Интерпретация спектров ядерного магнитного резонанса высокого разрешения с использованием методов машинного обучения

Новиков Валентин Владимирович

Научный руководитель: Чусов Денис Александрович, д.х.н., проф. ВШЭ

Аннотация

Содержание

Введение

Настоящее исследование находится на пересечении двух научных областей - химии и наук о данных. В области химии исследование относится к спектроскопии ядерного магнитного резонанса (ЯМР), в области наук о данных - к направлению генеративного искусственного интеллекта.

Спектроскопия ЯМР - один из видов спектроскопии, рутинно используемый в организациях химического профиля (химические НИИ, химические и биологические факультеты университетов, R&D отделы фармацевтических компаний и т.д.) для установления строения органических, элементоорганических, координационных и высокомолекулярных соединений. Результатом работы химиков-синтетиков являются новые химические соединения, и при использовании спектроскопия ЯМР возможно однозначно доказать их строение – задача, без которой невозможны ни публикация полученных данных, ни патентование, ни проведение дальнейших прикладных исследований. Настоящее исследование направлено на использование методов машинного обучения и генеративного искусственного интеллекта для решения обратной задачи спектроскопии ЯМР.

В результате проведения анализа химического вещества путем спектроскопии ЯМР получают спектр - набор пиков с такими характеристиками как «химический сдвиг» (значение по оси абсцисс, при котором наблюдается индивидуальный сигнал), «константы спин-спинового взаимодействия», определяющие форму каждого сигнала, и интегральная интенсивность сигнала.

В тех случаях, когда предполагаемое строение изучаемого вещества известно, перед химиком стоит задача подтверждения строения вещества путем отнесения каждого сигнала к какому-либо фрагменту изучаемой молекулы (прямая спектроскопическая задача). Данная задача решается относительно просто и во многих случаях может быть автоматизирована с высокой степенью достоверности получаемых отнесений. Тем не менее, очень часто в результате синтеза получают не то вещество, которое планировалось синтезировать, и тогда становится необходимо решить обратную спектроскопическую задачу – определить строение неизвестного соединения на основе спектральных данных. Известно, что решать обратную спектроскопическую задачу в любой спектроскопии – достаточно сложно. В данном случае для решения этой задачи требуется ручной анализ полученных спектров квалифицированным специалистом в области спектроскопии ЯМР. В масштабах одной научной организаций десятки и сотни часов времени специалистов тратятся еженедельно на определение строения новых соединений с использованием спектроскопии ЯМР, и этот процесс до сих пор не автоматизирован.

В научной литературе собраны сведения о спектрах ЯМР десятков миллионов химических соединений. Тем не менее, лишь часть этих спектров описана в

спектральных базах данных, что осложняет разработку подходов для решения обратной спектроскопической задачи на основе методов машинного обучения. С другой стороны, даже сбор и очистка таких данных не означает, что на их основе можно будет однозначным образом предложить метод решения обратной спектроскопической задачи, поскольку подходы для *de novo* генерации химических структур на основе спектральных данных практически отсутствуют.

Предлагаемое исследование поднимает **проблему** разработки новых генеративных подходов искусственного интеллекта, позволяющих на основе предложенных спектров ЯМР неизвестного химического соединения генерировать молекулярный граф, описывающий его строение.

Таким образом, **объектом исследования** является взаимосвязь между строением химического соединения и его спектрами ЯМР, а **предметом исследования** - методические основы *de novo* генерации химических структур на основе входных спектральных данных. Научная новизна исследования в первую очередь связана с отсутствие в научной литературе подходов к генерации молекулярных структур на основе спектров ЯМР, в которых обучение модели проходило на большом (более миллиона спектров) объеме **экспериментальных** данных.

Целью предлагаемого исследования является поиск универсальных подходов к автоматизированной интерпретации спектров ядерного магнитного резонанса (ЯМР) органических соединений. Успешное достижение поставленной цели позволит получить ответ на вопрос «Как на основе экспериментальных спектров ЯМР соединения без участия квалицированного специалиста получить молекулярный граф, отражающий структурную формулу данного вещества?»

Достижение глобальной цели исследования будет основано на выполнении следующих задач:

1. Сбор и подготовка набора данных, содержащего не менее миллиона экспериментальных спектров ЯМР на ядрах ^{13}C для органических соединений, на основе текстовых данных, представленных в научных публикациях.
2. Выбор подхода для эмбеддинга формул химических соединений, представленных в полученном наборе данных, для дальнейшего использования в машинном обучении.
3. Использование собранного набора данных и представленных в научной литературе подходов, основанных на применении графовых нейронных сетей, для обучения нейросети, решающей прямую спектроскопическую задачу (предсказание спектров ЯМР на ^{13}C на основе строения исходного соединения)
4. Разработка архитектуры и обучение на основе собранного набора данных генеративной нейросети, способной решать обратную спектроскопическую задачу, то есть генерировать набор молекулярных графов на основе введенных спектральных данных (спектров ЯМР на ядрах ^{13}C).

5. Тестирование предсказательной способности полученной генеративной нейросети, выявление классов химических соединений, для которых разработанные подходы демонстрируют наилучшую и наихудшую эффективность и итеративная доработка архитектуры и гиперпараметров модели для улучшения качества предсказания.

Глава 1. Обзор литературы

1.1. Магнитный резонанс и методы машинного обучения

Магнитный резонанс представляет собой современный физический метод исследования, широко применяемый в различных областях исследований, от синтетической химии до медицины. Этот метод основан на изучении магнитных свойств атомных ядер или электронов в молекулах и материалах. Основные направления в химии, использующие магнитный резонанс, включают ядерный магнитный резонанс (ЯМР), магнитно-резонансную томографию (МРТ) и электронный парамагнитный резонанс (ЭПР).

Ядерный магнитный резонанс (ЯМР) является одним из наиболее мощных и универсальных методов в современной химии[1] и биологии[2], позволяющий исследовать структуру, динамику и взаимодействия молекул. ЯМР используется для определения строение впервые синтезированных соединений различной природы, определения трехмерной структуры белков, нуклеиновых кислот и других макромолекул[3], а также для изучения метаболитов в различных биологических системах.

Магнитно-резонансная томография (МРТ) является невероятно ценным инструментом в диагностической медицине и биомедицинских исследованиях [4]. МРТ использует принципы ядерного магнитного резонанса, но в отличие от ЯМР, целью МРТ является создание подробных изображений внутренних структур человеческого тела. Этот метод позволяет безопасно и неинвазивно визуализировать мягкие ткани, кровеносные сосуды и органы, предоставляя ценную информацию для диагностики и лечения заболеваний.

Электронный парамагнитный резонанс (ЭПР) – это метод, используемый для исследования материалов и молекул, содержащих неспаренные электроны [5]. ЭПР широко применяется в химии, физике и биологии для изучения химических реакций, радикалов и структур металлоорганических соединений. Этот метод особенно ценен для понимания механизмов окислительно-восстановительных реакций (в том числе – биологических), изучения принципов действия катализаторов, определения пространственной структуры спин-меченых белков.

В целом, магнитный резонанс играет ключевую роль в современной химической науке, обеспечивая уникальные возможности для исследования структуры и функций молекул, а также для диагностики заболеваний в медицине. Тем не менее, применение всех трех основных магнитно-резонансных методов требует высокой квалификации соответствующего специалиста, поэтому использование методов машинного обучения может позволить снизить затраты времени высококвалифицированных научных сотрудников на выполнения типичных исследовательских задач.

Так, в области магнитно-резонансной томографии (МРТ), магнитное обучение находит применение в улучшении качества изображений и ускорении процесса сканирования. Алгоритмы глубокого обучения, например, могут быть использованы для устранения артефактов[6], улучшения разрешения изображений[7] и снижения уровня шума[8], что

важно для диагностики и исследований в медицине и биологии. Кроме того, методы машинного обучения нашли свое применение при разработке новых импульсных последовательностей и даже радиочастотных катушек для МРТ-томографов.[9]

Технологии магнитного обучения также применимы в области электронного парамагнитного резонанса (ЭПР)[10], где они могут быть использованы для автоматизации процессов сбора и анализа данных, обеспечивая более высокую точность и скорость определения параметров спиновых систем[11] и их взаимодействий[12].

В контексте ядерного магнитного резонанса (ЯМР), методы машинного обучения в первую очередь могут применяться для автоматизации процесса анализа спектров[13], обеспечивая более быстрое и точное определение химической структуры соединений [14,15]. Отдельно стоит упомянуть использование подходов машинного обучения для интерпретации данных ЯМР-порометрии в области нефтедобычи[16], анализа данных о временах магнитной релаксации[17], улучшения точности квантовохимических расчетов спектральных параметров[18], реконструкции данных многомерных импульсных ЯМР методик[19] и регистрации двумерных спектров ЯМР с наноразмерных образцов.[20]

Тем не менее, несмотря на прогресс, достигнутый в последние годы, одна из самых частых в исследовательской лаборатории задача интерпретации спектров ЯМР до сих пор не была решена. В практике большинства химических лабораторий за стадией регистрации спектров ЯМР идет этап их ручного анализа, целью которого является либо подтверждение соответствия спектральных данных ожидаемому строению изучаемого соединения, либо определению строения нового неизвестного соединения. Именно последняя задача является одной из наиболее трудозатратных стадией химического исследования, и ее автоматизация может привести к значительной экономии времени научных сотрудников.

1.2. Представление молекулярных структур для машинного обучения

Задача числового представления молекулярных структур

В основе применения подходов машинного обучения для решения химических задач лежит важнейшая задача представления молекулярных структур таким образом, чтобы алгоритмы машинного обучения могли эффективно обрабатывать эти данные. Представление молекулярных структур — это не просто техническая необходимость, а основополагающий аспект, который существенно влияет на производительность и надежность моделей МО в химии.

Молекулы с их разнообразной и сложной структурой создают уникальные проблемы для представления. В отличие от традиционных данных, используемых в машинном обучении, которые часто представлены в форме числовых или категориальных значений, молекулярные структуры требуют представления, отражающего их сложные геометрические и электронные свойства. Эта сложность требует разработки

специализированных методов представления, которые могут преобразовать молекулярные данные в форматы, подходящие для моделей машинного обучения.

Молекулярные структуры по своей сути сложны и характеризуются разнообразным атомным составом, различным типом возможных химических связей и трехмерной геометрией. Таким образом, однозначное представление молекулы является объектов в многомерном пространстве. При этом одной из основных проблем при представлении молекулярных структур является необходимость сохранения баланса между детальностью и сложностью молекулярного представления. Чрезмерно подробные представления могут привести к созданию моделей, которые являются весьма конкретными и лишенными возможности обобщения, в то время как чрезмерно упрощенные представления могут упускать важную информацию, необходимую для точных прогнозов. Кроме того, представление должно быть надежным, чтобы обрабатывать разнообразие молекулярных структур, встречающихся в различных химических базах данных и реальных приложениях.

Выбор молекулярного представления оказывает непосредственное влияние на прогнозирующую способность моделей машинного обучения. Эффективные представления могут повысить способность модели различать тонкие закономерности и взаимосвязи в данных, что приводит к более точным и надежным прогнозам.

Различные методы представления отражают разные аспекты молекулярных структур. Традиционные хемометрические методы, такие как генерация молекулярных фингерпринтов (от fingerprint - отпечатки пальцев), направлены на обнаружение присутствия или отсутствия определенных субструктур, в то время как более продвинутые методы, такие как представления на основе графов и молекулярные трансформеры, направлены на кодирование всей сложности молекулярного графа или даже трехмерной структуры молекулы. Каждый метод имеет свои сильные стороны и подходит для разных типов задач.

Молекулярные фингерпринты

Молекулярные фингерпринты (МФ, от англ. ”fingerprints” – «отпечатки пальцев») — важнейший инструмент в хемоинформатике, обеспечивающий компактное представление молекулярных структур [21]. Их задача - перевод сложного молекулярного графа в числовые векторы, которые можно легко обработать алгоритмами машинного обучения. Основная цель МФ — обеспечить эффективный поиск по сходству, кластеризацию и классификацию молекул, что является важными задачами в разработке лекарств, материаловедении и других химических приложениях.

МФ определяют наличие или отсутствие в молекуле определенных субструктур, функциональных групп или молекулярных фрагментов. Эта абстракция позволяет упрощенно, но информативно представить строение молекулы в числовом виде, обеспечивая возможность анализа больших химических баз данных. Кодируя молекулярные характеристики в машиночитаемом формате, МФ облегчают

идентификацию соединений со схожими свойствами, прогнозирование биологической активности и оптимизацию ведущих соединений при разработке лекарств [22].

Существуют различные типы МФ, каждый из которых предназначен для выявления различных аспектов молекулярных структур. Некоторые распространенные типы включают в себя:

- МФ на основе путей, кодирующие линейные последовательности атомов и связей внутри молекулы
- МФ на основе подструктур, предназначенные для обнаружения определенных заранее определенных подструктур или функциональных групп внутри молекулы
- Круговые МФ, содержащие информацию об атоме и его окрестностях в пределах заданного радиуса

Исчерпывающее описание различных МФ являлось темой многих обзоров[23] и монографий[24] в области хеминформатики, поэтому в настоящем обзоре в качестве примера рассмотрен только один из широко известных МФ, а именно МФ Моргана [25]. Они предназначены для выявления структурных особенностей молекулы путем итеративного рассмотрения окружения каждого атома. Алгоритм Моргана генерирует МФ следующим образом:

1. Инициализация. Каждому атому в молекуле присваивается первоначальный идентификатор, основанный на его атомном номере и других локальных свойствах.

2. Итерация: окружение каждого атома итеративно расширяется за счет рассмотрения атомов в пределах определенного радиуса. Во время каждой итерации идентификаторы обновляются, чтобы отражать растущее соседство.

3. Хеширование: идентификаторы хэшируются для создания двоичных векторов фиксированной длины. Каждый бит вектора указывает на наличие или отсутствие определенной подструктуры.

Выбор параметров радиуса и длины бита существенно влияет на представление. Больший радиус позволяет захватывать более расширенную структурную информацию, а более высокая длина в битах позволяет кодировать более уникальные подструктуры.

Поздней ряд ограничений МФ Моргана был преодолен путем разработки несколько более сложных вариантов круговых МФ, таких как ECFP (Extended-connectivity fingerprints) [26] и MAP4 (MinHashed atom-pair fingerprint)[22]. Тем не менее, важной особенностью молекулярных фингерпринтов является то, что в связи с природой хэш-функции полученные векторы МФ не могут быть использованы для обратного вычисления исходной молекулярной структуры. В связи с этим, несмотря на то, что МФ находят свое применения в хеминформатике, они полностью непригодны для решения обратной спектроскопической задачи, которая как раз подразумевает получение молекулярной структуры *de novo*.

Нотация SMILES

Альтернативой использованию молекулярных фингерпринтов для представления строения молекул является нотация SMILES (Simplified Molecular Input Line Entry System, упрощенная линейная система представления молекул), представляющая собой метод кодирования молекулярной структуры в виде однострочного текстового выражения. Основная идея SMILES заключается в том, чтобы представить структуру молекулы как последовательность символов, описывающих атомы и их связи, что позволяет легко обрабатывать химические структуры программным образом.

Основные принципы нотации SMILES:

1. **Атомы**: Каждый атом представляется его символом из таблицы Менделеева (например, C для углерода, O для кислорода). Водороды обычно не указываются явно, если только их количество не отличается от нормальной валентности атома.

2. **Связи**: Связи между атомами кодируются специальными символами. Одинарные связи обычно не обозначаются, двойные и тройные связи обозначаются символами '=' и '#'. Кольцевые структуры обозначаются числами, приписываемыми к атомам на концах "разорванной" связи кольца.

3. **Ветвление**: Ветвления в молекуле обозначаются круглыми скобками. Это позволяет записывать сложные молекулы с разветвленными структурами, не теряя последовательности описания основной цепи.

4. **Хиральность**: Стереохимическая информация о хиральности атомов может быть включена в SMILES с помощью символов '@' и '@@', что позволяет указывать абсолютную конфигурацию вокруг хирального центра.

5. **Ароматичность**: Ароматические кольца и связи обозначаются строчными буквами (например, ‘c’ для ароматического углерода по сравнению с ‘C’ для алифатического).

К плюсам нотации SMILES следует отнести то, что даже сложные молекулярные структуры таким образом могут быть представлены в виде относительно коротких и понятных строк, которые могут быть проанализированы программным образом, в том числе – с использованием методов машинного обучения. Тем не менее, использование SMILES имеет и ряд минусов. Во-первых, одна и та же молекула может быть корректно представлена с использованием нескольких различных строк SMILES, что усложняет их автоматический анализа и классификацию. Во-вторых, несмотря на то что существуют расширения SMILES для включения стереохимической информации, в некоторых случаях они могут быть недостаточными для точного описания всех аспектов молекулярной геометрии. Наконец, такое одномерное представление подразумевает потери значительного количества информации по сравнению с трехмерным графом, которым является молекулярное соединение. Тем не менее, несмотря на указанные недостатки, именно нотация SMILES чаще всего используется для предсказания спектральных данных и, во многих случаях, для решения обратной спектроскопической задачи.

Молекулярные графы для представления молекулярных структур

Молекулярные графы являются еще одним важным инструментом в хемоинформатике и молекулярном моделировании. Они позволяют визуализировать и анализировать молекулярные структуры, представляя атомы в виде узлов и химические связи в виде ребер графа.

В двумерных (2D) молекулярных графах структура молекулы представлена на плоскости, где каждый атом изображен точкой (узлом), а связи между атомами — линиями (рёбрами). Это позволяет легко визуализировать и анализировать структурные особенности, такие как функциональные группы и кольцевые системы. 2D графы часто используются для быстрого представления и сравнения молекул, а также в базах данных для удобного поиска по структуре. Важно, что именно такой способ представления формул молекулярных соединений является наиболее привычным химикам-синтетикам, которые в ежедневной работе сталкиваются с изображениями структурных формул, по сути являющимися именно двумерным графом.

Тем не менее, несмотря на свою распространенность, 2D молекулярные графы являются лишь упрощенным представлением молекулярной структуры. Трехмерные молекулярные графы обеспечивают более подробное представление молекул, включая информацию о стереохимии и пространственной ориентации атомов. В 3D графах, помимо узлов и рёбер, используются координаты x, y, и z для каждого атома, что позволяет моделировать реальную пространственную структуру молекулы. Эти модели используются для более детального учета межмолекулярных взаимодействий, в том числе - предсказания связывание лигандов с белками [27].

Недостатком использования молекулярных графов для автоматизированной обработки химической информации, особенно с применением подходов машинного обучения, является более высокая сложность соответствующих алгоритмов. В частности, хорошо себя зарекомендовало использование графовых нейронных сетей[28] (GNN, graph neural networks) для предсказания свойств молекулярных соединений [29,30].

1.3. Генерация спектра ЯМР по молекулярной структуре

Одномерный спектр ЯМР (по сравнению с ЯМР в двух и более измерениях [], которые в данном анализе не рассматриваются) в общем случае содержит три основных типа данных, которые могут быть связаны со строением молекулярного соединения: химический сдвиг (положение отдельного сигнала), константы спин-спинового взаимодействия (расщепление сигнала в так называемый мультиплет) и интегральная интенсивность сигнала. Тем не менее, для спектров на ядрах ^{13}C, анализ которых тут будет описываться в первую очередь важны химические сдвиги, потому что в большинстве случаев регистрируют и, соответственно, приводят данные в литературе о спектрах с гетероядерной развязкой от протонов $^{13}C\{^{1}H\}$, в которых отсутствует информация о расщеплении сигнала и его интенсивности. Предсказание химических сдвигов ядер в молекуле – задача, которая может быть решена несколькими способами [31], в частности – с использованием молекулярного моделирования при помощи теории функционала плотности[32]. Тем не менее, указанные расчеты являются достаточно

ресурсоемкими, особенно для систем, включающих в себя большое число ядер, в связи с этим в последнее время для этой цели пытаются использовать методы машинного обучения [33].

Первые такие попытки предсказания химических сдвигов основывались на классических методах машинного обучения [34], так и на использовании простых нейронных сетей [35], однако значительные улучшения были достигнуты при использовании в этих целях графовых нейросетей. Так, были разработаны методы быстрого предсказания химических сдвигов в спектрах ЯМР ^{1}H и ^{13}C [36]. Некоторые модели обладали крайне высокой эффективностью и была способны к предсказанию до пяти миллионов химических сдвигов в секунду [37], причем в ряде случае точность предсказания была сопоставима с точностью квантовохимических расчетов или даже превышала ее [38].

1.4. Генерация молекулярной структуры по спектральным данным

Определение строение химического соединения по спектральным данным всегда представляет собой более сложную задачу, чем предсказание формы спектра для молекулы, имеющей известное строение [39]. Решение обратной спектроскопической задачи в случае спектроскопии ЯМР в литературе представлено сравнительно бедно.

Поиск Монте-Карло по дереву

Одна из первых статей в этой области [40] описывала генерацию массива химических структур на основе комбинация метода Монте Карло с использованием рекуррентной нейронной сети. Затем применяли квантовохимическое DFT-моделирование для расчета спектров ^{1}H ЯМР, из результатов скоринга корректировали веса генерирующей сети и повторяли цикл до достижения сходимости входного и рассчитанного спектра. В результате был предложен рабочий подход для *de novo* генерации структур, который продемонстрировал хорошую точность для протонных спектров. С другой стороны, разработанный подход был апробирован протестировали только на девяти соединениях, и только для шести из них было получено верное решение. Дополнительной проблемой являлось высокая длительность процедуры генерации молекулярного графа, поскольку предложенный подход подразумевал использование времязатратных квантовохимических расчетов не только в ходе подготовки данных для обучения, но и непосредственно в цикле генерации, что исключает использование предложенного подхода для решения реальных задач.

Альтернативный подход к *de novo* генерации молекулярного графа на основе спектральных данных был предложен в работе Шридхарана с соавторами [41]. В данном случае обратная спектроскопическая задача была формализована как Марковский процесс принятия решения. Затем использовали комбинацию метода Монте Карло и графовой нейронной сети для итеративной генерации структуры соединения, соответствующего исходным спектрам. Несмотря на то, что разработанная модель показала достаточно высокую точность, процесс предсказания строения молекулы по-

прежнему занимал время, сопоставимое с временем, которое потребуется специалисту для интерпретации спектров вручную.

Развитие вышеописанного подхода было представлено в работе Девата с соавторами [42], в которой точность модели была заметно улучшена за счет включения в исходный датасет данных колебательной спектроскопии в инфракрасном (ИК) диапазоне. Важно отметить, что в зависимости от гиперпараметров модели, основанной, как и раньше, на Марковском процессе принятия решения, *de novo* генерация структуры занимала считанные минуты, однако набор данных был ограничен набором примерно пятидесяти тысяч молекулярных структур, причем как данные ИК спектроскопии, так и данных спектроскопии ЯМР на ядрах ^{13}C были получены в ходе квантовохимических расчетов. Таким образом, использованный для обучения модели датасет был ни хоть сколько-нибудь полным, ни точным, что естественным образом ограничивает ее применимость.

Использование трансформеров

Альтернативный подход был предложен в работе Яо с соавторами [43]. На основе базы спектров ^{13}C ЯМР была обучена модель CMGNet, позволяющая на основе брутто формулы, химических сдвигов ЯМР ^{13}C и информации о фрагментах, имеющихся в составе молекулы, генерировать библиотеку возможных соединений, которые могли бы обладать указанным спектров. Модель использовала BART (bidirectional and autoregressive transformer), обучение состояло из трех основных этапов. На этапе предварительного обучения использовали обширную библиотеку структур (360 миллионов) в формате SMILES (без спектральных данных), на этапе первичного обучения использовали библиотеку с 45.3 миллионами структур и спектров (рассчитанных с использованием квантовохимических DFT-расчетов), на финальном этапе дообучения использовали небольшую библиотеку экспериментальных спектров на 370 тысяч соединений. К недостаткам модели стоит отнести то, что обучение, как и ранее, проводили в основном на искусственно смоделированных данных в связи с наличием малого числа экспериментальных спектров ЯМР. В результате работы модели получали большую серию соединений-лидеров из десятков и сотен молекулярных графов, из которых затем нужно было выбирать верный. Кроме того, без сведений о брутто-формуле и строении молекулярных фрагментов метрики производительности модели падали практически вдвое.

Наконец, недавно[44] было предложено использовании базы из синтетических спектров ЯМР, полученных в широко используемой для анализа спектров ЯМР программе MestreNova [45], для обучения модели, основанной на архитектуре молекулярных трансформеров [46]. Таким образом, решение обратной спектроскопической задачи фактически было сведено к хорошо известной задаче машинного перевода, в которой на вход модели подавался спектр ЯМР, а на выходе получали закодированый молекулярный граф. Несмотря на воодушевляющие результаты, модель была протестирована только для очень сильно ограниченного набора исходных структур, спектральные данные для которых были получены искусственным образом с использованием большого числа приближений.

Таким образом, одним из основных ограничений всех спектроскопических генеративных моделей, описанных в литературе, является отсутствие достаточно большого набора экспериментальных спектров ЯМР, пригодных для обучения модели. В связи с этим ряд исследователей в последние годы предложил подходы для создания таких спектральных баз данных, в том числе – с использованием алгоритмов компьютерного зрения для распознавания изображения спектров ЯМР в том виде, в котором их приводят в научных публикациях [47].

1.5. Выводы

Таким образом, на основе анализа литературных данных можно сделать два основных вывода:

1) Налицо переход от случайного поиска возможных молекулярных структур, основанном на методе Монте Карло, к подходам, учитывающим «химическую логику» за счет использования практик, используемых для обработки естественного языка (в частности – архитектура трансформеров).

2) Наблюдается увеличение количества спектральных данных, используемых для обучения и валидации модели, однако до сих пор большая часть таких данных – синтетическая (получена путем квантовохимических расчетов или более примитивных подходов)

Таким образом, на основе обнаруженных трендов можно предложить следующую исследовательскую гипотезу: адаптация известных подходов генеративного искусственного интеллекта к области химии позволит на основе большого массива **экспериментальных** спектральных данных, приведенных в научных публикациях, построить генеративную модель, определяющую строение ранее не известного химического соединения на основе его спектров ЯМР.

Глава 2. Модели mol2spec

2.1. Введение

2.2. Предсказание спектра ЯМР на основе имеющегося молекулярного графа с использованием графовой нейронной сети с передачей сообщений

В данном подходе к предсказанию спектра ЯМР [Kwon с соавт, J. Chem. Inf. Model. 2020, 60, 2024–2030] молекулы представлены в виде графов, где узлы соответствуют тяжелым атомам, а рёбра — их связям. Для повышения эффективности графового представления:

- Атомы водорода учитываются имплицитно через признаки соседних узлов, что снижает сложность вычислений.
- Признаки узлов включают такие свойства, как тип атома, степень гибридизации, хиральность, степень окисления, количество связанных водородов, ароматичность и принадлежность к кольцам.
- Признаки рёбер учитывают тип связи, стереохимию, сопряжённость, принадлежность к одному кольцу и графовое расстояние между узлами.

Использование MPNN включает следующие этапы:

- Инициализация: Узлы и рёбра преобразуются в векторы признаков с использованием полносвязной нейронной сети.
- Передача сообщений: На каждом шаге графовой итерации обновляются векторы узлов за счёт агрегации информации от соседних узлов и рёбер.
- Обновление состояния: Для обновления векторов узлов используется рекуррентная нейронная сеть с пятью шагами передачи сообщений.
- Прогноз: После агрегации итоговый слой предсказывает химические сдвиги для каждого узла.

В связи с тем, что использованный датасет значительно превышал по размеру и структурному разнообразию стандартный датасет NMRShiftDB2, использованный авторами работы [Kwon с соавт, J. Chem. Inf. Model. 2020, 60, 2024–2030], в результате обучения удалось значительно повысить целевые метрики в предсказании химических сдвигов в спектрах ^{13}C ЯМР. При этом анализ показал, что во многих случаях основной вклад в усредненную ошибку вносится несколькими достаточно малораспространенными классами химических соединений, плохо представленными в обучающей выборке. Дальнейшее расширение датасета с высокой вероятностью позволит решить эту проблему.

2.3. Предсказание спектра ЯМР на основе имеющегося молекулярного графа с использованием графового трансформера

Архитектура модели:

Кодировщик признаков. Молекула представлена в виде неориентированного графа, где узлы соответствуют атомам, а рёбра — связям. Узлы и рёбра графа кодируются числовыми признаками, такими как тип атома (углерод, кислород и другие), гибридизация (например, sp, sp^2), тип связи (одинарная, двойная и так далее).

Позиционное кодирование. Для учёта структурных отношений между атомами используются вероятности случайных блужданий. Этот подход помогает одновременно захватывать локальные взаимодействия и глобальные зависимости в графе. Например, вероятности того, что атомы связаны через несколько шагов, учитываются в позиционном представлении графа.

Графовый трансформер. Узлы и связи преобразуются в новые представления с использованием обучаемых весов. Затем вычисляются коэффициенты внимания, которые определяют важность взаимодействия между атомами. Эти значения используются для обновления представлений узлов с учётом их взаимодействий с соседями.

Регрессионный блок. После обработки графа трансформером создаются финальные представления атомов, которые проходят через трехслойную полносвязную нейронную сеть для предсказания химических сдвигов. При этом фокусируется внимание только на тяжёлых атомах (например, углероде).

Несмотря на то, что описанная в литературе модель (Chen с соавт., J Cheminform, 2024, 16, 132) демонстрировала лучшую производимость по сравнению с MPNN, обучение ее на нашем датасете привело к получению достаточно посредственного качества предсказания спектра. Причины такого поведения и возможность их преодоление будет изучена в дальнейшем.

Глава 3. Модели spec2mol

3.1. Введение

3.2. Предсказание строение соединения (молекулярного графа) на основе его спектра ЯМР с использованием Монте-Карло поиска по дереву

В рамках данного подхода [S. Devata с соавт, Digital Discovery, 2024,3, 818-829] проблема предсказания структуры молекулы формулируется как Марковский процесс принятия решений (Markov Decision Process, MDP).
Генерация структур: Метод применяет алгоритм поиска Монте-Карло по дереву (Monte Carlo Tree Search, MCTS) для построения молекулярного графа. Каждое действие в MDP соответствует добавлению химической связи между двумя атомами.
Фичеризация графов: Узлы графа (атомы) кодируются с учетом таких характеристик, как тип атома, гибридизация, заряд, химический сдвиг из данных ^{13}C NMR, а рёбра (связи) — с учетом типа связи, сопряженности и принадлежности к кольцам.
Процесс предсказания структуры:

- На основе входной молекулярной формулы строится граф с атомами без связей.
- Итеративное построение графа: Алгоритм поиска Монте-Карло (Monte Carlo Tree Search, MCTS) последовательно добавляет связи между атомами. Каждое действие (добавление связи) выбирается на основе текущей политики, которая оптимизируется моделями.
- Управление построением молекулярной структуры:
    - Модель приоритета оценивает вероятность добавления различных типов связей (одинарной, двойной, тройной) между парами атомов.
    - Модель оценки вычисляет ценность текущего состояния графа, основанную на схожести с предполагаемой структурой и спектральными данными.
- Алгоритм завершает построение, когда добавление новых связей невозможно или текущая структура соответствует заданным критериям.

Несмотря на то, что для описанного авторами модели применения модель работает хорошо, ее использования для целей настоящей работы оказалось невозможным, поскольку она требует наличия во входных данных информации о мультиплетности сигнала в спектре ^{13}C ЯМР. Поскольку в подавляющем числе научных публикаций приводят данные только для спектров с гетероядерной развязкой спин-спинового взаимодействия с протонами, сбор соответствующего датасета на основе экспериментальных данных желаемого качества становится невозможным, а практическое применение такой модели будет затруднительным. Тем не менее, сам алгоритм постепенного достраивания молекулярной структуры, возможно, будет использован в дальнейшем для генерации полного молекулярного графа из фрагментов (см. далее)

3.3. Алгоритмическая сборка из молекуляных фрагментов, полученных на основании трансформерной модели

Данный метод был разработан на основе работы [Hu с соавт., ACS Cent. Sci., 2024, 10, 2162−2170], однако в отличии от мультимодельной архитектуры, призванной на основе спектра ЯМР генерировать сначала список подструктур, на основе которых вторая модель создавала итоговый молекулярный граф, в нашем случае лишь первая часть задачи выполнялась с использованием трансформерной модели, а комбинирование молекулярных фрагментов осуществлялось алгоритмически, за счет постепенно достраивания структуры доступными заместителями.

На данный момент, именно указанный подход показал максимальную эффективность в генерации молекулярного графа по спектральным данным. Стоит отметить, что лишь первая часть двухстадийного процесса, т.е. генерация молекулярных фрагментов, в настоящий момент реализована с использованием подходов машинного обучения. Сборка фрагментов в полную молекулярную структуру осуществляется путем алгоритмического поиска, направляемого сравнением полученного спектра с предсказанием, полученным с использованием MPNN. Ввиду высокой вычислительной сложности, предсказание строения больших молекул (более 20 неводородных атомов) занимает значительное время. Тем не менее, в дальнейшем планируется создание второй трансформерной модели для решения той же задачи.

Результаты и выводы

(пока – промежуточные!)

1. Предсказание спектра ЯМР с использованием MPNN
   Обучение графовой нейронной сети передачи сообщений на расширенном датасете позволило улучшить точность предсказания химических сдвигов. Однако основной вклад в ошибки связан с недостаточной представленностью редких классов соединений. Дальнейшее расширение датасета может решить эту проблему.
2. Графовый трансформер для предсказания спектров ЯМР
   Несмотря на высокую производительность трансформерной модели в литературе, её обучение на собственном датасете привело к посредственным результатам. Причины такого поведения требуют дальнейшего изучения и возможной адаптации модели.
3. Генерация структур методом поиска Монте-Карло по дереву
   Текущая реализация метода оказалась неприменимой из-за необходимости данных о мультиплетности сигнала ^{13}C спектра, которые практически всегда отсутствуют в публикациях. Тем не менее, сам подход пошагового построения структуры может быть полезен в задачах генерации молекулярного графа из фрагментов.
4. Двухстадийное предсказание молекулярного графа
   На данный момент данный подход демонстрирует максимальную эффективность. Первая стадия (генерация молекулярных фрагментов) реализована с использованием машинного обучения, а сборка структуры — алгоритмическим методом. Ограничения связаны с высокой вычислительной сложностью для больших молекул, но разработка второй трансформерной модели может решить эту задачу.

Таким образом, использование ряда подходов (MPNN, графовые трансформеры, алгоритмическая сборка) позволило добиться прогресса в предсказании структур молекул на основании их спектров ЯМР. Наибольшая эффективность достигнута в комбинированном подходе, который требует дальнейшей оптимизации для применения на более сложных молекулах.

Список литературы

[1] T. D. W. Claridge, *High-Resolution NMR Techniques in Organic Chemistry*, Elsevier, **2016**.

[2] L. Ferella, A. Rosato, P. Turano, J. Plavec, in *NMR of Biomolecules*, John Wiley & Sons, Ltd, **2012**, pp. 33–50.

[3] J. Cavanagh, *Protein NMR Spectroscopy: Principles and Practice*, Academic Press, **1996**.

[4] R. H. Hashemi, W. G. Bradley, C. J. Lisanti, *MRI: The Basics: The Basics*, Lippincott Williams & Wilkins, **2012**.

[5] D. Goldfarb, S. Stoll, *EPR Spectroscopy: Fundamentals and Methods*, John Wiley & Sons, **2018**.

[6] M. Manso Jimeno, K. S. Ravi, Z. Jin, D. Oyekunle, G. Ogbole, S. Geethanath, *Magnetic Resonance Imaging* **2022**, *89*, 42–48.

[7] I. Koktzoglou, R. Huang, W. J. Ankenbrandt, M. T. Walker, R. R. Edelman, *Magnetic Resonance in Medicine* **2021**, *86*, 335–345.

[8] S. U. Khan, N. Ullah, I. Ahmed, I. Ahmad, M. I. Mahsud, *Current Medical Imaging Reviews* **2019**, *15*, 243–254.

[9] G. Giovannetti, N. Fontana, A. Flori, M. F. Santarelli, M. Tucci, V. Positano, S. Barmada, F. Frijia, *Sensors* **2024**, *24*, 1954.

[10] G. Jeschke, *Journal of Magnetic Resonance* **2019**, *306*, 36–41.

[11] D. R. Davydov, D. O. Antonov, E. G. Kovaleva, *Appl Magn Reson* **2023**, *54*, 595–612.

[12] A. Ashuiev, A. Giorgia Nobile, D. Trummer, D. Klose, S. Guda, O. V. Safonova, C. Copéret, A. Guda, G. Jeschke, *Angewandte Chemie* **2024**, *136*, e202313348.

[13] D. Chen, Z. Wang, D. Guo, V. Orekhov, X. Qu, *Chemistry – A European Journal* **2020**, *26*, 10391–10401.

[14] C. Cobas, *Magnetic Resonance in Chemistry* **2020**, *58*, 512–519.

[15] Z. Zou, Y. Zhang, L. Liang, M. Wei, J. Leng, J. Jiang, Y. Luo, W. Hu, *Nat Comput Sci* **2023**, *3*, 957–964.

[16] H. Tamoto, R. dos S. Gioria, C. de C. Carneiro, *Journal of Petroleum Science and Engineering* **2023**, *220*, 111169.

[17] W. K. Peng, *Engineering Reports* **2021**, *3*, e12383.

[18] I. M. Novitskiy, A. G. Kutateladze, *J. Org. Chem.* **2022**, *87*, 4818–4828.

[19] X. Qu, Y. Huang, H. Lu, T. Qiu, D. Guo, T. Agback, V. Orekhov, Z. Chen, *Angewandte Chemie* **2020**, *132*, 10383–10386.

[20] X. Kong, L. Zhou, Z. Li, Z. Yang, B. Qiu, X. Wu, F. Shi, J. Du, *npj Quantum Inf* **2020**, *6*, 1–10.

[21] L. Xue, J. Bajorath, *Combinatorial chemistry & high throughput screening* **2000**, *3*, 363–372.

[22] A. Capecchi, D. Probst, J.-L. Reymond, *Journal of cheminformatics* **2020**, *12*, 1–15.

[23] L. David, A. Thakkar, R. Mercado, O. Engkvist, *J Cheminform* **2020**, *12*, 56.

[24] Ю. И. Нейн, М. Н. Иванцова, **2020**.

[25] H. L. Morgan, *J. Chem. Doc.* **1965**, *5*, 107–113.

[26] D. Rogers, M. Hahn, *J. Chem. Inf. Model.* **2010**, *50*, 742–754.

[27] W. Torng, R. B. Altman, *J. Chem. Inf. Model.* **2019**, *59*, 4131–4149.

[28] C. Merkwirth, T. Lengauer, *J. Chem. Inf. Model.* **2005**, *45*, 1159–1168.

[29] E. N. Feinberg, D. Sur, Z. Wu, B. E. Husic, H. Mai, Y. Li, S. Sun, J. Yang, B. Ramsundar, V. S. Pande, *ACS Cent. Sci.* **2018**, *4*, 1520–1530.

[30] K. Atz, F. Grisoni, G. Schneider, *Nat Mach Intell* **2021**, *3*, 1023–1032.

[31] M. W. Lodewyk, M. R. Siebert, D. J. Tantillo, *Chem. Rev.* **2012**, *112*, 1839–1862.

[32] M. Kaupp, M. Buhl, V. G. Malkin, *Calculation of NMR and EPR Parameters*, Wiley Online Library, **2004**.
[33] E. Jonas, S. Kuhn, N. Schlörer, *Magnetic Resonance in Chemistry* **2022**, *60*, 1021–1031.
[34] S. Kuhn, B. Egert, S. Neumann, C. Steinbeck, *BMC Bioinformatics* **2008**, *9*, 400.
[35] J. Meiler, W. Maier, M. Will, R. Meusinger, *Journal of Magnetic Resonance* **2002**, *157*, 242–252.
[36] Y. Guan, S. V. Shree Sowndarya, L. C. Gallegos, P. C. St. John, R. S. Paton, *Chem. Sci.* **2021**, *12*, 12012–12026.
[37] Z. Yang, M. Chakraborty, A. D. White, *Chem. Sci.* **2021**, *12*, 10802–10809.
[38] Y. Kwon, D. Lee, Y.-S. Choi, M. Kang, S. Kang, *J. Chem. Inf. Model.* **2020**, *60*, 2024–2030.
[39] B. Sridharan, M. Goel, U. Deva Priyakumar, *Chemical Communications* **2022**, *58*, 5316–5331.
[40] J. Zhang, K. Terayama, M. Sumita, K. Yoshizoe, K. Ito, J. Kikuchi, K. Tsuda, *Science and Technology of Advanced Materials* **2020**, *21*, 552–561.
[41] B. Sridharan, S. Mehta, Y. Pathak, U. D. Priyakumar, *J. Phys. Chem. Lett.* **2022**, *13*, 4924–4933.
[42] S. Devata, B. Sridharan, S. Mehta, Y. Pathak, S. Laghuvarapu, G. Varma, U. Deva Priyakumar, *Digital Discovery* **2024**, DOI 10.1039/D4DD00008K.
[43] L. Yao, M. Yang, J. Song, Z. Yang, H. Sun, H. Shi, X. Liu, X. Ji, Y. Deng, X. Wang, *Anal. Chem.* **2023**, *95*, 5393–5401.
[44] M. Alberts, F. Zipoli, A. C. Vaucher, **2023**, DOI 10.26434/chemrxiv-2023-8wxcz.
[45] M. R. Willcott, *J. Am. Chem. Soc.* **2009**, *131*, 13180–13180.
[46] "Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction | ACS Central Science," can be found under https://pubs.acs.org/doi/10.1021/acscentsci.9b00576, **n.d.**
[47] W. Jia, Z. Yang, M. Yang, L. Cheng, Z. Lei, X. Wang, *J. Chem. Inf. Model.* **2021**, *61*, 21–25.