ОПТИМИЗАЦИЯ АРХИТЕКТУРЫ ПОЛНОСВЯЗНОЙ НЕЙРОННОЙ СЕТИ ДЛЯ ЗАДАЧИ ОЦЕНИВАНИЯ ЛИПОФИЛЬНОСТИ ОРГАНИЧЕСКИХ СОЕДИНЕНИЙ

Пякилля Б.И.[1]
Томский политехнический университет, ИШИТР, ОАР, старший преподаватель,
e-mail: morphism@tpu.ru

Аннотация

В современной химии одним из ключевых параметров, определяющих свойства органических соединений, является липофильность. Традиционные методы оценивания сталкиваются с ограничениями, такими как высокие временные и ресурсные затраты. Таким образом, решением становится использование методов машинного обучения, способных предсказывать липофильность. Целью работы является оптимизация архитектуры нейронной сети, предназначенной для оценивания липофильности органических соединений.

Ключевые слова: моделирование, нейросеть, липофильность, хемоинформатика, машинное обучение

Введение

В настоящее время, в химии и фармакологии, липофильность органических соединений признается одним из важнейших факторов, влияющих на их характеристики и активность [1]. Однако традиционные подходы к её измерению, включающие в себя как экспериментальные, так и теоретические методы, часто ограничены из-за значительных затрат времени и ресурсов, а также из-за сложностей их применения к новым или комплексным структурам [2]. В этом контексте, методы машинного обучения, которые могут предсказывать липофильность на основе структурных данных молекул, представляют собой перспективное направление. Цель данного исследования - разработка и настройка архитектуры нейронной сети для повышения точности и эффективности в оценке липофильности органических соединений.

Важные исследования в области применения нейронных сетей и машинного обучения к задачам в области химии и фармакологии, в том числе для оценки липофильности, акцентируют внимание на методах глубокого обучения, например, на графовых нейронных сетях. Примером может служить исследование, опубликованное в Journal of Cheminformatics, где была представлена нейросетевая модель для прогнозирования липофильности и растворимости в воде, с каждым химическим соединением, представленным в виде математического графа [3]. Другие работы показывают применение глубоких нейросетевых моделей, с использованием так называемой архитектуры «трансформер», для предсказания молекулярных свойств [4]. Эти исследования демонстрируют возможности машинного обучения в решении сложных проблем в области химии и фармакологии и подчеркивают значимость выбора и настройки архитектуры модели для обеспечения высокой точности и эффективности предсказаний. Однако, одной из главных проблем вышеперечисленных работ и большинства современных методов, использующих графовые нейронные сети, является сложность в обработке и представлении химических структур в форме графов, что требует значительных вычислительных ресурсов и специализированных знаний в области хемоинформатики. Кроме того, эффективность таких моделей сильно зависит от точности и полноты входных данных, что становится серьёзным препятствием учитывая ограниченность химических данных в фармакологии.

Традиционные полносвязные нейронные сети (также известные как FC или Dense Networks), при эффективном выборе способа представления химических соединений, могут служить в качестве альтернативного метода для анализа липофильности. Выбор соответствующих признаков и архитектуры может значительно упрощать задачу для таких нейросетевых моделей, снижая потребности в объеме данных и достигая высокой точности в предсказаниях даже с использованием ограниченных наборов данных. В этой работе акцент сделан на нахождение идеальной архитектуры для полносвязной нейронной сети, исследуя, как различные изменения в архитектуре и гиперпараметрах влияют на способность сети предсказывать липофильность. Предложен подход к поэтапной оптимизации, начиная с определения исходной структуры сети и заканчивая тонкой настройкой гиперпараметров через метод автоматической оптимизации, такой как алгоритм поиска по сетке (Grid Search).

Основная часть

Для обучения нейронных сетей использовался публичный датасет липофильности, полученный из базы данных ChEMBL [5], включающей сведения о химических соединениях и их биологической активности. Липофильность, представляющая собой безразмерный логарифмический показатель, отражает способность молекулы связываться с жирами, и варьируется в интервале от -2 (высокая гидрофильность) до 5 (высокая липофильность) [6]. Набор данных имеет следующие характеристики:

• 4200 органических соединений.

• Среднее значение липофильности равняется 2,18.

• Стандартное отклонение липофильности равняется 1,20.

График распределения значений липофильности соединений изображен на рис. 1.

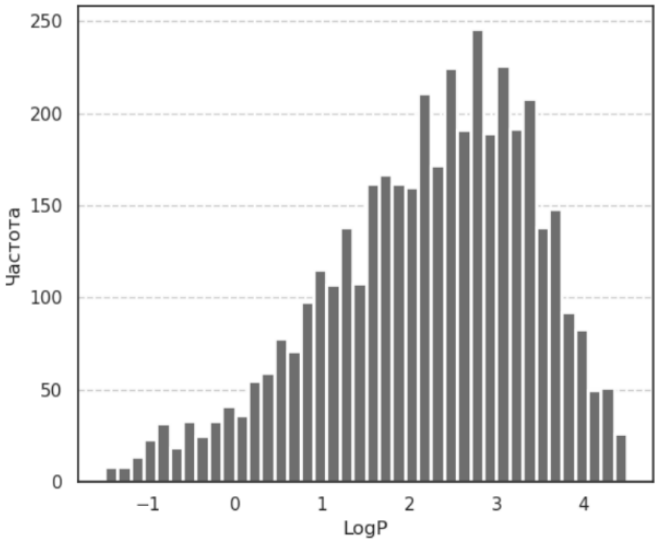

Рис. 1. График распределения липофильности в данных ChEMBL

Данные будут разделены на обучающий и тестовый наборы в пропорции 90:10 соответственно, причем обучающий набор будет применяться для тренировки и кросс-валидации нейронной сети, а тестовый — для проверки её эффективности на данных, с которыми модель не сталкивалась ранее. Этот метод деления данных широко распространен и рекомендуется для проверки способности модели адекватно работать с новыми данными [7]. Чтобы обеспечить воспроизводимость результатов, было установлено начальное значение генератора случайных чисел (random seed) в 42.

Для задачи предсказания липофильности, которая является задачей регрессии, оптимальным вариантом выбора метрики обычно является средняя квадратическая ошибка (Mean Squared Error, MSE) или корень из средней квадратической ошибки (Root Mean Squared Error, RMSE) [5].

В данной работе мы возьмем как основную метрику корень из средней квадратической ошибки из-за её лучшей интерпретации по сравнению с MSE. Такой выбор объясняется тем, что ошибка RMSE выражается в исходных единицах измерения целевой переменной LogP.

В хемоинформатике для количественного описания молекул обычно используются различные типы химических дескрипторов и отпечатков (fingerprints) [7]. Эти признаки помогают в анализе и сравнении молекулярных структур, а также в предсказании их свойств и биологической активности. Вот некоторые из наиболее часто используемых признаков:

• Молекулярные Дескрипторы. Молекулярный вес, количество водородных доноров и акцепторов, площадь поверхности молекулы, момент инерции и т. д.

• MACCS keys. Стандартный набор из 166 битов, представляющих наличие или отсутствие определенных химических структур или паттернов в молекуле.

• Extended Connectivity Fingerprints (ECFP). Отпечатки, основанные на структуре молекулы, которые учитывают окружение каждого атома.

•Continuous Distributed Description of Drug-like molecules (CDDD). Метод представления молекул, основанный на использовании глубокого обучения для преобразования молекулярных структур в непрерывное векторное пространство [8].

В данной работе за входные признаки, описывающие химические соединения, будут взяты Химческие отпечатки (ECFP), являющиеся стандартным выбором в большинстве задач хемоинформатики [7].

Процесс обучения

Процесс обучения нейронной сети связан не только с выбором пространства признаков, но с выбором таких условий, которые обеспечат наилучшее качество работы полученной модели на тестовом множестве т.е. позволят избежать переобучения. В качестве основной модели будет выбрана полносвязная нейронная сеть, входная размерность которой будет зависеть от выбранного пространства признаков, а количество скрытых слоев будет равняться двум, размерность же выходного слоя будет равняться одному нейрону в связи с необходимостью предсказания лишь одного значения, липофильности. Первый скрытый слой будет включать в себя 64 нейрона, а второй скрытый слой будет состоять из 32 нейронов. В виде активационной функции скрытых нейронов была выбрана ограниченная линейность (ReLU - Rectified Linear Unit), одним из главных преимуществ которой является уменьшение эффекта затухающего градиента, что часто встречается в глубоких сетях с сигмоидными или тангенциальными функциями активации. В качестве оптимизатора был выбран Adam (Adaptive Moment Estimation), преимуществом которого является адаптация скорости обучения для каждого параметра [9, 10].

Такая архитектура выбранной полносвязной нейронной сети обусловлена своей вычислительной простотой и задачей избежать переобучения, чего нельзя сказать о графовых нейронных сетях, которые имеют на порядок больше параметров для обучения. Наличие большего количества параметров для обучения требует большого количества обучающих данных, что часто является проблемой в хемоинформатике, где сбор данных является дорогостоящим процессом [7, 9].

Определение архитектуры нейронной сети

Для определения наилучшей архитектуры нейронной сети в смысле значения тестового RMSE и выбранных ECFP признаков, будем использовать поиск по сетке (Grid Search), где будем последовательно перебирать количество нейронов в обоих скрытых слоях, начиная с 1 нейрона до 256 [11]. Ограниченность данного диапазона связана с вычислительными затратами на обучение сети, возрастающими при увеличении количества нейронов.

Результаты поиска представлены на рис. 2 и как видно из него, наименьшее значение кросс-валидационной RMSE достигается при 229 нейронах в обоих скрытых слоях и равняется 0,786. Значение же тестовой RMSE равняется 0,687, что ниже, чем значение, получаемое при использовании базовой архитектуры (64 нейрона в первом скрытом слое, 32 нейрона во втором) на 10,3 %.

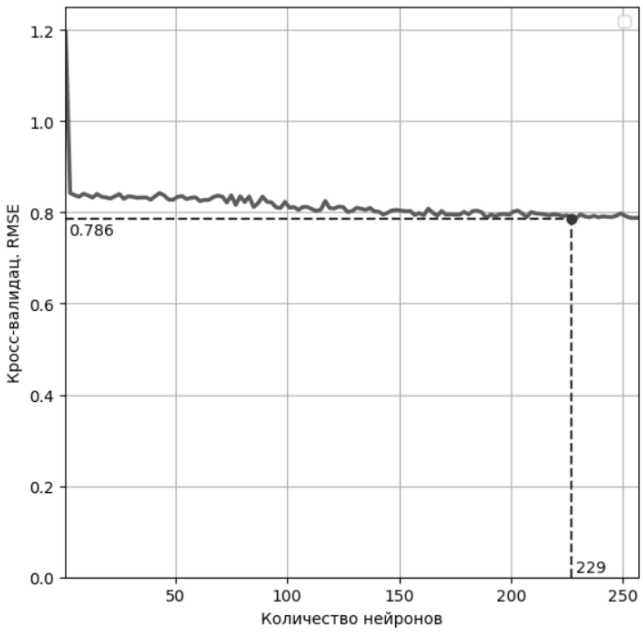

Рис. 2. График зависимости кросс-валидационной RMSE от количества нейронов в скрытых слоях

Заключение

Результаты работы, посвященной оптимизации архитектуры полносвязной нейронной сети для предсказания липофильности с использованием ECFP дескрипторов, демонстрируют важность подбора правильной архитектуры в смысле выбранной метрики, в данном случае среднеквадратичной ошибки. Оптимизация архитектуры нейронной сети с учетом ECFP признаков, особенно в части количества нейронов в скрытых слоях, позволила достигнуть высокой точности предсказаний. Данная модель показала результаты, превосходящие те, что были получены с использованием исходной неоптимизированной, с точки зрения архитектуры, модели. Это подчеркивает значимость тщательного подбора архитектуры модели для повышения ее предсказательной способности в химических исследованиях.

Список использованных источников

1. Wardecki D. Assessment of Lipophilicity Parameters of Antimicrobial and Immunosuppressive Compounds / D. Wardecki, M. Dołowy, K. Bober-Majnusz // Molecules. – 2023. – Vol. 28. – no. 6. – P. 1-14.
2. Integrating the Impact of Lipophilicity on Potency and Pharmacokinetic Parameters Enables the Use of Diverse Chemical Space during Small Molecule Drug Optimization / R. Miller, M. Madeira, H. Wood, W. Geissler, C. Raab, I. Martin // J. Med. Chem. – 2020. – Vol. 63. – no. 21. – P. 12156-12170.
3. Tang B. A self-attention-based message passing neural network for predicting molecular lipophilicity and aqueous solubility. / B. Tang, S. Kramer, M. Fang // J. Cheminform. – 2020. – Vol. 12. – no. 15. – P. 1-9.
4. Song Y. Double-head transformer neural network for molecular property prediction. / Y. Song, J. Chen, W. Wang // J. Cheminform. – 2023. – Vol. 15. – no. 27. – P. 1-16.
5. Wu Z. MoleculeNet: a benchmark for molecular machine learning / Z. Wu, B. Ramsundar, E. N. Feinberg // Chemical science. – 2018. – Т. 9. – №. 2. – P. 513-530.
6. Кольман Я., Рем К. Г. Наглядная биохимия. – Лаборатория знаний. – 2023. – 513 с.
7. Маджидов Т.И. Введение в хемоинформатику: учебное пособие. Ч. 1: Компьютерное представление химических структур. / Т.И. Маджидов, И.И. Баскин, А.А. Варнек. – Казань: Изд-во Казанского ун-та. – 2015. – 174 с.
8. Gómez-Bombarelli R. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. / R. Gómez-Bombarelli, N. W. Jennifer, D. Duvenaud // ACS Central Science. – 2018. – Vol. 4. – no. 2. – P. 268-276.
9. Bengio Y. Deep Learning. / Y. Bengio, I. Goodfellow, A. Courville // MIT Press. – 2016. – 800 p.
10. Stokes J.M. A Deep Learning Approach to Antibiotic Dis-covery / J.M. Stokes, K. Swanson, K. Yang // Cell, 2020. – Vol. 180. – no. 4. – P. 475-483.
11. Ali Y.A Hyperparameter Search for Machine Learning Algo-rithms for Optimizing Computational Complexity. // Pro-cesses. – 2023. – Vol. 11. – no. 2. – P. 1-21.