Применение моделей машинного обучения архитектуры BERT для предсказания показателей токсичности химических соединений

Макеев Иван Денисович
Кафедра химии и технологии биологически активных соединений имени Н.А. Преображенского, Москва
E-mail: makeev.i.d@yandex.ru
Соавторы: Павлова М.А.

Процесс открытия новых лекарств и изучения молекулярных основ заболеваний часто начинается с выявления потенциальных соединений, способных изменять активность биологических мишеней. Успех виртуального скрининга зависит от точности основополагающих прогностических моделей взаимодействия лекарств с мишенями (DTI). До недавнего времени не существовало достаточно точных неспецифических свойству вещества методов предсказания свойств, но сейчас у нас появился шанс обрести в арсенале хемоинформатика универсальный и одновременно точный инструмент для предсказания любых свойств химических соединений. Это стало возможным благодаря стремительному развитию архитектуры языковых моделей под названием BERT, представленной исследователями Google в 2018 году.

Актуальность данного исследования заключается в новом применении моделей ML на основе двунаправленных кодирующих представлений из трансформеров (BERT), таких как ChemBERTa, для прогнозирования показателей токсичности химических соединений [1]. Это отражает растущую тенденцию применения методов NLP в области хемоинформатики, где химические структуры отождествляются с текстовыми фразами и используются лингвистические подходы. Использование системы текстового представления молекул (SMILES) для предварительного обучения позволяет этим моделям улавливать сложные закономерности в молекулярных структурах и обучать значительно более сильные классификаторы и регрессионные модели поверх токенизированных представлений молекул.

Текущие модели предсказания токсичности D-MPNN, RF, SVM на Tox21 показывают результаты ROC 0.688, 0.724 и 0.708 соответственно [1]. Согласно гипотезе нашей исследовательской группы, эти значения можно повысить, используя архитектуру BERT.
Целью исследования является создание универсального и точного инструмента для предсказания различных химических и биологических свойств химических веществ, в том числе различных показателей токсичности.

Стратегия токенизации по умолчанию использует кодировщик байт-пар (BPE) из библиотеки токенизаторов HuggingFace [2]. BPE представляет собой гибрид между представлениями на уровне символов и слов, что позволяет работать с большими словарями. Очевидно, что редкие и неизвестные слова часто могут быть разложены на множество известных подслов, BPE находит наилучшее слово для сегментации путем итеративного и “жадного слияния” частых пар символов. [1, 3]

В задачах MoleculeNet ChemBERTa приближается, но не превосходит сильные базовые показатели D-MPNN [4]. Тем не менее, производительность ChemBERTa хорошо масштабируется с увеличением количества данных предварительного обучения. В среднем масштабирование от 100 тыс. до 10 млн привело к тому, что

ΔROC-AUC = +0,110 и ΔPRC-AUC = +0,059 [1]. Эти результаты свидетельствуют о том, что ChemBERTa обучается более надежным представлениям с

и способна использовать эту информацию при обучении последующим задачам [4-6].

Согласно гипотезе нашей исследовательской группы, эти значения можно повысить, незначительно увеличив размер датасета и количество параметров модели. Также модель

можно расширить на предсказание других не менее важных биологических и физико-химических свойств [7].

Наиболее важной перспективой применения CHEMBERT в химии является его способность сократить количество экспериментов, проводимых на животных. Традиционные методы определения химических свойств часто требуют проведения множества физических и биологических экспериментов, включая тесты на животных. Это не только вызывает этические и моральные вопросы, но и является дорогостоящим и времязатратным процессом. Другим важным аспектом является то, что ChemBERT может ускорить процесс открытия новых соединений, оптимизировать процессы синтеза и производства, а также повысить безопасность и экологическую устойчивость химических соединений [8].

Дальнейшие исследования в этой области позволят значительно сократить время на разработку препаратов-кандидатов. Кроме того, достаточно точные предсказания с использованием унифицированного подхода позволят использовать предсказания любых свойств молекул на лету, без обращения к внешним источникам и смогут заменить часть первичных in-vitro исследований [9,10]. Мы уже начали разрабатывать сервис для быстрого предсказания химических свойств этим методом.

Источники и литература

1) Seyone Chithrananda, Gabriel Grand, Bharath Ramsundar. ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction [Электронный ресурс] // arXiv.org. 2020. Дата обновления: 23.10.2020. URL: https://arxiv.org/abs/2010.09885 (дата обращения: 22.11.2023).
2) Subakti, A., Murfi, H., & Hariadi, N. (2022). The performance of BERT as data representation of text clustering. Journal of Big Data, 9(1). https://doi.org/10.1186/s40537-022-00564-9
3) Jiang, J., Zhang, R., Zhao, Z., Ma, J., Liu, Y., Yuan, Y., & Niu, B. (2022). MultiGran-SMILES: multi-granularity SMILES learning for molecular property prediction. Bioinformatics, 38(19). https://doi.org/10.1093/bioinformatics/btac550
4) Liu, J., Lei, X., Zhang, Y., & Pan, Y. (2023). The prediction of molecular toxicity based on BiGRU and GraphSAGE. Computers in Biology and Medicine, 153. https://doi.org/10.1016/j.compbiomed.2022.106524
5) Bu, Y., Gao, R., Zhang, B., Zhang, L., & Sun, D. (2023). CoGT: Ensemble Machine Learning Method and Its Application on JAK Inhibitor Discovery. ACS Omega, 8(14). https://doi.org/10.1021/acsomega.3c00160
6) Lang, A. S. (2023). Fine-Tuning ChemBERTa-2 for Aqueous Solubility Prediction. Annals of Chemical Science Research, 4(1). https://doi.org/10.31031/acsr.2023.04.000578
7) Nair, V. V., Pradeep, S. P., Nair, V. S., Pournami, P. N., Gopakumar, G., & Jayaraj, P. B. (2022). Deep Sequence Models for Ligand-Based Virtual Screening. Journal of Computational Biophysics and Chemistry, 21(2). https://doi.org/10.1142/S2737416522500107
8) Lee, J., Myeong, I. S., & Kim, Y. (2023). The Drug-Like Molecule Pre-Training Strategy for Drug Discovery. IEEE Access, 11. https://doi.org/10.1109/ACCESS.2023.3285811
9) Rahimovich, D. R., Abdiqayum O'G'li, S. R. zmat, & Qaxramon O'G'li, A. S. (2021). Predicting the activity and properties of chemicals based on RoBERTa. International Conference on Information Science and Communications Technologies: Applications, Trends and Opportunities, ICISCT 2021. https://doi.org/10.1109/ICISCT52966.2021.9670046

10) Kang, H., Goo, S., Lee, H., Chae, J. W., Yun, H. Y., & Jung, S. (2022). Fine-tuning of BERT Model to Accurately Predict Drug–Target Interactions. Pharmaceutics, 14(8). https://doi.org/10.3390/pharmaceutics14081710