

РАЗРАБОТКА ГЕНЕРАТИВНЫХ СЕТЕЙ ДЛЯ ПОИСКА ПОТЕНЦИАЛЬНЫХ ЛЕКАРСТВ ПРОТИВ ВИЧ

А. А. Малыщик

akim.malyschik@gmail.com;

*Научный руководитель — А. В. Тузиков, доктор физико-математических наук,
профессор*

Несмотря на значительный прогресс, достигнутый в разработке новых эффективных ингибиторов проникновения ВИЧ-1, в настоящее время нет лицензированных противовирусных препаратов, основанных на ингибировании критических взаимодействий белка gp120 оболочки ВИЧ-1 с клеточным рецептором CD4 [1]. В связи с этим исследования по разработке терапевтических средств, ингибирующих связывание белка gp120 с молекулой CD4, остаются чрезвычайно актуальными.

В настоящей работе была разработана генеративная модель нейронной сети, позволяющая получать SMILES-описания химических соединений, обладающих низкой энергией связывания с гликопротеином gp120. С помощью разработанной модели были сгенерированы соединения, для которых в дальнейшем была вычислена оценка энергии связывания методом молекулярного докинга.

В результате проведенного исследования были получены соединения с энергией связывания, сравнимой с энергиями известных на данный момент ингибиторов NBD11021 и NBD14010.

Ключевые слова: методы глубокого обучения; механизм внимания; трансформер, ВИЧ-1; белок gp120; ингибиторы проникновения ВИЧ-1; методы молекулярного моделирования; лекарственные препараты против ВИЧ

РАЗРАБОТКА МОДЕЛИ

Ставилась задача для входных данных получить их посимвольное представление в латентном пространстве (эмбеддинги), и, основываясь на этих эмбеддингах, посимвольно восстановить описание соединения в формате SMILES.

Для решения поставленной задачи было решено использовать архитектуру, основанную на модели BERT [2]. По сравнению с оригинальной архитектурой, было уменьшено число слоев энкодера (с 12 до 6), а также размеры латентного и промежуточного слоев. Перед началом обучения для каждого соединения из обучающей выборки 15% случайно выбранных символов заменялось на специальный токен [MASK], после чего соединения с замаскированными символами пропускались через энкодер, вычислялись эмбеддинги для символов. Полученные на выходе из энкодера эмбеддинги дополнительно обрабатывались полносвязным слоем, а после – функцией softmax.

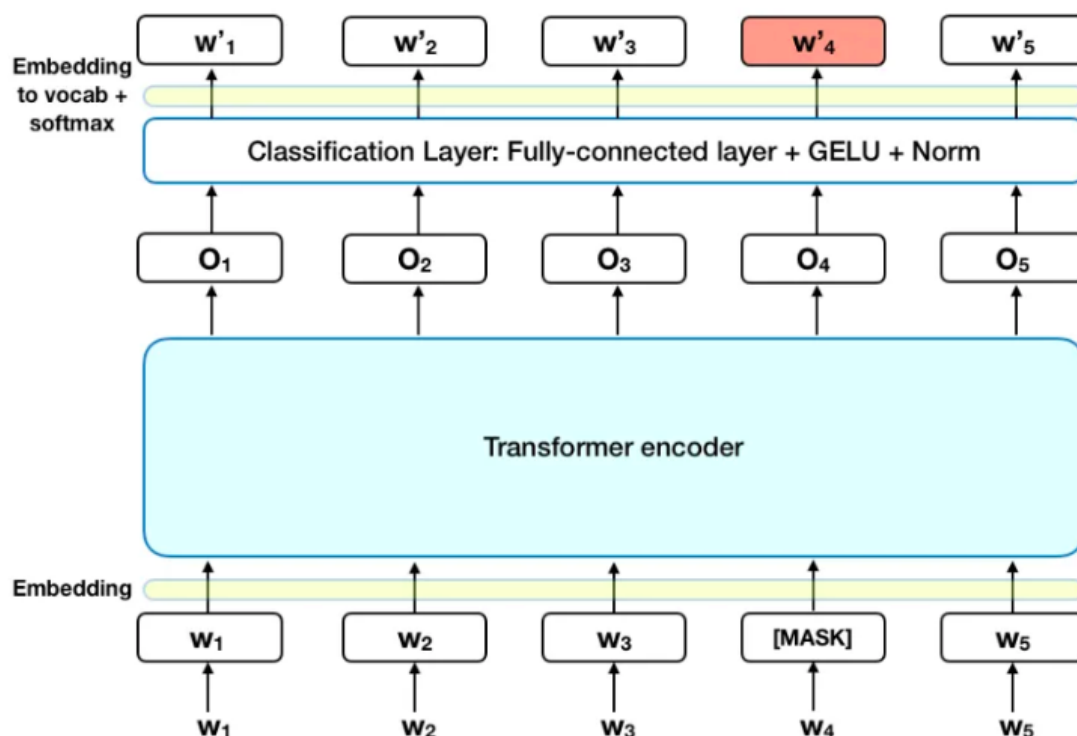


Рис. 1. Архитектура модели

ГЕНЕРАЦИЯ СОЕДИНЕНИЙ, РАСЧЕТ ЭНЕРГИИ СВЯЗЫВАНИЯ

Процесс генерации был организован следующим образом. Для соединений из выборки, полученной методом фармакофорного поиска, маскировалась часть символов, после чего соединения пропускались через нейросеть. На выходе энкодер сети BERT возвращал соединение с восстановленными символами. Повторяя данную операцию несколько раз, генерировались новые SMILES-описания. После генерации с помощью инструментария из пакета RDKit с открытым исходным кодом (URL: <http://www.rdkit.org/>) проверялась корректность и интерпретируемость сгенерированных соединений.

В результате генерации соединений, проверки валидности SMILES-дескрипторов, удаления дубликатов и соединений с некорректной трехмерной структурой, осталось 2364 SMILES-описания соединений. Для них был проведен молекулярный докинг с белком gp120. Трехмерная структура белка gp120 была выделена из комплекса этого гликопротеина с рецептором CD4 и антителом 17b (код 1GC1 из банка данных белков [3]). Атомы водорода добавлены к структуре белка gp120 с помощью программного пакета Chimera. Также была проведена оптимизация структуры в силовом поле UFF при помощи программы OpenBabel.

Молекулярный докинг лигандов из обучающего набора данных с белком gp120 выполнялся с помощью программы AutoDock Vina с учетом

конформационной подвижности лиганда. Ячейка для докинга представляла собой фрагмент белка gp120 с координатами x (19 Å; 39 Å), y (-20 Å; 0 Å), z (73 Å; 93 Å), включающий Phe43-полость гликопротеина. Для каждого лиганда генерировали девять моделей комплекса, лучших по значению оценочной функции. При этом охват конформационного пространства был задан равным 100.

АНАЛИЗ ПОЛУЧЕННЫХ РЕЗУЛЬТАТОВ

По результатам докинга было получено распределение энергий связывания для сгенерированных молекул (рис. 2)

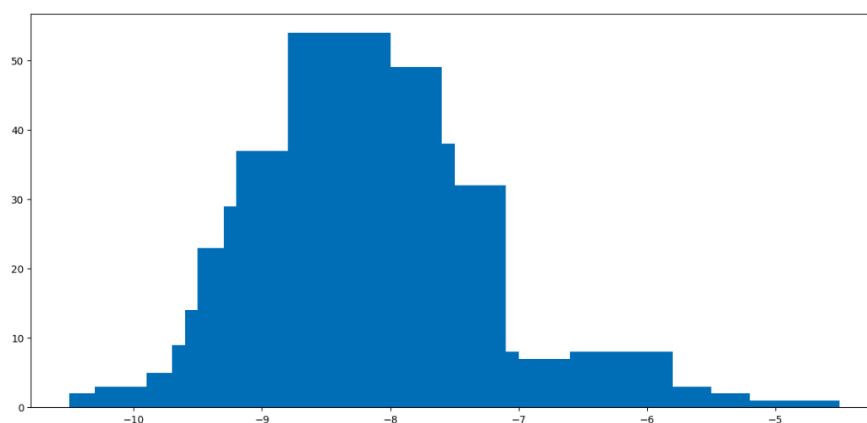


Рис.2. Распределение энергии сгенерированных соединений

Стоит отметить, что было сгенерировано несколько соединений с энергией связывания в диапазоне -10 ± 0.5 ккал/моль. Аналогичную энергию имеют известные на данный момент ингибиторы gp120 [4], то есть сгенерированные соединения представляют интерес для дальнейших исследований.

ЗАКЛЮЧЕНИЕ

Таким образом, в рамках работы был разработан алгоритм генерации SMILES-описаний химических соединений, позволяющий получать молекулы с низкой энергией связывания. Для первичного анализа сгенерированных соединений был проведен молекулярный докинг и оценена их энергия связывания.

В дальнейшем планируется сравнение лучших по энергии связывания сгенерированных соединений с ингибиторами NBD11021 и NBD14010, а также исследование полученных соединений методом молекулярной динамики. Некоторые из полученных соединений, в конечном итоге, могут быть синтезированы и протестированы на противовирусную активность, а разработанная модель может быть использована для поиска новых потенциальных ингибиторов проникновения ВИЧ-1.

Библиографические ссылки

1. *Jiang, S.* Small-molecule HIV-1 entry inhibitors targeting the epitopes of broadly neutralizing antibodies / S. Jiang, A. Tuzikov, A. Andrianov // *Cell Chemical Biology* 2022. Vol. 29, 5 P. 757-773
2. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [Electronic resource]. – Mode of access: <https://arxiv.org/abs/1810.04805>. Date of access: 17.03.2023.
3. Structure of an HIV gp120 envelope glycoprotein in complex with the CD4 receptor and a neutralizing human antibody / Kwong P. D. [et al.] // *Nature* 1998. Vol. 393 P. 648–659.
4. Synthesis, Antiviral Potency, in Vitro ADMET, and X-ray Structure of Potent CD4 Mimics as Entry Inhibitors That Target the Phe43 Cavity of HIV-1 gp120 / Curreli F [et al.] // *Journal of medicinal chemistry* 2017. Vol. 60, 7. P. 3124-3153