УДК 004.89

Б.И. Пякилля, В.И. Гончаров

Особенности разработки полносвязных нейросетей для решения задачи оценивания липофильности органических соединений

Оценка липофильности малых органических соединений играет ключевую роль в разработке и оптимизации новых лекарственных препаратов. К сожалению, экспериментальные методы требуют значительных временных и материальных затрат, включая использование лабораторного оборудования и реагентов. Кроме того, для получения достоверных результатов часто требуется ручная проверка и корректировка данных, что увеличивает трудоемкость процесса. В отличие от этого, компьютерные методы, такие как машинное обучение, предлагают более быстрые и менее ресурсоемкие способы оценки липофильности, которые позволяют эффективно обрабатывать большие объемы данных, адаптироваться к сложным зависимостям между структурой молекулы и ее липофильностью. Разработка нейросетевых моделей для решения задачи оценивания липофильности является непростой задачей в связи с недостаточным количеством экспериментальных данных и дороговизной их получения, а также высокими вычислительными затратами при использовании графовых нейросетевых моделей. В данной работе представлен анализ наиболее популярных способов описания химических структур в контексте поставленной задачи с целью их использования для построения полносвязных нейросетевых моделей, являющихся менее требовательными к объему обучающих данных. На основе проведенного анализа выбираются признаки, наилучшим образом описывающие органические соединения из открытого набора данных о липофильности, собранных из базы данных ChEMBL. Проводится поиск оптимальной архитектуры нейросетевой модели для выбранных в результате анализа признаков.

Ключевые слова: моделирование, нейросеть, липофильность, хемоинформатика.
DOI: 10.21293/1818-0442-2024-27-1-86-94

Оценка липофильности малых органических соединений играет ключевую роль в разработке и оптимизации новых лекарственных препаратов. Липофильность, представляющая собой способность молекулы распределяться между водной и липидной фазами, является одним из наиболее важных физико-химических параметров, влияющих на адсорбцию, распределение, метаболизм, выведение и токсичность (ADMET) биоактивных молекул. Этот параметр необходим для понимания механизма проникновения лекарственных средств через биологические мембраны, их связывания с белками крови и рецепторами, а также их общей биологической активности [1].

Поскольку традиционные методы оценки липофильности, такие как метод «тряски с флаконом», могут быть трудоемкими и ограниченными в применении, современные подходы, особенно методы машинного обучения и нейросетевые модели, представляют большой интерес для исследователей [2, 3]. Использование методов машинного обучения становится все более популярным благодаря их способности эффективно обрабатывать большие объемы данных, адаптироваться к сложным зависимостям между структурой молекулы и ее липофильностью, а также предоставлять более точные и надежные прогнозы липофильности на основе компьютерных методов. Эти технологии обеспечивают значительное ускорение процесса открытия и разработки новых лекарственных средств, позволяя исследователям оптимизировать молекулы с желаемыми свойствами еще до их синтеза в лаборатории.

В задаче оценки липофильности органических соединений с помощью методов машинного обучения, особенно нейросетевых моделей, был достигнут заметный прогресс. Например, исследование, опубликованное в журнале Journal of Cheminformatics [4], представляет собой работу, в которой была разработана и применена нейросетевая модель для предсказания липофильности и водорастворимости молекул, в которой каждое химическое соединение представлялось в виде математического графа. Также в работе [5] было описано использование глубоких нейросетевых моделей на основе архитектуры «трансформер» для прогнозирования молекулярных свойств. В работах [6, 7] представлено применение так называемого «переноса знания» с предобученных моделей, а также предложены новые архитектуры для нейросетевых моделей, использующих графовое представление химических соединений.

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.

Обычные полносвязные нейронные сети (Fully Connected, FC или Dense Networks) с правильным выбором представления химических соединений могут предложить альтернативный подход к оценке липофильности. Важным аспектом здесь является выбор подходящих химических дескрипторов или

признаков, которые могут включать физико-химические свойства, структурные характеристики и другие молекулярные параметры.

Правильно подобранные признаки могут значительно упростить задачу для полносвязных нейросетевых моделей, уменьшая требования к объёму данных и обеспечивая высокую точность предсказаний даже при работе с ограниченными датасетами.

Постановка задачи

Настоящая статья фокусируется на задачах и особенностях разработки эффективных нейросетевых моделей для оценки липофильности, что является ключевым элементом в разработке новых фармацевтических препаратов и химических соединений. Процесс создания и настройки таких моделей включает в себя несколько ключевых этапов [8, 9]:

1. Предварительная обработка данных. Включает очистку и нормализацию данных, преобразование химических структур в формат, пригодный для машинного обучения.

2. Отбор и оценивание значимости признаков, описывающих химические соединения. Выбор характеристик молекул, которые в наибольшей степени влияют на липофильность, а также исключение избыточных или нерелевантных данных.

3. Подготовка данных к обучению. Этап содержит задачи разбиения данных на обучающую и тестовую выборки, а также определение целевой переменной для модели.

4. Разработка и оптимизация модели нейросети. На этом этапе осуществляется выбор архитектуры нейросети, определяются размерности входных и выходных слоев, проводится подбор гиперпараметров, включая количество и структуру скрытых слоев, параметры активации и критерии остановки обучения.

Целью данного исследования является анализ влияния выбора признаков, описывающих химические структуры, на процесс разработки нейросетевой модели, оценивающей липофильность малых органических соединений.

В процессе разработки нейронных сетей часто применяются язык программирования Python [8, 10], а также созданные с его помощью инструменты для обработки массивов данных, хранящих химическую информацию. Учитывая широкое распространение таких инструментов, будут использованы следующие библиотекаи и API (интерфейс программирования приложений):

1. Открытая библиотека для обработки химических данных в хемоинформатике RDKit [10].

2. Библиотека Numpy, ориентированная на эффективную работу с многомерными массивами [12].

3. Библиотека Pandas, предлагающая функциональность для детального анализа табличных данных [13].

4. Библиотеки Matplotlib и Seaborn для построения 2D-диаграмм и графиков [14, 15].

5. TensorFlow, предоставляющий развитый API для разработки нейросетевых моделей [16].

Исходные данные

Для обучения нейросетей был выбран открытый набор данных о липофильности, собранных из базы данных ChEMBL, содержащей информацию о химических соединениях и их биологических активностях [17].

В наборе данных ChEMBL каждая молекула описана с помощью строки SMILES (Simplified Molecular Input Line Entry System), являющейся способом однозначного описания состава и структуры молекулы химического вещества с использованием строки символов ASCII. Строки SMILES используются как входная информация, на основе которой происходит в дальнейшем вычисление молекуляных признаков, предоставляющих количественную информацию о молекулах. Вместе с SMILES-нотацией для каждой молекулы указано значение липофильности, которое является мерой того, насколько химическое соединение способно растворяться в жирах, маслах и неполярных растворителях. Важной характеристикой набора данных ChEMBL является его разнообразие: он содержит множество различных типов молекул, что позволяет моделям машинного обучения учиться распознавать широкий спектр химических структур.

Липофильность является безразмерной логарифмической величиной и обычно лежит в диапазоне от –2 (крайне гидрофильная молекула) до 5 (крайне липофильная) [18].

Набор данных имеет следующие характеристики:

- 4 200 органических соединений.
- Среднее значение липофильности равняется 2,18.
- Стандартное отклонение липофильности равняется 1,20.

График распределения значений липофильности соединений изображен на рис. 1.

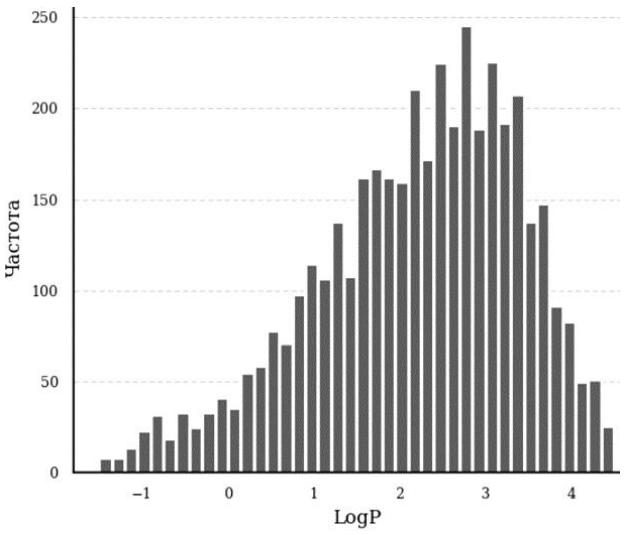

Рис. 1. График распределения липофильности в данных ChEMBL

Весь набор данных будем разделять случайным образом на обучающую и тестовую выборки в соот-

ношении 90:10, где первая будет использоваться непосредственно для обучения и кроссвалидации нейронной сети, а последняя – для её финального тестирования [9, 10]. Метод кросс-валидации позволяет оценить, как модель будет работать на данных, которые не использовались в процессе обучения, с помощью разбиения исходных обучающих данных случайным образом на несколько подмножеств. Данный процесс повторяется некоторое количество раз, и для каждой итерации одно из подмножеств используется в качестве валидационного набора, а оставшиеся подмножества – в качестве обучающего набора. Основная функция валидационного набора – предотвратить переобучение и помочь в выборе подходящих гиперпараметров, к которым могут относиться: степень регуляризации, скорость обучения или количество нейронов в скрытом слое нейронной сети.

Для воспроизводимости результатов кроссвалидации был зафиксирован random seed, равный 42. Random seed – это начальное значение генератора случайных чисел, которое используется в качестве исходной точки для создания последовательности случайных чисел. Фиксирование значения random seed обеспечивает, что разбиение данных, инициализация весов модели и другие элементы, использующие генератор случайных чисел, будут одинаковыми при каждом запуске модели.

Для задачи предсказания липофильности LogP, 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) [17]:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y-\hat{y})^2}\,, \qquad (1)$$

где y – экспериментальные значения липофильности LogP, $\hat{y}$ – предсказанные моделью значения липофильности.

В данной работе мы возьмем как основную метрику корень из средней квадратической ошибки (1) из-за её лучшей интерпретации по сравнению с MSE. Такой выбор объясняется тем, что ошибка RMSE выражается в исходных единицах измерения липофильности LogP, являющейся логарифмической величиной.

Выбор пространства признаков

В хемоинформатике для количественного описания молекул обычно используются различные типы химических дескрипторов и отпечатков (fingerprints) [19]. 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 [20].

Несмотря на то, что есть набор наиболее часто используемых признаков, нельзя заранее знать какой из вышеописанных признаков подходит наилучшим образом. Признаки должны быть выбраны таким образом, чтобы модель могла делать предсказания не только для молекул в обучающем наборе данных, но и для новых, ранее не встречавшихся соединений. В идеале признаки должны не только способствовать точности предсказаний, но и быть интерпретируемыми, чтобы обеспечить понимание того, какие молекулярные характеристики влияют на липофильность. Это особенно важно в фармацевтической индустрии для понимания механизмов действия лекарств. Как видим, это сложная и потому неоднозначная задача. Поэтому с целью ее конкретизации и, следовательно, снижения трудностей поиска решения ограничимся выбором признаков, обеспечивающих наилучшее качество предсказания модели на тестовых данных.

Отбор молекулярных признаков по значимости

Проверка молекулярных дескрипторов на корреляцию является важным шагом в процессе анализа данных в хемоинформатике и машинном обучении [21]. Это делается по причинам того, что в случаях, когда несколько признаков несут почти одинаковую информацию, это может привести к нестабильности коэффициентов модели и ухудшению ее способности к предсказанию на ранее не виданных молекулах. Кроме того, уменьшение размерности приводит к лучшей интепретации работы модели и может улучшить производительность модели, снизив вычислительные затраты.

Для оценки корреляции признаков будет использоваться линейный коэффициент корреляции Пирсона в связи со своей вычислительной простотой, что важно при работе с большим набором данных. Кроме того, в вычислительной химии молекулярные признаки типа площади поверхности полярности, количество водородных доноров и акцепторов и других структурных характеристик линейно зависят друг от друга. Вдобавок данный способ оценки корреляции легко интерпретировать в связи с наличием четкой шкалы от –1 до 1, где значения, близкие к 1 или –1, указывают на сильную положительную или отрицательную линейную зависимость соответственно

В качестве молекулярных дескрипторов будут взяты следующие признаки [19, 21, 22]:

- Молекулярный вес.
- Средний молекулярный вес.
- Количество донорно-акцепторных связей. 4 значения.
- Количество вращающихся химических связей, амидных связей.
- Количество всех атомов, тяжелых атомов, гетероатомов. 3 значения.
- Доля атомов углерода с sp3-гибридизацией.
- Количество алифатических и ароматических циклов, карбоциклов и гетероциклов. 8 значений.
- Количество спироатомов, мостиковых атомов, стереохимических центров. 4 значения.
- Топологическая площадь молекулы и площадь молекулы, доступная для растворения.
- Молекулярная липофильность и рефракция, вычисленные на основе метода Уилдмана–Криппена. 2 значения.
- Дескрипторы связности. 15 значений.

В итоге количество молекулярных признаков равняется 43 значениям.

Как отмечалось ранее, проверка молекулярных дескрипторов на корреляцию является важным шагом в процессе анализа данных. Кроме того, некоторые признаки, как молекулярный вес и средний молекулярный вес, могут быть взаимосвязаны. Для отбрасывания корреляции в данных была составлена корреляционная матрица всех 43 признаков.

Корреляционная матрица с коэффициентами корреляции Пирсона была составлена с помощью библиотек Pandas, matplotlib и seaborn.

На рис. 2 и 3 изображены корреляционные матрицы для первых 24 и оставшихся 19 признаков соответственно.

В представленных матрицах цвет ячейки, содержащей коэффициент, указывает на степень корреляции признаков: более темная ячейка означает более высокий коэффициент.

Были исключены следующие признаки, имеющие значение коэффициента корреляции выше 0,9:

1. CrippenMR. Молекулярная рефракция.
2. NumAtoms, NumHeavyAtoms. Количество всех атомов и количество тяжелых атомов.
3. NumSaturatedRings. Количество насыщенных циклов.
4. AMW. Средний молекулярный вес.
5. labuteASA, TPSA. Топологическая площадь молекулы и площадь молекулы доступная для растворения.
6. Phi, chi0n, chi0v, chi1n, chi1v, chi2n, chi2v, chi3n, chi3v, chi4n, chi4v, kappa1, kappa2, kappa3. Дескрипторы связности, исключая hallKierAlpha.

В итоге количество молекулярных признаков снизилось до 21 значения. Их корреляционная матрица представлена на рис. 4.

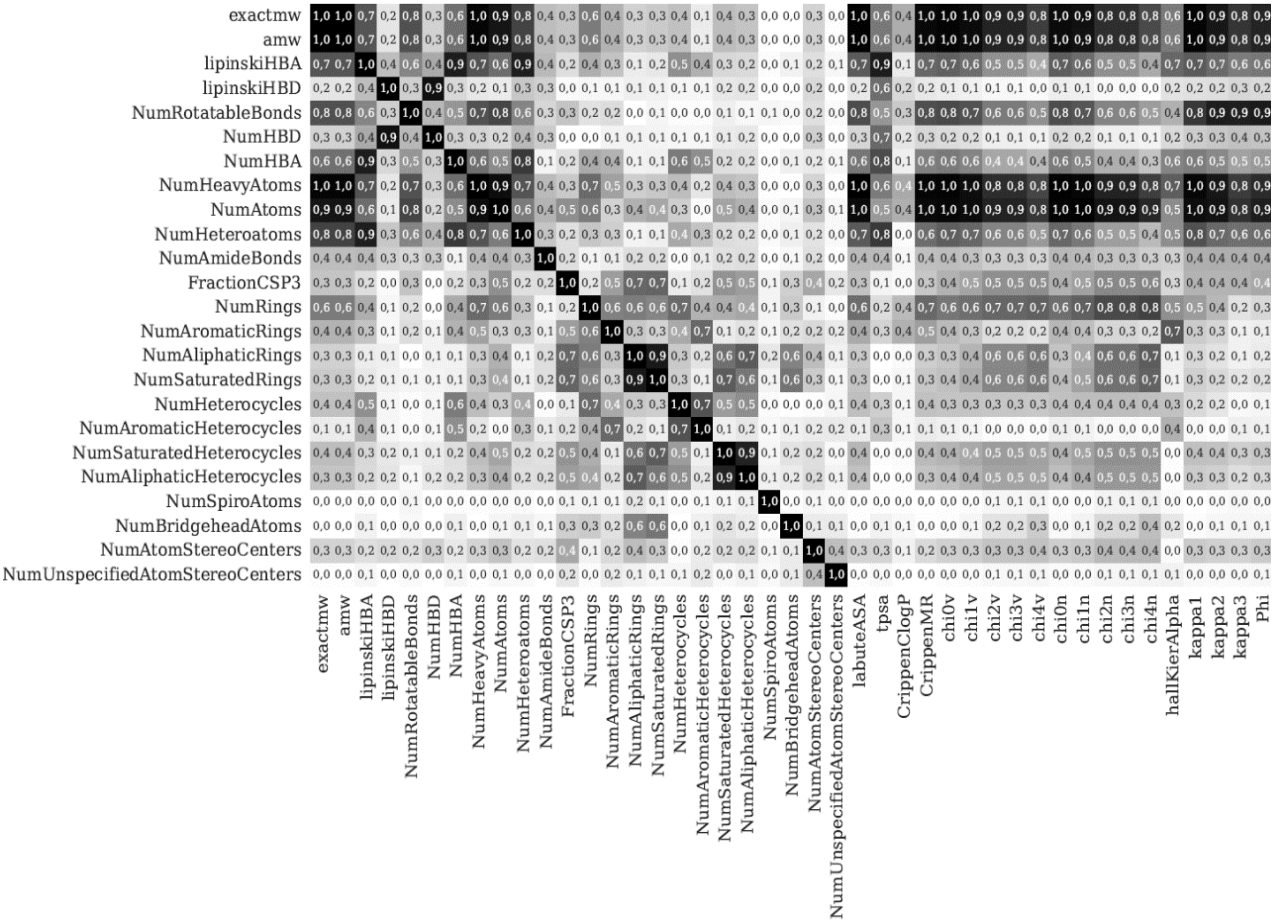

Рис. 2. Матрица корреляции первых 24 признаков

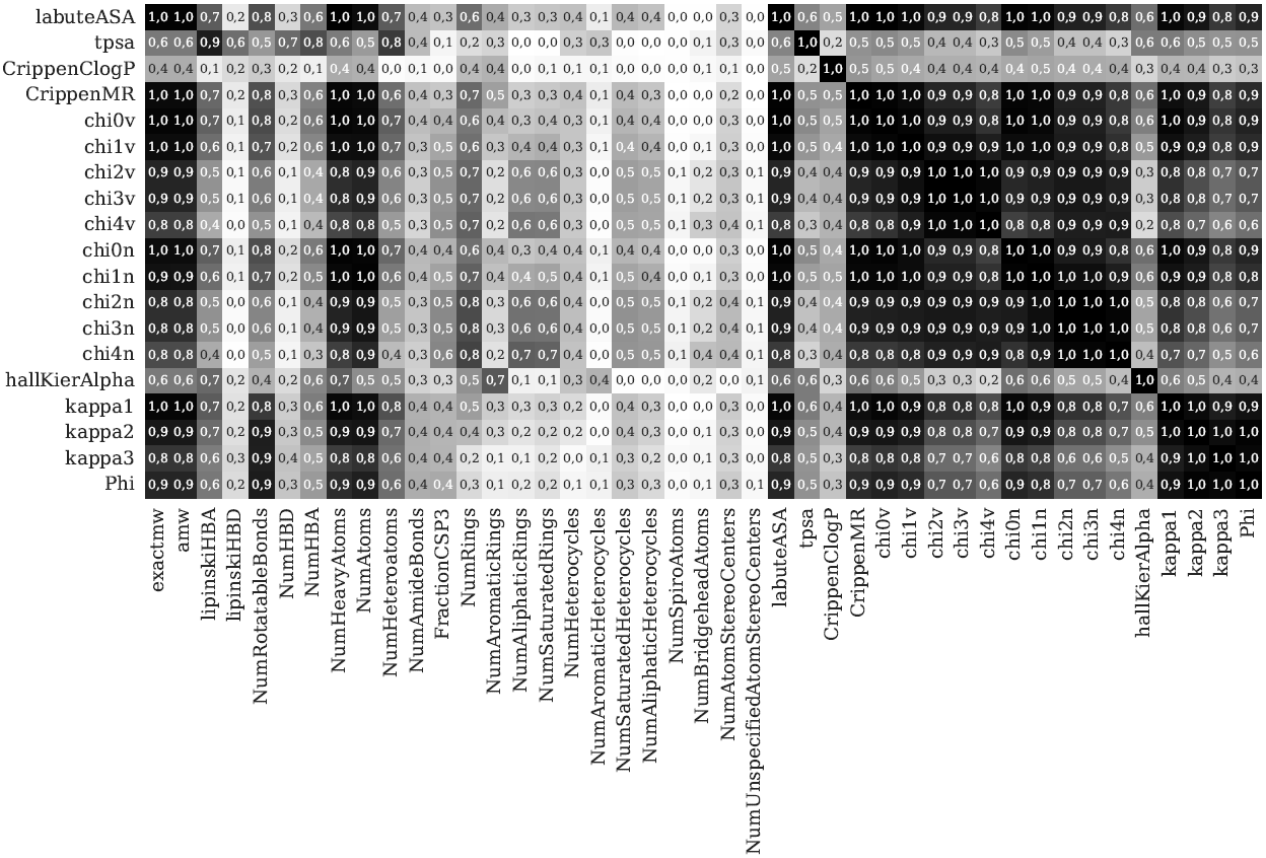

Рис. 3. Матрица корреляции 19 признаков

Химические отпечатки и ключи

Химические отпечатки (ECFP) и ключи (MACCS keys) являются популярными в хемоинформатике в связи с учетом ими химической структуры молекулы, включая информацию о её атомах и связях и легкости интерпретирования в случае с MACCS.

Оценка корреляции обычно не является стандартной практикой в случае данных признаков в связи с тем, что ECFP и MACCS обычно представляются в бинарном формате, где каждый бит отражает присутствие или отсутствие определенного химического фрагмента или паттерна. Также принцип генерации ECFP подразумевает, что каждый бит в отпечатке несет уникальную информацию, и корреляция между битами может быть не особенно информативной.

Размерности ECFP и MACCS равняются 2 048 и 166 значениям соответственно.

CDDD-признаки

CDDD-признаки (Continuous Distributed Description of Drug-like molecules) стали популярными в хемоинформатике в связи с представлением молекулы в виде непрерывных векторов, которые кодируют богатую информацию о молекулярной структуре и свойствах, что позволяет уловить более тонкие нюансы химических соединений.

В отличие от традиционных химических отпечатков, CDDD-признаки являются непрерывными и могут быть подвержены корреляционному анализу. Это позволяет выявить линейные зависимости между различными признаками и снизить размерность данных.

По умолчанию количество CDDD-признаков равняется 512, и изображение их корреляционной матрицы является непростой задачей. В результате вычисления коэффициентов корреляции было выявлено 16 признаков, которые были исключены. Таким образом, размерность была незначительно снижена до 496 значений. В отличие от вышеуказанных признаков интепретация CDDD затруднительна, так как каждый признак является некоторой точкой в 512-мерном пространстве и не связан напрямую с некоторым химическим фрагментом или паттерном.

Выбор основных признаков и параметров процесса 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 со-

стоять 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, 22].

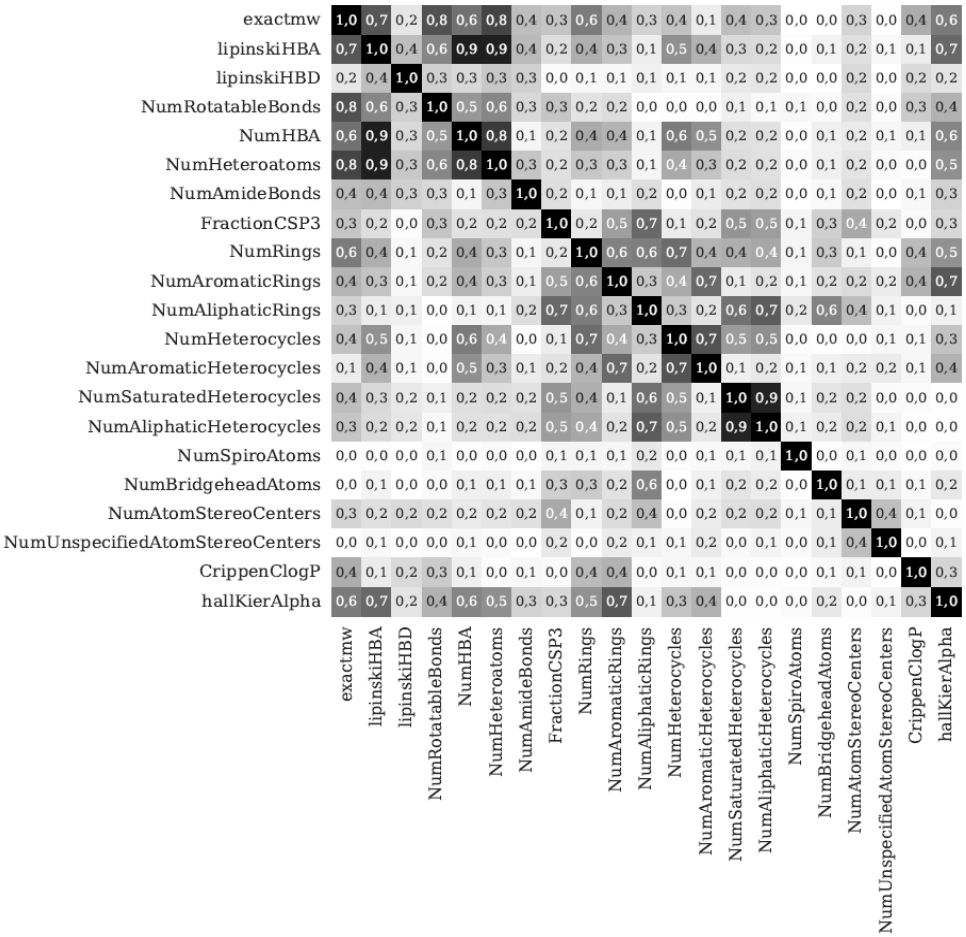

Рис. 4. Матрица корреляции оставшихся молекулярных признаков

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 [9, 19, 21].

В табл. 1 представлены получившиеся результаты при использовании различных признаков.

Как видно из результатов, отбрасывание коррелированных признаков для молекулярных и CDDD-признаков привело к увеличению значения валидационного и тестового RMSE. Причин у этого явления может быть несколько [9, 19, 23, 24].

1. **Потеря информации.** Коррелированные признаки, несмотря на их взаимную связь, могут содержать уникальную информацию, важную для предсказания целевой переменной. Удаление этих признаков приводит к потере этой информации, что может снизить точность модели.

Таблица 1

Результаты при использовании различных признаков

Вид признаков	Кроссвалидационное RMSE	Тестовое RMSE
На основе всех признаков		
Молекулярные признаки	0,855±0,025	0,832
ECFP 2048 bits	0,840±0,040	0,735
MACCS	0,836±0,010	0,808
CDDD	0,683±0,022	0,650
После отбрасывания коррелированных признаков		
Молекулярные признаки	0,880±0,020	0,858
CDDD	0,685±0,012	0,668

2. **Сложность модели.** Нейросетевые модели способны извлекать и комбинировать информацию из признаков сложным образом. Даже если признаки коррелируют между собой, сеть может научиться использовать эту корреляцию для улучшения предсказаний.

3. **Важность признаков.** В контексте молекулярных и химических свойств некоторые коррелированные признаки могут быть особенно важны для

предсказания липофильности. Отбрасывание этих признаков может исключить критически важные аспекты молекул, необходимые для точного предсказания.

Таким образом, наилучшие результаты получаются при использовании CDDD-признаков, без отбрасывания коррелированных.

Определение архитектуры нейронной сети

Для определения наилучшей архитектуры нейронной сети в смысле значения тестового RMSE и выбранных CDDD-признаков будем использовать поиск по сетке (Grid Search), где будем последовательно перебирать количество нейронов в обоих скрытых слоях, начиная с 1 нейрона до 256 [25]. Ограниченность данного диапазона связана с вычислительными затратами на обучение сети, возрастающими при увеличении количества нейронов.

Результаты поиска представлены на рис. 5. Как видно из него, наименьшее значение кросс-валидационной RMSE достигается при 227 нейронах в обоих скрытых слоях и равняется 0,652. Значение же тестовой RMSE равняется 0,610, что ниже, чем значение, получаемое при использовании базовой архитектуры, на 6,3%.

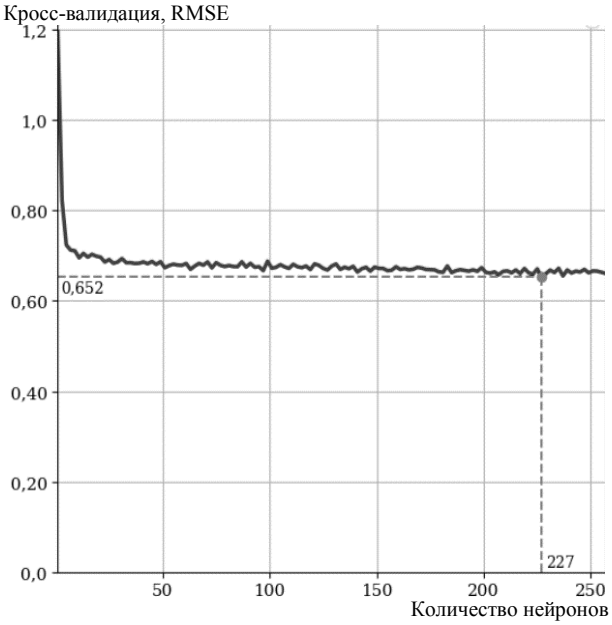

Рис. 5. График зависимости кроссвалидационной RMSE от количества нейронов в скрытых слоях

Сравнение с существующими решениями

Существующие решения в области машинного обучения для решения задачи предсказания липофильности представлены большей частью графовыми нейронными сетями, что является проблемой в случае недостатка обучающих данных либо вычислительных ресурсов. Предлагаемое же решение, как было отмечено ранее, является менее требовательным к объёму обучающему данных и вычислительных мощностей. Для сравнения предложенного решения с существующими использовались результаты, полученные авторами статьи, где использовался идентичный набор данных ChEMBL [17]. Предложенное решение сравнивалось со следующими моделями:

1. **MPNN** (Message Passing Neural Network): тип графовой нейронной сети, основанный на идее передачи сообщений между узлами в графе [26].

2. **Weave:** тип графовой сети, использующий глобальные свертки для вычисления признаков, описывающих структуру молекулы [27].

3. **GCN** (Graph Convolutional Network): классическая графовая нейронная сеть, использующая локальные свертки для вычисления молекулярных признаков [27].

4. **XGBoost** (eXtreme Gradient Boosting): модель градиентного бустинга на деревьях решений [28].

Сравнивая предложенное решение с существующими, важно отметить, что полученный результат тестовой RMSE ниже, чем лучший результат, полученный авторами вышеуказанной статьи, на 7%, который равняется 0,655 [17]. Результаты тестовых RMSE для всех моделей представлены в табл. 2.

Таблица 2

Результаты сравнения с аналогами

Модель	Тестовое RMSE
Предложенное решение	0,610
MPNN	0,757
Weave	0,734
XGBoost	0,799
GCN	0,655

Заключение

Результаты исследования, посвященного выбору признаков для предсказания липофильности с использованием молекулярных дескрипторов, ECFP, MACCS и CDDD, демонстрируют превосходство CDDD-признаков. Оптимизация архитектуры нейронной сети с учетом CDDD-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 модели авторов датасета по липофильности. Это подчеркивает значимость тщательного подбора признаков и архитектуры модели для повышения ее предсказательной способности в химических исследованиях.

Литература

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. Пякилля Б.И. Оценивание липофильности с помощью байесовских нейронных сетей / Б.И. Пякилля, В.И. Гончаров // Известия ТулГУ. Технические науки. – 2022. – № 9. – С. 288–292.

4. 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.

5. 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.

6. Wang Y. LogD7.4 prediction enhanced by transferring knowledge from chromatographic retention time, microscopic pKa and logP / Y. Wang, J. Xiong, F. Xiao // J. Cheminform. – 2023. – Vol. 15, No. 76. – P. 1–13.

7. Wieder O. Improved Lipophilicity and Aqueous Solubility Prediction with Composite Graph Neural Networks. / O. Wieder, M. Keunemann, M. Wieder, T. Seidel, C. Meyer, S. Bryant, T. Langer // Molecules. – 2021. – Vol. 26, No. 20. – P. 6185–6223.

8. Гафаров Ф.М. Искусственные нейронные сети и приложения: учеб. пособие / Ф.М. Гафаров, А.Ф. Галимянов. – Казань: Изд-во Казан. ун-та, 2018. – 121 с.

9. Bengio Y. Deep Learning. / Y. Bengio, I. Goodfellow, A. Courville. – MIT Press, 2016. – 800 p.

10. Жерон А. Прикладное машинное обучение с помощью Scikit-Learn и TensorFlow: концепции, инструменты и техники для создания интеллектуальных систем: пер. с англ. – СПб.: Альфа-книга, 2018. – 688 с.

11. Документация библиотеки RDkit: https://www.rdkit.org/docs/ (дата обращения: 26.03.2024).

12. Документация библиотеки Numpy: https://numpy.org/doc/stable/ (дата обращения: 26.03.2024).

13. Документация библиотеки Pandas: https://pandas.pydata.org/docs/reference/index.html (дата обращения: 26.03.2024).

14. Документация библиотеки Matplotlib: https://matplotlib.org/stable/index.html (дата обращения: 26.03.2024).

15. Документация библиотеки Seaborn: https://seaborn.pydata.org/api.html (дата обращения: 26.03.2024).

16. Документация библиотеки TensorFlow: https://www.tensorflow.org/api_docs (дата обращения: 26.03.2024).

17. Wu Z. MoleculeNet: a benchmark for molecular machine learning / Z. Wu, B. Ramsundar, E.N. Feinberg // Chemical science. – 2018. – T. 9, No. 2. – P. 513–530.

18. Waring M. Lipophilicity in drug discovery // Expert Opinion on Drug Discovery. – 2010. – Vol. 5, No. 3. – P. 235–248.

19. Маджидов Т.И. Введение в хемоинформатику: учеб. пособие. Ч. 1: Компьютерное представление химических структур / Т.И. Маджидов, И.И. Баскин, А.А. Варнек. – Казань: Изд-во Казан. ун-та, 2015. – 174 с.

20. 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.

21. Маджидов Т.И. Введение в хемоинформатику: учеб. пособие. Ч. 4: Методы машинного обучения / Т.И. Маджидов, И.И. Баскин, А.А. Варнек. – Казань: Изд-во Казан. ун-та, 2016. – 329 с.

22. Stokes J.M. A Deep Learning Approach to Antibiotic Discovery. / J.M. Stokes, K. Swanson, K. Yang // Cell. – 2020. – Vol. 180, No. 4. – P. 475–483.

23. Niazi S. Advances in Machine-Learning-Based Chemoinformatics: A Comprehensive Review / S. Niazi, Z. Mariam // Int. J. Mol. Sci. – 2023. – Vol. 24. – P. 11488–11503.

24. Rickert C. Efficiency-driven, correlation-based feature elimination strategy for small datasets / C. Rickert, M. Henkel, O. Lieleg // APL Mach. Learn. – 2023. – Vol. 1, No. 1. – P. 1–15.

25. Ali Y.A. Hyperparameter Search for Machine Learning Algorithms for Optimizing Computational Complexity // Processes. – 2023. – Vol. 11, No. 2. – P. 1–21.

26. Gilmer J. Neural message passing for quantum chemistry / J. Gilmer, S. Schoenholz, P. Riley, O. Vinyals // International conference on machine learning. PMLR. – 2017. – P. 1263–1272.

27. Kearnes S. Molecular graph convolutions: moving beyond fingerprints / S. Kearnes, K. McCloskey, M. Berndl, V. Pande, P. Riley // Journal of computer-aided molecular design. – 2016. – Vol. 30. – P. 595–608.

28. Chen T. Xgboost: A scalable tree boosting system / T. Chen, C. Guestrin // In Proceedings of the 22nd ACM International conference on knowledge discovery and data mining. – 2016. – P. 785–794.

Пякилля Борис Иванович
Ст. преп. отделения автоматизации и робототехники (ОАР) Инженерной школы информационных технологий и робототехники (ИШИТР) Томского политехнического университета (ТПУ)
Ленина пр-т, 30, г. Томск, Россия, 634050
ORCID: 0000-0003-1992-2753
Тел.: +7-913-860-01-76
Эл. почта: morphism@tpu.ru

Гончаров Валерий Иванович
Д-р техн. наук, проф.-консультант ОАР ИШИТР ТПУ
Ленина пр-т, 30, г. Томск, Россия, 634050
ORCID: 0000-0002-1249-1981
Тел.: +7-952-895-10-73
Эл. почта: gvi@tpu.ru

Piakillia B.I., Goncharov V.I.

Peculiarities of fully connected neural network design for estimating the lipophilicity of organic compounds

The assessment of lipophilicity of small organic compounds plays a crucial role in the development and optimization of new drugs. Unfortunately, experimental methods require significant time and resources, including the use of laboratory equipment and reagents. Additionally, manual verification and data adjustment often increase the process's labor intensity. In contrast, computational methods like machine learning offer faster and less resource-intensive ways to assess lipophilicity, allowing for efficient processing of large data volumes and adaptation to complex relationships between molecular structure and lipophilicity. Developing neural network models for lipophilicity assessment is challenging due to insufficient experimental data and high computational costs with graph neural network models. This work presents an analysis of popular methods for describing chemical structures for building fully connected neural network models, less demanding in training data volume. Based on this analysis, features best describing organic compounds from an open lipophilicity dataset collected from the ChEMBL database are selected. The search for the optimal neural network model architecture for the chosen features is conducted.

Keywords: modeling, neural network, lipophilicity, cheminformatics.
DOI: 10.21293/1818-0442-2024-27-1-86-94

References

1. Wardecki D., Dołowy M., Bober-Majnusz K. Assessment of Lipophilicity Parameters of Antimicrobial and Immunosuppressive Compounds. *Molecules*, 2023, vol. 28, no. 6, pp. 1–14.

2. Miller R., Madeira M., Wood H., Geissler W., Raab C., Martin I. Integrating the Impact of Lipophilicity on Potency and Pharmacokinetic Parameters Enables the Use of Diverse Chemical Space during Small Molecule Drug Optimization. *Journal of Medical Chemistry*, 2020, vol. 63, no. 21, pp. 12156–12170.

3. Piakillia B. I., Goncharov V.I. Ocenivanie lipofil'nosti s pomoshh'ju bajesovskih nejronnyh setej. [Lipophilicity estimation using Bayesian neural networks] *Izvestiia TulGu. Tehnicheskie Nauki*, 2022, no. 9, pp. 288–292 (in Russ.)

4. Tang B., Kramer S., Fang M. A self-attention based message passing neural network for predicting molecular lipophilicity and aqueous solubility. *Journal of Cheminformatics*, 2020, vol. 12, no. 15, pp. 1–9.

5. Song Y., Chen J., Wang W. Double-head transformer neural network for molecular property prediction. *Journal of Cheminformatics*, 2023, vol. 15, no. 27, pp. 1–16.

6. Wang Y., Xiong J., Xiao F. LogD7.4 prediction enhanced by transfer-ring knowledge from chromatographic retention time, mi-croscopic pKa and logP. *Journal of Cheminformatics*, 2023, vol. 15, no. 76, pp. 1–13.

7. Wieder O., Keunemann M., Wieder T., Seidel T., Meyer C., Bryant S., Langer T. Improved Lipophilicity and Aqueous Solubility Prediction with Composite Graph Neural Networks. *Molecules*, 2021, vol. 26, no. 20, pp. 6185–6223.

8. Gafarov F.M, Galimjanov A.F. Iskusstvennye nejronnye seti i prilozhenija: ucheb. posobie [Artificial Neural Networks and Applications: a study guide]. Kazan, Kazan Univ. Publ., 2018, 121 p. (in Russ.)

9. Bengio Y., Goodfellow I., Courville A. Deep Learning. *MIT Press*, 2016. 800 p.

10. Zheron A. Prikladnoe mashinnoe obuchenie s pomoshh'ju Scikit-Learn i TensorFlow: koncepcii, instrumenty i tehniki dlja sozdanija intellektual'nyh system [Applied Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques for Creating Intelligent Systems]. Saint-Petersburg. Alfa-kniga Publ., 2018. 688 p. (in Russ.)

11. Documentation RDkit: https://www.rdkit.org/docs/ (Accessed: March 26, 2024).

12. Documentation Numpy: https://numpy.org/doc/stable/ (Accessed: March 26, 2024).

13. Documentation Pandas: https://pandas.pydata.org/docs/reference/index.html (Accessed: March 26, 2024).

14. Documentation Matplotlib: https://matplotlib.org/stable/index.html (Accessed: March 26, 2024).

15. Documentation TensorFlow: https://www.tensorflow.org/api_docs (Accessed: March 26, 2024).

16. Documentation Seaborn: https://seaborn.pydata.org/api.html (Accessed: March 26, 2024).

17. Wu Z., Ransundar B., Feinberg E. MoleculeNet: a benchmark for molecular machine. *Chemical Science*, 2018, vol. 9, no. 2, pp. 513–530.

18. Waring M. Lipophilicity in drug discovery. *Expert Opinion on Drug Discovery*, 2010, vol. 5, no. 3, pp. 235–248.

19. Madzhidov T.I., Baskin I.I., Varnek A.A. Vvedenie v hemoinformatiku: uchebnoe posobie. Ch. 1: Komp'juternoe predstavlenie himicheskih struktur [Introduction to Chemoinformatics: A Study Guide. Part 1: Computer Representation of Chemical Structures]. Kazan, Kazan Univ. Publ., 2015, 174 p. (in Russ.).

20. Gómez-Bombarelli R., Jennifer W., Duvenaud D. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Central Science*, 2018, vol. 4, no. 2, pp. 268–276.

21. Madzhidov T.I., Baskin I.I., Varnek A.A. Vvedenie v hemoinformatiku: uchebnoe posobie. Ch. 4: Metody mashinnogo obuchenija. [Introduction to Cheminformatics: Study Guide. Part 4: Machine Learning Methods.] Kazan, Kazan Univ. Publ., 2016, 329 p. (in Russ.).

22. Stokes J.M., Swanson K., Yang K. A Deep Learning Approach to Antibiotic Discovery. *Cell*, 2020, vol. 180, no. 4, pp. 475–483.

23. Niazi S., Mariam Z. Advances in Machine-Learning-Based Chemoinformatics: A Comprehensive Review. *International Journal of Molecular Sciences*, 2023, vol. 24, pp. 11488–11503.

24. Rickert C., Henkel M., Lieleg O. Efficiency-driven, correlation-based feature elimination strategy for small datasets. *Machine Learning*, 2023, vol. 1, no. 1, pp. 1–15.

25. Ali Y.A Hyperparameter Search for Machine Learning Algorithms for Optimizing Computational Complexity. *Processes*, 2023, vol. 11, no. 2, pp. 1–21.

26. Gilmer J., Schoenholz S., Riley P., Vinyals O. Neural message passing for quantum chemistry. *International Conference on Machine Learning. PMLR*, 2017, pp. 1263–1272.

27. Kearnes S., McCloskey K., Berndl M., Pande V., Riley P. Molecular graph convolutions: moving beyond fingerprints *Journal of Computer-Aided Molecular Design*, 2016, vol. 30, pp. 595–608.

28. Chen T., Guestrin C. Xgboost: A scalable tree boosting system. *In Proceedings of the 22nd ACM International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

Boris I. Piakillia
Senior Lecturer, Division for Automation and Robotics of School of Computer Science & Robotics,
Tomsk Polytechnic University
ORCID: 0000-0002-1249-1981
30, Lenin pr., Tomsk, Russia, 634050
Phone: +7-913-860-01-76
Email: morphism@tpu.ru

Valery I. Goncharov
Doctor of Science in Engineering, Professor-consultant, Division for Automation and Robotics, School of Computer Science & Robotics, Tomsk Polytechnic University
ORCID: 0000-0002-1249-1981
30, Lenin pr., Tomsk, Russia, 634050
Phone: +7-952-895-10-73
Email: gvi@tpu.ru