УДК 51-76

ПРИМЕНЕНИЕ МНОГОЦЕЛЕВОГО ОБУЧЕНИЯ ДЛЯ ПРЕДСКАЗАНИЯ СИЛЫ СВЯЗЫВАНИЯ ЛИГАНДА С БЕЛКОМ-МИШЕНЬЮ

Виноградова Е.А. (Университет ИТМО)
Научный руководитель – Пац К.М. (Университет ИТМО), **PhD**, **проф**. **Молнар Ф.**
(Назарбаев Университет)

Разработка новых лекарств включает в себя поиск биологически активных соединений (лигандов) и их оптимизацию. Для поиска потенциальных лигандов среди огромного количества химических структур используется экспериментальный метод высокопроизводительного скрининга (*High-Throughput Screening, HTS*). Для оптимизации этого метода в настоящее время проводится предварительный виртуальный скрининг (*Virtual High-Throughput Screening, vHTS*), который позволяет экономить время и материальные ресурсы, избегая трудоемких экспериментов, которые заведомо могут оказаться бесполезными. В данной работе представлен новый метод решения одной из задач виртуального скрининга — задачи ранжирования молекул-кандидатов.

Введение. Ранжирование (*ranking*) является одной из задач виртуального скрининга. При ранжировании молекулы-кандидаты (*leads*) сортируются в соответствии с некоторым баллом. Кандидаты могут быть оценены по некоторым предсказываемым параметрам, таким как токсичность или сила связывания.

В настоящее время большинство методов прогнозирования силы связывания лиганда с белком (*binding affinity prediction*) используют подходы глубокого машинного обучения (*deep learning*). Однако исходные данные, используемые для создания таких моделей, поступают из множества экспериментов различного рода. Поэтому имеющиеся данные сильно отличаются друг от друга либо из-за вариаций в условиях анализа, либо из-за ошибок в базе данных. При этом количество доступных методов борьбы с шумом в таких данных в настоящий момент ограничено.

В данной работе представлен новый метод прогнозирования аффинности лигандов к мишеням, который позволяет справиться с шумом в исходных данных. Он использует парадигму многоцелевого обучения *(multi-task learning*). В этом подходе несколько задач решаются одновременно с помощью одной модели машинного обучения. Это хорошо зарекомендовавший себя подход, который широко используется в различных областях исследований, включая биохимию. Этот подход часто используется не только для обучения искусственной нейронной сети одновременному решению различных задач, но и для борьбы с шумом в данных. При таком подходе модель вынуждена обучаться более общему представлению данных, которое может быть использовано в любом контексте (для любой задачи); кроме того, использование многоцелевого подхода позволяет использовать все имеющиеся данные, а не концентрироваться на их подмножествах.

Основная часть. В настоящее время существует небольшое число моделей, использующих многоцелевое обучение для решения задачи прогнозирования силы связывания лигандов с мишенями. Две недавно разработанные модели, *Multi-PLI* (2021) и *GanDTI* (2021), акцентируют внимание на использовании многоцелевого подхода для одновременного решения задач классификации и регрессии (*joint-task learning*), при этом в процессе обучения разделяются различные меры силы связывания (*Kd, Ki, IC50, EC50*). В данном проекте, напротив, все меры аффинности (силы связывания) используются в обучении одновременно. Это позволяет использовать на порядок больше данных, что особенно важно, поскольку некоторые пары лиганд-белок неравномерно представлены для разных мер аффинности. Однако использование большего количества данных приводит к появлению большого

количества пропущенных значений в обучающем наборе. В подходе, предложенном в данном проекте, было также решено убирать (маскировать) недостающие данные при вычислении функции потерь, что является нечастым решением для данной проблемы.
Разработанная модель имеет 2 входа и 5 выходов, в зависимости от количества прогнозируемых значений. Модель принимает на вход информацию о структуре лигандов и белков в виде текстовых строк. На выходе формируется прогноз пяти значений для таких пар (лиганд-белок) — различные меры аффинности (*Kd, Ki, IC50, EC50*) и вероятность принадлежности к набору активных структур.
В данной работе для обучения и тестирования способности модели предсказывать силу связывания использовались данные, из базы данных *BindingDB*, содержащей записи экспериментальных результатов о силе связывания лигандов с целевыми белками.
Разработанная модель, а также базовые (*baseline*) модели (*DeepDTA, GraphDTA, DeepPurpose*) были обучены на обучающей выборке из этого набора данных. Базовые модели были обучены на текстовых последовательностях лигандов и белков-мишеней. Все модели были обучены предсказывать силу связывания. Базовые модели были обучены предсказывать константу диссоциации (*Kd*), а разработанная была обучена предсказывать все меры аффинности (*Kd, Ki, IC50, EC50*) одновременно.
Для тестирования разработанной модели на предмет качества ранжирования (*ranking task*) использовались наборы молекул, сгенерированных различными de novo алгоритмами (*Virtual Libraries, ReInvent v2, ReInvent v3, TransVAE, LigDream*). Во время тестирования разработанная и базовые модели использовались для ранжирования молекул по предсказанной силе связывания. Ожидаемое и предсказанное ранжирование молекул сравнивалось в качестве критерия оценки эффективности алгоритмов.
Таким образом, разработанная модель сравнивалась с известными базовыми моделями. Было установлено, что на тестовом наборе данных разработанная модель работает не хуже, чем базовые. Также было установлено, что в задаче ранжирования разработанная модель показывает лучшие результаты, чем базовые.

Выводы. Предлагаемая в данной работе модель не только предсказывает различные показатели биологической активности одновременно, но и способна обучаться на полном объёме имеющихся данных и заполнять недостающие значения в процессе обучения и предсказания. Согласно полученным результатам, использование многоцелевого подхода, делает разработанную модель более подходящей для оценки *de novo* генеративных моделей и сгенерированных молекул. Модель также сопоставима по производительности с эталонными моделями.