УДК 004.8

ПРИМЕНЕНИЕ ГЕНЕРАТИВНЫХ МОДЕЛЕЙ ДЛЯ ПРЕДСКАЗАНИЯ ЛЕКАРСТВЕННЫХ МОЛЕКУЛЯРНЫХ СОЕДИНЕНИЙ

Веселов Д.И., Андриянов Н.А.

ФГОБУ ВО «Финансовый университет при Правительстве Российской Федерации», Москва, e-mail: 201502@edu.fa.ru, naandriyanov@fa.ru

В статье рассматривается актуальная задача предсказания формы молекулярных соединений с лекарственными свойствами. Решение данной задачи требует применения генеративных моделей. В работе дается краткое описание и проводится исследование алгоритмов сэмплирования, вариационных автоэнкодеров (Variational Auto Encoder, VAE). Данные алгоритмы были реализованы в программной среде Python. Для обучения использовались данные, структурирующие информацию в виде строк SMILES. Это позволило рассматривать задачу как задачу генерации текста. В качестве основной метрики для сравнения алгоритмов рассмотрена метрика количественной оценки сходства с лекарственными соединениями (Quantitative Estimate of Druglikeness, QED). Наилучшие результаты при генерации в смысле метрики QED показали модели на базе сэмплирования и вариационный автоэнкодер. Несмотря на высокие метрики QED, алгоритм жадного поиска генерирует неадекватные структуры молекулярных соединений для дальнейших исследований. Другие алгоритмы показали адекватные результаты и могут быть использованы для генерации лекарственных молекулярных соединений. После генерации такие модели должны будут проверяться специалистами. В статье также были представлены непосредственные примеры реализации структур молекул, получаемые при моделировании с помощью разных алгоритмов.

Ключевые слова: вариационный автокодировщик, управляемый рекуррентный блок, GRU, VAE, генерация лекарств, QED

APPLICATION OF GENERATIVE MODELS FOR PREDICTION OF DRUG MOLECULAR COMPOUNDS

Veselov D.I., Andriyanov N.A.

Financial University under the Government of the Russian Federation, Moscow, e-mail: 201502@edu.fa.ru, naandriyanov@fa.ru

The article deals with the actual problem of predicting the shape of molecular compounds with medicinal properties. The solution of this problem requires the use of generative models. The paper gives a brief description and research of sampling algorithms, variational autoencoders (Variational Auto Encoder, VAE). These algorithms were implemented in the Python software environment. For training, data was used that structured information in the form of SMILES strings. This allowed us to consider the task as a text generation task. Quantitative Estimate of Druglikeness (QED) is considered as the main metric for comparing algorithms. The best generation results in terms of the QED metric were shown by sampling-based models and a variational autoencoder. Despite the high QED metrics, the greedy search algorithm generates inadequate structures of molecular compounds for further research. Other algorithms have shown adequate results and can be used to generate drug molecular compounds. After generation, such models will have to be checked by specialists. The article also presented direct examples of the implementation of molecular structures obtained by modeling using different algorithms.

Keywords: drug generator, variational autoencoder, gatted recurrent unit, GRU, VAE, drug generation, QED

Решения, основанные на технологиях математического моделирования, машинного и глубокого обучения, широко используются во всех сферах нашей жизни, и такие области знания, как биология, химия и медицина, здесь не исключение. Например, моделирование различных терапевтических воздействий на пациента [1–3] является важным шагом на пути к персонализированной медицине. Внедрение указанных технологий в медицину также заключается в диагностике заболеваний при помощи технологий компьютерного зрения [4–6], мониторинга состояния пациентов, предсказания течения болезней, а особый класс моделей – генеративные модели [7] – позволяют создавать новые объекты. В рамках биологии, химии и хемоинформатики такими новыми объектами и структурами являются молекулярные соединения, в частности лекарственные. В данной статье будет рассмотрена тема генерации лекарственных молекулярных соединений при помощи нейронных сетей, которые будут обучены на специальном наборе данных, представляющих из себя набор известных химических молекул.

Следует отметить, что открытие и разработка нового лекарственного средства – это чрезвычайно длительный, дорогостоящий, сложный и неэффективный процесс, который занимает в среднем 10–15 лет. Несмотря на достижения в области технологий и очень хорошее понимание биологических систем, в последние два десятилетия в фармацевтической промышленности на-

блюдается все большее снижение производительности исследований и разработок из-за растущих затрат, в то время как абсолютное число вновь одобренных лекарств постоянно сокращается из-за постоянно растущих регуляторных препятствий и возрастающих трудностей в поиске следующего препарата. Таким образом, процесс создания лекарственного препарата становится дорогостоящим и трудоемким [8], а большинство новых одобренных лекарств – низкомолекулярные препараты.

С другой стороны, успехи генеративных моделей в компьютерном зрении и обработке текстов позволяют рассчитывать на то, что их применение в химии и биологии будет способствовать ускорению процесса разработки новых лекарственных соединений. Например, трансформерная модель AlphaFold2 [9] значительно увеличила качество предсказания протеиновых структур по сравнению с известными ранее моделями. Однако такие модели имеют огромное количество параметров, что делает затруднительным понимание их работы и возможности их обучения на стандартных средствах вычислительной техники. Другие генеративные модели, такие как вариационные автоэнкодеры (VAEs) [10], генеративные состязательные сети (GANs) [7] и рекуррентные нейронные сети (RNN) [11], специально разработаны для изучения скрытых представлений молекул и генерации большого количества кандидатов на лекарства для дальнейшего скрининга.

Нейронные сети широко используются для создания миллионов *de novo* молекул в известном химическом пространстве. Эти глубокие генеративные модели обычно настраиваются с помощью LSTM или GRU, которые обучаются на специальном представлении молекул SMILES – Simplified Molecular Input Line Entry System (упрощенная система строкового представления молекулярных соединений). В исследовании [12] авторы представляют модель нейронной сети, Generative Examination Networks, основанной на двунаправленной RNN с конкатенированные подмодели для обучения и генерации молекулярных строк SMILES.

Таким образом, анализ литературы показывает, что основными генеративными алгоритмами являются модели GAN и VAE. Однако таким алгоритмам, как сэмплирование с температурой, top-k sampling [13], уделяется недостаточное внимание. В данной статье было предпринято решение воспользоваться именно такими генеративными алгоритмами и сравнить их с алгоритмами VAE и GAN.

Основными целями исследования являются:

1) моделирование лекарственных молекулярных соединений на базе искусственных нейронных сетей и проверка адекватности генерируемых моделей по метрике qed;

2) показать, что с задачей генерации потенциальных лекарственных молекулярных соединений наравне с моделями gan, vae справляются генеративные алгоритмы сэмплирования.

Материалы и методы исследования

Задача генерации лекарственных молекулярных соединений при помощи строк SMILES в терминах нейронной сети является задачей обработки естественного языка (Natural Language Processing, NLP), в которой химическое пространство молекул, кодированных SMILES строкой, является языковой моделью. То есть для решения данной задачи требуется решить задачу NLP: генерация нового текста.

В данной работе в качестве набора данных будет использоваться датасет, состоящий из 250 000 химических молекул, которые кодированы по специальным правилам. Такие закодированные молекулы называются SMILES. SMILES – это популярный метод описания молекул с помощью текстовых строк. Такое представление описывает атомы и связи молекулы одновременно точно и достаточно интуитивно понятно. Например, строка «*OCCc1c(C)[n+](cs1)Cc2cnc(C)nc2N*» описывает важный питательный элемент тиамин, также известный как витамин B1.

В качестве основных генеративных алгоритмов использовались следующие: сэмплирование с температурой; Top-K Sampling (сэмплирование K-верхних); жадный поиск (Greedy search); VAE – вариационный автокодировщик.

Рассмотрим сэмплирование с температурой [14]. Само по себе случайное сэмплирование потенциально может сгенерировать совершенно произвольное слово. Чтобы избежать данного явления, вводится понятие “temperature” (t), для увеличения вероятности получения наиболее вероятных слов. Обычно берётся диапазон $0 < t \leq 1$. Условная вероятность следующего состояния описывается выражением

$$P(x_i \mid x_1, x_2, ..., x_{i-1}) = \frac{\exp(u_i / t)}{\sum_{j=1}^{i-1} \exp(u_j / t)}, \quad (1)$$

где u – вектор, содержащий значения каждого токена в словаре.

Генерация следующего слова будет производиться с распределением *p'=softmax(log(p)/t)*. Тогда при $t = 1$ получаем $p' = p$. При больших t сэмплирование происходит равновероятно. Однако при малых t всегда выбирается самый вероятный токен. То есть семплирование с температурой – это общий вид разных видов семплирования, в разной степени учитывающих предсказания модели. Это нужно, чтобы лавировать между уверенностью модели и разнообразием. Можно поднимать температуру, чтобы генерировать более разнообразные тексты, или опускать её, чтобы генерировать тексты, в которых модель в среднем более уверена.

Fan и соавт. [15] в 2018 г. представили простую, но очень мощную модель сэмплирования, называемую top-K sampling. В выборке Top-K фильтруются K наиболее вероятных следующих слов, и сумма вероятности перераспределяется только между этими K-следующими словами.

Жадный поиск (Greedy search) выбирает токен с наибольшей вероятностью в качестве следующего слова

$$w_t = \arg\max_w P(w_i \mid w_1, w_2, ..., w_i)$$

на каждом временном шаге t.

Последним предложенным решением стало создание модели вариационного автокодировщика (VAE). Данная модель была предложена в 2013 г. в статье [16]. Данная модель учится отображать некоторый объект в заданное скрытое пространство (latent space) и обратно. Ключевое отличие VAE от обычного автокодировщика заключается в наличии вариационного вывода. Данный метод используется для аппроксимации распределений, который использует процесс оптимизации по параметрам, чтобы найти наилучшее приближение среди данного семейства распределений. Структурно модель VAE состоит из следующих категорий: кодировщик, скрытое пространство, декодировщик. Модель VAE обладает уникальным свойством: скрытое пространство является непрерывным. Данное свойство помогает выполнять случайные преобразования и интерполяцию. Непрерывность достигается тем, что на выходе из кодировщика появляется два вектора: вектор средних значений и вектор стандартных отклонений.

Вариационные автокодировщики не лишены проблем и минусов. Первой является то, что для обучения двух нейронных сетей с помощью алгоритма обратного распространения ошибок нужно контролировать все этапы обучения распространения ошибки. Поскольку декодер не является детерминированным (оценка его вывода требует оценки по многомерному гауссову распределению). Второй проблемой является требовательность к ресурсам для обучения. Третьей проблемой вариационных автокодировщиков является чувствительность к наборам данных: если набор данных не идентифицирован должным образом, VAE не сможет изучить какое-либо соответствующее распределение вероятностей в наборе данных и впоследствии не сможет сгенерировать новые объекты на должном уровне.

В основе таких моделей лежит блок GRU, представленный в 2014 г. в статье [17]. По эффективности и по качеству обучения данный вид нейронной сети схож с известной LSTM, однако из-за того, что GRU имеет на один "gate" меньше, данный блок имеет меньше параметров и потенциально может быстрее обучаться и сходиться. Представим архитектуру GRU с помощью математической модели

$$\begin{aligned} z_t &= \sigma_g(W_z x_t + U_z h_{t-1} + b_z), \\ r_t &= \sigma_g(W_r x_t + U_r h_{t-1} + b_r), \\ h_t &= z_t \circ h_{t-1} + (1 - z_t) \circ \sigma_h(W_h x_t + U_h[r_t \circ h_{t-1}] + b_h), \end{aligned} \quad (2)$$

где $\circ$ – это произведение Адамара; σ_g и σ_h – это две функции активации на основе сигмоиды и гиперболического тангенса соответственно; x_t – это входной вектор; h_t – это выходной вектор; z_t – это вектор вентиля обновления; r_t – это вектор вентиля сброса; W,U,b – это матрицы переменных и вектор свободных весов.

В качестве метрики будем использовать специальную химическую метрику QED [18]. В некоторых недавних публикациях о генеративных моделях для определения пригодности молекулы используют вычисляемые молекулярные свойства. Методика QED сравнивает распределение набора свойств, рассчитанных для молекулы, с распределениями тех же свойств в продаваемых лекарствах. Показатель совпадения варьируется от 0 до 1. При этом молекулы с показателем около 1 считаются наиболее похожими на лекарства. В качестве оценки мы рассчитаем QED для сгенерированных молекул и отбросим те, у которых показатели меньше 0,5. Данная метрика вычисляется в соответствии с выражением

$$QED_w = \exp\left(\frac{\sum_{i=1}^{n} w_i \ln(d_i)}{\sum_{i=1}^{n} w_i}\right). \quad (3)$$

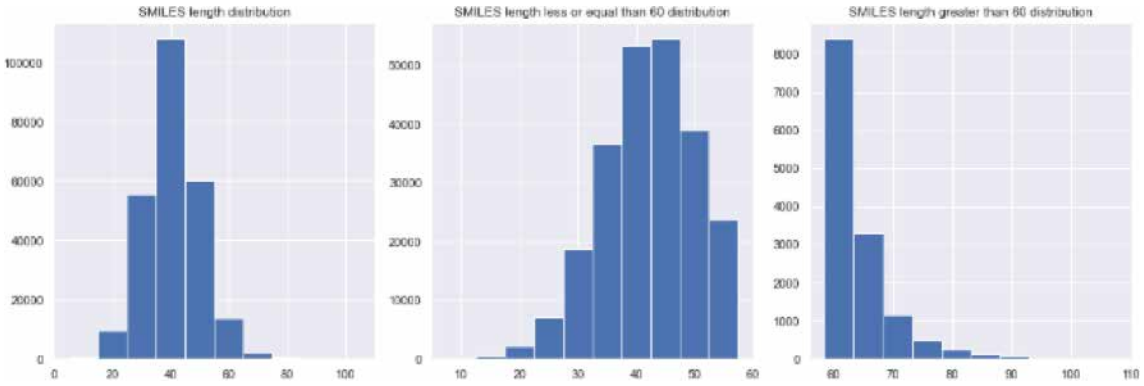


Рис. 1. Распределение атомов в молекулах

В данном уравнении d – это индивидуальная функция желательности, w – вес, применяемый к каждой функции, а n – количество молекулярных дескрипторов. Обычно функции желательности определяются произвольно. Как правило, это убывающие или возрастающие монотонные функции или функции «горба» в определенных диапазонах параметров и точках перегиба. Оптимальный набор весов – это тот, который максимизирует информационное содержание, которое может быть измерено путем вычисления энтропии Шеннона.

Результаты исследования и их обсуждение

Данные были взяты из открытых источников [19] и насчитывают порядка 250 000 молекулярных соединений, кодированных в SMILE структуру. Распределения атомов представлены на рис. 1.

Как можно видеть, больше всего молекул имеют в своем составе от 30 до 50 молекул, и лишь незначительная часть имеет больше 90. Максимальное же количество атомов в молекуле – 110.

Согласно рис. 2, в данном наборе данных можно заметить, что большая часть молекул имеют QED около 0,8, что говорит о достаточной схожести с лекарственными молекулами. Однако следует заметить, что лекарством может являться и молекулярное соединение с достаточно низким QED.

Обучение модели исследования V-GRU на основе GRU проводилось на полном наборе данных (250 000 молекул) с разбивкой выборки обучения на тестовую и валидационную с соотношением 0,8:0,2. Данная модель имеет 6 слоев GRU, 2 полносвязных слоя. Общее число параметров – 2 369 839. Процесс обучения представлен на рис. 3. Можно заметить, что данная нейронная сеть достаточно хорошо обучается на тренировочном наборе данных. Функция потерь уменьшается, а точность увеличивается. Эффекта переобучения нет.

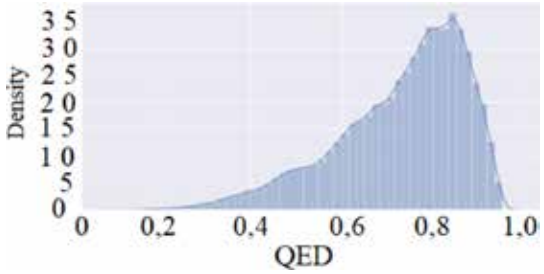


Рис. 2. Распределение свойства QED в наборе данных

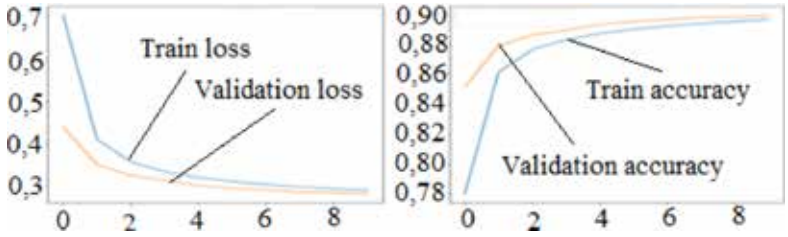
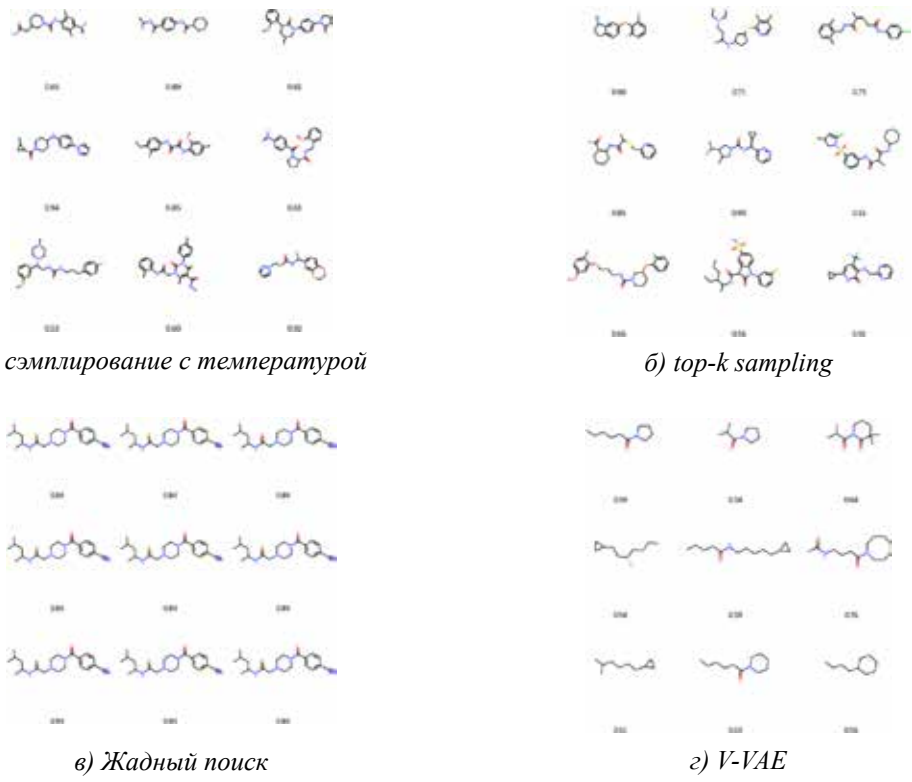


Рис. 3. Процесс обучения модели V-GRU



а) сэмплирование с температурой *б) top-k sampling*

в) Жадный поиск *г) V-VAE*

Рис. 4. Примеры генерируемых SMILES

В связи с ограниченностью ресурсов, обучение V-VAE проводилось не на полном наборе данных. Модель имеет слои энкодера и декодера. Всего 864 932 параметра. Для обучения были взяты 75000 молекул.

Результаты генерации представлены на рис. 4.

Таким образом, алгоритм жадного поиска не справляется с данной задачей, генерируя однотипные структуры.

Заключение

В данной работе мы предложили четыре алгоритма для создания потенциальных лекарственных молекулярных соединений.

В результате проделанной работы было сгенерировано некоторое множество лекарственных молекулярных соединений, которые были оценены метрикой QED. Данные соединения являются кандидатами на лекарства. Оставшаяся работа по проверке данных соединений на эффективность и безопасность входит в компетенцию научных сотрудников лабораторий фармацевтических компаний.

Таким образом, основная гипотеза данного исследования выполняется: мы показали, что с задачей генерации лекарственных молекулярных соединений наравне с моделями GAN, VAE справляются генеративные алгоритмы сэмплирования.

В качестве дальнейших направлений исследований можно рассмотреть применение более инновационных алгоритмов с целью достижения желаемых результатов: GAN, квантовый GAN, WGAN, CVAE или другие алгоритмы сэмплирования: beam search или nucleus sampling. Также в спецификации модели можно учесть уязвимый белок болезни, на который должно воздействовать желаемое лекарственное соединение.

Список литературы

1. Широканев А.С., Андриянов Н.А., Ильясова Н.Ю. Разработка векторного алгоритма по технологии CUDA для трехмерного моделирования процесса лазерной коагуляции сетчатки // Компьютерная оптика. 2021. Т. 45. № 3. С. 427–437. DOI: 10.18287/2412-6179-CO-828.

2. Shirokanev A., Ilyasova N., Andriyanov N., Zamytskiy E., Zolotarev A., Kirsh D. Modeling of Fundus Laser Exposure for Estimating Safe Laser Coagulation Parameters in the Treatment of Diabetic Retinopathy. Mathematics. 2021. Vol. 9. P. 967. DOI: 10.3390/math9090967.

3. Поляков М.В., Хоперсков А.В. Математическое моделирование пространственного распределения радиационного поля в биоткани: определение яркостной температуры для диагностики // Вестник Волгоградского государственного университета. 2016. Т. 36. № 5. С. 73–84.

4. Gad A.F. Practical Computer Vision Applications Using Deep Learning with CNNs. Apress. 2019. [Электронный ре-

cypc]. URL: https://www.pdfdrive.com/practical-computer-vision-applications-using-deep-learning-with-cnns-with-detailed-examples-in-python-using-tensorflow-and-kivy-d185770768.html (дата обращения: 05.03.2022).

5. Roffman D., Hart G., Girardi M., Ko C., Deng J. Predicting non-melanoma skin cancer via a multi-parameterized artificial neural network. Scientific reports. 2018. Vol. 8. P. 1701.

6. Zhu Y., Wang Q., Xu M. Application of convolutional neural network in the diagnosis of the invasion depth of gastric cancer based on conventional endoscopy. Gastrointestinal Endoscopy. 2019. Vol. 89. No. 4. P. 806–815.

7. Goodfellow I., Pouget-Abadie J., Mirza M., Xu B., Warde-Farley D., Ozair S., Courville A., Bengio Y. Generative Adversarial Networks. arXiv preprint. 2014. [Электронный ресурс]. URL: https://arxiv.org/abs/1406.2661 (дата обращения: 07.03.2022).

8. Grow C., Gao K., Nguyen D., Wei G.W. Generative network complex (gnc) for drug discovery // arXiv preprint. 2019. [Электронный ресурс]. URL: https://arxiv.org/abs/1910.14650 (дата обращения: 07.03.2022).

9. Jumper J., Evans R., Pritzel A., Green T., Figurnov M., Ronneberger O., Tunyasuvunakool K., Bates R., Žídek A., Potapenko A., Bridgland A., Meyer C., Kohl S., Ballard A., Cowie A., Romera-Paredes B., Nikolov S., Jain R., Adler J., Back T., Petersen S., Reiman D., Clancy E., Zielinski M., Steinegger M., Pacholska M., Berghammer T., Bodenstein S., Silver D., Vinyals O., Senior A., Kavukcuoglu K., Kohli P., Hassabis D. Highly accurate protein structure prediction with AlphaFold. Nature. 2021. Vol. 596. P. 583–589.

10. Batool M., Ahmad B., Choi S. A structure-based drug discovery paradigm. International journal of molecular sciences. 2019. Vol. 20. P. 13–21.

11. Sherstinsky A. Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) Network. arXiv preprint. 2018. [Электронный ресурс]. URL: https://arxiv.org/abs/1808.03314 (дата обращения: 07.03.2022).

12. Chenthamarakshan V., Das P., Padhi I. Target-Specific and Selective Drug Design for COVID-19 Using Deep Generative Models. arXiv preprint. 2020. [Электронный ресурс]. URL: https://arxiv.org/abs/2004.01215 (дата обращения: 07.03.2022).

13. Holtzman A., Buys J., Du L., Forbes M., Choi Y. The curious case of neural text degeneration. arXiv preprint. 2019. [Электронный ресурс]. URL: https://arxiv.org/abs/1904.09751 (дата обращения 07.03.2022).

14. Ackley D., Hinton G., Sejnowski T. A learning algorithm for Boltzman nmachines. Cognitive science. 1985. Vol. 9. No. 1. P. 147–169.

15. Fan A., Lewis M., Dauphin Y. Hierarchical Neural Story Generation. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. 2018. Vol. 1. P. 10–18.

16. Kingma D., Welling M. Auto-Encoding Variational Bayes. arXiv preprint. 2013. [Электронный ресурс]. URL: https://arxiv.org/abs/1312.6114 (дата обращения: 07.03.2022).

17. Junyoung C., Caglar G., Kyung C., Yoshua H., Yoshua B. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. arXiv preprint. 2014. [Электронный ресурс]. URL: https://arxiv.org/abs/1412.3555 (дата обращения: 07.03.2022).

18. Bickerton R., Paolini G., Besnard J., Muresan S., Hopkins A. Quantifying the Chemical Beauty of Drugs. Nature Chemistry. 2012. Vol. 4. P. 90–98.

19. Sterling T., Irwin J. Zinc 15-ligand discovery for everyone. Chem Inf Model. Vol. 55. No. 11. P. 2324–2337.