УДК 004.94+547.7+616-006+616-085

ГЕНЕРАТИВНЫЕ МОДЕЛИ ГЛУБОКОГО ОБУЧЕНИЯ ДЛЯ РАЗРАБОТКИ НОВЫХ ПОТЕНЦИАЛЬНЫХ ЛЕКАРСТВЕННЫХ ПРЕПАРАТОВ ПРОТИВ КОРОНАВИРУСА SARS-CoV-2

Н.А. Шульдов[1], А.М. Юшкевич[1], К.В. Фурс[1], А.В. Тузиков[1], А.М. Андрианов[2]

[1]*Объединенный институт проблем информатики, Национальная академия наук Беларуси, 220012, ул. Сурганова, 6, Минск, Республика Беларусь,* tuzikov@newman.bas-net.by
[2]*Институт биоорганической химии, Национальная академия наук Беларуси, ул. ак. Купревича, 5/2, 220141, г. Минск, Республика Беларусь*, alexande.andriano@yandex.ru

Разработана генеративная нейронная сеть глубокого обучения для компьютерного конструирования потенциальных ингибиторов основной протеазы SARS-CoV-2, играющей важную роль в репликации и транскрипции коронавируса. Проведено обучение и тестирование нейронной сети на наборе химических соединений, содержащих функциональные группы, способные обеспечить эффективные взаимодействия этих молекул с молекулярной мишенью. Показано, что использование нейронной сети совместно с методами молекулярного моделирования формирует продуктивную платформу для генерации новых молекул с заданными фармакологическими свойствами, перспективных для создания эффективных противовирусных препаратов.

Ключевые слова: Методы машинного обучения; глубокое обучение; генеративные нейронные сети; коронавирус SARS-CoV-2; основная протеаза; виртуальный скрининг; молекулярный докинг; противовирусные препараты.

GENERATIVE DEEP LEARNING MODELS FOR THE DEVELOPMENT OF NOVEL POTENTIAL DRUGS AGAINST SARS-CoV-2 CORONAVIRUS

N.A. Shuldau[a], A.M. Yushkevich[a], K.V. Furs[a], A.V. Tuzikov[a], A.M. Andrianov[b]

[a]*United Institute of Informatics Problems, National Academy of Sciences of Belarus, Surganov str., 6, Minsk, Republic of Belarus*
[b]*Institute of Bioorganic Chemistry, National Academy of Sciences of Belarus, ac. Kuprevich str., 5/2, Surganov str., Minsk, Republic of Belarus,* alexande.andriano@yandex.ru

A generative deep learning neural network has been developed for the computer-aided design of potential inhibitors of the SARS-CoV-2 main protease, which plays an important role in coronavirus replication and transcription. The neural network was trained and tested on a set of chemical compounds containing functional groups capable of providing effective

interactions of these molecules with the molecular target. The use of the neural network in conjunction with molecular modeling methods was shown to form a productive platform for the generation of novel molecules with desired pharmacological properties that are promising for the design of effective antiviral drugs.

Keywords: Machine learning methods; deep learning; generative neural networks; coronavirus SARS-CoV-2; main protease; antiviral drugs.

Введение

В последние годы генеративные модели глубокого обучения нашли широкое применение в исследованиях по разработке лекарств *de novo* [1]. Благодаря огромному прогрессу методов глубокого обучения в настоящее время разработаны генеративные модели с различной архитектурой и разными методами обучения, использующие разные типы и структуры данных. Применение генеративных моделей глубокого обучения уже показало их способность генерировать молекулы, которые могут быть синтезированы, активны *in vitro*, стабильны и проявляют активность *in vivo* в моделях, связанных с различными заболеваниями [1]. Однако, несмотря на то что генеративные модели глубокого обучения становятся все более распространенными в био- и хемоинформатике, их потенциал в этой области еще не раскрыт в полном объеме. В связи с этим разработка и применение генеративных методов глубокого обучения для компьютерного конструирования потенциальных лекарственных препаратов имеют большое научное и практическое значение.

Цель исследования: Разработать генеративные модели глубокого обучения для рационального дизайна новых потенциальных ингибиторов основной протеазы (M^{Pro}; Main Protease) SARS-CoV-2 – фермента, критически важного для репликации и транскрипции вируса и, поэтому, представляющего перспективную мишень для конструирования эффективных противовирусных препаратов [2].

Для решения поставленной задачи были выполнены исследования, включающие:

1) разработку и реализацию архитектуры генеративных моделей глубокого обучения, позволяющих генерировать новые высокоаффинные ингибиторы M^{Pro} SARS-CoV-2;

2) формирование обучающей библиотеки малых молекул, содержащих элементы структуры, способные обеспечить специфические и эффективные взаимодействия потенциальных лигандов с каталитическтим сайтом M^{Pro} SARS-CoV-2;

3) обучение и тестирование нейронной сети на соединениях из сформированной обучающей библиотеки;

4) оценку результатов обучения и работы нейронной сети в двух разных режимах генерации.

1. Материалы и методы

Архитектура нейронных сетей. В результате проведенных исследований были разработаны две генеративные модели глубокого обучения: 1) автоэнкодер, основанный на SMILES-представлениях молекул, который обучали без учителя с использованием ячеек долгой краткосрочной памяти (LSTM; Long Shot-Term Memory); 2) автоэнкодер, основанный на описаниях SMILES, который обучали с частичным привлечением учителя с использованием технологии LSTM. Во второй модели значение свободной энергии связывания химических соединений с молекулярной мишенью использовали как дополнительный параметр в латентном слое для обучения на основе результатов докинга соединений из тренировочного набора и как желаемое значение аффинности их связывания с M^{Pro} в режиме генерации новых соединений. Первая (эмбеддинговая) модель состоит из энкодера, латентного слоя, также являющегося входным слоем для гауссовского шума, и декодера. Эта модель получает на входной слой векторизованную матрицу описаний SMILES, которая проходит через слой LSTM, имеющий выходную размерность 64 нейрона. Особенность этой модели определяется тем, что сами выходные данные из ячеек LSTM энкодера не используются, а вместо этого берутся скрытые векторы и векторы состояний ячеек. Эти векторы конкатенируются вместе, образуя 128-ми элементные векторы, которые пропускаются через полносвязный слой с выходной размерностью 32 нейрона. Данные на выходе из полносвязного слоя попадают на латентный слой, состоящий из 32 нейронов, и являются эмбеддингами SMILES в контексте автоэнкодера. Далее эмбеддинги подаются параллельно на два полносвязных слоя с выходной размерностью 64 нейрона каждый, создавая начальные скрытые векторы и векторы состояния ячеек для слоя LSTM в декодере, имеющего такую же размерность, как и LSTM слой в энкодере. Существует также входной слой для декодера, используемый в качестве входных данных для LSTM, который в режиме обучения получает ту же векторизованную матрицу SMILES, что и вход энкодера, и, как обычная генеративная модель LSTM, предсказывает следующий символ. В режиме генерации на вход декодера подается символ начала строки “!”, запуская процесс посимвольной генерации новой строки SMILES. В этом режиме эмбеддинги используются для прогнозирования начальных состояний декодера LSTM, и они определяют, какая строка SMILES будет сгенерирована. Во всех полносвязных

слоях (кроме последнего слоя) используется функция активации ReLu, а для последнего слоя – функция активации softmax.

Подготовка обучающего набора данных. Для формирования обучающего набора данных методами фармакофорного моделирования, виртуального скрининга и молекулярного докинга идентифицировали набор из 342 102 молекул, способных имитировать структурно-функциональные свойства известных ингибиторов основной протеазы коронавируса SARS-CoV-2 и генетически родственного ему вируса атипичной пневмонии SARS-CoV. Химические структуры этих соединений преобразовывали в представления SMILES, каждое из которых было векторизовано в матрицу в соответствии с максимальной длиной и размером словаря SMILES-символов с добавленными символами начала и конца SMILES-строки. Обнаруженные в результате виртуального скрининга химические соединения интегрировали в молекулярную библиотеку для обучения разработанной генеративной нейронной сети. Подготовленная обучающая молекулярная библиотека объемом в 342 102 соединения и соответствующие им значения свободной энергии связывания с M^{Pro} SARS-CoV-2 сформировали набор данных для обучения и тестирования нейронной сети. Молекулярная библиотека была разделена на тренировочный, валидационный и тестовый наборы, включавшие соответственно 70%, 15% и 15% от общего числа содержащихся в ней соединений.

Обучение моделей. Обе разработанные модели были составлены слой за слоем с использованием высокоуровневого интерфейса пакета TensorFlow 2.1 (https://www.tensorflow.org/). Модели прошли 150 эпох обучения с дополнительным использованием функций обратного вызова "Уменьшить скорость обучения на плато" и "Ранняя остановка", чтобы помочь им сойтись к лучшему локальному минимуму, а также избежать переобучения. В качестве оптимизатора использовали метод стохастической оптимизации градиентного спуска – метод Адама – с начальным значением скорости обучения 0,005; при этом была выбрана категориальная функция потерь кросс-энтропии.

Генерация соединений и их анализ. В проведенном исследовании были рассмотрены два режима генерации соединений. В первом режиме генерацию выполняли из случайных чисел, взятых из нормального распределения с параметрами, полученными с помощью распределения тестовых данных на латентном слое для каждой компоненты вектора. Для этого режима процесс генерации в случае энергетической модели предполагал задание априорного значения свободной энергии связывания для аппроксимации генерируемых молекул. При этом были проведены вычислительные эксперименты для различных пороговых значений энергии.

Основная идея второго режима генерации заключалась в использовании выборки лучших лигандов из тестового набора с целью добавления шума к их эмбеддинговым представлениям. Этот подход должен был изменить модифицированный лиганд и, в случае энергетической модели, также увеличить аффинность связывания молекулы к целевому белку, вынуждая автоэнкодер генерировать более перспективные соединения.

2. Результаты и обсуждение

Обе модели были протестированы с использованием каждого из двух режимов генерации. Полученные результаты оценивали на основе значений свободной энергии связывания, предсказанных методом молекулярного докинга, а также путем их сравнения с величинами, рассчитанными для 34 152 эталонных соединений, взятых из тестового набора данных. Эти соединения использовали в вычислительных экспериментах в качестве положительного контроля при оценке способности нейронной сети генерировать из этих молекул более перспективные соединения. Эталонные соединения выбирали случайным образом из тестового набора данных в различных режимах генерации.

Сравнение моделей и режимов генерации. Для визуализации результатов эффективности работы двух разработанных моделей в двух режимах генерации, для каждой их комбинации было построено кумулятивное распределение свободной энергии связывания (рис.). Этот метод сравнения позволяет наблюдать общее качество сгенерированных экспериментальных выборок и проводить анализ распределения в целом без учета их количества. Из-за небольшого числа генерируемых соединений, демонстрирующих значения свободной энергии связывания < –6 ккал/моль (менее 2,8% от всех генерируемых соединений), результаты генерации были отфильтрованы согласно этой пороговой величине, использованной ранее для получения обучающего набора данных. Для наглядности представления результатов сравнительного анализа для каждого эксперимента выбирали определенные подмножества. В случае эмбеддинговой модели были отобраны все соединения, сгенерированные из гауссовского шума (эксперимент I), и все соединения, сгенерированные с использованием эталонных соединений, которые были изменены с его помощью (эксперимент II). Последний режим генерации включал лишь эталонные соединения со значениями свободной энергии связывания < –9 ккал/моль. Для энергетической модели были выбраны три вычислительных эксперимента. Первый эксперимент (эксперимент III) генерировал соединения из гауссовского шума с использованием передаваемых в «энергетический» нейрон пороговых значений свободной энергии связывания, заданных в

интервале от –9 ккал/моль до –11 ккал/моль с шагом 1 ккал/моль. Во втором эксперименте (эксперимент IV) использовали эталонные соединения, у которых соответствующие эмбеддинги изменяли путем добавления гауссовского шума, а пороговые значения энергии связывания понижали с шагом 0,5 ккал/моль и 1,0 ккал/моль относительно величин, полученных в результате докинга соответствующих эталонных соединений.

Наконец, последний эксперимент (эксперимент V) также включал эталонные соединения, однако в этом случае изменение эмбеддингов путем добавления шума не применяли, при этом в «энергетический» нейрон передавали улучшенное значение энергии.

Как видно из графика кумулятивного распределения свободной энергии связывания (рис.), модели из экспериментов II, IV и V генерировали больше соединений с более низкими значениями энергии, поскольку соответствующие им линии на графике расположены ниже, чем в других экспериментах. Сдвиг кумулятивной кривой вправо относительно других линий для каждой доли фиксированного набора данных f означает, что остальная часть набора данных 1–f для кривой, сдвинутой вправо, содержит больше соединений с более низкими значениями энергии. Например, при f = 0,8, можно наблюдать, что для экспериментов I и III лучшие 20% сгенерированных соединений демонстрируют значения свободной энергии связывания выше –8,0 ккал/моль, а для экспериментов II и IV соответствующие значения немного ниже –8,0 ккал/моль. В то же время результаты эксперимента V значительно ближе к величине –8,5 ккал/моль.

Таким образом, сравнение моделей показало, что попытка генерировать новые соединения на основе уже существующих лигандов дает лучшие результаты по сравнению с их генерацией из гауссовского шума (рис.). Кроме того, энергетической модели автоэнкодера удалось использовать введенный в латентный слой «энергетический» нейрон и получить больше соединений с более низкими значениями свободной энергии связывания с целевым белком, чем эмбеддинговой модели (рис.). При этом добавление гауссовского шума уменьшает преимущества использования «энергетических» нейронов. Обобщая все полученные данные, можно сделать вывод, что архитектура модели автоэнкодера, использованная в эксперименте V, является наиболее перспективной из всех рассмотренных вариантов.

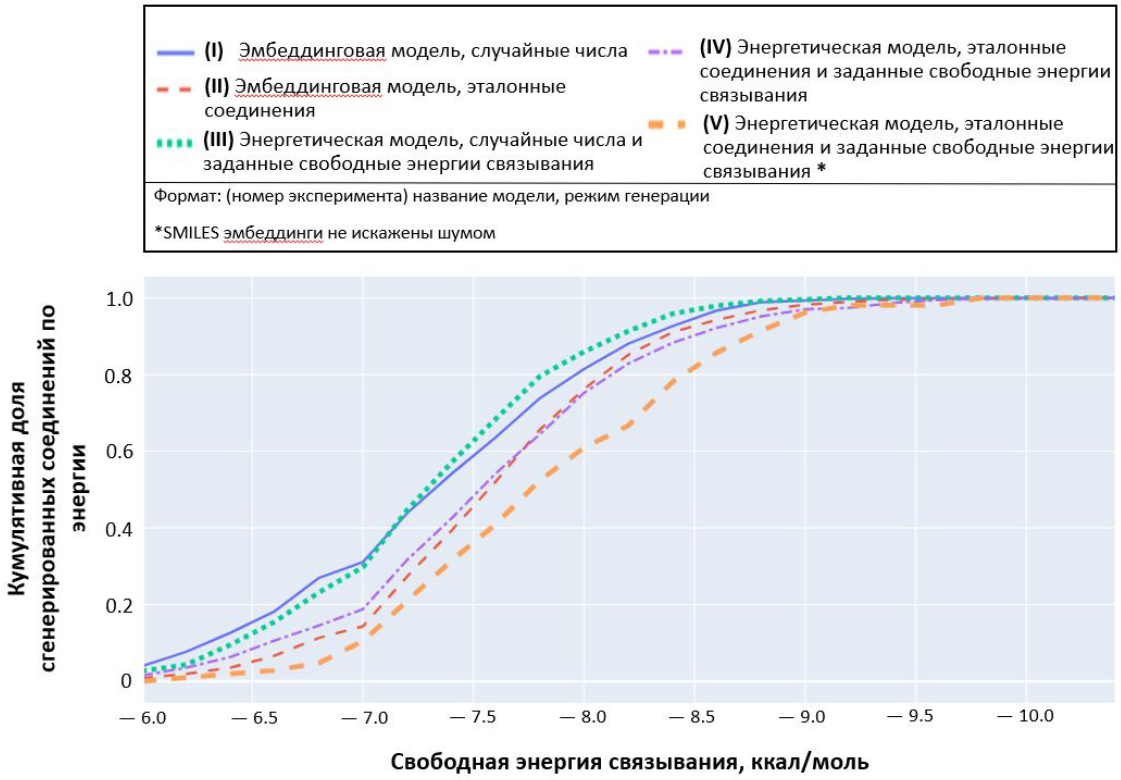

Рис. – Кумулятивные распределения свободной энергии связывания

Заключение

С помощью технологий глубокого обучения разработана генеративная нейронная сеть для компьютерного дизайна новых потенциальных ингибиторов, способных блокировать M^{Pro} SARS-CoV-2. Использование разработанной нейронной сети совместно с методом молекулярного докинга продемонстрировало ее значительный потенциал для заполнения неисследованных областей химического пространства новыми молекулами с заданными свойствами, что подтверждается полученными результатами, согласно которым из 4805 сгенерированных соединений только одно присутствовало в исходном наборе данных. Разработанные модели глубокого обучения в сочетании с традиционными методами молекулярного моделирования могут стать продуктивной основой для идентификации новых перспективных соединений против COVID-19, терапевтическое действие которых основано на блокаде основной протеазы SARS-CoV-2, критически важной для репликации и транскрипции вируса [2].

Работа поддержана Белорусским республиканским фондом фундаментальных исследований (проекты Ф21КОВИД-002, Х21COVID-003,

Ф21АРМГ-001) и Союзом международных научных организаций ANSO (ANSO-CR-PP-2021-04).

Библиографические ссылки

1. Lipinski C.F., Maltarollo V.G., Oliveira P.R., da Silva A.B.F., Honorio, K.M. Advances and Perspectives in Applying Deep Learning for Drug Design and Discovery // Front. Robotics and AI. 2019. № 6. P. 108.
2. Pillaiyar T., Manickam M., Namasivayam V., Hayashi Y., Jung S.H. An overview of Severe Acute Respiratory Syndrome-Coronavirus (SARS-CoV) 3CL protease inhibitors: peptidomimetics and small molecule chemotherapy // J. Med.Chem. 2016. № 59(14). P. 6595–6628. doi: 10.1021/acs.jmedchem.5b01461.