы или симметричные функциональные мотивы), модель получает управляющую информацию (conditioning information). На основе этих входных данных модель удаляет шум из координат остова и предсказывает экспериментально реализуемую структуру белка [70]. Авторы обучили RFdiffusion на разнообразных наборах данных, включая структуры из PDB и модели, предсказанные AF2. Что наиболее важно, RFdiffusion является экспериментально подтверждённой генеративной моделью, способной создавать широкий спектр белков, удовлетворяющих различным свойствам. Экспериментальная валидация показала, что RFdiffusion может генерировать растворимые белки со сложной архитектурой высшего порядка и любой желаемой симметрией [79].

Таким образом, диффузионные подходы к докингу и дизайну белков (DiffDock, RFdiffusion) знаменуют переход от дискретных, эвристических методов к непрерывному, управляемому генеративному моделированию. Экспериментальная подтверждаемость RFdiffusion особенно важна: она показывает, что *in silico* дизайн может выходить за рамки теоретической демонстрации и приводить к физически реализуемым структурам.

Трансферное обучение в биологии: анализ транскриптомных данных

Как обсуждалось ранее, трансферное обучение является мощным методом решения задач с ограниченными данными, таких как моделирование омиксных сетей. Для точного описания реакции клеток в различных условиях (например, при патологии или в ответ на химическое воздействие) требуются обширные данные. Нехватка таких наборов данных была частично компенсирована развитием технологий, таких как массовое (bulk) и одноклеточное секвенирование РНК (scRNA-seq). Однако взаимосвязи в биологических регуляторных сетях сильно зависят от контекста, что делает трансферное обучение идеальным подходом для моделирования биологических процессов при ограниченном размере выборки (например, при малом количестве клинических образцов от пациентов).

Geneformer – это модель на основе механизма внимания (attention), состоящая из шести блоков-энкодеров трансформера, которая использует трансферное обучение для получения контекстно-зависимых предсказаний (например, чувствительность к дозе гена) на основе тканеспецифичных одноклеточных транскриптомов [80]. Модель прошла предварительное обучение (pretraining) на базе данных Genecorpus-30M, содержащей около 29,9 миллиона одноклеточных транскриптомов из широкого спектра тканей, организованных в виде таблицы со строками (признаки) и столбцами (клетки) [80]. Атрибуты признаков включают аннотации Ensembl: идентификатор гена, его название и тип (например, микроРНК, митохондриальный ген или ген, кодирующий белок). Атрибуты клеток включают уникальный идентификатор Genecorpus-30M, основной орган, конкретный орган(ы), а также рассчитанные для каждой клетки метрики (например, общее количество прочтений и доля митохондриальных прочтений).

В процессе предварительного обучения каждый одноклеточный транскриптом кодируется с помощью рангового преобразования (rank value encoding), где гены

ранжируются по уровню их экспрессии в клетке и нормализуются по их экспрессии в Genecorpus-30M. Этот метод понижает приоритет генов «домашнего хозяйства» (housekeeping genes) с конститутивно высокой экспрессией (они нормализуются к более низкому рангу), тогда как гены, оказывающие более сильное влияние на состояние клетки (например, факторы транскрипции, которые могут экспрессироваться на относительно низком уровне), получают более высокий ранг. После этого модель дообучается (fine-tuned) для решения конкретной задачи.

Авторы применили Geneformer для предсказания динамики хроматина, дозы гена и иерархии в генных сетях с целью идентификации потенциальных терапевтических мишеней. В одном из случаев для дообучения Geneformer использовали ограниченное число одноклеточных транскриптомных образцов, полученных от пациентов без сердечной недостаточности или с гипертрофической/дилатационной кардиомиопатией [81]. Для валидации работы модели авторы ингибировали мишени, предсказанные Geneformer, в модели кардиомиопатии *in vitro*, что привело к улучшению состояния сердечной ткани. В контексте установления МД Geneformer может быть эффективным инструментом для предсказания мишеней ЛС на основе одноклеточных транскриптомов, полученных в норме и при патологии.

Таким образом, применение трансформеров к одноклеточным транскриптомным данным (Geneformer) демонстрирует, что принципы, разработанные в области обработки естественного языка, успешно переносятся на биологические регуляторные сети. Ранговое кодирование экспрессии позволяет модели фокусироваться на биологически значимых изменениях, игнорируя «шум» конститутивно экспрессируемых генов.

Стадия 3. Трансляционные исследования

На поздних стадиях доклинической разработки ключевой задачей является оптимизация перспективных лекарственных веществ в более жизнеспособные кандидатные ЛС [82]. Традиционно оптимизация по схеме «хит-лид» (hit-to-lead) включает детальное изучение взаимосвязи «структура–активность», в ходе которого модифицируют ключевые фрагменты молекулы и оценивают влияние на активность [83]. Такие структурные модификации позволяют оптимизировать свойства, включая растворимость, ADMET-профиль (абсорбция, распределение, метаболизм, экскреция и токсичность), активность и селективность действия на молекулярную мишень [82]. Эти характеристики учитываются не только на начальном этапе оптимизации хитов, но и тщательно контролируются на протяжении всей доклинической разработки [83].

Технологии МО могут быть применены для прогнозирования лекарственноподобных свойств и токсичности молекул, что позволяет в ходе доклинических исследований лучше оценивать перспективы успеха кандидатных ЛП в клинических испытаниях [84–86]. Стратегия комплексной оптимизации свойств с применением МО может стать рациональным подходом, позволяющим учитывать клинически значимые параметры уже на ранних этапах разработки, а не использовать его лишь как инструмент для доработки кандидата, выбранного исключительно по критерию фармакологической активности [84]. Применение набора моделей МО для получения согласованных прогнозов в отношении желаемой фармакологической активности, токсичности и других свойств может позволить исследователям отдавать приоритет благоприятным клиническим характеристикам уже на этапе отбора хитов, даже если соответствующие соединения не обладают максимальной активностью в отношении целевой мишени.

Прогнозирование растворимости

В период с 1990-х по 2017 год доля кандидатов, потерпевших неудачу в клинических испытаниях из-за неудовлетворительных лекарственноподобных свойств, снизилась с 30–40% до 10–15% [87, 88]. Это улучшение можно связать с выявлением химических свойств, коррелирующих с благоприятным фармакокинетическим профилем, а именно с правилом Липински (правилом пяти) [87, 89].

Одно из правил Липински касается липофильности, которую можно количественно описать с помощью коэффициента logP. LogP является важным детерминантом мембранной проницаемости, растворимости, биодоступности и метаболической стабильности [89]. Экспериментальные методы его измерения (метод «взбал-

тывания в колбе» и метод медленного перемешивания) трудоемки и требуют много времени [90]. Clogp – это вычислительный метод оценки logP, при котором суммируются заданные константы для молекулярных фрагментов с введением поправок на межмолекулярные взаимодействия [91]. С 1980-х годов clogP остается наиболее часто используемым методом оценки липофильности; однако он имеет ограничения для соединений с определенными структурными особенностями и характеризуется наличием систематических ошибок для схожих молекул [92]. В контексте разработки ЛС даже небольшие систематические ошибки могут определить судьбу соединения, что подчеркивает важность точного расчета logP.

Для преодоления этих ограничений стали применяться методы МО. В исследовании [93] были обучены три архитектуры машинного обучения (SVM, MLP, XGBoost) для прогнозирования значений logP и коэффициента распределения logD молекул на основе молекулярных дескрипторов и времени удерживания при жидкостной хроматографии, которое коррелирует с logP и logD. Модели обучали на данных по 2070 соединениям из базы данных METLIN, для которых значения logP были доступны из ChEMBL [93–95]. Молекулы были представлены в виде 125-признаковых векторов RDKit, при этом время удерживания добавлено в качестве дополнительного признака. Модель MLP показала наилучшую производительность по метрике средней абсолютной ошибки (mean absolute error) [93]. При сравнении с существующими моделями прогнозирования logP, не использующими время удерживания, MLP снизила уровень ошибок на 20–30% [93, 96, 97]. Однако использование этого признака может иметь ограниченную применимость, поскольку требует получения экспериментальных данных [93].

Прогнозирование токсичности

Токсичность является причиной неудачи в клинических испытаниях для 30% кандидатных ЛП, причем в большинстве случаев это связано со связыванием с белком hERG (human ether-à-go-go-related gene) и вызванной этим сердечно-сосудистой токсичностью [85, 87, 98].

В статье 2023 года исследователи обучили ряд моделей МО для прогнозирования кардиотоксичности, опосредованной hERG. Авторы собрали набор данных для 4556 соединений со значениями IC50 для блокады канала hERG, определенные с помощью метода патч-кламп (фиксации потенциала). Данные были бинаризированы с использованием двух порогов: соединения с IC50 ниже 1 мкМ считались блокаторами hERG, а соединения с IC50 выше 10 мкМ – неблокирующими. Алгоритм k-ближайших соседей (k-NN), метод опорных векторов (SVM), случайного леса (RF) и многослойный перцептрон (MLP) были обучены на четырех различных молекулярных представлениях: отпечатках Моргана, отпечатках пар атомов, ключах MACCS и отпечатках TopoTorsion. Авторы также обучили графовую сверточную нейронную сеть, в которой признаки формируются в процессе обучения непосредственно из графового представления каждой молекулы [85]. Графовая сверточная сеть показала наилучшую производительность согласно метрике AUC-ROC [99]. Хотя такие модели могут быть полезны для оценки риска блокады hERG на ранних этапах разработки ЛС, авторы не провели экспериментальной валидации на новых соединениях [85], что оставляет открытым вопрос об их обобщающей способности.

ADMET-AI. В 2023 году была представлена платформа ADMET-AI – интерактивная платформа МО, в основе которой лежат многозадачные модели, построенные на базе библиотеки Chemprop (графовые нейронные сети с использованием дескрипторов RDKit) [100]. Авторы обучили две такие модели на 41 наборе данных ADMET из Therapeutics Data Commons (TDC): одну регрессионную модель на 10 наборах и одну классификационную на 31 наборе. ADMET-AI реализована в виде веб-интерфейса, где пользователи могут вводить SMILES-строки и получать прогнозы по растворимости, пероральной биодоступности, токсичности и безопасности в отношении hERG, среди других параметров.

ADMET-AI также сравнивает введенные молекулы с референсными препаратами из базы данных DrugBank, показывая, какой процентиль в распределении значений каждого свойства среди препаратов DrugBank занимает тестируемое соединение. Это помогает пользователям интерпретировать прогнозы модели. Такой доступный веб-интерфейс позволяет иссле-

дователям без вычислительного бэкграунда легко применять методы МО [100]. Важно отметить, что ADMET-AI представляет собой одну из наиболее доступных и эффективных многозадачных платформ для прогнозирования ADMET-свойств. Согласно лидерборду Therapeutics Data Commons (TDC), она показывает высокие результаты на широком спектре задач; однако для узкоспециализированных эндпоинтов (например, hERG, CYP-ингибирование) существуют модели (например, графовые сверточные сети), демонстрирующие более высокие метрики [100].

ADMET-AI обладает уникальными возможностями для быстрого расчета обширного набора ADMET-свойств, что необходимо для ранжирования соединений с учетом множества параметров. Однако для более целенаправленного анализа, сфокусированного на отдельных ADMET-характеристиках, целесообразно провести углубленный обзор существующих моделей, представленных в рейтинге TDC. Некоторые из этих моделей демонстрируют более высокую производительность, чем ADMET-AI, при оценке по отдельным задачам.

Прогнозирование клинической токсичности

Поскольку этические ограничения не позволяют массово собирать данные о клинической токсичности, серьезной проблемой в ее прогнозировании остается моделирование сложной взаимосвязи между данными о токсичности, полученными *in vitro, in vivo* и в клинических исследованиях. В недавней работе эта проблема была решена путем обучения полносвязных нейронных сетей (FFNN) на объединенных данных всех трех типов [86]. Авторы использовали данные *in vitro* из базы Tox21Challenge, данные *in vivo* из Реестра токсических эффектов химических веществ и данные клинической токсичности из набора ClinTox проекта MoleculeNet [86, 101, 102]. Молекулы представлялись в виде отпечатков Моргана или предобученных SMILES-векторных представлений, извлеченных из GRU-автоэнкодера, обученного переводить неканонические SMILES-строки в канонические. Используя в качестве входных данных эти латентные представления, авторы обеспечили фиксацию наиболее значимых признаков каждой молекулы.

Многозадачные прямые нейронные сети, обученные на SMILES-представлениях, показали более высокую точность по сравнению с однозадачными моделями и существующими моделями из MoleculeNet. Исследователи применили метод контрастивного объяснения (contrastive explainability) для выявления токсикофоров – наименьших структурных единиц, влияющих на классификацию, – в обучающей выборке. Многие выявленные субструктуры соответствовали известным токсикофорам, что указывает на способность модели к обучению релевантным признакам. Однако некоторые идентифицированные элементы представляли собой отдельные атомы (например, углерод или кислород), которые сами по себе нетоксичны [86].

Таким образом, интеграция прогнозных моделей ADMET на ранних этапах разработки позволяет сместить фокус от максимизации биологической активности к многопараметрической оптимизации, что исторически снизило долю неудач, связанных с неудовлетворительными лекарственноподобными свойствами. Однако прогнозирование клинической токсичности остается наиболее сложной задачей, требующей не только объединения разнородных данных, но и интерпретируемых методов, способных идентифицировать истинные токсикофоры.

Перспективы

МО – не панацея. Однако при грамотном применении эти методы способны расширить возможности по поиску, изучению и применению новых ЛС. Современные исследования в области разработки ЛС привели к накоплению огромных массивов данных [70, 103, 104], созданию мощной вычислительной инфраструктуры [105], а также формированию открытых и закрытых партнерств для разработки открытых наборов данных по химической биологии [19] и созданию доступных методов тестирования новых соединений [106]. Поскольку методы МО применимы к широкому спектру типов данных, химических пространств и биологических систем, подход к разработке ЛС с использованием МО может усилить способность ученых решать сложные задачи на стыке химии, биологии и медицины [19, 107–110].

Помимо рассмотренного в этой статье применения МО в разработке ЛС на осно-

ве малых молекул, эти модели также используются для создания биологических ЛП [111, 112] и исследования природных соединений [113].

Необходимость развития инфраструктуры

Для повсеместного внедрения подходов на основе МО необходимо продолжать развивать и поддерживать централизованные репозитории данных с контролем качества, на которых можно обучать и валидировать робастные модели [114]. Эти репозитории должны учитывать вариабельность, независимость, стабильность и динамичность, присущую данным из различных источников и экспериментальных систем, а также решать задачи, порождаемые наборами данных, охватывающими уникальные химические пространства. Для повышения способности моделей к генерализации на различных независимых наборах данных необходимы совместные усилия по созданию экспериментально прозрачных, открытых и в определенной степени химически сопоставимых ресурсов данных.

Интерпретируемость и регуляторные требования

Повышение интерпретируемости моделей крайне важно для понимания их прогностических механизмов и подтверждения их способности выявлять химически и биологически значимые признаки, а не опираться на ложные корреляции в данных. С регуляторной точки зрения потребность в интерпретируемости возрастает, поскольку новые нормативные требования предполагают наличие четкого обоснования для выводов, сделанных моделями МО, особенно в контексте клинических исследований [115].

Использование исключительно *in silico* критериев (clogP, QED, уникальность, валидность) для оценки соединений, предсказанных с помощью МО, имеет серьезные ограничения, что подчеркивает критическую важность лабораторной экспериментальной проверки полученных соединений [47]. Крайне необходимо поощрять сотрудничество между специалистами по данным (для разработки моделей) и экспериментаторами (для тестирования молекул в лаборатории). Такое сотрудничество не только позволяет совершенствовать существующие модели с помощью дополнительных данных (активное обучение), но и создает разнообразный набор надежных моделей, применимых в различных областях.

Существующие инициативы

Такие инициативы уже появляются. В частности, конкурс «Критическая оценка вычислительных экспериментов по поиску соединений-хитов» (Critical Assessment of Computational Hit-finding Experiments, CACHE) предлагает специалистам находить новые лиганды для ряда биологически релевантных мишеней [99, 116]. Впоследствии предполагаемые хиты проходят тщательную валидацию *in vitro*, а все соответствующие данные публикуются в открытом доступе, способствуя развитию дальнейших работ по моделированию [99]. Существует множество библиотек МО с открытым исходным кодом (например, Chemprop [117]), которые снижают барьеры для входа специалистов, не являющихся экспертами в МО, и позволяют им начать использовать эти методы.

Междисциплинарное взаимодействие

Успешное внедрение МО в разработку ЛС требует не просто совместной работы, а формирования исследователей нового типа – специалистов, владеющих одновременно концептуальным аппаратом химической биологии и инструментарием data science. Для химических биологов это означает освоение принципов построения и валидации прогностических моделей; для специалистов по данным – углубление в биологическую интерпретируемость предсказаний. Такой двусторонний трансфер знаний способен кардинально снизить барьеры между дисциплинами и ускорить трансляцию вычислительных подходов в реальную лабораторную практику.

Ниже представлена сравнительная характеристика ключевых моделей и платформ, рассмотренных в настоящем обзоре. Систематизация по типу задачи, ключевым особенностям и наличию экспериментальной валидации позволяет оценить текущий уровень зрелости различных подходов (таблица 1).

На основе литературного обзора авторами статьи сделана инфографика, представленная на рисунках 1 и 2. На инфографике (рисунок 1) отражены три стадии, на каждой – ключевые методы МО, их входные данные и результат; внизу – общие вы-

зовы (интерпретируемость, экспериментальная валидация и развитие открытых баз данных). На инфографике (рисунок 2) все аналогии являются лишь упрощениями и не претендуют на математическую строгость.

Таблица 1. – Сравнительная таблица моделей машинного обучения в ранней разработке лекарственных средств

Категория	Модель/ Платформа	Тип задачи	Ключевая особенность	Экспериментальная валидация (пример)
Поиск хитов	D-MPNN	Виртуальный скрининг	Направленная передача сообщений по ребрам, учет стереохимии	+ (галицин, абауцин)
	Deep Docking	Виртуальный скрининг	Гибридный подход: докинг + нейросеть, скрининг миллиардных библиотек	+ (ингибиторы Mpro SARS-CoV-2)
	JT-VAE/ JAEGER	Генерация молекул	Декомпозиция на деревья соединений	+ (противомалярийные агенты)
	MolGPT	Генерация молекул	Трансформер (GPT), управляемая генерация по условию, интерпретируемость через saliency maps	–
Установление МД	AlphaFold2	Предсказание структуры белка	MSA + Evoformer + структурный модуль, высокая точность	+ (CDK20)
	ESMFold	Предсказание структуры белка	Языковая модель (без MSA на инференсе), высокая скорость	– (нет прямого примера в статье, но модель валидирована на орфанных белках)
	RFdiffusion	Дизайн белков	Диффузия + RoseTTAFold, управляемая генерация по условиям	+ (дизайн растворимых симметричных белков, экспериментальное подтверждение *in vitro*)
	Geneformer	Анализ транскриптомов	Трансформер, ранговое кодирование экспрессии, трансферное обучение	+ (кардиомиопатия)
	DiffDock	Молекулярный докинг	Диффузионная генеративная модель, SE(3)-эквивариантность, совместное предсказание поз и оценка достоверности	– (валидация на ретроспективных данных PDBbind, проспективные эксперименты не проводилась)
Трансляционные исследования	ADMET-AI	Прогнозирование ADMET	Многозадачная модель (Chemprop-RDKit), веб-интерфейс, сравнение с DrugBank	– (только *in silico*)*
	FFNN (токсичность)	Прогнозирование токсичности	Многозадачное обучение (*in vitro*, *in vivo*, клинические данные)	+ (выявление известных токсикофоров в обучающей выборке; новые молекулы не синтезировались)

Примечания: «+» в колонке «Экспериментальная валидация» означает наличие подтверждения *in vitro* или *in vivo* для хотя бы одного соединения, предсказанного моделью. «–» означает отсутствие экспериментальной валидации в рамках опубликованного исследования. *Для ADMET-AI модель обучалась на проверенных наборах данных Therapeutics Data Commons, но валидации на новых синтезированных соединениях не проводилось.

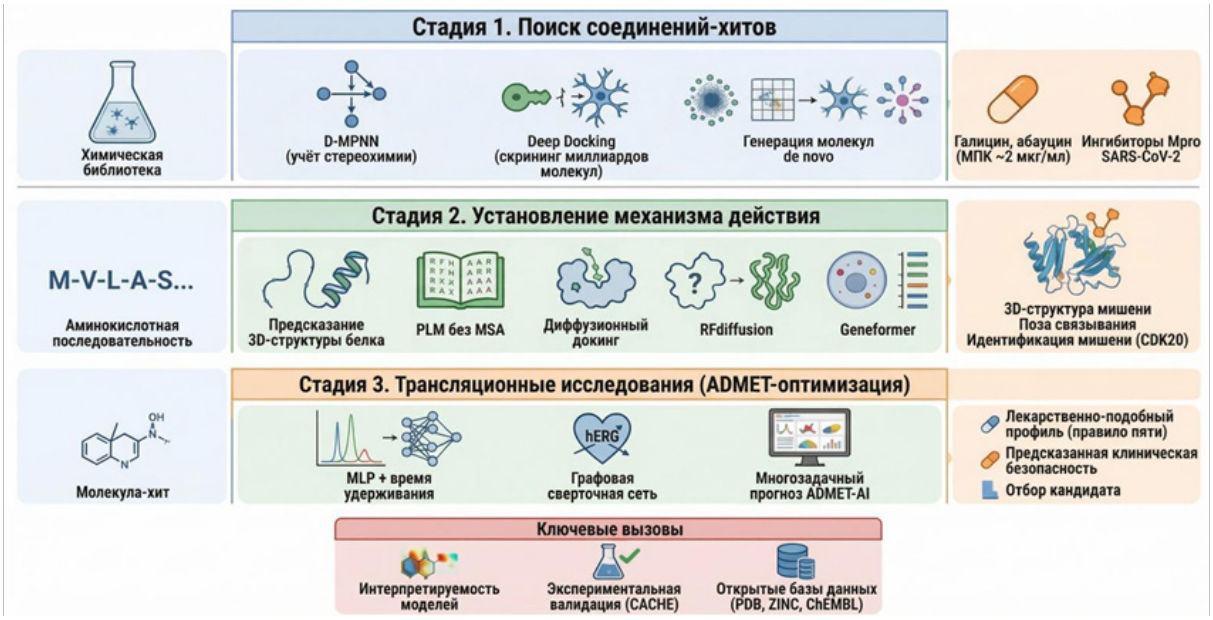

Рисунок 1. – Инфографика, отражающая сквозной конвейер: от поиска «хита» до трансляционных исследований

Рисунок 2. – Инфографика, отражающая упрощенное объяснение на основе аналогий работы основных архитектур, упоминающихся в статье

Глоссарий

Термин	Определение
D-MPNN (Directed Message Passing Neural Network)	Графовая нейронная сеть, в которой информация передается по направленным ребрам (связям) между атомами. Это позволяет более эффективно учитывать стереохимическую информацию и избегать повторной обработки информации.
Deep Docking	Гибридный метод виртуального скрининга, объединяющий физико-химический докинг (расчет энергии связывания с помощью молекулярного докинга) с нейросетевым прогнозированием. Сначала выполняется докинг для 1% молекул, затем обученная нейросеть предсказывает оценки для остальных 99%, что ускоряет скрининг миллиардных библиотек.
Вариационный автоэнкодер (VAE)	Генеративная модель, состоящая из энкодера (сжимает молекулу в непрерывное латентное пространство) и декодера (восстанавливает молекулу из латентного вектора). После обучения декодер может генерировать новые молекулы, семплируя из латентного пространства.

Нормализующие потоки (Normalizing Flows)	Генеративные модели, использующие обратимые преобразования между простым (латентным) и сложным (наблюдаемым) распределениями. Позволяют точно вычислять вероятность сгенерированных образцов.
Диффузионные модели (Diffusion Models)	Генеративные модели, которые сначала итеративно добавляют шум к данным (прямая диффузия), а затем обучаются восстанавливать исходные данные, удаляя шум шаг за шагом (обратная диффузия). Применяются для генерации молекул в 3D и докинга.
Трансформер (Transformer)	Архитектура нейронной сети, основанная на механизме внимания. Позволяет обрабатывать последовательности (например, аминокислотные последовательности или SMILES-строки) параллельно, улавливать долгосрочные зависимости и обеспечивать интерпретируемость через карты внимания.
MSA (Multiple Sequence Alignment)	Множественное выравнивание аминокислотных последовательностей, используемое для поиска эволюционно консервативных участков и предсказания структуры белков.
Языковые модели белка (PLM)	Модели на основе трансформера, предварительно обученные на миллионах белковых последовательностей. Они «выучивают» эволюционные закономерности и не требуют MSA на этапе инференса. Примеры: ESMFold, OmegaFold.
DiffDock	Диффузионная модель для молекулярного докинга, которая рассматривает стыковку лиганда в белок как задачу генеративного моделирования. Обратный процесс диффузии применяется к трансляциям, вращениям и торсионным углам лиганда.
RFdiffusion	Диффузионная модель для дизайна белков, построенная на архитектуре RoseTTAFold. Генерирует координаты остова (backbone) белков с заданными свойствами (симметрия, связывание с мишенью) и позволяет затем подбирать аминокислотные последовательности, сворачивающиеся в эти структуры. Прошла экспериментальную валидацию (например, дизайн растворимых симметричных белков).
Geneformer	Трансформерная модель для анализа одноклеточных транскриптомов. Использует ранговое кодирование экспрессии генов и трансферное обучение для предсказания функций генов и идентификации терапевтических мишеней.
ADMET-AI	Веб-платформа на основе многозадачных нейросетей (Chemprop-RDKit) для прогнозирования абсорбции, распределения, метаболизма, экскреции и токсичности (ADMET) молекул. Позволяет сравнивать свойства соединений с зарегистрированными ЛП.
logP / clogP	Коэффициент распределения октанол-вода, характеризующий липофильность молекулы. clogP – расчетное значение, получаемое суммированием фрагментарных констант (по методу Leo-Hansch). Важный параметр для прогнозирования мембранной проницаемости и биодоступности.
hERG	Белок, кодируемый геном human ether-à-go-go-related gene, формирующий калиевый канал. Блокада этого канала некоторыми ЛС может вызывать сердечную аритмию (удлинение интервала QT). Прогнозирование hERG-токсичности – критический этап ранней разработки.
QSAR (Quantitative Structure-Activity Relationship)	Количественный анализ взаимосвязи «структура–активность». Методы, позволяющие прогнозировать биологическую активность молекул на основе их химической структуры. В статье обсуждаются подходы с использованием ГО (deep QSAR).
CACHE (Critical Assessment of Computational Hit-finding Experiments)	Международный конкурс, в рамках которого специалисты по вычислительным методам предсказывают активные соединения для заданных мишеней, а затем полученные хиты проходят экспериментальную валидацию *in vitro*.
Модели химического языка (Chemical Language Models, CLM)	Модели, которые используют, как правило, рекуррентные нейронные сети (RNN) и обучаются в самоконтролируемом режиме (self-supervised learning) на больших наборах SMILES-строк, предсказывая следующий токен. Это позволяет им выучивать синтаксис и распределение свойств без ручной разметки.

SE(3)-трансформер	Нейросетевая архитектура на механизме внимания, специально разработанная для работы с трехмерными данными (координатами атомов). Ее ключевое свойство – эквивариантность относительно вращений и переносов: если повернуть или сдвинуть входные данные, предсказания модели преобразуются согласованным образом. Это позволяет модели корректно обрабатывать пространственную структуру молекул и белков без необходимости дополнительного обучения на всех возможных ориентациях.
Эквивариантность	Свойство модели, при котором применение преобразования (например, поворота) к входным данным вызывает такое же предсказуемое преобразование выходных данных. Пример: если на вход поданы координаты атомов лиганда, а модель предсказывает координаты связывания (аффинное смещение), то при повороте лиганда выходные координаты поворачиваются так же.

Ограничения обзора

Авторы признают ряд ограничений, которые следует учитывать при интерпретации результатов настоящего обзора.

Языковое ограничение. Поиск литературы проводился только на английском и русском языках, что могло привести к исключению релевантных публикаций на других языках.

Публикационная предвзятость. В обзор включались главным образом статьи из рецензируемых журналов, что потенциально смещает выборку в сторону положительных результатов, поскольку исследования с негативными или нулевыми результатами публикуются реже и преимущественно в специализированных журналах с низким импакт-фактором.

Неполнота охвата препринтов. В обзор не включались публикации только в виде препринтов (bioRxiv, ChemRxiv, arXiv) без последующего рецензирования, что могло привести к задержке в освещении наиболее актуальных, но ещё не прошедших рецензирование результатов.

Хронологические ограничения. Временные рамки поиска (2015–2024 гг.) были выбраны для охвата наиболее активного периода развития методов ГО в медицинской химии. Отметим, что некоторые ключевые исторические работы, опубликованные до 2015 года, QSAR-модели и первые применения нейронных сетей в медицинской химии не вошли в обзор, что может ограничить полноту исторического контекста.

Отсутствие количественного мета-анализа. Ввиду гетерогенности рассматриваемых задач (разные типы данных, метрики оценки, наборы данных) проведение формального мета-анализа не представлялось возможным; обзор носит описательный характер.

При оценке моделей МО в задачах ранней разработки ЛС необходимо учитывать ряд систематических ограничений. К ним относятся: утечка информации между обучающей и тестовой выборками при случайном разбиении (random split) – для химических данных предпочтительны разбиения по молекулярным каркасам (scaffold split) либо по времени (time split); смещение публикационных датасетов в сторону активных соединений и недостаточное представительство истинно неактивных молекул; ограниченная переносимость моделей между химическими сериями и между лабораториями вследствие различий в протоколах испытаний. Игнорирование этих факторов систематически завышает заявляемые метрики качества и затрудняет независимую воспроизводимость.

ЗАКЛЮЧЕНИЕ

1. Методы МО становятся неотъемлемым инструментом на всех этапах ранней разработки ЛС – от поиска соединений-хитов (графовые нейросети, генеративные модели) до оптимизации ADMET-профиля и прогнозирования токсичности (ADMET-AI, мультимодальные подходы).

2. Генеративные модели (вариационные автоэнкодеры, нормализующие потоки, диффузионные модели, модели химического языка, трансформеры) позволяют эффективно исследовать химическое пространство, создавая новые молекулы с заданными свойствами и значительно расширяя границы доступных химических библиотек. Эволюция этих методов идет по пути увеличения структурной валидности (от SMILES к графам и 3D-координатам) и внедрения механизмов управления свойствами.

3. Модели предсказания структуры белков (AlphaFold2, RoseTTAFold, языковые модели белка) и докинга (DiffDock, RFdiffusion) открывают новые возможно-

сти для установления МД и структурно-ориентированного дизайна ЛП, особенно для мишеней, структурная информация о которых ранее отсутствовала. Переход от MSA-зависимых подходов к языковым моделям белка позволяет обрабатывать сиротские белки и гипервариабельные участки.

4. Трансферное обучение и модели на основе одноклеточных транскриптомов (Geneformer) позволяют преодолевать ограниченность данных в биологических и клинических исследованиях, обеспечивая контекстно-зависимые предсказания терапевтических мишеней. Ранговое кодирование экспрессии смещает фокус с конститутивных генов на регуляторно значимые.

5. Для успешного внедрения МО в практику разработки ЛС необходимы: развитие открытых репозиториев данных высокого качества; повышение интерпретируемости моделей (карты значимости, выявление рационалов и токсикофоров); экспериментальная валидация *in silico* предсказаний (инициативы типа CACHE); активное междисциплинарное сотрудничество между специалистами по данным и химическими биологами. Сравнительный анализ рассмотренных моделей представлен в таблице 1.

6. Несмотря на впечатляющие успехи, методы МО не заменяют, а дополняют традиционные экспериментальные подходы. Ключевые вызовы остаются: точное прогнозирование клинической токсичности, генерация стабильных крупных молекул, интерпретируемость «черных ящиков» и регуляторное признание вычислительных моделей.

SUMMARY

A. S. Maksimov, Zh. M. Dergacheva

APPLICATION OF MACHINE LEARNING METHODS IN EARLY DRUG DEVELOPMENT

This article presents a systematic review of machine learning (ML) methods application at an early drug development. Three key phases are considered: top-connections search, mechanism of action elucidation, and translational research. Graph neural networks (D-MPNN) and hybrid approaches (Deep Docking) are shown to substantially enhance virtual screening efficiency, while generative models–variational autoencoders normalizing flows, diffusion models and transformers–enable targeted de novo design of molecules with the defined properties leaving beyond existing chemical libraries. In the field of protein structure prediction, the evolution from AlphaFold2 and RoseTTAFold to protein language models (OmegaFold, ESMFold) highlighting their ability to overcome limitations associated with the lack of evolutionary information is analyzed. Diffusion-based approaches to molecular docking (DiffDock) and protein design (RFdiffusion) demonstrating the transition from heuristic methods to continuous generative modeling are discussed. Special attention is given to transfer learning exemplified by Geneformer for single-cell transcriptome analysis, as well as to the integration of ADMET predictive models (ADMET-AI) for multi-parameter optimization of drug candidates. Key challenges are identified: model interpretability, experimental validation of in silico predictions, regulatory acceptance of computational approaches, and the need for high-quality open data repositories. It is concluded that ML methods complement rather than replace traditional experimental approaches, forming the basis for a new interdisciplinary field at the intersection of chemical biology and data science.

Keywords: machine learning, early drug development, virtual screening, generative models, graph neural networks, protein structure prediction, AlphaFold, protein language models, molecular docking, diffusion models, ADMET prediction, toxicity, transfer learning, model interpretability.

ЛИТЕРАТУРА

1. Wouters, O. J. Estimated research and development investment needed to bring a new medicine to market, 2009–2018 / O. J. Wouters, M. McKee, J. Luyten // JAMA. – 2020. – Vol. 323, N 9. – P. 844–853. – DOI: 10.1001/jama.2020.1166.

2. Target identification and mechanism of action in chemical biology and drug discovery / M. Schenone, V. Dančik, B. K. Wagner, P. A. Clemons // Nature chemical biology. – 2013. – Vol. 9, N 4. – P. 232–240. – DOI: 10.1038/nchembio.1199.

3. Ashenden, S. K. The Era of Artificial Intelligence, Machine Learning, and Data Science in the Pharmaceutical Industry / S. K. Ashenden. – Cambridge : Academic Press, 2021. – 264 p. – Chap. 6.

4. Smietana, K. Trends in clinical success rates / K. Smietana, M. Siatkowski, M. Møller //

Nature reviews. Drug discovery. – 2016. – Vol. 15, N 6. – P. 379–380. – DOI: 10.1038/nrd.2016.85.

5. Harrison, R. K. Phase II and phase III failures: 2013–2015 / R. K. Harrison // Nature reviews. Drug discovery. – 2016. – Vol. 15, N 12. – P. 817–818. – DOI: 10.1038/nrd.2016.184.

6. Dowden, H. Trends in clinical success rates and therapeutic focus / H. Dowden, J. Munro // Nature reviews. Drug discovery. – 2019. – Vol. 18, N 7. – P. 495–496. – DOI: 10.1038/d41573-019-00074-z.

7. Computer vision for autonomous vehicles: problems, datasets and state of the art / J. Janai, F. Güney, A. Behl, A. Geiger // Foundations and Trends in Computer Graphics and Vision. – 2020. – Vol. 12, N 1–3. – P. 1–308. – DOI: 10.1561/0600000079.

8. Machine learning and natural language processing in psychotherapy research: alliance as example use case / S. B. Goldberg, N. Flemotomos, V. R. Martinez [et al.] // Journal of counseling psychology. – 2020. – Vol. 67, N 4. – P. 438–448. – DOI: 10.1037/cou0000382.

9. Peterson, A. A. Small-molecule discovery through DNA-encoded libraries / A. A. Peterson, D. R. Liu // Nature reviews. Drug discovery. – 2023. – Vol. 22, N 9. – P. 699–722. – DOI: 10.1038/s41573-023-00713-6.

10. Machine learning on DNA-encoded library count data using an uncertainty-aware probabilistic loss function / K. S. Lim, A. G. Reidenbach, B. K. Hua [et al.] // Journal of chemical informaition and modeling. – 2022. – Vol. 62, N 10. – P. 2316–2331. – DOI: 10.1021/acs.jcim.2c00041.

11. Machine-learning-based data analysis method for cell-based selection of DNA-encoded libraries / R. Hou, C. Xie, Y. Gui [et al.] // ACS omega. – 2023. – Vol. 8, N 21. – P. 19057–19071. – DOI: 10.1021/acsomega.3c02152.

12. Applications of single-cell RNA sequencing in drug discovery and development / B. Van de Sande, J. S. Lee, E. Mutasa-Gottgens [et al.] // Nature reviews. Drug discovery. – 2023. – Vol. 22, N 6. – P. 496–520. – DOI: 10.1038/s41573-023-00688-4.

13. scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data / F. Yang, W. Wang, F. Wang [et al.] // Nature machine intelligence. – 2022. – Vol. 4, N 10. – P. 852–866. – DOI: 10.1038/s42256-022-00534-z.

14. Deep transfer learning of cancer drug responses by integrating bulk and single-cell RNA-seq data / J. Chen, X. Wang, A. Ma [et al.] // Nature communications. – 2022. – Vol. 13, N 1. – Art. 6494. – DOI: 10.1038/s41467-022-34277-7.

15. A multi-scale convolutional neural network for phenotyping high-content cellular images / W. J. Godinez, I. Hossain, S. E. Lazic [et al.] // Bioinformatics. – 2017. – Vol. 33, N 13. – P. 2010–2019. – DOI: 10.1093/bioinformatics/btx069.

16. A deep learning approach to antibiotic discovery / J. M. Stokes, K. Yang, K. Swanson [et al.] // Cell. – 2020. – Vol. 180, N 4. – P. 688–702. – DOI: 10.1016/j.cell.2020.01.021.

17. Convolutional embedding of attributed molecular graphs for physical property prediction / C. W. Coley, R. Barzilay, W. H. Green [et al.] // Journal of chemical information and modeling. – 2017. – Vol. 57, N 8. – P. 1757–1772. – DOI: 10.1021/acs.jcim.6b00601.

18. Deep learning identifies synergistic drug combinations for treating COVID-19 / W. Jin, J. M. Stokes, R. T. Eastman [et al.] // Proceedings of the National Academy of Sciences of the United States of America. – 2021. – Vol. 118, N 39. – Art. e2105070118. – DOI: 10.1073/pnas.2105070118.

19. Applications of machine learning in drug discovery and development / J. Vamathevan, D. Clark, P. Czodrowski [et al.] // Nature reviews. Drug discovery. – 2019. – Vol. 18, N 6. – P. 463–477. – DOI: 10.1038/s41573-019-0024-5.

20. Rogers, D. Extended-connectivity fingerprints / D. Rogers, M. Hahn // Journal of chemical information and modeling. – 2010. – Vol. 50, N 5. – P. 742–754. – DOI: 10.1021/ci100050t.

21. Database fingerprint (DFP): an approach to represent molecular databases / E. Fernández-De Gortari, C. R. García-Jacas, K. Martinez-Mayorga, J. L. Medina-Franco // Journal of cheminformatics. – 2017. – Vol. 9, N 1. – Art. 9. – Doi: 10.1186/s13321-017-0195-1.

22. Deep learning-guided discovery of an antibiotic targeting Acinetobacter baumannii / G. Liu, D. B. Catacutan, K. Rathod [et al.] // Nature chemical biology. – 2023. – Vol. 19, N 11. – P. 1342–1350. – DOI: 10.1038/s41589-023-01349-8.

23. Analyzing learned molecular representations for property prediction / K. Yang, K. Swanson, W. Jin [et al.] // Journal of chemical information and modeling. – 2019. – Vol. 59, N 8. – P. 3370–3388. – DOI: 10.1021/acs.jcim.9b00237.

24. The Drug Repurposing Hub: a next-generation drug library and information resource / S. M. Corsello, J. A. Bittker, Z. Liu [et al.] // Nature medicine. – 2017. – Vol. 23, N 4. – P. 405–408. – DOI: 10.1038/nm.4306.

25. Discovery of a structural class of antibiotics with explainable deep learning / F. Wong, E. J. Zheng, J. A. Valeri [et al.] // Nature. – 2024. – Vol. 626, N 7997. – P. 177–185. – DOI: 10.1038/s41586-023-06887-8.

26. A practical guide to large-scale docking / B. J. Bender, S. Gahbauer, A. Luttens [et al.] // Nature protocols. – 2021. – Vol. 16, N 10. – P. 4799–4832. – DOI: 10.1038/s41596-021-00597-z.

27. Artificial intelligence-enabled virtual screening of ultra-large chemical libraries with deep docking / F. Gentile, J. C. Yaacoub, J. Gleave [et al.] // Nature protocols. – 2022. – Vol. 17,

N 3. – P. 672–697. – DOI: 10.1038/s41596-021-00659-2.

28. Integrating QSAR modelling and deep learning in drug discovery: the emergence of deep QSAR / A. Tropsha, O. Isayev, A. Varnek [et al.] // Nature reviews. Drug discovery. – 2024. – Vol. 23, N 2. – P. 141–155. – DOI: 10.1038/s41573-023-00832-0.

29. Supercomputer-based ensemble docking drug discovery pipeline with application to Covid-19 / A. Acharya, R. Agarwal, M. B. Baker [et al.] // Journal of chemical information and modeling. – 2020. – Vol. 60, N 12. – P. 5832–5852. – DOI: 10.1021/acs.jcim.0c01010.

30. A critical overview of computational approaches employed for COVID-19 drug discovery / E. N. Muratov, R. Amaro, C. H. Andrade [et al.] // Chemical Society reviews. – 2021. – Vol. 50, N 16. – P. 9121–9151. – DOI: 10.1039/d0cs01065k.

31. Sterling, T. ZINC 15 – ligand discovery for everyone / T. Sterling, J. J. Irwin // Journal of chemical information and modeling. – 2015. – Vol. 55, N 11. – P. 2324–2337. – DOI: 10.1021/acs.jcim.5b00559.

32. Non-covalent SARS-CoV-2 Mpro inhibitors developed from in silico screen hits / G. G. Rossetti, M. A. Ossorio, S. Rempel [et al.] // Scientific reports. – 2022. – Vol. 12, N 1. – Art. 2505. – DOI: 10.1038/s41598-022-06306-4.

33. Reymond, J. L. The chemical space project / J. L. Reymond // Accounts of chemical research. – 2015. – Vol. 48, N 3. – P. 722–730. – DOI: 10.1021/ar500432k.

34. Automatic chemical design using a data-driven continuous representation of molecules / R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud [et al.] // ACS central science. – 2018. – Vol. 4, N 2. – P. 268–276. – DOI: 10.1021/acscentsci.7b00572.

35. Anstine, D. M. Generative models as an emerging paradigm in the chemical sciences / D. M. Anstine, O. Isayev // Journal of the American Chemical Society. – 2023. – Vol. 145, N 16. – P. 8736–8750. – DOI: 10.1021/jacs.2c13467.

36. Jin, W. Junction tree variational autoencoder for molecular graph generation / W. Jin, R. Barzilay, T. Jaakkola // arXiv. – 2018. – 18 Feb. – URL: https://arxiv.org/abs/1802.04364 (date of access: 16.03.2026).

37. Design of potent antimalarials with generative chemistry / W. J. Godinez, E. J. Ma, A. T. Chao [et al.] // Nature machine intelligence. – 2022. – Vol. 4, N 2. – P. 180–186. – DOI: 10.1038/s42256-022-00448-w.

38. Walters, W. P. Assessing the impact of generative AI on medicinal chemistry / W. P. Walters, M. Murcko // Nature biotechnology. – 2020. – Vol. 38, N 2. – P. 143–145. – DOI: 10.1038/s41587-020-0418-2.

39. Deep learning tools to accelerate antibiotic discovery / A. Cesaro, M. Bagheri, M. Torres [et al.] // Expert opinion on drug discovery. – 2023. – Vol. 18, N 11. – P. 1245–1257. – DOI: 10.1080/17460441.2023.2250721.

40. Rezende, D. J. Variational inference with normalizing flows / D. J. Rezende, S. Mohamed // International Conference on Machine Learning : proceedings of the 32nd International Conference on Machine Learning / ed.: F. Bach, D. Blei. – 2015. – Vol. 37. – P. 1530–1538.

41. Shekhovtsov, A. VAE approximation error: ELBO and exponential families / A. Shekhovtsov, D. Schlesinger, B. Flach // arXiv. – 2021. – 18 Feb. – URL: https://arxiv.org/abs/2102.09310 (date of access: 16.03.2026).

42. GraphAF: a flow-based autoregressive model for molecular graph generation / C. Shi, M. Xu, Z. Zhu [et al.] // arXiv. – 2020. – 27 Jan. – URL: https://arxiv.org/abs/2001.09382 (date of access: 16.03.2026).

43. Equivariant diffusion for molecule generation in 3D / E. Hoogeboom, V. G. Satorras, C. Vignac, M. Welling // International Conference on Machine Learning : proceedings of the 39th International Conference on Machine Learning. – 2022. – Vol. 162. – P. 8867–8887.

44. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules / D. Weininger // Journal of Chemical Information and Computer Sciences. – 1988. – Vol. 28, N 1. – P. 31–36. – DOI: 10.1021/ci00057a005.

45. Grisoni, F. Chemical language models for de novo drug design: challenges and opportunities / F. Grisoni // Current opinion in structural biology. – 2023. – Vol. 79. – Art. 102527. – DOI: 10.1016/j.sbi.2023.102527.

46. Flam-Shepherd, D. Language models can learn complex molecular distributions / D. Flam-Shepherd, K. Zhu, A. Aspuru-Guzik // Nature communications. – 2022. – Vol. 13, N 1. – Art. 3293. – DOI: 10.1038/s41467-022-30839-x.

47. Chemical language models enable navigation in sparsely populated chemical space / M. A. Skinnider, R. Greg Stacey, D. S. Wishart, L. J. Foster // Nature machine intelligence. – 2021. – Vol. 3, N 9. – P. 759–770. – DOI: 10.1038/s42256-021-00368-1.

48. Generative molecular design in low data regimes / M. Moret, L. Friedrich, F. Grisoni [et al.] // Nature machine intelligence. – 2020. – Vol. 2, N 3 – P. 171–180. – DOI: 10.1038/s42256-020-0160-y.

49. De novo design of Nurr1 agonists via fragment-augmented generative deep learning in low-data regime / M. Ballarotto, S. Willems, T. Stiller [et al.] // Journal of medicinal chemistry. – 2023. – Vol. 66, N 12. – P. 8170–8177. – DOI: 10.1021/acs.jmedchem.3c00485.

50. Leveraging molecular structure and bioactivity with chemical language models for de novo drug design / M. Moret, I. P. Angona,

L. Cotos [et al.] // Nature communications. – 2023. – Vol. 14, N 1. – Art. 114. – DOI: 10.1038/s41467-022-35692-6.

51. Combining generative artificial intelligence and on-chip synthesis for de novo drug design / F. Grisoni, B. J. H. Huisman, A. L. Button [et al.] // Science advances. – 2021. – Vol. 7, N 24. – Art. eabg3338. – DOI: 10.1126/sciadv.abg3338.

52. De novo design of bioactive small molecules by artificial intelligence / D. Merk, L. Friedrich, F. Grisoni, G. Schneider // Molecular informatics. – 2018. – Vol. 37, N 1/2. – Art. 1700153. – DOI: 10.1002/minf.201700153.

53. Attention is all you need / A. Vaswani, N. Shazeer, N. Parmar [et al.] // arXiv. – 2017. – 12 June. – URL: https://arxiv.org/abs/1706.03762 (date of access: 16.03.2026).

54. MolGPT: molecular generation using a transformer-decoder model / V. Bagal, R. Aggarwal, P. K. Vinod, U. D. Priyakumar // Journal of chemical information and modeling. – 2021. – Vol. 62, N 9. – P. 2064–2076. – DOI: 10.1021/acs.jcim.1c00600.

55. GuacaMol: benchmarking models for de novo molecular design / N. Brown, M. Fiscato, M. H. S. Segler, A. C. Vaucher // Journal of chemical information and modeling. – 2019. – Vol. 59, N 3. – P. 1096–1108. – DOI: 10.1021/acs.jcim.8b00839.

56. Molecular sets (MOSES): a benchmarking platform for molecular generation models / D. Polykovskiy, A. Zhebrak, B. Sanchez-Lengeling [et al.] // Frontiers in pharmacology. – 2020. – Vol. 11. – Art. 565644. – DOI: 10.3389/fphar.2020.565644.

57. Autonomous chemical research with large language models / D. A. Boiko, R. MacKnight, B. Kline, G. Gomes // Nature. – 2023. – Vol. 624. – P. 570–578. – DOI: 10.1038/s41586-023-06792-0.

58. Leveraging large language models for predictive chemistry / K. M. Jablonka, P. Schwaller, A. Ortega-Guerrero, B. Smit // Nature machine intelligence. – 2024. – Vol. 6, N 2. – P. 161–169. – DOI: 10.1038/s42256-023-00788-1.

59. Born, J. Regression Transformer enables concurrent sequence regression and generation for molecular language modelling / J. Born, M. Manica // Nature machine intelligence. – 2023. – Vol. 5, N 4. – P. 432–444. – DOI: 10.1038/s42256-023-00639-z.

60. Neural scaling of deep chemical models / N. C. Frey, R. Soklaski, S. Axelrod [et al.] // Nature machine intelligence. – 2023. – Vol. 5, N 11. – P. 1297–1305. – DOI: 10.1038/s42256-023-00740-3.

61. Grechishnikova, D. Transformer neural network for protein-specific de novo drug generation as a machine translation problem / D. Grechishnikova // Scientific reports. – 2021. – Vol. 11, N 1. – Art. 321. – DOI: 10.1038/s41598-020-79682-4.

62. Highly accurate protein structure prediction with AlphaFold / J. Jumper, R. Evans, A. Pritzel [et al.] // Nature. – 2021. – Vol. 596, N 7873. – P. 583–589. – DOI: 10.1038/s41586-021-03819-2.

63. Structure of the decoy module of human glycoprotein 2 and uromodulin and its interaction with bacterial adhesin FimH / A. Stsiapanava, C. Xu, S. Nishio [et al.] // Nature structural & molecular biology. – 2022. – Vol. 29, N 3. – P. 190–193. – DOI: 10.1038/s41594-022-00729-3.

64. Cryo-EM structures of human hepatitis B and woodchuck hepatitis virus small spherical subviral particles / H. Liu, X. Hong, J. Xi [et al.] // Science advances. – 2022. – Vol. 8, N 31. – Art. eabo4184. – DOI: 10.1126/sciadv.abo4184.

65. AlphaFold accelerates artificial intelligence powered drug discovery: efficient discovery of a novel CDK20 small molecule inhibitor / F. Ren, X. Ding, M. Zheng [et al.] // Chemical science. – 2023. – Vol. 14, N 6. – P. 1443–1452. – DOI: 10.48550/arXiv.2201.09647.

66. Structural comparison and drug screening of spike proteins of ten SARS-CoV-2 variants / Q. Yang, X. Lian, A. A. S. Syed [et al.] // Research (Washington, D.C.). – 2022. – Vol. 2022. – Art. 9781758. – DOI: 10.34133/2022/9781758.

67. Highly accurate protein structure prediction and drug screen of monkeypox virus proteome / Q. Yang, D. Xia, A. A. S. Syed [et al.] // The Journal of infection. – 2023. – Vol. 86, N 1. – P. 66–117. – DOI: 10.1016/j.jinf.2022.08.006.

68. Chemistry42: an AI-driven platform for molecular design and optimization / Y. A. Ivanenkov, D. Polykovskiy, D. Bezrukov [et al.] // Journal of Chemical Information and Modeling. – 2023. – Vol. 63, N 3. – P. 695–701. – DOI: 10.1021/acs.jcim.2c01191.

69. Accurate prediction of protein structures and interactions using a three-track neural network / M. Baek, F. DiMaio, I. Anishchenko [et al.] // Science. – 2021. – Vol. 373, N 6557. – P. 871–876. – DOI: 10.1126/science.abj8754.

70. The Protein Data Bank / H. M. Berman, J. Westbrook, Z. Feng [et al.] // Nucleic acids research. – 2000. – Vol. 28, N 1. – P. 235–242. – DOI: 10.1093/nar/28.1.235.

71. Van Wart, H. E. The cysteine switch: a principle of regulation of metalloproteinase activity with potential applicability to the entire matrix metalloproteinase gene family / H. E. Van Wart, H. Birkedal-Hansen // Proceedings of the National Academy of Sciences of the United States of America. – 1990. – Vol. 87, N 14. – P. 5578–5582. – DOI: 10.1073/pnas.87.14.5578.

72. Michaud, J. M. A language model beats AlphaFold2 on orphans / J. M. Michaud, A. Madani, J. S. Fraser // Nature biotechnology. – 2022. – Vol. 40, N 11. – P. 1576–1577. – DOI:

10.1038/s41587-022-01466-0.

73. High-resolution de novo structure prediction from primary sequence / R. Wu, F. Ding, R. Wang [et al.] // bioRxiv. – 2022. – 21 July. – DOI:10.1101/2022.07.21.500999.

74. A method for multiple-sequence-alignment-free protein structure prediction using a protein language model / X. Fang, F. Wang, L. Liu [et al.] // Nature machine intelligence. – 2023. – Vol. 5, N 10. – P. 1087–1096. – DOI: 10.1038/s42256-023-00721-6.

75. Large language models generate functional protein sequences across diverse families / A. Madani, B. Krause, E. R. Greene [et al.] // Nature biotechnology. – 2023. – Vol. 41, N 8. – P. 1099–1106. – DOI: 10.1038/s41587-022-01618-2.

76. Evolutionary-scale prediction of atomic-level protein structure with a language model / Z. Lin, H. Akin, R. Rao [et al.] // Science. – 2023. – Vol. 379, N 6637. – P. 1123–1130. – DOI: 10.1126/science.ade2574.

77. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches / B. E. Suzek, Y. Wang, H. Huang [et al.] // Bioinformatics. – 2015. – Vol. 31, N 6. – P. 926–932. – DOI: 10.1093/bioinformatics/btu739.

78. DiffDock: diffusion steps, twists, and turns for molecular docking / G. Corso, H. Stärk, B. Ling [et al.]. // arXiv. – 2022. – 4 Oct. – URL: https://arxiv.org/abs/2210.01776 (date of access: 16.03.2026).

79. De novo design of protein structure and function with RFdiffusion / J. L. Watson, D. Juergens, N. R. Bennett [et al.] // Nature. – 2023. – Vol. 620, N 7976. – P. 1089–1100. – DOI: 10.1038/s41586-023-06415-8.

80. Transfer learning enables predictions in network biology / C. V. Theodoris, L. Xiao, A. Chopra [et al.] // Nature. – 2023. – Vol. 618, N 7965. – P. 616–624. – DOI: 10.1038/s41586-023-06139-9.

81. Single-nucleus profiling of human dilated and hypertrophic cardiomyopathy / M. Chaffin, I. Papangeli, B. Simonson [et al.] // Nature. – 2022. – Vol. 608, N 7921. – P. 174–180. – DOI: 10.1038/s41586-022-04817-8.

82. Principles of early drug discovery / J. P. Hughes, S. S. Rees, S. B. Kalindjian, K. L. Philpott // British journal of pharmacology. – 2011. – Vol. 162, N 6. – P. 1239–1249. – DOI: 10.1111/j.1476-5381.2010.01127.x.

83. Goodnow, R. A. Hit and lead identification: integrated technology-based approaches / R. A. Goodnow // Drug discovery today. Technologies. – 2006. – Vol. 3, N 4. – P. 367–375. – DOI: 10.1016/j.ddtec.2006.12.009.

84. Transformer-based deep learning method for optimizing ADMET properties of lead compounds / L. Yang, C. Jin, G. Yang [et al.] // Physical chemistry chemical physics. – 2023. – Vol. 25, N 3. – P. 2377–2385. – DOI: 10.1039/d2cp05332b.

85. In silico prediction of hERG blockers using machine learning and deep learning approaches / Y. Chen, X. Yu, W. Li [et al.] // Journal of applied toxicology. – 2023. – Vol. 43, N 10. – P. 1462–1475. – DOI: 10.1002/jat.4477.

86. Accurate clinical toxicity prediction using multi-task deep neural nets and contrastive molecular explanations / B. Sharma, V. Chenthamarakshan, A. Dhurandhar [et al.] // Scientific reports. – 2023. – Vol. 13, N 1. – Art. 4908. – DOI: 10.1038/s41598-023-31169-8.

87. Why 90% of clinical drug development fails and how to improve it? / D. Sun, W. Gao, H. Hu, S. Zhou // Acta Pharm Sin B. – 2022. – Vol. 12, N 7. – P. 3049–3062. – DOI: 10.1016/j.apsb.2022.02.002.

88. Kola, I. Can the pharmaceutical industry reduce attrition rates? / I. Kola, J. Landis // Nature reviews. Drug discovery. – 2004. – Vol. 3, N 8. – P. 711–716. – DOI: 10.1038/nrd1470.

89. Lipinski, C. A. Lead- and drug-like compounds: the rule-of-five revolution / C. A. Lipinski // Drug discovery today. Technologies. – 2004. – Vol. 1, N 4. – P. 337–341. – DOI: 10.1016/j.ddtec.2004.11.007.

90. A robust, viable, and resource sparing HPLC-based log P method applied to common drugs / A. L. Coutinho, R. Gristofoletti, F. Wu [et al.] // International journal of pharmaceutics. – 2023. – Vol. 644. – Art. 123325. – DOI: 10.1016/j.ijpharm.2023.123325.

91. Faller, B. Computational approaches to determine drug solubility / B. Faller, P. Ertl // Advanced drug delivery reviews. – 2007. – Vol. 59, N 7. – P. 533–545. – DOI: 10.1016/j.addr.2007.05.005.

92. Comparison of log P and log D correction models trained with public and proprietary data sets / I. Aliagas, A. Gobbi, M. L. Lee, B. D. Sellers // Journal of computer-aided molecular design. – 2022. – Vol. 36, N 4. – P. 253–262. – DOI: 10.1007/s10822-022-00450-9.

93. Win, Z. M. Using machine learning to predict partition coefficient (log P) and distribution coefficient (log D) with molecular descriptors and liquid chromatography retention time / Z. M. Win, A. M. Y. Cheong, W. S. Hopkins // Journal of chemical information and modeling. – 2023. – Vol. 63, N 7. – P. 1906–1913. – DOI: 10.1021/acs.jcim.2c01373.

94. The METLIN small molecule dataset for machine learning-based retention time prediction / X. Domingo-Almenara, C. Guijas, E. Billings [et al.] // Nature communications. – 2019. – Vol. 10, N 1. – Art. 5811. – DOI: 10.1038/s41467-019-13680-7.

95. ChEMBL: a large-scale bioactivity database for drug discovery / A. Gaulton, L. J. Bellis,

A. P. Bento [et al.] // Nucleic acids research. – 2012. – Vol. 40, N Database issue. – P. D1100–D1107. – DOI: 10.1093/nar/gkr777.

96. Datta, R. Efficient lipophilicity prediction of molecules employing deep-learning models / R. Datta, D. Das, S. Das // Chemometrics and intelligent laboratory systems. – 2021. – Vol. 213. – Art. 104309. – DOI: 10.1016/j.chemolab.2021.104309.

97. Prasad, S. A deep learning approach for the blind log P prediction in SAMPL6 challenge / S. Prasad, B. R. Brooks // Journal of computer-aided molecular design. – 2020. – Vol. 34, N 5. – P. 535–542. – DOI: 10.1007/s10822-020-00292-3.

98. Cardiac safety assays / J. Heijman, N. Voigt, L. G. Carlsson, D. Dobrev // Current opinion in pharmacology. – 2014. – Vol. 15. – P. 16–21. – DOI: 10.1016/j.coph.2013.11.004.

99. CACHE (Critical Assessment of Computational Hit-finding Experiments): a public-private partnership benchmarking initiative to enable the development of computational methods for hit-finding / S. Ackloo, R. Al-Awar, R. E. Amaro [et al.] // Nature reviews. Chemistry. – 2022. – Vol. 6, N 4. – P. 287–295. – DOI: 10.1038/s41570-022-00363-z.

100. ADMET-AI: a machine learning ADMET platform for evaluation of large-scale chemical libraries / K. Swanson, P. Walther, J. Leitz [et al.] // Zenodo. – 2023. – 9 Dec. – DOI: 10.5281/zenodo.10372930.

101. MoleculeNet: a benchmark for molecular machine learning / Z. Wu, B. Ramsundar, E. N. Feinberg [et al.] // Chemical science. – 2017. – Vol. 9, N 2. – P. 513–530. – DOI: 10.1039/c7sc02664a.

102. Tox21Challenge to build predictive models of nuclear receptor and stress response pathways as mediated by exposure to environmental chemicals and drugs / R. Huang, M. Xia, D. T. Nguyen [et al.] // Frontiers in environmental science. – 2016. – Vol. 3. – Art. 85. – DOI: 10.3389/fenvs.2015.00085.

103. ZINC-22 – a free multi-billion-scale database of tangible compounds for ligand discovery / B. I. Tingle, K. G. Tang, M. Castanon [et al.] // Journal of chemical information and modeling. – 2023. – Vol. 63, N 4. – P. 1166–1176. – DOI: 10.1021/acs.jcim.2c01253.

104. Quantum chemistry structures and properties of 134 kilo molecules / R. Ramakrishnan, P. O. Dral, M. Rupp, O. A. von Lilienfeld // Scientific data. – 2014. – Vol. 1. – Art. 140022. – DOI: 10.1038/sdata.2014.22.

105. From computer-aided drug discovery to computer-driven drug discovery / L. Frye, S. Bhat, K. Akinsanya, R. Abel // Drug discovery today. Technologies. – 2021. – Vol. 39. – P. 111–117. – DOI: 10.1016/j.ddtec.2021.08.001.

106. High-throughput screening technology in industrial biotechnology / W. Zeng, L. Guo, S. Xu [et al.] // Trends in biotechnology. – 2020. – Vol. 38, N 8. – P. 888–906. – DOI: 10.1016/j.tibtech.2020.01.001.

107. Sarkar, N. Practical applications of machine learning for anti-infective drug discovery / N. Sarkar, J. M. Stokes // Medicinal Chemistry Reviews / ed.: J. Rudolph, J. J. Bronson. – Washington : ACS, 2023. – Vol. 58. – P. 345–375.

108. Applications of machine learning in microbial natural product drug discovery / A. Arnold, J. Alexander, G. Liu, J. M. Stokes // Expert opinion on drug discovery. – 2023. – Vol. 18, N 11. – P. 1259–1272. – DOI: 10.1080/17460441.2023.2251400.

109. Artificial intelligence for natural product drug discovery / M. W. Mullowney, K. R. Duncan, S. S. Elsayed [et al.] // Nature reviews. Drug discovery. – 2023. – Vol. 22, N 11. – P. 895–916. – DOI: 10.1038/s41573-023-00774-7.

110. Exploiting machine learning for end-to-end drug discovery and development / S. Ekins, A. C. Puhl, K. M. Zorn [et al.] // Nature materials. – 2019. – Vol. 18, N 5. – P. 435–441. – DOI: 10.1038/s41563-019-0338-z.

111. Designing anticancer peptides by constructive machine learning / F. Grisoni, C. S. Neuhaus, G. Gabernet [et al.] // ChemMedChem. – 2018. – Vol. 13, N 13. – P. 1300–1302. – DOI: 10.1002/cmdc.201800204.

112. Chen, J. xDeep-AcPEP: deep learning method for anticancer peptide activity prediction based on convolutional neural network and multitask learning / J. Chen, H. H. Cheong, S. W. I. Siu // Journal of chemical information and modeling. – 2021. – Vol. 61, N 8. – P. 3789–3803. – DOI: 10.1021/acs.jcim.1c00181.

113. Walker, A. S. A machine learning bioinformatics method to predict biological activity from biosynthetic gene clusters / A. S. Walker, J. Clardy // Journal of chemical information and modeling. – 2021. – Vol. 61, N 6. – P. 2560–2571. – DOI: 10.1021/acs.jcim.0c01304.

114. MELLODDY: cross-pharma federated learning at unprecedented scale unlocks benefits in QSAR without compromising proprietary information / W. Heyndrickx, L. Mervin, T. Morawietz [et al.] // Journal of chemical information and modeling. – 2024. – Vol. 64, N 7. – P. 2331–2344. – DOI: 10.1021/acs.jcim.3c00799.

115. A perspective on explanations of molecular prediction models / G. P. Wellawatte, H. A. Gandhi, A. Seshadri, A. D. White // Journal of chemical theory and computation. – 2023. – Vol. 19, N 8. – P. 2149–2160. – DOI: 10.1021/acs.jctc.2c01235.

116. Crowdsourced mapping of unexplored target space of kinase inhibitors / A. Cichońska, B. Ravikumar, R. J. Allaway [et al.] // Nature communications. – 2021. – Vol. 12, N 1. – Art. 3307. – DOI: 10.1101/2019.12.31.891812.

117. Ketkar, N. Deep Learning with

Python / N. Ketkar. – Berkeley, CA : Apress, 2017. – P. 97–111.

REFERENCES

1. Wouters OJ, McKee M, Luyten J. Estimated research and development investment needed to bring a new medicine to market, 2009–2018. JAMA. 2020;323(9):844–53. doi: 10.1001/jama.2020.1166

2. Schenone M, Dančik V, Wagner BK, Clemons PA. Target identification and mechanism of action in chemical biology and drug discovery. Nat Chem Biol. 2013;9(4):232–40. doi: 10.1038/nchembio.1199

3. Ashenden SK. The Era of Artificial Intelligence, Machine Learning, and Data Science in the Pharmaceutical Industry, Cambridge, UK: Academic Press; 2021. 264 p. Chapter 6

4. Smietana K, Siatkowski M, Møller M. Trends in clinical success rates. Nat Rev Drug Discov. 2016;15(6):379–80. doi: 10.1038/nrd.2016.85

5. Harrison RK. Phase II and phase III failures: 2013–2015. Nat Rev Drug Discov. 2016;15(12):817–8. doi: 10.1038/nrd.2016.184

6. Dowden H, Munro J. Trends in clinical success rates and therapeutic focus. Nat Rev Drug Discov. 2019;18(7):495–6. doi: 10.1038/d41573-019-00074-z

7. Janai J, Güney F, Behl A, Geiger A. Computer vision for autonomous vehicles: problems, datasets and state of the art. Foundations and Trends in Computer Graphics and Vision. 2020;12(1–3):1–308. doi: 10.1561/0600000079

8. Goldberg B, Flemotomos N, Martinez VR, Tanana MJ, Kuo PB, Pace BT, et al. Machine learning and natural language processing in psychotherapy research: alliance as example use case. J Couns Psychol. 2020;67(4):438–48. doi: 10.1037/cou0000382

9. Peterson AA, Liu DR. Small-molecule discovery through DNA-encoded libraries. Nat Rev Drug Discov. 2023;22(9):699–722. doi: 10.1038/s41573-023-00713-6

10. Lim KS, Reidenbach AG, Hua BK, Mason JW, Gerry CJ, Clemans PA, et al. Machine learning on DNA-encoded library count data using an uncertainty-aware probabilistic loss function. J Chem Inf Model. 2022;62(10):2316–31. doi: 10.1021/acs.jcim.2c00041

11. Hou R, Xie C, Gui Y, Li G, Li X. Machine-learning-based data analysis method for cell-based selection of DNA-encoded libraries. ACS Omega. 2023:8(21):19057–71. doi: 10.1021/acsomega.3c02152

12. Van de Sande B, Lee JS, Mutasa-Gottgens E, Naughton B, Bacon W, Manning J, et al. Applications of single-cell RNA sequencing in drug discovery and development. Nat Rev Drug Discov. 2023;22(6):496–520. doi: 10.1038/s41573-023-00688-4

13. Yang F, Wang W, Wang F, Fang Y, Tang D, Huang J, et al. scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. Nat Mach Intell. 2022;4(10):852–66. doi: 10.1038/s42256-022-00534-z

14. Chen J, Wang X, Ma A, Wang QE, Liu B, Li L, et al. Deep transfer learning of cancer drug responses by integrating bulk and single-cell RNA-seq data. Nat Commun. 2022;13(1 Art 6494). doi: 10.1038/s41467-022-34277-7

15. Godinez WJ, Hossain I, Lazic SE, Davies JW, Zhang X. A multi-scale convolutional neural network for phenotyping high-content cellular images. Bioinformatics. 2017;33(13):2010–9. doi: 10.1093/bioinformatics/btx069

16. Stokes JM, Yang K, Swanson K, Jin W, Cubillos-Ruiz A, Donghia NM, et al. A deep learning approach to antibiotic discovery. Cell. 2020;180(4):688–702. doi: 10.1016/j.cell.2020.01.021

17. Coley CW, Barzilay R, Green WH, Jaakkola TS, Jensen KF. Convolutional embedding of attributed molecular graphs for physical property prediction. J Chem Inf Model. 2017;57(8):1757–72. doi: 10.1021/acs.jcim.6b00601

18. Jin W, Stokes JM, Eastman RT, Itkin Z, Zakharov AV, Collins JJ, et al. Deep learning identifies synergistic drug combinations for treating COVID-19. Proc Natl Acad Sci USA. 2021;118(39 Art e2105070118). doi: 10.1073/pnas.2105070118

19. Vamathevan J, Clark D, Czodrowski P, Dunham I, Ferran E, Lee G, et al. Applications of machine learning in drug discovery and development. Nat Rev Drug Discov. 2019;18(6):463–77. doi: 10.1038/s41573-019-0024-5

20. Rogers D, Hahn M. Extended-connectivity fingerprints. J Chem Inf Model. 2010;50(5):742–54. doi: 10.1021/ci100050t

21. Fernández-De Gortari E, García-Jacas CR, Martinez-Mayorga K, Medina-Franco JL. Database fingerprint (DFP): an approach to represent molecular databases. J Cheminform. 2017;9(1 Art 9). doi: 10.1186/s13321-017-0195-1

22. Liu G, Catacutan DB, Rathod K, Swanson K, Jin W, Mohammed JC, et al. Deep learning-guided discovery of an antibiotic targeting Acinetobacter baumannii. Nat Chem Biol. 2023;19(11):1342–50. doi: 10.1038/s41589-023-01349-8

23. Yang K, Swanson K, Jin W, Coley C, Eiden P, Gao H, et al. Analyzing learned molecular representations for property prediction. J Chem Inf Model. 2019;59(8):3370–88. doi: 10.1021/acs.jcim.9b00237

24. Corsello SM, Bittker JA, Liu Z, Gould J, McCarren P, Hirschman JE, et al. The Drug Repurposing Hub: a next-generation drug library and

information resource. Nat Med. 2017;23(4):405–8. doi: 10.1038/nm.4306

25. Wong F, Zheng EJ, Valeri JA, Donghia NM, Anahtar MN, Omori S, et al. Discovery of a structural class of antibiotics with explainable deep learning. Nature. 2024;626(7997):177–85. doi: 10.1038/s41586-023-06887-8

26. Bender BJ, Gahbauer S, Luttens A, Lyu J, Webb CM, Stein RM, et al. A practical guide to large-scale docking. Nat Protoc. 2021;16(10):4799–832. doi: 10.1038/s41596-021-00597-z

27. Gentile F, Yaacoub JC, Gleave J, Fernandez M, Ton AT, Ban F, et al. Artificial intelligence-enabled virtual screening of ultra-large chemical libraries with deep docking. Nat Protoc. 2022;17(3):672–97. doi: 10.1038/s41596-021-00659-2

28. Tropsha A, Isayev O, Varnek A, Schneider G, Cherkasov A. Integrating QSAR modelling and deep learning in drug discovery: the emergence of deep QSAR. Nat Rev Drug Discov. 2024;23(2):141–55. doi: 10.1038/s41573-023-00832-0

29. Acharya A, Agarwal R, Baker MB, Baudry J, Bhowmik D, Boehm S, et al. Supercomputer-based ensemble docking drug discovery pipeline with application to Covid-19. J Chem Inf Model. 2020;60(12):5832–52. doi: 10.1021/acs.jcim.0c01010

30. Muratov EN, Amaro R, Andrade CH, Brown N, Ekins S, Fourches D, et al. A critical overview of computational approaches employed for COVID-19 drug discovery. Chem Soc Rev. 2021;50(16):9121–51. doi: 10.1039/d0cs01065k

31. Sterling T, Irwin JJ. ZINC 15 – ligand discovery for everyone. J Chem Inf Model. 2015;55(11):2324–37. doi: 10.1021/acs.jcim.5b00559

32. Rossetti GG, Ossorio MA, Rempel S, Kratzel A, Dionellis VS, Barriot S, et al. Non-covalent SARS-CoV-2 Mpro inhibitors developed from in silico screen hits. Sci Rep. 2022;12(1 Art 2505. doi: 10.1038/s41598-022-06306-4

33. Reymond JL. The chemical space project. Acc Chem Res. 2015;48(3):722–30. doi: 10.1021/ar500432k

34. Gómez-Bombarelli R, Wei JN, Duvenaud D, Hernandez-Lobato JM, Sanchez-Lengeling B, Sheberla D, et al. Automatic chemical design using a data-driven continuous representation of molecules. ACS Cent Sci. 2018;4(2):268–76. doi: 10.1021/acscentsci.7b00572

35. Anstine DM, Isayev O. Generative models as an emerging paradigm in the chemical sciences. J Am Chem Soc. 2023;145(16):8736–50. doi: 10.1021/jacs.2c13467

36. Jin W, Barzilay R, Jaakkola T. Junction tree variational autoencoder for molecular graph generation. arXiv. 2018 Feb 18. URL: https://arxiv.org/abs/1802.04364 (date of access: 16.03.2026)

37. Godinez WJ, Ma EJ, Chao AT, Pei L, Skewes-Cox P, Canham SM, et al. Design of potent antimalarials with generative chemistry. Nat Mach Intell. 2022;4(2):180–6. doi: 10.1038/s42256-022-00448-w

38. Walters WP, Murcko M. Assessing the impact of generative AI on medicinal chemistry. Nat Biotechnol. 2020;38(2):143–5. doi: 10.1038/s41587-020-0418-2

39. Cesaro A, Bagheri M, Torres M, Wan F, de la Fuente-Nunez C. Deep learning tools to accelerate antibiotic discovery. Expert Opin Drug Discov. 2023;18(11):1245–57. doi: 10.1080/17460441.2023.2250721

40. Rezende DJ, Mohamed S. Variational inference with normalizing flows. In: Bach F, Blei D, editors. International Conference on Machine Learning: proceedings of the 32nd International Conference on Machine Learning. 2015. vol. 37. p. 1530–8

41. Shekhovtsov A, Schlesinger D, Flach B. VAE approximation error: ELBO and exponential families. arXiv. 2021 Feb 18. URL: https://arxiv.org/abs/2102.09310 (date of access: 16.03.2026)

42. Shi C, Xu M, Zhu Z, Zhang W, Zhang M, Tang J. GraphAF: a flow-based autoregressive model for molecular graph generation. arXiv. 2020 Jan 27. URL: https://arxiv.org/abs/2001.09382 (date of access: 16.03.2026)

43. Hoogeboom E, Satorras VG, Vignac C, Welling M. Equivariant diffusion for molecule generation in 3D. In: International Conference on Machine Learning: proceedings of the 39th International Conference on Machine Learning. 2022. vol. 162. p. 8867–87

44. Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J Chem Inf Comput Sci. 1988;28(1):31–6. doi: 10.1021/ci00057a005

45. Grisoni F. Chemical language models for de novo drug design: challenges and opportunities. Curr Opin Struct Biol. 2023;79(Art 102527). doi: 10.1016/j.sbi.2023.102527

46. Flam-Shepherd D, Zhu K, Aspuru-Guzik A. Language models can learn complex molecular distributions. Nat Commun. 2022;13(1 Art 3293). doi: 10.1038/s41467-022-30839-x

47. Skinnider MA, Greg Stacey R, Wishart DS, Foster LJ. Chemical language models enable navigation in sparsely populated chemical space. Nat Mach Intell. 2021;3(9):759–70. doi: 10.1038/s42256-021-00368-1

48. Moret M, Friedrich L, Grisoni F, Merk D, Schneider G. Generative molecular design in low data regimes. Nat Mach Intell. 2020;2(3):171–80. doi: 10.1038/s42256-020-0160-y

49. Ballarotto M, Willems S, Stiller T, Nawa F, Marschner JA, Grisoni F, et al. De novo design of Nurr1 agonists via fragment-augmented gen-

erative deep learning in low-data regime. J Med Chem. 2023;66(12):8170–7. doi: 10.1021/acs.jmedchem.3c00485

50. Moret M, Angona IP, Cotos L, Yan S, Atz K, Brunner C, et al. Leveraging molecular structure and bioactivity with chemical language models for de novo drug design. Nat Commun. 2023;14(1 Art 114). doi: 10.1038/s41467-022-35692-6

51. Grisoni F, Huisman BJH, Button AL, Moret M, Atz K, Merk D, et al. Combining generative artificial intelligence and on-chip synthesis for de novo drug design. Sci Adv. 2021;7(24 Art eabg3338). doi: 10.1126/sciadv.abg3338

52. Merk D, Friedrich L, Grisoni F, Schneider G. De novo design of bioactive small molecules by artificial intelligence. Mol Inform. 2018;37(1-2 Art 1700153). doi: 10.1002/minf.201700153

53. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. arXiv. 2017 June 12. URL: https://arxiv.org/abs/1706.03762 (date of access: 16.03.2026)

54. Bagal V, Aggarwal R, Vinod PK, Priyakumar UD. MolGPT: molecular generation using a transformer-decoder model. J Chem Inf Model. 2021;62(9):2064–76. doi: 10.1021/acs.jcim.1c00600

55. Brown N, Fiscato M, Segler MHS, Vaucher AC. GuacaMol: benchmarking models for de novo molecular design. J Chem Inf Model. 2019;59(3):1096–108. doi: 10.1021/acs.jcim.8b00839

56. Polykovskiy D, Zhebrak A, Sanchez-Lengeling B, Golovanov S, Tatanov O, Belyaev S, et al. Molecular sets (MOSES): a benchmarking platform for molecular generation models. Front Pharmacol. 2020;11(Art 565644). doi: 10.3389/fphar.2020.565644

57. Boiko DA, MacKnight R, Kline B, Gomes G. Autonomous chemical research with large language models. Nature. 2023;624:570–8. doi: 10.1038/s41586-023-06792-0

58. Jablonka KM, Schwaller P, Ortega-Guerrero A, Smit B. Leveraging large language models for predictive chemistry. Nat Mach Intell. 2024;6(2):161–9. doi: 10.1038/s42256-023-00788-1

59. Born J, Manica M. Regression Transformer enables concurrent sequence regression and generation for molecular language modeling. Nat Mach Intell. 2023;5(4):432–44. doi: 10.1038/s42256-023-00639-z

60. Frey NC, Soklaski R, Axelrod S, Samsi S, G'omez-Bombarelli R, Coley CW, et al. Neural scaling of deep chemical models. Nat Mach Intell. 2023;5(11):1297–305. doi: 10.1038/s42256-023-00740-3

61. Grechishnikova D. Transformer neural network for protein-specific de novo drug generation as a machine translation problem. Sci Rep. 2021;11(1 Art 321). doi: 10.1038/s41598-020-79682-4

62. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596(7873):583–9. doi: 10.1038/s41586-021-03819-2

63. Stsiapanava A, Xu C, Nishio S, Han L, Yamakawa N, Carroni M, et al. Structure of the decoy module of human glycoprotein 2 and uromodulin and its interaction with bacterial adhesin FimH. Nat Struct Mol Biol. 2022;29(3):190–3. doi: 10.1038/s41594-022-00729-3

64. Liu H, Hong X, Xi J, Menne S, Hu J, Wang JC. Cryo-EM structures of human hepatitis B and woodchuck hepatitis virus small spherical subviral particles. Sci Adv. 2022;8(31 Art eabo4184). doi: 10.1126/sciadv.abo4184

65. Ren F, Ding X, Zheng M, Korzinkin M, Cai X, Zhu W, et al. AlphaFold accelerates artificial intelligence powered drug discovery: efficient discovery of a novel CDK20 small molecule inhibitor. Chem Sci. 2023;14(6):1443–52. doi: 10.48550/arXiv.2201.09647

66. Yang Q, Lian X, Syed AAS, Fahira A, Zheng C, Zhu Z, et al. Structural comparison and drug screening of spike proteins of ten SARS-CoV-2 variants. Research (Wash D C). 2022;2022(Art 9781758). doi: 10.34133/2022/9781758

67. Yang Q, Xia D, Syed AAS, Wang Z, Shi Y. Highly accurate protein structure prediction and drug screen of monkeypox virus proteome. J Infect. 2023;86(1):66–117. doi: 10.1016/j.jinf.2022.08.006

68. Ivanenkov YA, Polykovskiy D, Bezrukov D, Zagribelnyy B, Aladinskiy V, Kamya P, et al. Chemistry42: an AI-driven platform for molecular design and optimization. J Chem Inf Model. 2023;63(3):695–701. doi: 10.1021/acs.jcim.2c01191

69. Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, et al. Accurate prediction of protein structures and interactions using a three-track neural network. Science. 2021;373(6557):871–6. doi: 10.1126/science.abj8754

70. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TM, Weissig H, et al. The Protein Data Bank. Nucleic Acids Res. 2000;28(1):235–42. doi: 10.1093/nar/28.1.235

71. Van Wart HE, Birkedal-Hansen H. The cysteine switch: a principle of regulation of metalloproteinase activity with potential applicability to the entire matrix metalloproteinase gene family. Proc Natl Acad Sci U S A. 1990;87(14):5578–82. doi: 10.1073/pnas.87.14.5578

72. Michaud JM, Madani A, Fraser JS. A language model beats AlphaFold2 on orphans. Nat Biotechnol. 2022;40(11):1576–7. doi: 10.1038/s41587-022-01466-0

73. Wu R, Ding F, Wang R, Shen R, Zhang X, Luo S, et al. High-resolution de novo structure prediction from primary sequence. bioRxiv. 2022 July 21. doi:10.1101/2022.07.21.500999

74. Fang X, Wang F, Liu L, He J, Lin D, Xiang Y, et al. A method for multiple-sequence-alignment-free protein structure prediction using a protein language model. Nat Mach Intell. 2023;5(10):1087–96. doi: 10.1038/s42256-023-00721-6

75. Madani A, Krause B, Greene ER, Subramanian S, Mohr BP, Holton JM, et al. Large language models generate functional protein sequences across diverse families. Nat Biotechnol. 2023;41(8):1099–106. doi: 10.1038/s41587-022-01618-2

76. Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science. 2023;379(6637):1123–30. doi: 10.1126/science.ade2574

77. Suzek BE, Wang Y, Huang H, McGarvey PB, Wu CH. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics. 2015;31(6):926–32. doi: 10.1093/bioinformatics/btu739

78. Corso G, Stärk H, Ling B, Barzilay R, Jaakkola T. DiffDock: diffusion steps, twists, and turns for molecular docking. arXiv. 2022 Oct 4. URL: https://arxiv.org/abs/2210.01776 (date of access: 16.03.2026)

79. Watson JL, Juergens D, Bennett NR, Trippe BL, Yit J, Eisenach HE, et al. De novo design of protein structure and function with RFdiffusion. Nature;2023;620(7976):1089–100. doi: 10.1038/s41586-023-06415-8

80. Theodoris CV, Xiao L, Chopra A, Chaffin MD, Al Sayed ZR, Hill MC, et al. Transfer learning enables predictions in network biology. Nature. 2023;618(7965):616–24. doi: 10.1038/s41586-023-06139-9

81. Chaffin M, Papangeli I, Simonson B, Akkad AD, Hill MC, Arduini A, et al. Single-nucleus profiling of human dilated and hypertrophic cardiomyopathy. Nature. 2022;608(7921):174–80. doi: 10.1038/s41586-022-04817-8

82. Hughes JP, Rees SS, Kalindjian SB, Philpott KL. Principles of early drug discovery. Br J Pharmacol. 2011;162(6):1239–49. doi: 10.1111/j.1476-5381.2010.01127.x

83. Goodnow RA. Hit and lead identification: integrated technology-based approaches. Drug Discov Today Technol. 2006;3(4):367–75. doi: 10.1016/j.ddtec.2006.12.009

84. Yang L, Jin C, Yang G, Bing Z, Huang L, Niu Y, et al. Transformer-based deep learning method for optimizing ADMET properties of lead compounds. Phys Chem Chem Phys. 2023;25(3):2377–85. doi: 10.1039/d2cp05332b

85. Chen Y, Yu X, Li W, Tang Y, Liu G. In silico prediction of hERG blockers using machine learning and deep learning approaches. J Appl Toxicol. 2023;43(10):1462–75. doi: 10.1002/jat.4477

86. Sharma B, Chenthamarakshan V, Dhurandhar A, Pereira S, Hendler JA, Dordick JS, et al. Accurate clinical toxicity prediction using multi-task deep neural nets and contrastive molecular explanations. Sci Rep. 2023;13(1 Art 4908). doi: 10.1038/s41598-023-31169-8

87. Sun D, Gao W, Hu H, Zhou S. Why 90% of clinical drug development fails and how to improve it? Acta Pharm Sin B. 2022;12(7):3049–62. doi: 10.1016/j.apsb.2022.02.002

88. Kola I, Landis J. Can the pharmaceutical industry reduce attrition rates? Nat Rev Drug Discov. 2004;3(8):711–6. doi: 10.1038/nrd1470

89. Lipinski C. A. Lead- and drug-like compounds: the rule-of-five revolution. Drug Discov Today Technol. 2004;1(4):337–41. doi: 10.1016/j.ddtec.2004.11.007

90. Coutinho AL, Gristofoletti R, Wu F, Shoyaib AA, Dressman J, Polli JE. A robust, viable, and resource sparing HPLC-based log P method applied to common drugs. Int J Pharm. 2023;644(Art 123325). doi: 10.1016/j.ijpharm.2023.123325

91. Faller B, Ertl P. Computational approaches to determine drug solubility. Adv Drug Deliv Rev. 2007;59(7):533–45. doi: 10.1016/j.addr.2007.05.005

92. Aliagas I, Gobbi A, Lee ML, Sellers BD. Comparison of log P and log D correction models trained with public and proprietary data sets. J Comput Aided Mol Des. 2022;36(4):253–62. doi: 10.1007/s10822-022-00450-9

93. Win ZM, Cheong AMY, Hopkins WS. Using machine learning to predict partition coefficient (log P) and distribution coefficient (log D) with molecular descriptors and liquid chromatography retention time. J Chem Inf Model. 2023;63(7):1906–13. doi: 10.1021/acs.jcim.2c01373

94. Domingo-Almenara X, Guijas C, Billings E, Montenegro-Burke JR, Uritboonthai W, Aisporna AE, et al. The METLIN small molecule dataset for machine learning-based retention time prediction. Nat Commun. 2019;10(1 Art 5811). doi: 10.1038/s41467-019-13680-7

95. Gaulton A, Bellis LJ, Bento AP, Gambers J, Davies M, Hersey A, et al. ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res. 2012;40(Database issue):D1100–7. doi: 10.1093/nar/gkr777

96. Datta R, Das D, Das S. Efficient lipophilicity prediction of molecules employing deep-learning models. Chemometr Intell Lab Syst. 2021;213(Art 104309). doi: 10.1016/j.chemolab.2021.104309

97. Prasad S, Brooks BR. A deep learning approach for the blind log P prediction in SAMPL6 challenge. J Comput Aided Mol Des.

2020;34(5):535–42. doi: 10.1007/s10822-020-00292-3

98. Heijman J, Voigt N, Carlsson LG, Dobrev D. Cardiac safety assays. Curr Opin Pharmacol. 2014;15:16–21. doi: 10.1016/j.coph.2013.11.004

99. Ackloo S, Al-Awar R, Amaro RE, Arrowsmith CH, Azevedo H, Batey RA, et al. CACHE (Critical Assessment of Computational Hit-finding Experiments): a public-private partnership benchmarking initiative to enable the development of computational methods for hit-finding. Nat Rev Chem. 2022;6(4):287–95. doi: 10.1038/s41570-022-00363-z

100. Swanson K, Walther P, Leitz J, Mukherjee S, Wu JC, Shivnaraine RV, et al. ADMET-AI: a machine learning ADMET platform for evaluation of large-scale chemical libraries. Zenodo. 2023 Dec 9. doi: 10.5281/zenodo.10372930

101. Wu Z, Ramsundar B, Feinberg EN, Gomes J, Geniesse C, Pappu AS, et al. MoleculeNet: a benchmark for molecular machine learning. Chem Sci. 2017;9(2):513–30. doi: 10.1039/c7sc02664a

102. Huang R, Xia M, Nguyen DT, Zhao T, Sakamuru S, Zhao J, et al. Tox21Challenge to build predictive models of nuclear receptor and stress response pathways as mediated by exposure to environmental chemicals and drugs. Front Environ Sci. 2016;3(Art 85). doi: 10.3389/fenvs.2015.00085

103. Tingle BI, Tang KG, Castanon M, Gutierrez JJ, Khurelbaatar M, Dandarchuluun C, et al. ZINC-22 – a free multi-billion-scale database of tangible compounds for ligand discovery. J Chem Inf Model. 2023;63(4):1166–76. doi: 10.1021/acs.jcim.2c01253

104. Ramakrishnan R, Dral PO, Rupp M, von Lilienfeld OA. Quantum chemistry structures and properties of 134 kilo molecules. Sci Data. 2014;1(Art 140022). doi: 10.1038/sdata.2014.22

105. Frye L, Bhat S, Akinsanya K, Abel R. From computer-aided drug discovery to computer-driven drug discovery. Drug Discov Today Technol. 2021;39:111–7. doi: 10.1016/j.ddtec.2021.08.001

106. Zeng W, Guo L, Xu S, Chen J, Zhou J. High-throughput screening technology in industrial biotechnology. Trends Biotechnol. 2020;38(8):888–906. doi: 10.1016/j.tibtech.2020.01.001

107. Sarkar N, Stokes JM. Practical applications of machine learning for anti-infective drug discovery. In: Rudolph J, Bronson JJ, editors. Medicinal Chemistry Reviews. Washington, USA: ACS; 2023. vol. 58. p. 345–75

108. Arnold A, Alexander J, Liu G, Stokes JM. Applications of machine learning in microbial natural product drug discovery. Expert Opin Drug Discov. 2023;18(11):1259–72. doi: 10.1080/17460441.2023.2251400

109. Mullowney MW, Duncan KR, Elsayed SS, Garg N, van der Hooft JJJ, Martin NI, et al. Artificial intelligence for natural product drug discovery. Nat Rev Drug Discov. 2023;22(11):895–916. doi: 10.1038/s41573-023-00774-7

110. Ekins S, Puhl AC, Zorn KM, Lane TR, Russo DP, Klein JJ, et al. Exploiting machine learning for end-to-end drug discovery and development. Nat Mater. 2019;18(5):435–41. doi: 10.1038/s41563-019-0338-z

111. Grisoni F, Neuhaus CS, Gabernet G, Muller AT, Hiss JA, Schneider G. Designing anticancer peptides by constructive machine learning. ChemMedChem. 2018;13(13):1300–2. doi: 10.1002/cmdc.201800204

112. Chen J, Cheong HH, S. Siu SWI. xDeepAcPEP: deep learning method for anticancer peptide activity prediction based on convolutional neural network and multitask learning. J Chem Inf Model. 2021;61(8):3789–803. doi: 10.1021/acs.jcim.1c00181

113. Walker AS, Clardy J. A machine learning bioinformatics method to predict biological activity from biosynthetic gene clusters. J Chem Inf Model. 2021;61(6):2560–71. doi: 10.1021/acs.jcim.0c01304

114. Heyndrickx W, Mervin L, Morawietz T, Sturm N, Friedrich L, Zalewski A, et al. MELLODDY: cross-pharma federated learning at unprecedented scale unlocks benefits in QSAR without compromising proprietary information. J Chem Inf Model. 2024;64(7):2331–44. doi: 10.1021/acs.jcim.3c00799

115. Wellawatte GP, Gandhi HA, Seshadri A, White AD. A perspective on explanations of molecular prediction models. J Chem Theory Comput. 2023;19(8):2149–60. doi: 10.1021/acs.jctc.2c01235

116. Cichońska A, Ravikumar B, Allaway RJ, Wan F, Park S, Isayev O, et al. Crowdsourced mapping of unexplored target space of kinase inhibitors. Nat Commun. 2021;12(1 Art 3307). doi: 10.1101/2019.12.31.891812

117. Ketkar N. Deep Learning with Python. Berkeley, CA: Apress; 2017. p. 97–111

Адрес для корреспонденции:
210009, Республика Беларусь,
г. Витебск, пр. Фрунзе, 27,
УО «Витебский государственный ордена Дружбы народов медицинский университет»,
кафедра фармацевтической и токсикологической химии,
тем. моб.: +375 29 711 50 83,
chemistryandron@mail.ru
Дергачёва Ж. М.

Поступила 23.03.2026 г.