

НОВЫЕ ИНТЕРПРЕТИРУЕМЫЕ ЭМБЕДДИНГИ ПЕПТИДОВ НА ОСНОВЕ СВОЙСТВ АМИНОКИСЛОТ ДЛЯ УЛУЧШЕННОГО ПРЕДСКАЗАНИЯ АКТИВНОСТИ

Нам Е.В. (ИТМО), Дин Е. (ИТМО)

Научный руководитель – кандидат химических наук, Серов Н.С. (ИТМО)

Введение. Пептиды – короткие аминокислотные последовательности, проявляющие высокую эффективность действия и селективность по отношению к мишеням. Они находят широкое применение в биотехнологии и медицине, например, для терапии онкологических и генетических заболеваний [1]. Однако разработка новых пептидов химическими методами связана со значительными затратами времени и ресурсов [2]. В подобных условиях релевантно применение методов машинного обучения (МО) и искусственного интеллекта (ИИ), и они уже применяются для исследований белков, например AlphaFold [3]. К сожалению, для прикладных задач предсказательная способность моделей все еще мала, дополнительной сложностью является узкая специализация разрабатываемых подходов – модели МО и нейронные сети, обученные для решения одной задачи, с малой вероятностью продемонстрируют высокие метрики на другой. Эти проблемы связаны с применяемыми дескрипторами и эмбедингами – они недостаточно эффективны для отражения комплексных свойств пептидов. В связи с этим целью данной работы является разработка новых интерпретируемых эмбедингов пептидов, которые были бы эффективны при обучении моделей машинного и глубокого обучения для решения различных задач.

Основная часть. Для получения эмбедингов мы разработали нейронную сеть с архитектурой сверточного автоэнкодера (САЭ). Принцип действия подобных моделей заключается в снижении размерности пространства признаков до заданных значений. В данном случае сжатие происходило вдоль оси длины аминокислотной последовательности для получения из матрицы признаков числового вектора, однозначно характеризующего целевой пептид. В качестве начальных данных для составления матриц признаков мы использовали дескрипторы аминокислот из библиотеки RDKit – суммарно 43, а также 3 дескриптора из результатов квантово-химических расчетов на основе теории функционала плотности [4]. Эти данные были нормализованы и использованы для обучения САЭ.

Эмбединги, полученные с помощью разработанной модели, были валидированы *in silico* на 4 различных датасетах по активности пептидов – антибактериальной [5], противовоспалительной [6], антидиабетической [7] и антиоксидантной [8]. Сравнение проводилось с 3 распространенными методами энкодинга, а именно ван хот, тример и блоссум энкодингами. Разработанные эмбединги продемонстрировали метрики, сравнимые с таковыми у бенчмарк подходов. Также была изучена их интерпретируемость: результирующие элементы эмбедингов связаны с соответствующими исходными свойствами аминокислот, что позволяет предполагать свойства, наиболее значимые для решения поставленных задач.

Выводы. Разработана модель глубокого обучения с архитектурой сверточного автоэнкодера для получения эмбедингов пептидов. Проведена *in silico* валидация результатов, полученные метрики сравнимы с бенчмарк методами. Изучена интерпретируемость эмбедингов модели, продемонстрирована связь исходных свойств аминокислот с результирующими элементами эмбедингов, и их значимость для решения поставленных в валидации задач.

Список использованных источников:

1. Wang, L., Wang, N., Zhang, W. *et al.* Therapeutic peptides: current applications and future directions. *Sig Transduct Target Ther* **7**, 48 (2022).
2. Emonts, J., Buyel, J.F. An overview of descriptors to capture protein properties – Tools and perspectives in the context of QSAR modeling, *Comput. Struct. Biotechnol. J*, V. 21, P. 3234-3247 (2023).
3. Abramson, J., Adler, J., Dunger, J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024). <https://doi.org/10.1038/s41586-024-07487-w>
4. Hu, X., Lenz-Himmer, MO. & Baldauf, C. Better force fields start with better data: A data set of cation dipeptide interactions. *Sci Data* **9**, 327 (2022). <https://doi.org/10.1038/s41597-022-01297-3>
5. Cao, Q.; Ge, C.; Wang, X. *et al.* 2023. Designing Antimicrobial Peptides Using Deep Learning and Molecular Dynamic Simulations. *Briefings in Bioinformatics* 24(2). bbad058. doi.org/10.1093/bib/bbad058.
6. Raza, A.; Uddin, J.; Almuhaimeed, A. *et al.* 2023. AIPs-SnTCN: Predicting Anti-Inflammatory Peptides Using FastText and Transformer Encoder-Based Hybrid Word Embedding with Self-Normalized Temporal Convolutional Networks. *Journal of chemical information and modeling* 63(21): 6537-6554. doi.org/10.1021/acs.jcim.3c01563.
7. Chen X, Huang J, He B. 2022. AntiDMPpred: a web service for identifying anti-diabetic peptides. *PeerJ* 10:e13581 <https://doi.org/10.7717/peerj.13581>
8. Qin, D.; Jiao, L.; Wang, R., et al. 2023. Prediction of Antioxidant Peptides Using a Quantitative Structure–Activity Relationship Predictor (AnOxPP) Based on Bidirectional Long Short-Term Memory Neural Network and Interpretable Amino Acid Descriptors. *Computers in Biology and Medicine* 154(106591). doi.org/10.1016/j.combiomed.2023.106591.