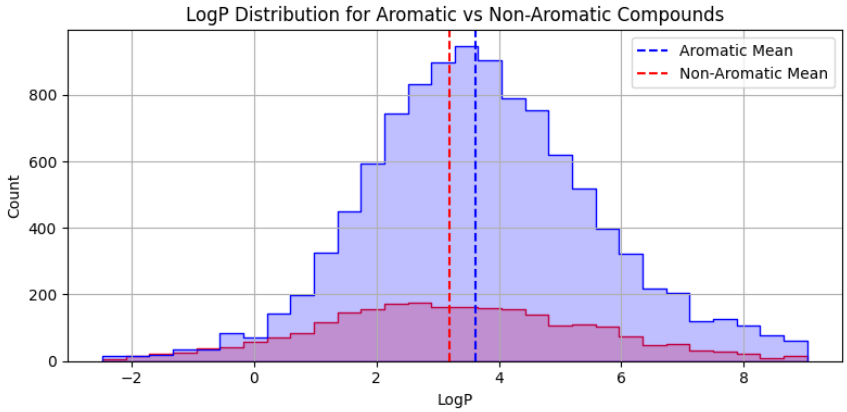
Элемент 119. ИИ в химии.

Отчет команды “Группа симметрии”

Исследование данных

Перед тем, как начать работу над построением модели, мы решили поближе посмотреть на предоставленные данные. Построив гистограмму распределения LogP, мы обнаружили наличие выбросов – молекул с очень большими или очень маленькими LogP. Посмотрев на них внимательнее, мы поняли, что их LogP не соотносится с их структурой, поэтому мы отбросили хвосты до 0.01 и после 0.99 квантиля.

Далее, мы отметили что большая часть молекул содержит ароматические фрагменты, и решили сравнить гистограммы распределения ароматических и неароматических молекул. Оказалось, что медианное значение LogP ароматических соединений значимо выше, чем неароматических (p-value в тесте Манна-Уитни 10^-23), а дисперсия (3.4) меньше как дисперсии неароматических (4.7), так и дисперсии по всем молекулам (3.8). Исходя из этого, мы решили, что имеет смысл обучить отдельную модель на ароматических молекулах – меньшая дисперсия может способствовать тому, что модели будет легче обучиться. Мы опробовали такой подход и он действительно сработал.



Мы также смотрели на зависимости LogP от различных дескрипторов и искали корреляции. Например, мы заметили, что LogP имеет тенденцию уменьшаться с ростом отношения числа гетероатомов к общему числу атомов в молекуле (коэффициент корреляции -0.4).

Восстановление SMILES

Среди обучающих данных имеется 450 SMILES, которые не удается перевести в молекулу. Для того, чтобы использовать их для обучения модели, нужно их восстановить. Для этого мы пробовали отдельную модель на искусственно сгенерированных ошибках и предсказывать символы, которые нужно поменять. Однако, таким образом удалось восстановить лишь 40% SMILES.

Лучше показал себя следующий подход: для каждого неправильного SMILES искать ближайших соседей среди правильных, которые ближе всего к ним по расстоянию Левенштейна. Как видно из распределения, предаставленного ниже, большинство неправильных SMILES имеет соседей на расстоянии 0 или 1 (расстояние считается только по

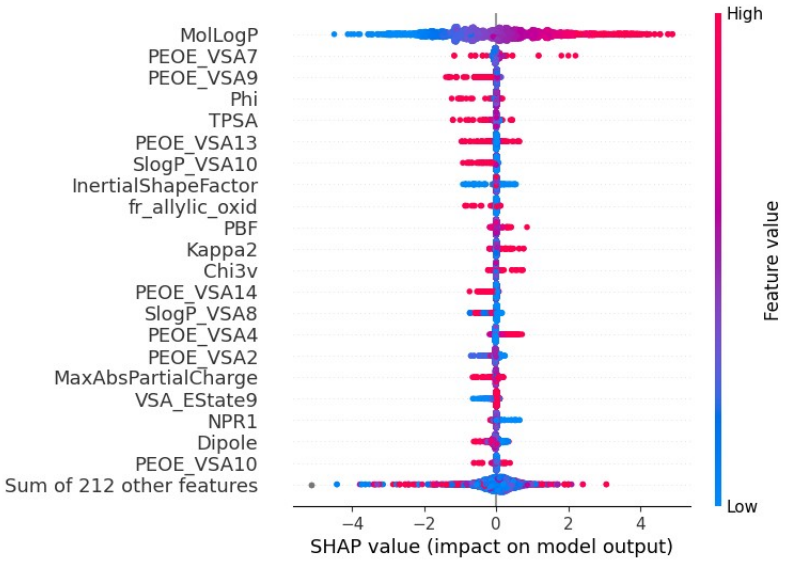
корректным токенам). Такие SMILES мы заменяем на своих соседей, предполагая, что символ, в котором они отличаются - либо ошибка, либо не вносит значительного вклада в LogP. Таким образом удается восстановить около 70% неправильных SMILES.

Квантово-химические фичи

Одной из первых нашей идей было использовать фичи, полученные с помощью простейшего квантово-химического метода, встроенного в rdkit – расширенного метода Хюккеля. Для этого мы генерировали конформеры для каждой молекулы, оптимизировали их в силовом поле, и рассчитывали энергии HOMO, LUMO, их разницу, электронную энергию, а также дипольный момент для полученных конформеров. Так как LogP должен сильно зависеть от величины дипольного момента, а также энергий граничных орбиталей, мы посчитали такие фичи очень перспективными.

Бейслайн

В качестве бейслайна мы взяли градиентный бустинг с использованием всех возможных дескрипторов rdkit с добавлением квантово-химических фичей. Эта модель показала себя не очень хорошо, достигнув RMSE на валидации, равного 0.94. Кроме этой модели, мы также пробовали использовать RandomForest и различные комбинации дескрипторов и фингерпринтов, но не значительного улучшения результатов добиться не удалось. Мы визуализировали значения SHAP, чтобы определить важность фичей. Довольно ожидаемо, наиболее важным оказалось значение MolLogP. После него по важности идут VSA дескрипторы, тесно связанные с MolLogP, а также различные 3D дескрипторы, TPSA и после них дипольный момент.



После того, как мы построили такую простейшую модель, нам стало понятно, что простых дескрипторов недостаточно, и необходимо дать модели больше информации о структуре молекулы, об ее атомах, связях, и их взаимодействиях. В связи с этим, мы решили строить нашее дальнейшее решение на нейросетевых моделях, умеющих работать с молекулами.

Chemprop

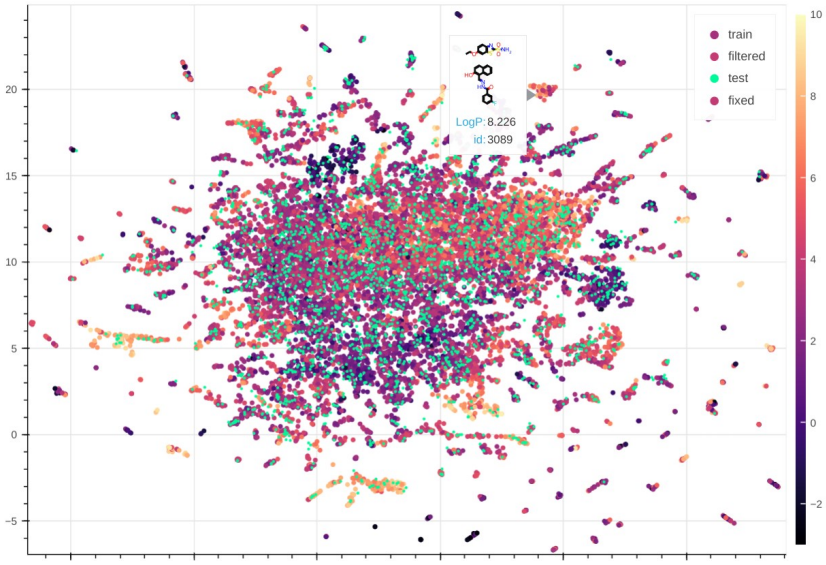
В качестве следующего шага, мы опробовали довольно популярную модель Chemprop. Это графовая нейросеть с механизмом message-passing с удобным API на python. Применив эту модель к нашим данным, мы достигли RMSE около 0.62 – хорошо, но хотелось бы лучше. Мы перешли к поиску других вариантов, но продолжили использовать Chemprop для того, чтобы проверять возникающие гипотезы – благо, она работает довольно быстро.

Uni-Mol

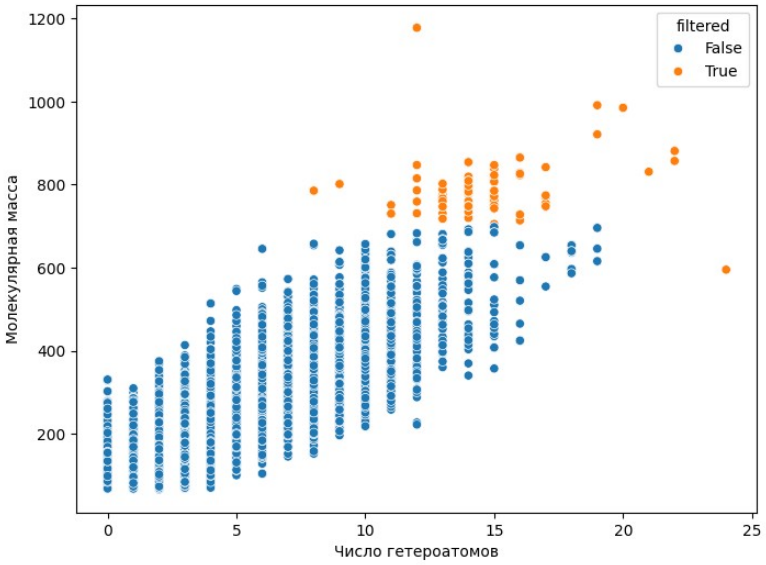
Продолжив поиски, мы вышли на модель Uni-Mol. Она отличается от Chemprop тем, что основана на архитектуре SE(3)-эквивариантного трансформера с механизмом self-attention и использует 3D представления молекул, а не двумерный граф. Учитывая замеченную нами важность 3D дескрипторов, такая архитектура показалась нам предпочтительной. Кроме того, Uni-Mol занимает лидирующие позиции во многих бенчмарках по предсказанию свойств молекул.

Визуализация химического пространства

Используя предобученную модель Uni-Mol, мы получили эмбеддинги для всех молекул. Затем, понизив размерность с помощью UMAP, мы построили интерактивную карту химического пространства с изображениями молекул. На этой карте нам сразу бросилось в глаза большое количество маленьких кластеров молекул, отдаленных от центра. В одном из таких кластеров мы обнаружили SMILES, содержащие по несколько молекул. С помощью фильтрации по фрагментам молекул, мы убрали часть из этих маленьких кластеров как выбросы, так как они слишком непохожи на остальные молекулы. Это привело к повышению качества предсказаний.



Кроме того, мы визуализировали зависимости между различными свойствами молекул и также искали на них выбросы. В результате, мы остановились на графике молекулярная масса-количество гетероатомов, по которому мы отбросили молекулы с массой больше 600 и числом гетероатомов больше 20, так как они довольно редки.



Обработка дубликатов

Мы заметили, что в обучающих данных содержится довольно много повторяющихся SMILES с различными значениями LogP. Для того, чтобы использовать эти данные, мы применяли различные стратегии – брали среднее, медиану, но прироста точности эти способы не давали. Поэтому мы решили использовать модель Chemprop, чтобы предсказать наиболее вероятные значения LogP для этих молекул, и выбирали среди дубликатов то значение, которое было ближе к предсказанному. Это позволило нам дополнительно снизить RMSE.

Итоговый пайплайн

В результате наши исследований, мы пришли к следующему решению. После предобработки датасета, мы выделяем из него ароматические молекулы, а также “малые” молекулы с молекулярной массой менее 600 и числом гетероатомов менее 13. Затем мы обучаем три модели Uni-Mol на полном датасете, на ароматических молекулах и на малых молекулах, используя 5-fold кросс-валидацию. После этого каждая модель предсказывает значение LogP на соответствующей части теста. Затем предсказания объединяются в порядке приоритета ароматические>малые>остальные.

Предложения по улучшению

На этом наши идеи не закончились. Для дальнейшего улучшения точности мы попробовали внедрить в Uni-Mol физико-химические дескрипторы и квантово-химические фичи. Так как такого рода модификации не были предусмотрены разработчиками модели, мы написали отдельный класс. К сожалению, на тестирование этой модели у нас не хватило времени и ресурсов, но мы считаем такой подход перспективным для дальнейшего исследования. Нам удалось найти статью, где подобный подход успешно применяется к графовым нейросетям: https://repositum.tuwien.at/bitstream/20.500.12708/191402/1/Brasoveanu-2023-Extending%20Graph%20Neural%20Networks%20with%20Global%20Features-am.pdf

Выводы

В ходе решения мы опробовали большое число различных моделей. Лучше всего себя показала модель Uni-Mol, использующая для предсказания 3D представление молекул. Большое внимание было уделено восстановлению неправильных SMILES, для этого был придуман собственный способ, оказавшийся довольно эффективным. Также были испробованы различные стратегии работы с дубликатами, что тоже дало вклад в итоговый результат. Больше всего помогло разделение датасета на отдельные классы молекул и обучение ансамбля моделей на этих классах, а также обработка выбросов. Большую пользу в понимании структуры данных принесла визуализация химичского пространства.