НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
ноябрь–декабрь 2018 Том 18 № 6 ISSN 2226-1494 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
November–December 2018 Vol. 18 No 6 ISSN 2226-1494 http://ntv.ifmo.ru/en

УДК 004.855.5: 004.032.26

ПОДКРЕПЛЕННЫЙ ПОСЛЕДОВАТЕЛЬНОСТЬ-К-ПОСЛЕДОВАТЕЛЬНОСТИ КОНКУРЕНТНЫЙ АВТОЭНКОДЕР ДЛЯ ГЕНЕРАЦИИ МАЛЫХ ОРГАНИЧЕСКИХ МОЛЕКУЛЯРНЫХ СТРУКТУР

Е.О. Путин[a]

[a] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация
Адрес для переписки: putin.evgeny@gmail.com

Информация о статье
Поступила в редакцию 20.09.18, принята к печати 25.10.18
doi: 10.17586/2226-1494-2018-18-6-1084-1090
Язык статьи – русский

Ссылка для цитирования: Путин Е.О. Подкрепленный последовательность-к-последовательности конкурентный автоэнкодер для генерации малых органических молекулярных структур // Научно-технический вестник информационных технологий, механики и оптики. 2018. Т. 18. № 6. С. 1084–1090. doi: 10.17586/2226-1494-2018-18-6-1084-1090

Аннотация
Исследованы современные модели глубокого обучения для генерации целевых малых органических молекулярных структур. Исследования проводились на двух выборках размером в 250 000 лекарственно-подобных молекулярных соединений из базы ZINC и 23 000 активных ингибиторов киназ, собранных вручную из открытой базы ChemBL. Предложена модель глубокой нейронной сети, основанная на концепциях конкурентного обучения и обучения с учителем. Модель контролирует молекулярную восстанавливаемость генерируемых структур за счет использования конкурентный seq2seq автоэнкодера и внешнего генератора. Наличие внешнего генератора обеспечивает гибкость модели в выборе архитектуры, а также позволяет подавать на вход условия для генерации. Сравнительные эксперименты показали, что предложенная модель превзошла ближайших конкурентов в экспериментах с предобучением и дообучением с точки зрения генерации валидных и уникальных молекулярных структур. Дополнительный химический анализ генерируемых структур демонстрирует лучшее качество генерации предлагаемой модели в сравнении с другими моделями конкурентами. **Практическая значимость.** Предложенная модель может быть использована для разработки новых лекарственных препаратов медицинскими химиками в качестве умного помощника.

Ключевые слова
машинное обучение, глубокое обучение, обучение с подкреплением, генеративные конкурентные нейронные сети, дизайн и разработка лекарств

Благодарности
Работа выполнена при финансовой поддержке Правительства Российской Федерации, грант 074-U01 и РФФИ, грант 16-37-60115-мол_а_дк.

REINFORCED SEQ2SEQ ADVERSARIAL AUTOENCODER FOR DE NOVO MOLECULAR DESIGN

E.O. Putin[a]

[a] ITMO University, Saint Petersburg, 197101, Russian Federation
Corresponding author: putin.evgeny@gmail.com

Article info
Received 20.09.18, accepted 25.10.18
doi: 10.17586/2226-1494-2018-18-6-1084-1090
Article in Russian

For citation: Putin E.O. Reinforced seq2seq adversarial autoencoder for de novo molecular design. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2018, vol. 18, no. 6, pp. 1084–1090 (in Russian). doi: 10.17586/2226-1494-2018-18-6-1084-1090

Abstract
Subject of Research. The modern models of deep training for generation of target small organic molecules are studied. The studies were carried out on two datasets of 250,000 drug-like molecular compounds from the ZINC database and 23,000 kinase molecular structures collected manually from the open accessed ChemBL database. **Method.** We propose the model of a deep neural network based on the concepts of adversarial learning and reinforcement learning. The model controls the molecular validity of the generated structures through the use of a recurrent seq2seq autoencoder and an external generator. The presence of an external generator gives the model flexibility in the choice of architecture, and also allows for the input

conditions for the generation. **Main Results.** Comparative experiments have shown that the proposed model is better than its closest competitors in experiments with pre- and post-training in terms of generating valid and unique molecular structures. Additional chemical analysis of generated structures demonstrates the best quality of the introduced model in comparison with the other competitor models. **Practical Relevance.** The proposed model can be used by medical chemists as an intelligent assistant for development of new drugs.

Keywords
machine learning, deep learning, reinforcement learning, generative adversarial networks, drug design and development

Acknowledgements
This work was financially supported by the Government of the Russian Federation, Grant 074-U01, and the Russian Foundation for Basic Research, Grant 16-37-60115 mol_a_dk.

Введение

Ранние этапы разработки новых лекарственных препаратов (drug design and discovery, DDD) основаны на трех важнейших научных дисциплинах: *in silico*-моделировании, которое включает в себя дизайн лекарства и его виртуальную оценку, комбинаторном органическом синтезе и высокопроизводительном биологическом скрининге (high throughput screening, HTS) [1]. Огромное количество новых лекарственных молекул с разнообразной структурой было обнаружено с помощью этого кумулятивного подхода.

Однако DDD-процесс крайне долгий и дорогой. Так, например, от разработки до вывода препарата на рынок может потребоваться 10 лет и 2,6 млрд долларов [2]. Более того, по оценкам [3], запатентовано до 70 % всех возможных структурных модификаций молекулы-кандидата (соединение, предназначенное для тестирования на животных или людях), поэтому, как правило, медицинский химик обычно сталкивается с очень узким химическим пространством вокруг запатентованной молекулы-кандидата.

Помимо этого, пространство всех органических синтетически доступных лекарственно-подобных молекул оценивается от 10^{60} до 10^{100} соединений [4]. Поэтому обход такого пространства с его виртуальной оценкой, и тем более с выполнением HTS, является вычислительно невозможной задачей. Таким образом, возникает потребность в разработке новых компьютерных моделей и методов, способных создавать молекулярную структуру лекарства и проводить ее виртуальную оценку, обеспечивая необходимое разнообразие и новизну генерируемых структур.

В последнее время методы машинного обучения, и в частности глубокие нейронные сети, достигли значительного прогресса в распознавании образов, обработке естественного языка [5], биомедицине [6], биоинформатике [7, 8] и во многих процессах современного DDD [9–14]. Так, задача генерации новых молекулярных соединений с заданными структурами и свойствами может быть адаптирована как подход, основанный на данных (data-driven) для порождения новых качественных молекул, нацеленных на конкретную белковую мишень или класс мишеней.

Однако современные модели должны иметь возможность порождать интересные и привлекательные по структупе молекулярные соединения, быть легко синтезируемыми и удовлетворять целевым закономерностям и критериям медицинской химии. К таким критериям можно отнести разработку лекарственно-подобных молекул, не содержащих токсичные подфрагменты.

Цель настоящей работы заключается в разработке новой модели Reinforced Seq2seq Adversarial Autoencoder (RSAAE) глубокой нейронной сети, основанной на объединении конкурентного обучения и обучения с подкреплением [15–17] для генерации малых молекулярных органических структур. Предлагаемая архитектура использует рекуррентный автокодировщик на основе парадигмы seq2seq [18], это позволяет отображать молекулярное соединение в латентное пространство (скрытый слой нейронной сети), а также получать молекулярную структуру по точке в латентном пространстве. Такой подход дает возможность контролировать долю валидных с точки зрения валентности молекулярных структур, накладывать на латентное пространство дополнительные ограничения (такие как условие [19]), выполнять предобучение и перенос знаний на небольшие выборки целевых молекулярных структур. Более того, возможно проводить анализ (интерполировать, кластеризовать, оптимизировать) латентного пространства, что может привести к повышению эффективности генерируемых молекулярных структур.

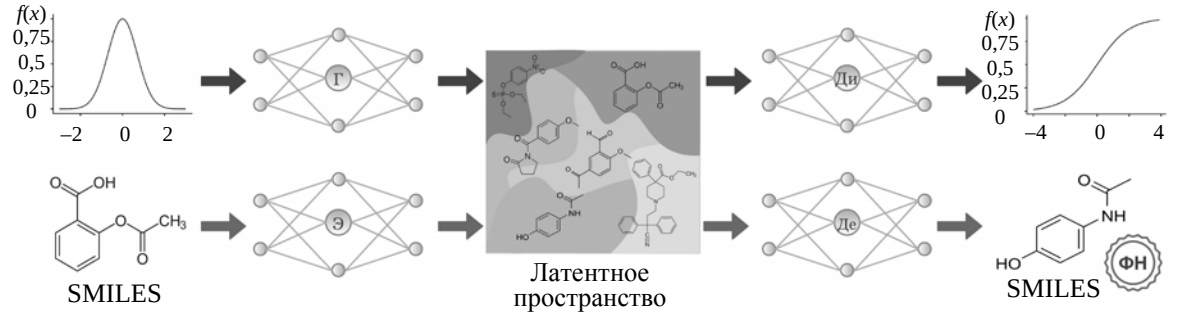
Модель RSAAE

Архитектура предлагаемой модели представлена на рисунке ($f(x)$ слева обозначает гауссиану, а справа – сигмоиду, x – переменная). Модель состоит из пяти компонентов: энкодера Э, декодера Де, генератора Г, дискриминатора Ди и блока обучения с подкреплением. Энкодер и декодер образуют seq2seq автоэнкодер, который предобучается с использованием функции награды, а генератор и дискриминатор реализуют парадигму конкурентного обучения. Совместно генератор и декодер задают процесс генерации новых молекулярных структур.

Энкодер отображает оригинальные молекулярные структуры, представленные в строковом формате SMILES (Simplified Molecular-Input Line-Entry System) [20], в точки латентного пространства. Декодер, в свою очередь, отображает точки из латентного пространства в молекулы. Обучаясь вместе, энкодер и декодер

образуют рекуррентный автоэнкодер seq2seq [18], который выучивает общие молекулярные зависимости и отвечает за валидность (осуществляет проверку валентности) молекулярных структур.

Однако так как автоэнкодер seq2seq является исключительно дискриминирующей моделью, порождать новые молекулы он не способен. Это в предлагаемой модели RSAAE обеспечивает конкурентное обучение, реализуемое через генеративные конкурентные нейронные сети (generative adversarial network, GAN) [19]. Парадигма GAN подразумевает наличие двух игроков – генератора и дискриминатора. Задача генератора генерировать правдоподобные объекты таким образом, чтобы «обмануть» дискриминатор (т.е. чтобы последний не смог отличить сгенерированные объекты от тренировочных). С другой стороны, задача дискриминатора состоит в том, чтобы эффективно отличать сгенерированные примеры от тренировочных. Таким образом, генератор и дискриминатор задают минимаксную игру, в которой теоретически достигается равновесие Нэша.



Архитектура модели RSAAE

В случае RSAAE генератор по входному шуму из стандартного нормального распределения порождает точки латентного пространства. Дискриминатор, принимая точки латентного пространства от энкодера или генератора, определяет, является ли точка (молекула) реальной или сгенерированной.

Пятый компонент RSAAE отвечает за то, чтобы генерируемые молекулярные структуры обладали конкретными целевыми свойствами (заданными пользователем при запуске обучения модели). Это достигается за счет использования обучения с подкреплением, при котором по выходным молекулам с декодера вычисляется объектная функция награды (ФН) [16, 17]. Примером такой функции может быть какой-либо критерий медицинской химии, например, критерий того, что молекула является лекарственно-подобной. За счет использования ФН, которой должны соответствовать генерируемые молекулярные структуры в процессе обучения модели, достигается возможность генерировать молекулы с заданными свойствами.

Так как выход с декодера является дискретным (SMILES-строка), то стандартный метод распространения ошибки не может быть использован. В таких случаях применяются методы policy gradient, а именно в модели RSAAE используется алгоритм REINFORCE [21].

Необходимо отметить, что в генеративных конкурентных автоэнкодерах (adversarial autoencoder, AAE) [22] декодер и генератор – это одно и то же, в то время как RSAAE разделяет декодер и генератор, т.е. использует внешний генератор, что обеспечивает гибкость выбора архитектуры генератора, а также позволяет подавать на вход генератору помимо шума некоторое условие [19].

С другой стороны, RSAAE отличается от моделей ORGANIC [15], RANC [16], ATNC [17] использованием рекуррентного seq2seq автоэнкодера, что дает возможность контролировать восстанавливаемость, проводить эффективное предобучение на больших наборах данных и дообучение на небольших выборках (что крайне актуально для DDD), осуществлять перенос знаний, анализировать латентное пространство и накладывать на него различные ограничения.

Эффективное предобучение – крайне важное преимущество модели RSAAE перед ORGANIC, RANC, ATNC. Оно позволяет обучить RSAAE на огромном наборе молекулярных соединений, например, на всей базе ZINC в 35 млн молекул и выучить общие молекулярные зависимости для разных хемотипов молекулярных структур с контролируемой точностью валидности, т.е. покрыть все возможное молекулярное пространство. Далее можно предобученную модель дообучить на специфической задаче (например, генерации различных активных ингибиторов киназ).

Для модели RSAAE может быть реализовано несколько алгоритмов обучения. Например, можно обучать все компоненты модели совместно с нуля либо делать это поочередно. Однако самый лучший с точки зрения стабильности обучения модели алгоритм представляет собой предобучение seq2seq автоэнкодера с подкрепляющим блоком, и далее – обучение генератора и дискриминатора.

Экспериментальное исследование модели RSAAE

Так как RSAAE расширяет и развивает модели ORGANIC, RANC, ATNC, целью экспериментов было сравнение этих моделей. Эксперименты выполнялись на NVIDIA Titan X Pascal с 256 RAM с

одинаковыми настройками и гиперпараметрами [16, 17]. Всего было проведено два эксперимента с функцией награды по правилу пяти Липинского [15]:

1) Kin – обучение моделей осуществлялось «с нуля» на небольшой целевой выборке молекул, в качестве которой использовалось собранное вручную подмножество (набор данных Kin) из 23 000 лекарственно-подобных молекул, относящихся к классу киназной химии, доступных в базе ChemBL [23]. При этом средняя длина SMILES-строк молекул в наборе данных Kin составила 54 символа;
2) ZINC+Kin – предобучение моделей в этом случае осуществлялось на большой выборке молекулярных структур, а дообучение – на целевой выборке Kin. При этом в качестве набора тренировочных образцов для экспериментов с предобучением использовалось подмножество лекарственно-подобных молекул из базы ZINC [24]. Обучающая выборка составила 250 000 молекулярных структур, средняя длина SMILES-строк молекул в наборе данных ZINC составила 44 символа.

Для оценки качества генеративных моделей в работе применены следующие математические статистики: процент валидных (valid) молекул из общего числа сгенерированных SMILES-строк (валидность проверялась функцией MolFromSmiles из библиотеки RDKit[1]), процент уникальных (Unique) молекулярных структур из общего числа валидных молекул, средняя длина (Length) сгенерированных SMILES-строк из числа валидных молекул (табл. 1).

Чтобы охарактеризовать генерируемые наборы молекулярных структур, каждая модель определяла [16, 17]:

1) внутреннее разнообразие (diversity) молекулярного пространства;
2) число молекулярных структур, не прошедших медицинские химические фильтры (MCF);
3) число уникальных гетероциклов (hetero). Кроме того, был проведен кластерный анализ каждого сгенерированного набора молекул [16, 17], в ходе которого рассчитывались число кластеров (Clusters) и средний размер кластера (Cluster size). При этом очевидно, что чем больше число кластеров, тем меньше средний размер кластера, а чем больше внутреннее разнообразие, тем больше число кластеров.

Модель	Kin			ZINC+Kin		
	Valid, %	Unique, %	Length	Valid, %	Unique, %	Length
ORGANIC	82	16	31	80	19	28
RANC	68	48	52	63	42	50
ATNC	**72**	**75**	**52**	65	68	50
RSAAE	70	73	50	**82(+12)**	**87(+14)**	**53**

Таблица 1. Сравнение моделей по валидности, уникальности и средней длине генерируемых молекул

Как видно из табл. 1, модель ORGANIC показала себя хуже всех: средняя длина генерируемых молекул почти в два раза меньше по обоим экспериментам, чем средняя длина в тренировочном наборе Kin, число полученных уникальных молекул в три раза меньше, чем у ее ближайшего конкурента RANC. Кроме того, из табл. 1 видно, что в обоих экспериментах средняя длина генерируемых молекулярных структур у RANC и ATNC одинакова. Однако ATNC опередила RANC по проценту валидности и уникальности генерируемых молекул. Очевидно, это произошло из-за того, что в RANC отсутствует механизм фильтрации и предварительной оценки молекул, который есть в ATNC.

Также стоит отметить, что в первом эксперименте, обучаясь с нуля, модель RSAAE уступала модели ATNC (вероятнее всего, из-за небольшого числа тренировочных примеров). Однако в эксперименте с дообучением RSAAE существенно превзошла все остальные модели, достигая 82 % валидности и 87 % уникальности генерируемых молекулярных структур, и при этом почти полностью соответствуя средней длине SMILES-строк в наборе данных Kin.

Более того, во втором эксперименте другие модели теряли в метриках (по сравнению с первым), в то время как RSAAE достигла прироста 12 и 14 % по валидности и уникальности генерируемых молекулярных структур. Очевидно, это происходило за счет контролируемого (благодаря использованию seq2seq автоэнкодера) латентного пространства, с которым оперировала данная модель (в частности, выучивая общие закономерности построения SMILES-строк молекул RSAAE, может эффективно переносить знания и дообучаться на более специфических выборках). При этом стоит заметить, что наборы данных Kin и ZINC сильно различаются (во-первых, средняя длина SMILES-строк в Kin на 10 больше чем в ZINC, а во-вторых, в Kin содержится исключительно киназная химия, которой в наборе ZINC всего 10–15 %).

Таким образом, эксперимент с дообучением демонстрирует преимущества RSAAE перед другими моделями и показывает возможность модели RSAAE обобщать знания и переносить их на новые специфические наборы данных.

Химические статистики по двум экспериментам представлены в табл. 2.

[1] http://www.rdkit.org

Модель	Kin				
	Diversity, %	MCF, %	Hetero	Clusters	Cluster size
ORGANIC	84	35	4109	441	16
RANC	85	9	2398	604	13
ATNC	**85**	**8**	**2670**	**658**	**11**
RSAAE	85	11	2552	627	12
	ZINC+Kin				
ORGANIC	86	39	3750	398	17
RANC	87	10	2533	563	14
ATNC	87	8	3054	602	12
RSAAE	**89**	**6**	**3847**	**696**	**10**

Таблица 2. Сравнение моделей ORGANIC, RANC, ATNC, RSAAE по химическим статистикам

Как видно из табл. 2, в обоих экспериментах ORGANIC сильно уступала остальным моделям по разнообразию, проценту молекулярных структур, не прошедших медицинские химические фильтры, числу кластеров и среднему размеру кластера. Очевидно, это было связано с проблемой совершенного дискриминатора данной модели [16] и неспособностью ORGANIC к эффективному дообучению.

Интересно отметить, что в эксперименте Kin RANC и RSAAE показали близкие результаты. В то время как во втором эксперименте RSAAE превосходила все остальные модели, демонстрируя 89 % внутреннего разнообразия, MCF=6 % и clusters=696. Более того, в эксперименте ZINC+Kin RSAAE достигала улучшения в +4 % по diversity, –5% по MCF, +69 по clusters. Это еще раз доказывает возможность модели RSAAE к эффективному обобщению и переносу знаний на новые наборы данных.

Таким образом, сравнительные результаты наглядно демонстрируют возможности предложенной модели RSAAE: в экспериментах с дообучением RSAAE превосходила остальные модели как по математическим, так и по химическим параметрам. Следует отметить, что в DDD-процессе наиболее широко используются генеративные модели с дообучением. Это связано с тем, что для многих белковых мишеней известно сравнительно мало активных молекулярных структур (так, например, в среднем на каждый белок из открытой базы ChemBL приходится 1000–2000 активных молекул). И так как в силу небольшого количества тренировочных примеров эффективное дообучение таких моделей как ORGANIC, RANC, ATNC невозможно, то использование RSAAE является лучшим решением.

Заключение

Предложена архитектура нейронной сети RSAAE, используемой для генерации малых органических молекулярных структур, основанная на концепциях конкурентного обучения и обучения с подкреплением. По сравнению с ORGANIC, RANC, ATNC модель RSAAE добавляет отдельно обучаемый внешний генератор и использует seq2seq автоэнкодер, который контролирует валидность молекулярных структур. Результаты экспериментальных исследований на наборах данных Kin и ZINC позволяют сделать следующие выводы:

- благодаря использованию seq2seq автоэнкодера и внешнего генератора модель RSAAE эффективно выучивает общие молекулярные закономерности в формате SMILES. Это позволяет данной модели осуществлять качественное предобучение на больших выборках данных, таким образом обобщая знания. А эксперименты с дообучением демонстрируют возможность RSAAE производить перенос знаний на другие специфичные наборы данных. При этом молекулы, генерируемые с помощью RSAAE, сохраняют среднюю длину тренировочных SMILES-строк;
- молекулы, сгенерированные RSAAE, имеют наименьший, по сравнению с остальными моделями, процент молекулярных структур, не прошедших медицинские химические фильтры, наибольшее внутреннее молекулярное разнообразие и наибольшее число кластеров. И так как это крайне важные параметры для современного DDD-процесса, модель RSAAE можно считать перспективной стартовой точкой в автоматической генерации органических молекулярных структур, которая уже сейчас может быть использована как умный помощник медицинскими химиками или биоинформатиками.

Литература

1. Holenz J. (eds) Lead Generation: Methods and Strategies. John Wiley & Sons, 2016. V. 2.
2. 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. V. 47. P. 20–33. doi: 10.1016/j.jhealeco.2016.01.012
3. Ivanenkov Y.A. et al. Small-molecule inhibitors of hepatitis C virus (HCV) non-structural protein 5A (NS5A): a patent review (2010-2015) // Expert Opinion on herapeutic Patents. 2017. V. 27. N 4. P. 401–414. doi:

References

1. Holenz J. (eds) *Lead Generation: Methods and Strategies.* John Wiley & Sons, 2016, vol. 2.
2. 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, pp. 20–33. doi: 10.1016/j.jhealeco.2016.01.012
3. Ivanenkov Y.A. et al. Small-molecule inhibitors of hepatitis C virus (HCV) non-structural protein 5A (NS5A): a patent review (2010-2015). *Expert Opinion on Therapeutic Patents*, 2017, vol. 27, no. 4, pp. 401–414. doi:

10.1080/13543776.2017.1272573
4. Schneider G., Fechner U. Computer-based de novo design of drug-like molecules // Nature Reviews Drug Discovery. 2005. V. 4. N 8. P. 649–663. doi: 10.1038/nrd1799
5. LeCun Y., Bengio Y., Hinton G. Deep learning // Nature. 2015. V. 521. N 7553. P. 436–444. doi: 10.1038/nature14539
6. Mamoshina P., Vieira A., Putin E., Zhavoronkov A. Applications of deep learning in biomedicine // Molecular Pharmaceutics. 2016. V. 13. N 5. P. 1445–1454. doi: 10.1021/acs.molpharmaceut.5b00982
7. Min S., Lee B., Yoon S. Deep learning in bioinformatics // Briefings in Bioinformatics. 2017. V. 18. N 5. P. 851–869.
8. Pastur-Romay L., Cedron F. et al. Deep artificial neural networks and neuromorphic chips for big data analysis: pharmaceutical and bioinformatics applications // International Journal of Molecular Sciences. 2016. V. 17. N 8. P. 1313. doi: 10.3390/ijms17081313
9. Zhang L., Tan J., Han D., Zhu H. From machine learning to deep learning: progress in machine intelligence for rational drug discovery // Drug Discovery Today. 2017. V. 22. N 11. P. 1680–1685. doi: 10.1016/j.drudis.2017.08.010
10. Gawehn E., Hiss J.A., Schneider G. Deep learning in drug discovery // Molecular Informatics. 2016. V. 35. N 1. P. 3–14.
11. Gupta A., Muller A.T., Huisman B.J.H. et al. Generative recurrent networks for de novo drug design // Molecular Informatics. 2018. V. 37. N 1-2. doi: 10.1002/minf.201880141
12. Yuan W. et al. Chemical space mimicry for drug discovery // Journal of Chemical Information and Modeling. 2017. V. 57. N 4. P. 875–882. doi: 10.1021/acs.jcim.6b00754
13. Korotcov A., Tkachenko V., Russo D.P., Ekins S. Comparison of deep learning with multiple machine learning methods and metrics using diverse drug discovery data sets // Molecular Pharmaceutics. 2017. V. 14. N 12. P. 4462–4475. doi: 10.1021/acs.molpharmaceut.7b00578
14. Olivecrona M., Blaschke T., Engkvist O., Chen H. Molecular de-novo design through deep reinforcement learning // Journal of Cheminformatics. 2017. V. 9. N 1. P. 48. doi: 10.1186/s13321-017-0235-x
15. Sanchez-Lengeling B., Outeiral C., Guimaraes G.L., Aspuru-Guzik A. Optimizing distributions over molecular space. An objective-reinforced generative adversarial network for inverse-design chemistry (ORGANIC) // ChemRxiv. Preprint. 2017. doi: 10.26434/chemrxiv.5309668.v3
16. Putin E., Asadulaev A., Ivanenkov Y., Aladinskiy V. et al. Reinforced adversarial neural computer for de novo molecular design // Journal of Chemical Information and Modeling. 2018. V. 58. N 6. P. 1194–1204. doi: 10.1021/acs.jcim.7b00690
17. Putin E., Asadulaev A., Vanhaelen Q., Ivanenkov Y. et al. Adversarial threshold neural computer for molecular de novo design // Molecular Pharmaceutics. 2018. V. 15. N 10. P. 4386–4397. doi: 10.1021/acs.molpharmaceut.7b01137
18. Sutskever I., Vinyals O., Le Q.V. Sequence to sequence learning with neural networks // Advances in Neural Information Processing Systems. 2014.
19. Goodfellow I., Pouget-Abadie J., Mirza M. et al. Generative adversarial nets // Advances in Neural Information Processing Systems. 2014. P. 2672–2680.
20. Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules // Journal of Chemical Information and Computer Sciences. 1988. V. 28. N 1. P. 31–36. doi: 10.1021/ci00057a005
21. Williams R.J. Simple statistical gradient-following algorithms for connectionist reinforcement learning // Machine Learning. 1992. V. 8. N 3-4. P. 229–256. doi: 10.1007/bf00992696
22. Makhzani A., Shlens J., Jaitly N. et al. Adversarial autoencoders // arXiv preprint. 2015. arXiv:1511.05644
23. Gaulton A., Bellis L.J., Bento A.P. et al. ChEMBL: a large-scale bioactivity database for drug discovery // Nucleic Acids Research. 2011. V. 40. N D1. P. D1100-D1107. doi: 10.1093/nar/gkr777
24. Irwin J.J., Shoichet B.K. ZINC − A free database of commercially available compounds for virtual screening // Journal of Chemical Information and Modeling. 2005. V. 45. N 1. P. 177–182. doi: 10.1021/ci049714+

10.1080/13543776.2017.1272573
4. Schneider G., Fechner U. Computer-based de novo design of drug-like molecules. *Nature Reviews Drug Discovery*, 2005, vol. 4, no. 8, pp. 649–663. doi: 10.1038/nrd1799
5. LeCun Y., Bengio Y., Hinton G. Deep learning. *Nature*, 2015, vol. 521, no. 7553, pp. 436–444. doi: 10.1038/nature14539
6. Mamoshina P., Vieira A., Putin E., Zhavoronkov A. Applications of deep learning in biomedicine. *Molecular Pharmaceutics*, 2016, vol. 13, no. 5, pp. 1445–1454. doi: 10.1021/acs.molpharmaceut.5b00982
7. Min S., Lee B., Yoon S. Deep learning in bioinformatics. *Briefings in Bioinformatics*, 2017, vol. 18, no. 5, pp. 851–869.
8. Pastur-Romay L., Cedron F. et al. Deep artificial neural networks and neuromorphic chips for big data analysis: pharmaceutical and bioinformatics applications. *International Journal of Molecular Sciences*, 2016, vol. 17, no. 8, p. 1313. doi: 10.3390/ijms17081313
9. Zhang L., Tan J., Han D., Zhu H. From machine learning to deep learning: progress in machine intelligence for rational drug discovery. *Drug Discovery Today*, 2017, vol. 22, no. 11, pp. 1680–1685. doi: 10.1016/j.drudis.2017.08.010
10. Gawehn E., Hiss J.A., Schneider G. Deep learning in drug discovery. *Molecular Informatics*, 2016, vol. 35, no. 1, pp. 3–14.
11. Gupta A., Muller A.T., Huisman B.J.H. et al. Generative recurrent networks for de novo drug design. *Molecular Informatics*, 2018, vol. 37, no. 1-2. doi: 10.1002/minf.201880141
12. Yuan W. et al. Chemical space mimicry for drug discovery. *Journal of Chemical Information and Modeling*, 2017, vol. 57, no. 4, pp. 875–882. doi: 10.1021/acs.jcim.6b00754
13. Korotcov A., Tkachenko V., Russo D.P., Ekins S. Comparison of deep learning with multiple machine learning methods and metrics using diverse drug discovery data sets. *Molecular Pharmaceutics*, 2017, vol. 14, no. 12, pp. 4462–4475. doi: 10.1021/acs.molpharmaceut.7b00578
14. Olivecrona M., Blaschke T., Engkvist O., Chen H. Molecular de-novo design through deep reinforcement learning. *Journal of Cheminformatics*, 2017, vol. 9, no. 1, p. 48. doi: 10.1186/s13321-017-0235-x
15. Sanchez-Lengeling B., Outeiral C., Guimaraes G.L., Aspuru-Guzik A. Optimizing distributions over molecular space. An objective-reinforced generative adversarial network for inverse-design chemistry (ORGANIC). *ChemRxiv. Preprint*, 2017. doi: 10.26434/chemrxiv.5309668.v3
16. Putin E., Asadulaev A., Ivanenkov Y., Aladinskiy V. et al. Reinforced adversarial neural computer for de novo molecular design. *Journal of Chemical Information and Modeling*, 2018, vol. 58, no. 6, pp. 1194–1204. doi: 10.1021/acs.jcim.7b00690
17. Putin E., Asadulaev A., Vanhaelen Q., Ivanenkov Y. et al. Adversarial threshold neural computer for molecular de novo design. *Molecular Pharmaceutics*, 2018, vol. 15, no. 10, pp. 4386–4397. doi: 10.1021/acs.molpharmaceut.7b01137
18. Sutskever I., Vinyals O., Le Q.V. Sequence to sequence learning with neural networks. *Advances in Neural Information Processing Systems*, 2014.
19. Goodfellow I., Pouget-Abadie J., Mirza M. et al. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 2014, pp. 2672–2680.
20. Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 1988, vol. 28, no. 1, pp. 31–36. doi: 10.1021/ci00057a005
21. Williams R.J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 1992, vol. 8, no. 3-4, pp. 229–256. doi: 10.1007/bf00992696
22. Makhzani A., Shlens J., Jaitly N. et al. Adversarial autoencoders. *arXiv preprint*, 2015, arXiv:1511.05644
23. Gaulton A., Bellis L.J., Bento A.P. et al. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Research*, 2011, vol. 40, no. D1, pp. D1100-D1107. doi: 10.1093/nar/gkr777
24. Irwin J.J., Shoichet B.K. ZINC − A free database of commercially available compounds for virtual screening. *Journal of Chemical Information and Modeling*, 2005, vol. 45, no. 1, pp. 177–182. doi: 10.1021/ci049714+

Автор

Путин Евгений Олегович – инженер-программист, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, Scopus ID: 57189310406, ORCID ID: 0000-0002-3012-9708, putin.evgeny@gmail.com

Author

Evgeniy O. Putin – software engineer, ITMO University, Saint Petersburg, 197101, Russian Federation, Scopus ID: 57189310406, ORCID ID: 0000-0002-3012-9708, putin.evgeny@gmail.com