УДК 004.94+547.7::616-006+616-085

ГЕНЕРАТИВНЫЙ ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ ДЛЯ РАЗРАБОТКИ ЛЕКАРСТВ DE NOVO: НОВЫЕ РУБЕЖИ В ОБЛАСТИ МОЛЕКУЛ

Адылова Ф.Т., Давронов Р.Р.

*fatadilova@gmail.com

Институт математики им. В.И. Романовского АН РУз,

100174, Узбекистан, г. Ташкент, ул. Университетская, д. 9.

Методы, основанные на искусственном интеллекте, могут значительно улучшить традиционный дорогостоящий процесс разработки лекарств, учитывая тот факт, что различные генеративные модели уже широко используются в химии. Генеративные модели для разработки лекарств de novo, сосредоточены на создании новых биологических соединений полностью с нуля, что представляет собой многообещающее направление в будущем. Быстрое развитие в этой области в сочетании с присущей процессу разработки лекарств сложностью создает непростые условия для исследователей. В рамках темы создания малых молекул мы определяем множество подзадач и приложений, выделяя важные наборы данных, контрольные показатели, архитектуру моделей и сравниваем производительность лучших моделей. В обзоре представлены ключевые достижения в этой области, включая появление квантовых вычислений, которые обещают дальнейшее ускорение применения глубокого QSAR для поддержки автоматизированного проектирования лекарств в области молекул.

Ключевые слова: генеративные модели, биологические соединения, малые молекулы, наборы данных, контрольные показатели, архитектура моделей, квантовые вычисления, QSAR.

Цитирование: *Адылова Ф.Т., Давронов Р.Р.* Генеративный искусственный интеллект для разработки лекарств de novo: новые рубежи в области молекул // Проблемы вычислительной и прикладной математики. – 2024. – №2(56). – С. 85-98.

1 Введение

Мотивацией и материалом к написанию данного обзора послужили два исследования [1, 2], любезно присланные мне авторами, за что ещё раз выражаю им свою благодарность.

Современные подходы к вычислению количественного отношения структуры молекулы и её активности, известного как QSAR-моделирование, можно описать как применение методов статистики и машинного обучения для нахождения эмпирических зависимостей вида $A_i = k(D_1, D_2, ..., D_n)$, где A_i – это биологическая активность (или другие представляющие интерес свойства) молекул, структурные свойства соединений, $D_1, D_2, ..., D_n$, называемые молекулярными дескрипторами, рассчитываются или измеряются экспериментально, k – некоторое математическое преобразование, которое должно быть применено к дескрипторам для вычисления значений свойств всех молекул, для которых выявляется взаимосвязь.

В зависимости от того, как характеризуется молекулярная структура, модели QSAR были классифицированы как 1D, 2D, 3D,4D (рис. 1).

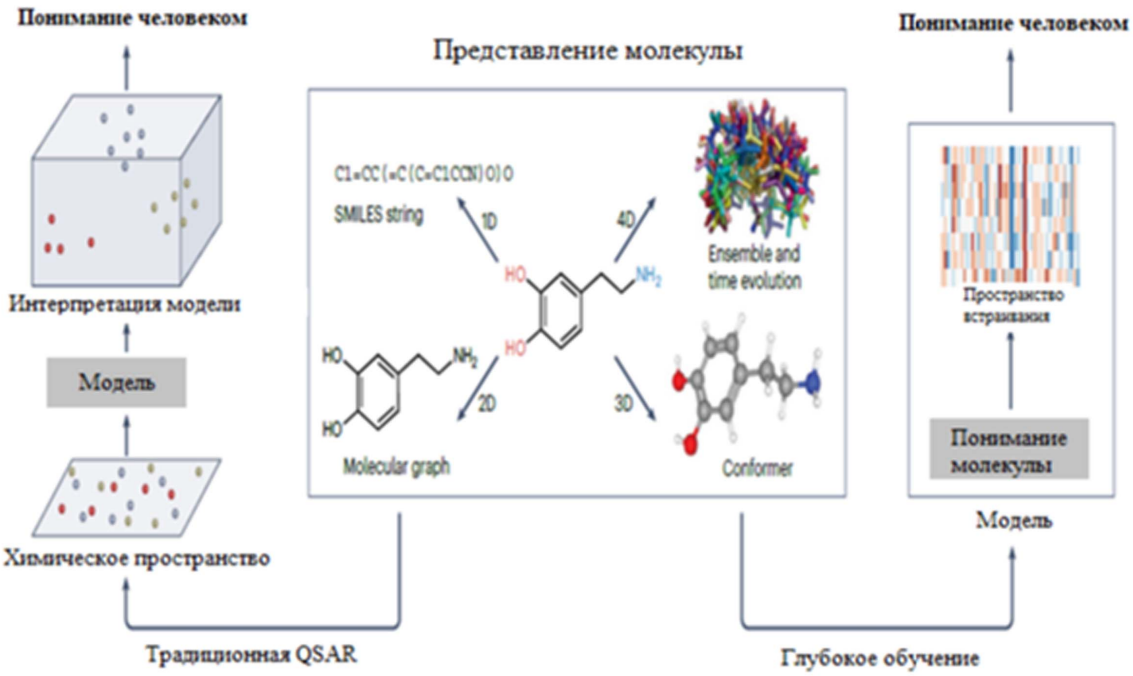

Рис. 1 Сравнение традиционной и глубокой моделей QSAR. Обе модели используют аналогичные входные данные, представляющие молекулы в 1D-4D. Традиционные методы QSAR (показаны слева) требуют вычисления явных числовых дескрипторов из соответствующих молекулярных представлений, которые используются для различных задач машинного обучения (вычисления дескрипторов и машинное обучение разделены), тогда как подходы глубокого QSAR (показаны справа) изучают молекулярные представления как часть оптимизации модели, проводимой в скрытом химическом пространстве.

С ростом объёма данных и сложности методов моделирования QSAR постепенно перешло от простой статистики к применению методов многомерного статистического анализа. Сегодня в QSAR-моделировании широко применяются методы искусственного интеллекта (ИИ), включая наиболее мощные из них, – языковые модели (LLM).

Разработка новых лекарственных препаратов (дизайн de novo) исследует неизвестное химическое пространство и генерирует с нуля кандидатов, похожих на лекарственные средства, что в сочетании с огромным количеством потенциальных соединений (до $10^{23} - 10^{60}$), делает традиционную разработку небольших лекарственных средств трудоемкой и дорогостоящей [3, 4]. При использовании традиционных методов доклинические испытания могут стоить сотни миллионов долларов и занимать от 3 до 6 лет [5].

Биотехнологические компании, ориентированные на ИИ, имеют более 150 низкомолекулярных лекарственных препаратов на стадии разработки и 15 - в фазе клинических испытаний, при этом использование процесса, основанного на ИИ, расширяется каждый год почти на 40% [6]. Актуальность темы привлекает внимание к динамике развития этого направления ИИ в химии, что отражено в недавних обзорах [7, 8].

Подход к генерации малых молекул является введением в генеративный ИИ разработки лекарств, который позволяет выделить взаимосвязи между параллельными изменениями в методах представления входных данных, определяет появление ар-

хитектур, например, эквивариантных графовых нейронных сетей (equivariant graph neural networks EGNNS), и решает другие проблемы, с которыми сталкиваются при проектировании молекул [2].

2 Методы

Основные модели генеративного ИИ в разработке лекарств

Исторически известные подходы включают в себя генеративные состязательные сети (GAN) [9], вариационные автоэнкодеры (VAE) [10] и потоковые модели [11]. Совсем недавно в качестве многообещающих альтернатив появились диффузионные модели [12].

Вариационные автоэнкодеры (VAE) – тип порождающей модели, которая расширяет типичную структуру кодера-декодера, представляя каждый скрытый атрибут с использованием распределения, а не отдельного значения. Формально VAE можно записать как:

$$q_\varphi(z|x) = \mathcal{N}(z; \mu_\varphi(x), \sigma^2_\varphi(x)I). \tag{1}$$

Интуитивно понятно, что каждый x будет сопоставлен некоторому среднему значению $\mu_\varphi(x)$ и дисперсии $\sigma^2_\varphi(x)$, которые описывают соответствующее нормальное распределение. Декодер записывается как $p_\theta(x|z)$, где z – случайно выбранная точка из скрытого распределения $N(\mu_\varphi(x), \sigma^2_\varphi(x)I)$ и отображается в точку x в процессе декодирования.

Потери VAE вычисляются с использованием двух подходов: потери при восстановлении, и потери при дивергенции Кульбака–Лейблера (KL). Потери при восстановлении измеряют разницу между исходной истиной и восстановленным выходом декодера, часто выражаемую с использованием потерь перекрестной энтропии:

$$\mathcal{L}_{\text{recon}} = -1 * \int q_\varphi(z|x) \log p_\theta(x|z) dz. \tag{2}$$

Дивергенция KL измеряет разницу между двумя распределениями вероятностей. Для VAE расхождение KL вычисляется между закодированным распределением и стандартным нормальным распределением. Это можно рассматривать как “регуляризацию”, поскольку процедура побуждает кодировщик отображать элементы в более центральную область с перекрывающимися распределениями, тем самым улучшая непрерывность во всем скрытом пространстве. Формально потеря KL может быть выражена следующим образом, где k представляет k – е измерение в скрытом пространстве:

$$\mathcal{L}_{KL} = D_{KL}(q_\varphi(z|x)||\mathcal{N}(0, I)) = -\frac{1}{2} \sum_k \left(1 + \log(\sigma^2_k) - \mu^2_k - \sigma^2_k\right). \tag{3}$$

Тогда общая функция потерь может быть записана в виде

$$\mathcal{L} = \mathcal{L}_{\text{recon}} + \beta \mathcal{L}_{KL}, \tag{4}$$

где β может быть скорректирована так, чтобы сбалансировать потери при восстановлении и потери Кульбака–Лейблера.

Генеративные состязательные сети (GAN) используют “конкурирующие” нейронные сети для взаимного улучшения. Две нейронные сети — генератор и дискриминатор — соревнуются в игре с нулевой суммой. Генератор (G) создает экземпляры (например, химические структуры потенциальных лекарств) из случайного шума

(z), отобранного из предыдущего распределения $p_z(z)$, чтобы имитировать обучающие выборки, в то время как дискриминатор (D) стремится различать синтетические данные и обучающие выборки. Процесс обучения включает в себя оптимизацию функции потерь:

$$\min_G \max_D \mathbb{E}_x[\log D(x;\theta_d)] + \mathbb{E}_{z\sim p(z)}[\log(1 - D(G(z;\theta_g);\theta_d))]. \tag{5}$$

Здесь $\mathbb{E}_x[\log D(x;\theta_d)]$ представляет вероятность, применяемую дискриминатором к правильной выборке, в то время как $\mathbb{E}_{z\sim p(z)}[\log(1 - D(G(z;\theta_g);\theta_d))]$ представляет отрицательную вероятность, применяемую дискриминатором к некорректной выборке. Эта функция даёт более высокое значение, когда дискриминатор точно классифицирует выборки; таким образом, дискриминатор стремится максимизировать эту функцию, в то время как генератор стремится её минимизировать.

Потоковые генерирующие модели [11] генерируют данные в соответствии с целевым распределением $x \sim p(x)$ путем применения цепочки преобразований к простому скрытому распределению, часто гауссову, $z_0 \sim p_0(z_0)$. Это преобразование применяет обратимую функцию $f : z_0 \to x$, такую, что:

$$x = f(z;\theta) \Rightarrow z \Rightarrow f^{-1} = f^{-1}(x;\theta), \tag{6}$$

где обученная модель запоминает параметры θ. Поскольку f обратимо и, следовательно, изученное отображение биективно, z имеет ту же размерность, что и x. Часто f является составной функцией, где $f(x) = f_K \circ ... \circ f_1(x)$ это позволяет моделировать более сложные распределения вероятностей. Поскольку каждая функция обратима, апостериорная вероятность может быть легко вычислена, т.е. логарифмическая вероятность отдельной точки x может быть записана в терминах ее скрытой переменной z:

$$\log p(x) = \log p_0(z) + \log \left| \det \frac{\partial f}{\partial z} \right|. \tag{7}$$

Эта функция используется для обучения параметров θ, чтобы максимизировать вероятность наблюдения данных. Различные модели основаны на этой предпосылке для представления сложных распределений данных и фиксации взаимосвязей в последовательных данных.

Диффузионные модели [12] выполняют фиксированную процедуру обучения, постепенно добавляя гауссовский шум к данным в течение ряда временных шагов. Определяется два этапа модели: процесс добавления шума (прямой) и процесс его удаления (обратный).

В прямом процессе каждый шаг $x_t + 1$ может быть представлен в виде цепи Маркова и состоит из $x_t - 1$ и небольшого количества гауссовского шума:

$$x_{t+1} = \sqrt{(1 - \beta_t x_t)} + \sqrt{\beta_t}\varepsilon, \quad \varepsilon \sim \mathcal{N}(0, I). \tag{8}$$

Здесь x_t – данные в момент времени t, а β_t обозначает график шума. Дисперсия β_t уменьшается в прямом процессе, так что после многих шагов мы имеем $p\left(x_t \mid x_0\right) \approx N(0,1)$. В обратном процессе цель состоит в восстановлении данных по шуму. В этом процессе изучается функция шумоподавления, часто моделируемая нейронной сетью:

$$x_{t-1} = f_\theta\left(x_t, t\right) + \sqrt{\beta_t \varepsilon}, \quad \varepsilon \sim \mathcal{N}(0, I), \tag{9}$$

где $f_\theta(x_t, t)$ – функция шумоподавления, параметризованная через θ. Для обучения диффузионной модели добавленный шум аппроксимируется на каждом шаге; функция потерь минимизирует разницу между истинным шумом и шумом, предсказанным моделью:

$$L_t = \mathbb{E}_{t\sim[1,T],x_0,\varepsilon_t}\left[\left\|\varepsilon_t - \varepsilon_\theta\left(x_t, t\right)\right\|^2\right]. \tag{10}$$

Здесь $t \sim U1, T$ означает, что временной шаг t выбирается равномерно случайным образом из множества $1, 2, ..., T$, а $\varepsilon_\theta(x_t, t)$ представляет шум, предсказанный моделью, параметризованной с помощью θ, T, – последний временной шаг модели. Как только нейронная сеть обучена, можно собрать выборку из распределения шума и повторить обратный процесс для генерации новых данных.

Диффузионные модели можно разделить на три подтипа. Вероятностные модели диффузии с шумоподавлением (Denoising diffusion probabilistic model, DDPM) [13] фокусируются на итеративном процессе шумоподавления, где нейронная сеть изучает обратную функцию, которая постепенно удаляет добавленный шум. Основанные на оценке генеративные модели (Score-based generative models, SGM) [14] вместо этого оценивают градиент (функцию оценки) распределения данных на каждом временном шаге, используя его для вычитания ошибки из текущего представления. В работе [13] демонстрируют, что SGM эквивалентен DDPM во время обучения, но при другой параметризации. Наконец, модели, основанные на стохастических дифференциальных уравнениях (SDE) [15], расширяют SGM и DDPM за счет изучения процессов диффузии в непрерывном времени. В то время как математическая основа процесса диффузии, как правило, основана на непрерывных данных, методы работы J.Austin [16] обеспечивают более плавную реализацию с использованием дискретных форм данных, таких, как, например, молекулярные графы.

Графовые нейронные сети (GNN) не являются генерирующими моделями, но представляют собой важные компоненты крупных генерирующих методов. GNN особенно важны для обработки данных, структурированных графами. Как только данные преобразованы в граф $G = (V, E)$ с узлами V и ребрами E, GNN учится выполнять эмбеддинг (вложение) посредством передачи сообщения и агрегирования. Для каждой пары узлов v_i, v_j, GNN получает "сообщение" m_{ij} на основе существующих объектов и координат на уровне l (a_{ij} обозначает запись (i, j) в матрице смежности A):

$$\mathbf{m}_{ij} = \varphi_e\left(\mathbf{h}_i^l, \mathbf{h}_j^l, \mathbf{x}_i^l, \mathbf{x}_j^l, a_{ij}\right). \tag{11}$$

Тогда сообщение, полученное любым узлом Vi, представляет собой совокупность сообщений от его соседей:

$$\mathbf{m}_i = \sum_{j\in\mathcal{N}(i)} \mathbf{m}_{ij}. \tag{12}$$

Наконец, другая сеть учится комбинировать старые вложения и позиции с сообщениями для создания новых вложений и позиций на уровне $l + 1$:

$$\mathbf{h}_i^{l+1}, \mathbf{x}_i^{l+1} = \varphi_h\left(\mathbf{h}_i^l, \mathbf{x}_i^l, \mathbf{m}_i\right). \tag{13}$$

Нейронные сети с эквивариантным графом (Equivariant Graph Neural Networks, EGNN) предназначены для генерации 3D-структур, при этом эквивариантность является полезным индуктивным отклонением, которое можно включить в модели глубокого обучения. Две 3D-структуры следует рассматривать одинаково, если они различаются только при серии поворотов, отражений и перемещений. Для некоторого условного распределения $p(y|x)$ оно эквивариантно действию поворотов и отражений, когда $p(Ry) = Rp(y)$ (альтернативно, $p(y|x) = p(Ry|Rx)$) для преобразования R. Распределение инвариантно, если $p(Ry) = p(y)$.

Satorras et al. [17] предлагают рассматривать EGNN, как простую корректировку традиционной структуры GNN для сохранения эквивариантности. Этапы окончательного агрегирования сообщений и обновления встраивания эквивалентны этапам GNN. Здесь сохраняется равная вариантность, поскольку при передаче сообщений учитывается только относительные положения/расстояния. Следовательно, вращение, трансляция или отражение всех атомов приведет к эквивалентным функциям.

Известны другие модели для конкретных приложений и задач, такие как трансформеры, модели на основе энергии (Energy-Based, EBM), BERT и т.п. Понятно, что генерируемые молекулы должны быть (1) действительными, (2) стабильными и (3) уникальными для фармацевтического применения, означающего свойство молекулы хорошо связываться с различными биологическими мишенями.

В вычислительной химии известны две области генерации моделей: область генерации молекул, не зависящих от мишени, где целью является генерирование действительных наборов молекул без учета какой-либо биологической мишени, и область генерации молекул, зависящих от мишени (генерация лиганда),где фокусируются на генерации молекул для специфических белковых структур. В этих областях для обучения и тестирования моделей входные данные молекул могут быть отформатированы разными способами, в зависимости от доступной информации или желаемого результата. Молекулы могут быть выражены в формате 1D с помощью упрощенной системы линейного ввода молекулярных данных (SMILES), в 2D с использованием графов связности для представления связей или в 3D с использованием встраивания облаков точек в узлы графа [18]. В данной работе мы будем рассматривать первую из обозначенных выше областей.

3 Обсуждение

Генерация молекулы, не зависящей от мишени

Задача: при отсутствии входных данных сгенерировать набор новых, валидных и стабильных молекул. Глубокое обучение может помочь в изучении абстрактных характеристик существующих действительных соединений и эффективном генерировании новых молекул с более высокой вероятностью их развитие этого направления показано на рисунке 2.

Чтобы изучить эти абстрактные ограничения, модели должны быть обучены на больших наборах существующих валидных стабильных молекул. Для этого чаще всего используют следующие наборы данных:

- QM9 [19] - Quantum Machines 9, содержит небольшие стабильные молекулы, извлеченные из более крупной базы данных GDB-17;
- GEOM-Drug [20]-геометрический ансамбль молекул, содержащий более сложные молекулы, похожие на лекарственные средства, часто используемый для тестирования масштабируемости за пределами более простых молекул QM9.

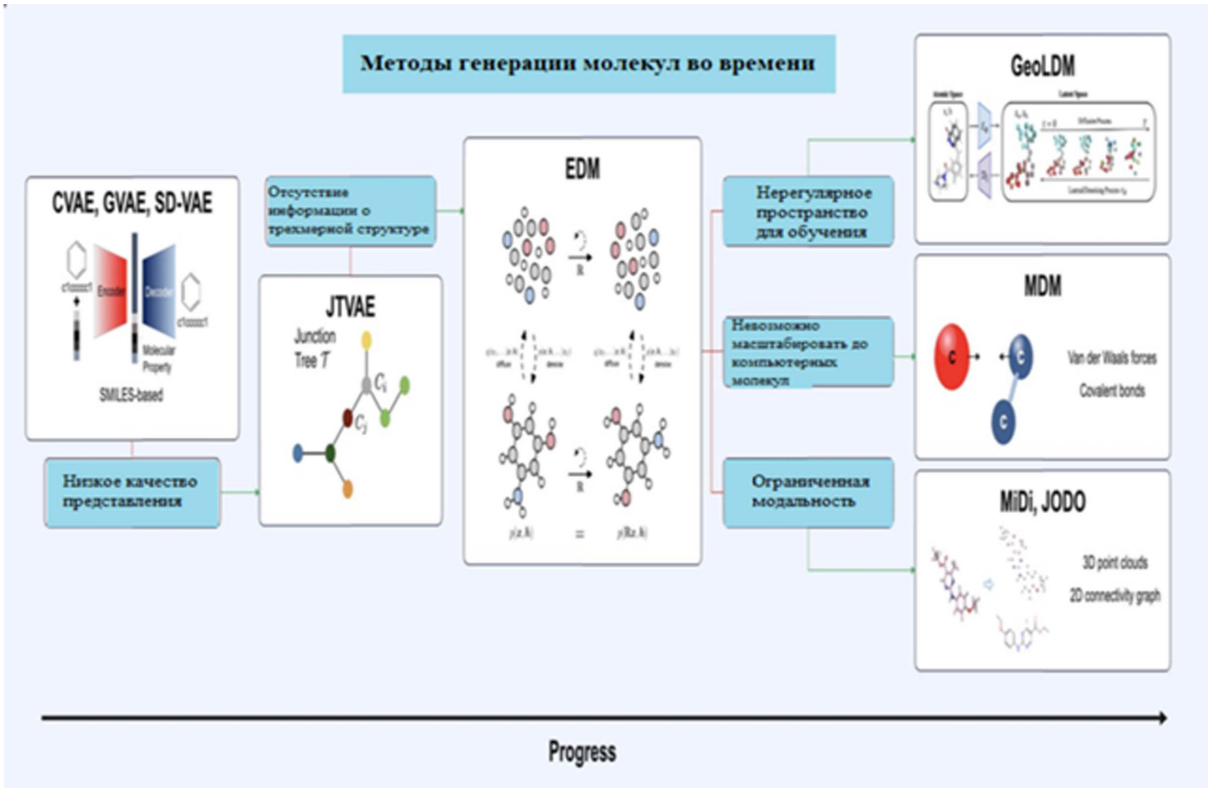

Рис. 2 Прогресс в разработке молекул, не зависящих от мишени. Недостатки предыдущих моделей выделены красным цветом, а новые модели устраняют эти недостатки за счет вариантов дизайна [2, 22, 23, 26, 28–30].

Обобщением постановки задачи является безусловная генерация молекул: после обучения на наборе существующих молекул генерируют новый набор валидных, стабильных молекул на обучающих наборах QM9 или GEOM-Drug, которые оценивают по следующим критериям:

- Стабильность атомов - процент атомов с правильной валентностью;
- Стабильность молекул - процент молекул, все атомы которых стабильны;
- Валидность - процент стабильных молекул, которые считаются валидными и часто оценивается RDKit;
- Уникальность - процент действительных молекул, которые являются уникальными (не дублируются);
- Новизна - процент молекул, не содержащихся в обучающем наборе данных;
- QED - количественная оценка сходства с лекарственным средством, комбинация различных молекулярных свойств, которые в совокупности оценивают вероятность использования молекулы в фармацевтических целях [21].

Модели часто оцениваются на основе условной генерации молекул: учитывая желаемое химическое свойство, модели генерируют молекулы, соответствующие этому свойству. Чтобы проверить эту способность, сеть классификатора свойств φ обучается на половине набора данных QM9, в то время как модель обучается на другой половине. Затем на сгенерированных моделью молекулах оценивается φ и вычисляется средняя абсолютная погрешность между целевым значением свойства и оцененным его значением. Ниже приведены шесть проверяемых в этом случае молекулярных свойств:

• α – поляризуемость, или тенденция молекулы приобретать электрический дипольный момент при воздействии внешнего электрического поля, измеряемая в кубическом радиусе Бора (Bohr3);

• eHOMO – наибольшая энергия занимаемой молекулярной орбиты, измеряемая в миллиэлектронвольтах (МэВ);

• eLUMO – энергия наименьшей незанятой молекулярной орбиты, измеряемая в (МэВ);

• $\triangle\varepsilon$ – Разница между eHOMO и eLUMO, измеряемая в (МэВ);

• μ – дипольный момент, измеренный в debyes (D);

• Cv – молярная теплоемкость при 298,15 K, измеренная в кал/моль K.

За последние несколько лет подходы к задаче генерации молекул перешли от одномерных строк данных к двумерным графам связности, а затем к трехмерным геометрическим структурам и, наконец, к включению как 2D, так и трехмерной информации [22–25].

Новая волна моделей, основанных на диффузии, работает с трехмерными облаками точек, используя преимущества эквивариантности и демонстрируя превосходную производительность.

Модель EDM [26] обеспечила начальную основу для применения диффузии, применив стандартный процесс диффузии к эквивариантной GNN с атомами, представленными в виде узлов с переменными как для скалярных объектов, так и для трехмерных координат. В то время как модели авторегрессии требуют произвольного упорядочения атомов, методы, основанные на диффузии, не являются последовательными и не нуждаются в таком упорядочении, что снижает уровень сложности и, тем самым, повышают эффективность.

Многие последующие модели сравнивали с EDM в качестве базовой линии в задаче генерации молекул, стремясь улучшить производительность путем добавления дополнительных соображений и корректировок. GCDM [27] реализует переход между геометрическим глубоким обучением и диффузией, используя геометрию полной сети персептронов для внедрения геометрической передачи сообщений на основе внимания.

Хотя EDM и GCDM уже продемонстрировали значительные улучшения производительности, обе модели по-прежнему испытывают трудности как с масштабируемостью крупных молекул, так и с разнообразием генерируемых молекул. В MDM [28] рассмотрели проблему масштабируемости, указав на отсутствие учета межатомных связей в EDM и GCDM. MDM отдельно определяет ребра графа для ковалентных связей и для ван-дер-ваальсовых сил (в зависимости от порога физического расстояния τ), чтобы обеспечить тщательный учет межатомных сил и локальных ограничений. Кроме того, в MDM решили проблему разнообразия, введя дополнительную переменную шума, контролирующую распределение на каждом этапе диффузии.

В то время как предыдущие модели диффузии работали непосредственно в пространстве сложных атомарных признаков, GeoLDM [29] применяет VAE для отображения структур молекул в скрытом пространстве меньшей размерности. Это пространство имеет более плавное распределение и меньшую размерность, что приводит к более высокой эффективности и масштабируемости больших молекул. Кроме того, здесь улучшается условная генерация, поскольку заданные химические свойства более четко определены в скрытых пространствах. В то время как предыдущие модели изучались исключительно на основе 2D или 3D-представлений, новая волна моделей признает необходимость и того, и другого: двумерная структура связи молекулы

необходима для определения типов связей и сбора информации о химических свойствах и синтезе, в то время как трехмерная конформация имеет решающее значение для ее взаимодействия и сродства к связыванию с другими молекулами. Совместно изучая и генерируя оба представления, модели могут максимизировать объем химически значимой информации и получать молекулярные образцы более высокого качества. Совместная 2D- и 3D-диффузионная модель (JODO) [30] использует геометрическое графическое представление для сбора как трехмерной пространственной информации, так и информации о связности, применяя оценки SDE к этому совместному представлению и одновременно предлагая преобразователь диффузионного графа для параметризации модели прогнозирования данных с целью предотвращения потери корреляции после независимого добавления шума к каждому из отдельных каналов.

MiDi [31] использует аналогичное графическое представление, но вместо этого применяет DDPM. В нем предлагается "смягченная" EGNN, которая улучшает классическую архитектуру EGNN, используя наблюдение о том, что трансляционная инвариантность не требуется в подпространстве с нулевым центром масс.

Как показано в таблице 1, методы, основанные на диффузии, демонстрируют значительные улучшения по сравнению с предыдущими методами, все они обеспечивают стабильность атомов более чем на 98,5%.

Таблица 1 Обзор соответствующих моделей генерации. Все показатели бенчмаркинга представлены самостоятельно и оцениваются с помощью набора данных QM9. Для моделей с несколькими вариациями была выбрана версия с наибольшей производительностью. Представляет текущее значение SOTA [2].

Модель	**Тип модели**	**Датасет**	**Схаб. атомов (%)**	**Схаб. Молекул (%)**	**Валид (%)**	**Валид. /Уник. (%)**
G-SchNet	SchNet	QM9	95.7	68.1	85.5	80.3
E-NF	EGNN-Flow	QM9	85	4.9	40.2	39.4
EDM	EGNN_-Diffusion	QM9, GEOM-Drugs	98.7	82.0	91.9	90.7
GCDM	EGNN, Diffusion	QM9	98.7	85.7	94.8	93.3
MDM	EGNN, VAE, Diffusion	QM9, EOM-Drugs	99.2	89.6	98.6	94.6
JODO	EGNN, Diffusion	QM9, EOM-Drugs	99.2	93.4	99.0	96.0
MiDi**	EGNN, Diffusion	QM9, EOM-Drugs	99.8	97.5	97.9	97.6
GerLDM***	VAE, Diffusion	QM9, EOM-Drugs	98.9	89.4	93.8	92.7

Таблица 2 Модели генерации молекул, оцененные в задаче условной генерации молекул [13].

Задача	α	$\Delta\varepsilon$	$\varepsilon_{\text{Homo}}$	$\varepsilon_{\text{LUMO}}$	μ	C_V
Единицы измерения	Bohr^3	meV	meV	meV	D	$\frac{\text{cal}}{\text{mol K}}$
EDM	2.76	655	356	584	1.111	1.101
GCDM	1.97	602	344	479	0.844	0.689
MDM	1.1591	44	19	40	1.177	1.647
GeoLDM	2.37	587	340	522	1.108	1.025

Из таблицы 2 видно, что MDM и GCDM превосходно справляются с задачами условной генерации, причем первая модель достигает наилучшей производительности в четырех из шести задач, а вторая превосходит две оставшиеся. В целом, текущие модели демонстрируют высокую производительность в наборе данных QM9, но есть возможности для улучшения при работе с более сложными молекулами, обнаруженными в наборе данных GEOM-Drugs.

Глубокое обучение и вычисления в квантовой механике Высокие требования к вычислениям в квантовой механике традиционно ограничивали их применимость в QSAR-моделировании и компьютерной разработке лекарств (Computer Aided Drug Design, CADD) в целом. Разработка быстрых, точных и универсальных приближений в квантовой механике уже давно находится в центре внимания вычислительной химии, которая недавно обогатилась за счет использования подходов глубокого обучения. Одним из таких типов моделей являются атомистические нейронные сети (neural network potentials, NNPs), исследованные в [32, 33]. NNP могут предсказывать энергии и другие свойства молекул в квантовой механике, обобщая с тем же уровнем точности, что и теория функционала плотности, на большом наборе органических молекул, при этом на шесть порядков быстрее. Активное обучение позволяет автоматически отбирать области химического пространства, где машинное обучение не позволяет точно предсказать потенциальную энергию, тем самым уменьшая размер набора данных, необходимого для обучения до 90% по сравнению с методами наивной случайной выборки [34].

Примечательно, что процесс обучения для NNP аналогичен тем, которые используются для обучения моделей глубокого QSAR, за исключением того, что целевое свойство, такое как энергия, вычисляется с использованием методов полной квантовой механики, а не измеряется, как в задаче прогнозирования биологической активности. Имеющиеся ограничения NNP были сняты в прорывной разработке первого универсального атомистического NNP для органических молекул, известного как ANI-1 [35].Результаты ANI-1 [34] приближаются к 'химической точности' (погрешность $\sim$1 ккал/моль) относительно эталонных данных квантовой механики для множества применений. Было обнаружено, что даже ранняя версия потенциала ANI-1 является более точной, чем полуэмпирические методы квантовой механики с жесткой привязкой, и при этом работает намного быстрее.

Недавно разработанная модель AIMNet, которая имеет пересмотренную на успехе ANI [36] архитектуру, была мотивирована теорией атомов в молекулах (atoms in molecules AIM), утверждающей, что функция распределения электронной плотности может быть использована для разделения молекулы на взаимодействующие атомы. Быстрые и точные NNP, которые реализованы с помощью новейшего метода

AIMNet, обещают повысить точность функций оценки и разработать новые подходы к глубокой стыковке, а также сделать методы молекулярного моделирования высокоэффективными и точными. Растет число исследований, в которых квантовые вычисления уже были оценены в случае их применения в обычном машинном обучении или приложениях глубокого обучения в рамках QSAR и CADD, включая обнаружение мишеней, сворачивание белков, характеристику целевого сайта **?**, генеративное молекулярное моделирование [38], стыковку и уточнение силового поля [39], оптимизацию хитов [40], оценку риска токсичности [41], а также сопоставление молекул, и поиск по сходству в химическом пространстве [42].

Появление гибридных архитектур CADD, интегрирующих алгоритмы моделирования больших данных в специализированное оборудование или сочетающих классическое оборудование и оборудование для конкретных задач, такое как зашумленные промежуточные и масштабные квантовые компьютеры или платформы GPU, является нарастающей тенденцией в разработке лекарств.

4 Заключение

Кратко остановимся на проблемах и будущих направлениях. В области генерации молекул есть следующие вызовы:

- Сложность - модели генерируют много допустимых и стабильных молекул при обучении на простом наборе данных QM9, но испытывают трудности при обучении на более сложном наборе данных GEOMDrugs;
- Объяснимость-все обсуждаемые методы являются стандартными и абстрактными; существующие модели не раскрывают такие аспекты, как “важные” атомы или структуры, а объяснимый ИИ в генерации молекул в целом ещё не развит.

Хотя еще слишком рано, чтобы методы глубокого обучения и, в частности, глубокий QSAR позволили разработать одобренные лекарства [43], появляется все больше свидетельств того, что эти методы ускорили стадии доклинических исследований для низкомолекулярных лекарственных препаратов-кандидатов. После 2020 года, когда Exscientia объявила, что ее первый препарат-кандидат, разработанный с помощью ИИ, вступил в первую фазу клинических испытаний, несколько компаний выступили с аналогичными объявлениями. Примечательно, что Exscientia сообщила, что на завершение этапа предварительных исследований, предшествующего испытанию, ушло всего 12 месяцев. https://www.cas.org/resources/cas-insights/drug-discovery/ai-designed-drug-candidates Аналогично, Insilico Medicine сообщила, что им потребовалось 30 месяцев, чтобы разработать новый клинический препарат против фиброза первой фазы, разработанный ИИ, начиная с обнаружения новой мишени. https://insilico.com/phase1 .

Обе компании используют подходы к глубокому обучению QSAR, как часть своих вычислительных платформ. Эти недавние успехи свидетельствуют о том, что область глубокого QSAR начинает выходить на "плато продуктивности" [44]. Продолжающаяся разработка и использование методов глубокого QSAR должны во все большей степени способствовать ускоренному открытию низкомолекулярных лекарственных препаратов-кандидатов, что может быть особенно важно перед лицом новых и часто непредсказуемых угроз, исходящих от возникающих инфекционных заболеваний, таких как COVID-19.

В целом, генеративный ИИ показал большие перспективы в области разработки лекарств, и продолжение исследований в этой области может привести к захватывающим достижениям в будущем.

Литература

[1] *Tropsha A. et al.* Integrating QSAR modelling and deep learning in drug discovery: the emergence of deep QSAR // Nature Reviews Drug Discovery. – 2024. – Vol. 23. – P. 141–155. – doi: http://dx.doi.org/10.1038/s41573-023-00832-0.

[2] *Tang X. et al.* A Survey of Generative AI for de novo Drug Design: New Frontiers in Molecule and Protein Generation // arXiv. – 2024. – https://arxiv.org/html/2402.08703v1.

[3] *Wang M. et al.* Deep learning approaches for de novo drug design: An overview // Current Opinion in Structural Biology. – 2022. – Vol. 72. – P. 135-144. – doi: http://dx.doi.org/10.1016/j.sbi.2021.10.001.

[4] *Liu X. et al.* Computational approaches for de novo drug design: past, present, and future // Methods Mol Biol.. – 2021. – Vol. 2190. – P. 139-165. – doi: http://dx.doi.org/10.1007/978-1-0716-0826-5_6.

[5] *DiMasi J.A., Grabowski H.G., Hansen R.W.* Innovation in the pharmaceutical industry: new estimates of r&d costs // Journal of health economics. – 2016. – Vol. 47. – P. 20-33.

[6] *Jayatunga M. et al.* AI in small-molecule drug discovery: a coming wave? // Nature Reviews Drug Discovery. – 2022. – Vol. 21. – P. 175-176.

[7] *Zhang M. et al.* A survey on graph diffusion models: Generative ai in science for molecule, protein and material // arXiv. – 2023. – https://arxiv.org/abs/2304.01565.

[8] *Guo Z. et al.* Diffusion models in bioinformatics: A new wave of deep learning revolution in action // arXiv. – 2023. – https://arxiv.org/abs/2302.10907.

[9] *Goodfellow I. et al.* Generative adversarial nets // arXiv. – 2014. – https://arxiv.org/abs/1406.2661.

[10] *Kingma D., Welling M.* Auto-encoding variational bayes // arXiv. – 2013. – https://arxiv.org/abs/1312.6114.

[11] *Rezende D., Mohamed Sh.* Variational inference with normalizing flows // arXiv. – 2015. – https://arxiv.org/abs/1505.05770.

[12] *Yang L. et al.* Diffusion models: A comprehensive survey of methods and applications // arXiv. – 2022. – https://arxiv.org/abs/2209.00796.

[13] *Ho J., Jain A., Abbeel P.* Denoising diffusion probabilistic models // Advances in neural information processing systems. – 2020. – Vol. 33. – P. 6840-6851.

[14] *Song Y., Ermon S.* Generative modeling by estimating gradients of the data distribution // arXiv. – 2019. – https://arxiv.org/abs/1907.05600.

[15] *Song Y. et al.* Score based generative modeling through stochastic differential equations // arXiv. – 2020. – https://arxiv.org/abs/2011.13456.

[16] *Austin J.* Structured denoising diffusion models in discrete state-spaces // Advances in Neural Information Processing Systems. – 2021. – Vol. 34. – P. 17981-17993

[17] *Satorras V.G., Hoogeboom E., Welling M.* E(n) Equivariant Graph Neural Networks // arXiv. – 2021. – https://arxiv.org/abs/2102.09844.

[18] *Tang X. et al.* Mollm: A unified language model for integrating biomedical text with 2d and 3d molecular representations // bioRxiv. – 2023. – doi: http://dx.doi.org/10.1101/2023.11.25.568656.

[19] *Ramakrishnan R. et al.* Quantum chemistry structures and properties of 134 kilo molecules // Scientific data. – 2014. – No. 1(1). – P. 1-7.

[20] *Axelrod S., Gomez-Bombarelli R.* Geom, energy-annotated molecular conformations for property prediction and molecular generation // Scientific Data. – 2022. – No. 9(1). - P. 185

[21] *Bickerton G.R. et al.* Quantifying the chemical beauty of drugs // Nature chemistry. – 2012. – No. 4(2). – P. 90–98.

[22] *G´omez-Bombarelli R. et al.* Automatic chemical design using a data-driven continuous representation of molecules // ACS central science. – 2018. – No. 4(2). – P. 268–276.

[23] *Jin W., Barzilay R., Jaakkola T.* Junction tree variational autoencoder for molecular graph generation // arXiv. – 2018. – https://arxiv.org/abs/1802.04364.

[24] *Satorras V.G. et al.* E(n) Equivariant normalizing flows // arXiv. – 2021. – https://arxiv.org/abs/2105.09016.

[25] *Gebauer N., Gastegger M., Schutt K.* Symmetry-adapted generation of 3d point sets for the targeted discovery of molecules // arXiv. – 2019. – https://arxiv.org/abs/1906.00957.

[26] *Hoogeboom E. et al.* Equivariant diffusion for molecule generation in 3d // arXiv. – 2022. – https://arxiv.org/abs/2203.17003.

[27] *Alex Morehead, Jianlin Cheng* Geometry-complete diffusion for 3d molecule generation // arXiv. – 2023. – https://arxiv.org/abs/2302.04313.

[28] *Huang L. et al.* Mdm: Molecular diffusion model for 3d molecule generation // arXiv. – 2022. – https://arxiv.org/abs/2209.05710.

[29] *Xu M. et al.* Geometric latent diffusion models for 3d molecule generation // arXiv. – 2023. – https://arxiv.org/abs/2305.01140.

[30] *Huang H. et al.* Learning joint 2d & 3d diffusion models for complete molecule generation // arXiv. – 2023. – https://arxiv.org/abs/2305.12347.

[31] *Vignac C. et al.* Midi: Mixed graph and 3d denoising diffusion for molecule generation // arXiv. – 2023. – https://arxiv.org/abs/2302.09048.

[32] *Zubatiuk T., Isayev O.* Development of multimodal machine learning potentials: toward a physics-aware artificial intelligence // Acc. Chem. Res. . – 2021. – Vol. 54. – P. 1575-1585.

[33] *Behler J.* Four generations of high-dimensional neural network potentials // Chem. Rev. . – 2021. – Vol. 121. – P. 10037-10072.

[34] *Smith J.S. et al.* Less is more: sampling chemical space with active learning // J. Chem. Phys. – 2018. – Vol. 148. – 241733.

[35] *Smith J.S., Isayev O., Roitberg A.E.* ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost // Chem. Sci. – 2017. – Vol. 8. – P. 3192-3203.

[36] *Zubatyuk R. et al.* Accurate and transferable multitask prediction of chemical properties with an atoms-in-molecules neural network // ScienceAdvances . – 2019. – Vol. 5. – doi: http://dx.doi.org/10.1126/sciadv.aav6490.

[37] *Li J. et al.* Drug discovery approaches using quantum machine learning // arXiv. – 2021. – https://arxiv.org/abs/2104.00746.

[38] *Romero J., Olson J.P., Aspuru-Guzik A.* Quantum autoencoders for efficient compression of quantum data // Quantum Sci. Technol. – 2017. – Vol. 2. – 045001.

[39] *Xiang W. et al.* Application of High Throughput Technologies in the Development of Acute Myeloid Leukemia Therapy: Challenges and Progress // Int. J. Mol. Sci.. –2022. – Vol. 23. – 2863. – doi: http://dx.doi.org/10.3390/ijms23052863.

[40] *Cavasotto C.N.* Binding free energy calculation using quantum mechanics aimed for drug lead optimization // Methods Mol. Biol. – 2020. – Vol. 2114. – P. 257-268.

[41] *Heinen, S. et al* Predicting toxicity by quantum machine learning. // J. Phys. Commun. – 2020. – Vol. 4. – 125012.

[42] *Outeiral, C. et al* The prospects of quantum computing in computational molecular biology. // Wiley Interdiscip. Rev. Comput. Mol. Sci. – 2015. – Vol. 11. – e1481.

[43] *Jayatunga M.K.P. et al.* AI in small-molecule drug discovery: a coming wave? // Nature Reviews Drug Discovery. – 2022. – Vol. 21. – P. 175-176.

[44] Innovations in and around generative AI dominate and have transformative impact / Jackie Wiles. – 2022. – URL: https://shorturl.at/qBFN3.

Поступила в редакцию 15.04.2024

UDC 004.94+547.7::616-006+616-085

GENERATIVE AI FOR DE NOVO DRUG DESIGN: NEW CHALLENGES IN MOLECULE

* ***Adilova F.T., Davronov R.R.***

*fatadilova@gmail.com

V.I. Romanovskiy Institute of Mathematics UzAS,
9, University str., Tashkent, 100174 Uzbekistan.

Artificial intelligence-based methods can significantly improve the traditional expensive drug development process, given the fact that various generative models are already widely used in chemistry. Generative models for de novo drug design are focused on creating new biological compounds completely from scratch, which represents a promising direction in the future. The rapid development in this field, combined with the inherent complexity of the drug design process, creates difficult conditions for researchers. Within the framework of the topic of creating small molecules, we define many subtasks and applications, highlighting important datasets, benchmarks, model architecture and comparing the performance of the best models. The review presents key advances in this field, including the advent of quantum computing, which promises to further accelerate the application of deep QSAR to support computer-aided drug design in the field of molecules.

Keywords: generative models, biological compounds, small molecules, datasets, benchmarks, model architecture, quantum computing, QSAR.

Citation: Adilova F.T., Davronov R.R. 2024. Generative AI for de novo drug design: new challenges in molecule. *Problems of Computational and Applied Mathematics.* 2(56): 85-98.

№ 2(56) 2024

ISSN 2181-8460

HISOBLASH VA AMALIY MATEMATIKA MUAMMOLARI

ПРОБЛЕМЫ ВЫЧИСЛИТЕЛЬНОЙ И ПРИКЛАДНОЙ МАТЕМАТИКИ

PROBLEMS OF COMPUTATIONAL AND APPLIED MATHEMATICS

RAQAMLI TEXNOLOGIYALAR VA
SUN'IY INTELLEKTNI RIVOJLANTIRISH
ILMIY-TADQIQOT INSTITUTI

ПРОБЛЕМЫ ВЫЧИСЛИТЕЛЬНОЙ И ПРИКЛАДНОЙ МАТЕМАТИКИ

№ 2(56) 2024

Журнал основан в 2015 году.
Издается 6 раз в год.

Учредитель:
Научно-исследовательский институт развития цифровых технологий и искусственного интеллекта.

Журнал зарегистрирован в Агентстве информации и массовых коммуникаций при Администрации Президента Республики Узбекистан.
Регистрационное свидетельство №0856 от 5 августа 2015 года.

ISSN 2181-8460, eISSN 2181-046X

Адрес редакции:
100125, г. Ташкент, м-в. Буз-2, 17А.
Тел.: +(99871) 263-41-98.
E-mail: journals@airi.uz.
Сайт: www.pvpm.uz.

Дизайн и компьютерная вёрстка:
Шарипов Х.Д.

Отпечатано в типографии НИИ РЦТИИ.
Подписано в печать 30.04.2024 г.
Формат 60x84 1/8. Заказ №2. Тираж 100 экз.

Содержание

Алимов Х.Т., Паровик Р.И.
Некоторые аспекты численного анализа дробной математической модели Макшерри для описания искусственной ЭКГ 7
Халджигитов А.А., Адамбаев У.Э., Джумаёзов У.З., Хасанова З.З.
Новые модельные уравнения в деформациях для анизотропных тел 17
Нуралиев Ф.М., Султанов Б.Ж., Дауытова Ж.К.
Математическое моделирование процессов деформированного состояния сетчатых пластин со сложной формой 30
Равшанов Н., Таштемирова Н., Каршиев Д.А.
Моделирование процесса распространения аэрозольных частиц в пограничном слое атмосферы с учетом их поглощения и захвата растительным покровом 41
Назирова Э.Ш., Неъматов А., Исмаилов Ш., Артикбаева Г.
Математическое моделирование фильтрации газа с учетом изменения пористости породы в зависимости от давления 58
Равшанов Н., Холматова И.И., Курбонов Н.М., Исламов Ю.Н.
Математическое моделирование процесса подземного выщелачивания с учетом изменения гидродинамических параметров пористой среды 72
Адылова Ф.Т., Давронов Р.Р.
Генеративный искусственный интеллект для разработки лекарств de novo: новые рубежи в области молекул 85
Мадатов Х.А.
Математическая модель автоматического определения несущественных слов текстов на узбекском языке 99
Сулюкова Л.Ф., Ёркулов Б.А.
Методика оценки имеющегося уровня информационной безопасности образовательной информационной системы 106