ДОКЛАДЫ РОССИЙСКОЙ АКАДЕМИИ НАУК. МАТЕМАТИКА, ИНФОРМАТИКА, ПРОЦЕССЫ УПРАВЛЕНИЯ, 2022, том 508, с. 70–72

ПЕРЕДОВЫЕ ИССЛЕДОВАНИЯ В ОБЛАСТИ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА И МАШИННОГО ОБУЧЕНИЯ

УДК 004.8

ГЕОМЕТРИЧЕСКОЕ ГЛУБОКОЕ ОБУЧЕНИЕ ДЛЯ ДИЗАЙНА КАТАЛИЗАТОРОВ И МОЛЕКУЛ

© 2022 г. Р. Ю. Лукин[1,*], Р. А. Григорьев[2]

Представлено академиком РАН Г.И. Савиным
Поступило 28.10.2022 г.
После доработки 31.10.2022 г.
Принято к публикации 03.11.2022 г.

Применение глубокого обучения для поиска катализаторов является важной задачей для решения вызванных глобальным потеплением проблем хранения энергии и преобразования парниковых газов в более ценные продукты. В нашей работе мы представляем несколько графовых нейронных сетей (GNN), включая сверхточные архитектуры и архитектуры передачи сообщений с физически информированными атрибутами узлов и ребер для атомистических систем. Мы демонстрируем улучшение прогнозов энергии адсорбции в наборе данных OC20 с использованием предложенной нами архитектуры в терминах средней абсолютной ошибки прогнозируемой энергии и энергии в пределах пороговых показателей. Предлагаемые архитектуры устойчивы к переобучению и могут быть использованы для прогнозирования экспериментальных и квантово-химических свойств широкого спектра материалов и молекул. Мы предлагаем использовать две архитектуры GNN (EdgeUpdateNet и OFMNet) вместе с расширенным методом описания узлов и ребер. Мы представляем отпечатки ребер как элементы матриц межатомного взаимодействия (матрица Кулона, матрица суммы Эвальда, синусоидальная матрица). Для отпечатков пальцев узлов мы используем элементы матрицы орбитального поля (OFM), однократного представления электронного состояния атомов с окружающими атомными орбиталями. Кроме того, мы предлагаем и реализуем представление каталитически активных атомов в виде подграфа. Предлагаемые методы и архитектуры демонстрируют повышение точности прогнозирования энергии адсорбции. Особенно значительные улучшения наблюдаются в примерах, не относящихся к предметной области, как для адсорбатов, так и для катализаторов. Возможности обобщения и экстраполяции на примеры предлагаемых архитектур вне предметной области также делают предлагаемые GNNs пригодными для использования при скрининге катализаторов в обширном химическом пространстве.

Ключевые слова: графовые нейросети, глубокое обучение, квантовая химия, катализ, хемоинформатика

DOI: 10.31857/S2686954322070153

1. ВВЕДЕНИЕ

Компьютерное химическое моделирование и экспериментальные измерения – это два традиционных метода, которые широко применяются в области материаловедения. Однако использование этих двух методов для ускорения поиска материалов и проектирования затруднено, поскольку они отнимают много времени и неэффективны. В большинстве случаев для расчетов электронной структуры использовалась теория функционала плотности (DFT). Несмотря на широкое применение машинного обучения (ML) ко многим проблемам, возникшим за последние годы при прогнозировании свойств молекул и материалов с использованием машинного обучения, возрастает роль подходов к глубокому обучению. Достижения в области атомистического машинного обучения оказывают большое влияние во многих областях, включая материаловедение, катализ и разработку лекарств. Алгоритмы ML – это обучаемый оценщик между входными представлениями материалов и молекул и выходными данными, представляющими интерес для физических или химических свойств. Эти модели использовались для прогнозирования широкого спектра свойств для различных классов материалов, включая точечные дефекты для нейроморфных вычислений, солнечных фотокатализаторов и, особенно, гете-

[1] *Лаборатория в сфере развития продукта ИИ для новых материалов Исследовательского центра в сфере искусственного интеллекта, Университет Иннополис, Иннополис, Россия*

[2] *Исследовательский центр в сфере искусственного интеллекта, Университет Иннополис, Иннополис, Россия*

*****E-mail: r.lukin@innopolis.ru*

рогенных, и гомогенных, катализаторов. Обычно первым шагом всех основанных на ML подходов к прогнозированию свойств материала является представление материала. Кристаллические структуры могут быть закодированы различными способами, включая объекты (часто называемые дескрипторами), 3D-графики с пространственной информацией и соответствующими атрибутами узлов и ребер. В случае дескрипторов основными требованиями являются: низкая вычислительная стоимость, различимость между подобными структурами и способность кодировать как пространственную, так и информационную, электронную структуру и композицию. Архитектурами нейронных сетей на основе графов установлено, что они эффективны для больших наборов данных о молекулах и материалах из-за субпространственного числа обучаемых параметров, в то время как подходы, основанные на дескрипторах, являются склонными к переобучению.

2. МЕТОДОЛОГИЯ ИССЛЕДОВАНИЯ

Чтобы решить проблему с данными, Ulissi et al. разработали набор данных OC20, состоящий из 1 281 040 плотностей. Релаксации функциональной теории (DFT) (~264 890 000 одиночные точечные оценки) по широкому спектру материалов, поверхностей, и адсорбаты (химия азота, углерода и кислорода). По сравнению с ранее опубликованными наборами данных о катализаторах набор данных OC20 позволяет нам обобщать различные промежуточные продукты, образующиеся в ходе реакций восстановления CO2 и N2. Помимо создания и обмена набором данных, авторы предлагают три связанные проблемы предметной области в качестве открытого соревнования: 1) предсказать энергию и силу для данного состояния (S2EF), 2) предсказать соседнее расслабленное состояние с учетом начального состояния (IS2RS) и 3) предсказывают релаксированную энергию адсорбции при заданном начальном состоянии (IS2RE). Набор данных разделен на обучающую, тестовую и валидационную выборку. Валидационная выборка моделирует химическое пространство, используемое при скрининге катализаторов, включающая в себя как адсорбаты, так и катализаторы, состав и структура которых ранее не была представлена в обучающей выборке. Для практических целей наиболее подходящей задачей является предсказание DFT-энергии связывания на основе эвристически сгенерированных структур катализатор-адсорбат.

Ядерные заряды и положения атомов не являются подходящим входным представлением атомистического, не вращательно или поступательно инвариантного по сравнению с уравнением Шредингера. В настоящее время во многих подходах используются представления объектов, специфичные для атомов или геометрии, а также архитектуры ML на основе ядра или нейронных сетей (NN). Недавние исследования сосредоточены на характеристике атомных систем в абстрактных представлениях, таких как квантово-механические свойства, полученные в результате расчетов электронной структуры с низкими затратами, и использовании новых методов графовых нейронных сетей для повышения эффективности обучения и обобщаемости.

Было разработано несколько методов ML для прогнозирования энергий корреляции высокого уровня (связанных кластеров) на основе квантово-механических характеристик на уровне среднего поля (например, теория ВЧ или ДПФ) расчет электронной структуры. Например, авторы OrbNet использовали архитектуру графовой нейронной сети для прогнозирования высококачественных энергий электронной структуры на основе характеристик, полученных с помощью недорогих/минимальных методов электронной структуры среднего поля.

Различные графовые нейронные сети недавно добились больших успехов в предсказании квантово-механических свойств молекул. В предыдущих работах MPNNS с соответствующими функциями сообщения, обновления и вывода продемонстрировали полезное индуктивное смещение для прогнозирования молекулярных свойств, превосходя несколько сильных базовых показателей и устраняя необходимость в сложном проектировании функций. Важной задачей является разработка MPNN, которые могут эффективно обобщаться на более крупные графы, чем те, которые появляются в обучающем наборе, или, по крайней мере, работать с контрольными показателями, предназначенными для выявления проблем с обобщением по размерам графа. Обобщение на большие размеры молекул кажется особенно сложным при использовании пространственной информации. Межатомные взаимодействия (ковалентные и нековалентные) являются определяющим фактором в каталитических процессах; в частности, их необходимо учитывать при моделировании энергии адсорбции. Поэтому мы считаем важным ввести дескрипторы, описывающие межатомные электростатические взаимодействия, в качестве атрибутов ребер. В то время как орбитально-орбитальные взаимодействия вместе с информацией об электронном состоянии атомов могут быть универсальными атрибутами узла. Также при выборе признаков важны несколько требований, таких как изотропность пространства, изометрия пространства, инвариантность относительно перестановки атомных индексов, непрерывность и вычислительная дешевизна. Из-за вычислительных ограничений мы не используем гибридные подходы DFT/GNN. В качестве матриц межатомных сил мы выбрали куло-

новскую матрицу в качестве простого дескриптора электростатического взаимодействия между ядрами и синусоидальными и Матрицы сумм Эвальда как потенциалы электростатических взаимодействий в периодических (или псевдопериодических) системах. Как представление матрицы орбитального поля (OFM), основанной на распределении электронов валентной оболочки центрального и окружающих атомов.

3. ОСНОВНЫЕ РЕЗУЛЬТАТЫ, ВЫВОДЫ

В нашей работе мы представили несколько графовых нейронных сетей (GNN), включая сверточные архитектуры и архитектуры передачи сообщений с физически информированными атрибутами узлов и ребер для атомистических систем. Мы демонстрируем улучшение прогнозов энергии адсорбции в наборе данных OC20 с использованием предложенной нами архитектуры с точки зрения средней абсолютной ошибки прогнозируемой энергии и энергии в пределах пороговых показателей. Предлагаемые архитектуры устойчивы к переобучению и могут быть использованы для прогнозирования экспериментальных и квантово-химических свойств широкого спектра материалов и молекул.

Мы представили две архитектуры GNN (EdgeUpdateNet и OFMNet) с атрибутами в виде элементов матриц межатомного взаимодействия (матрица Кулона, матрица суммы Эвальда, Синусоидальная матрица). Для узловых отпечатков пальцев мы используем элементы матрицы орбитального поля (OFM), однократного представления электронного состояния атомов с окружающими атомными орбиталями. Разработанные модели, использующие физически информированные отпечатки пальцев, показывают улучшение точности прогнозирования энергии адсорбции. Особенно значительные улучшения наблюдаются в примерах, не относящихся к предметной области, как для адсорбатов, так и для катализаторов. Возможности обобщения и экстраполяции предлагаемых архитектур также делают GNNS пригодными для использования при скрининге катализаторов в обширном химическом пространстве. Код с используемыми сценариями, моделями и конфигурациями доступен по адресу https://github.com/AI4Materials-lab/catgnns.