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

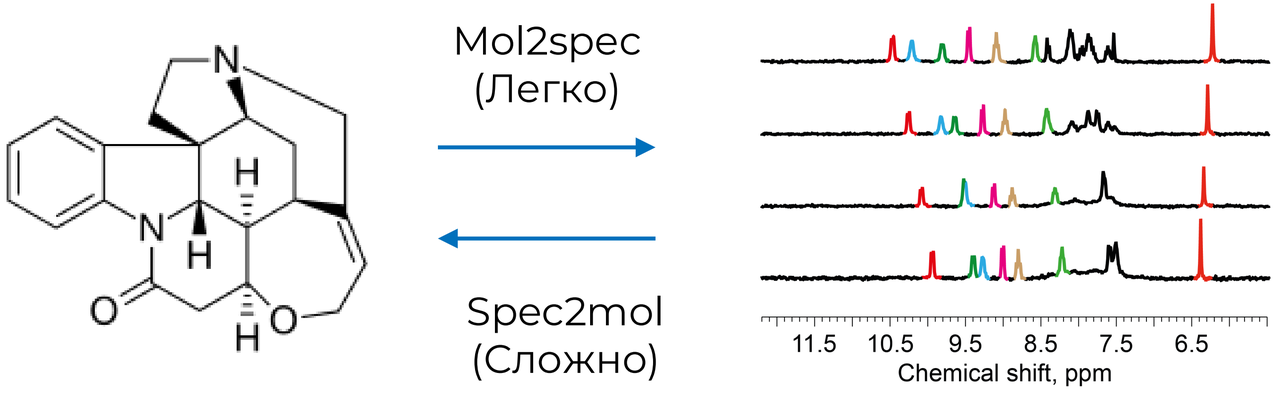
Выполнил: Новиков Валентин Владимирович, д.х.н.

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

# Проблема, новизна, актуальность

АКТУАЛЬНОСТЬ:

- Спектроскопия ЯМР – основной метод подтверждения строения органических соединений
- Используется не только в академических исследованиях, но и в фарминдустрии, нефтехимии, медицине и т.д.
- Ручная интерпретация спектров требует участия квалифицированного специалиста
- При автоматизации химических исследований часто именно характеризация соединений (в том числе – ЯМР) является «бутылочным горлышком»




НОВИЗНА:

- Подходы к решению прямой спектроскопической задачи (mol2spec) существуют и неплохо работают
- Обратную спектроскопическую задачу ЯМР (spec2mol) в общем случае решать без участия человека до сих не умеют

ПРОБЛЕМА:

Существующие генеративные подходов искусственного интеллекта, не позволяют на основе только предложенных спектров ЯМР неизвестного химического соединения генерировать молекулярный граф, описывающий его строение.

# Цель и задачи исследования

Целью предлагаемого исследования является поиск универсальных подходов для автоматизированной интерпретации спектров ЯМР. 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?»

Объект исследования: взаимосвязь между строением химического соединения и его спектрами ЯМР

Предмет исследования: методические основы de novo генерации химических структур на основе входных спектральных данных.

## ЗАДАЧИ ИССЛЕДОВАНИЯ:

1. Сбор и подготовка набора данных, содержащего не менее миллиона экспериментальных спектров ЯМР 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.

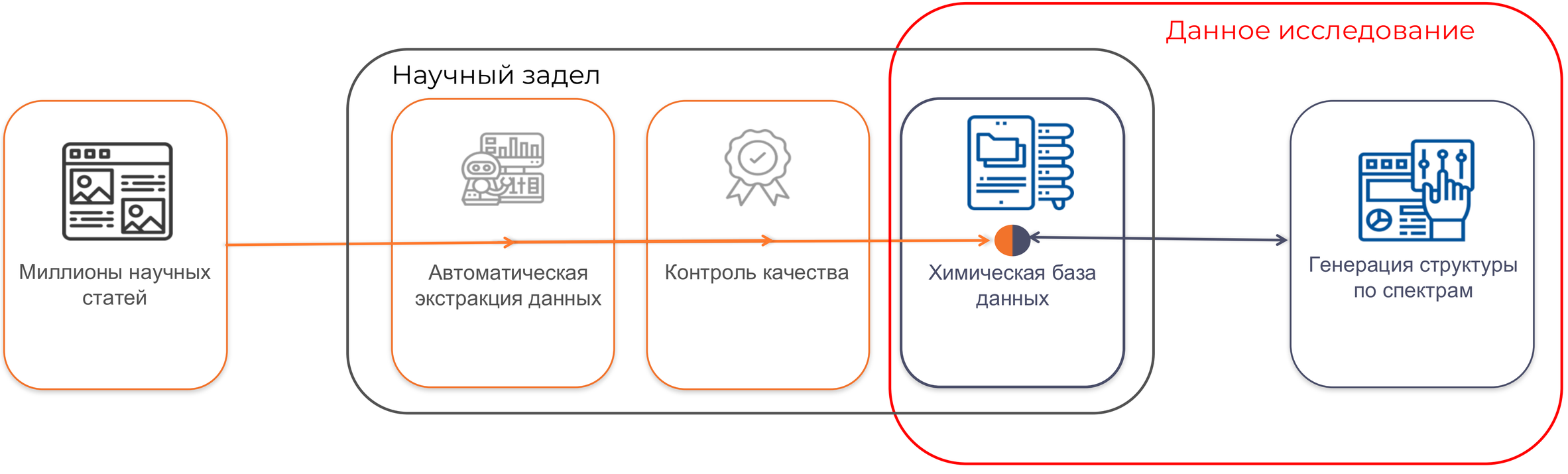
3. Использование собранного набора данных и представленных в научной литературе подходов, основанных на применении графовых нейронных сетей, для обучения нейросети, решающей прямую спектроскопическую задачу (предсказание спектров ЯМР на основе строения исходного соединения)

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.

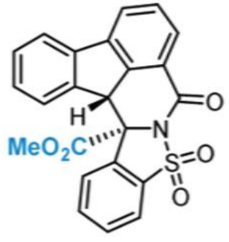
5. Тестирование предсказательной способности полученной генеративной нейросети, выявление классов химических соединений, для которых разработанные подходы демонстрируют наилучшую и наихудшую эффективность и итеративная доработка архитектуры и гиперпараметров модели для улучшения качества предсказания.

# Исследовательская гипотеза

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.



# Реальный пример спектральных данных в статье




3b

**3b**, Prepared according to the general procedure in 0.1 mmol scale, 82% yield, 34 mg, white solid, m.p. 240-243°C, flash column with a petroleum ether/ethyl acetate eluent (gradient 3:1). **$^{1}$H NMR** (600 MHz, $CDCl_3$) $\delta$ 8.24 (d, $J$ = 7.8 Hz, 1H), 8.03 (d, $J$ = 7.8 Hz, 1H), 7.96 (t, $J$ = 7.2 Hz, 2H), 7.93 – 7.88 (m, 2H), 7.84 – 7.81 (m, 2H), 7.63 (t, $J$ = 7.2 Hz, 1H), 7.56 (t, $J$ = 7.2 Hz, 1H), 7.49 (t, $J$ = 7.2 Hz, 1H), 4.53 (s, 1H), 3.24 (s, 3H); **$^{13}$C {$^{1}$H} NMR** (151 MHz, $CDCl_3$): $\delta$ 166.4, 161.7, 143.8, 142.0, 140.8, 139.9, 135.3, 134.1, 131.6, 131.0, 130.2, 128.9, 128.2, 126.3, 126.2, 125.8, 125.1, 124.9, 122.5, 121.6, 116.7, 73.3, 53.6, 52.8; **HPLC:** CHIRALPAK IA, *n*-hexane/isopropanol = 80/20, flow rate = 1.0 mL/min, UV = 254 nm, $t_R$ = 23.7 min (major), 26.5 min (minor), 96% ee. **Optical rotation**: $[\alpha]^{25}_{D}$ = +185 (c = 1.0 in EtOAc); **HRMS (ESI) m/z**: $[M+Na]^+$ Calcd for $C_{23}H_{15}NNaNO_5S^+$ 440.0563; Found 440.0562.

Спектры ЯМР $^{1}$H часто перекрываются и зависят от использованной частоты спектрометра ЯМР. В связи с этим, было принято решения использовать данные спектроскопии ЯМР на ядрах $^{13}$C.

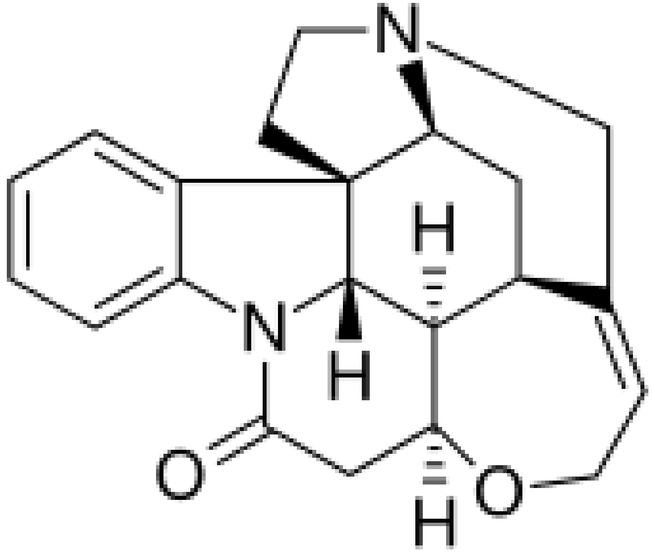
# Методология исследования

Mol2Spec

1) Графовая нейронная сеть.
2) Трансформерная архитектура

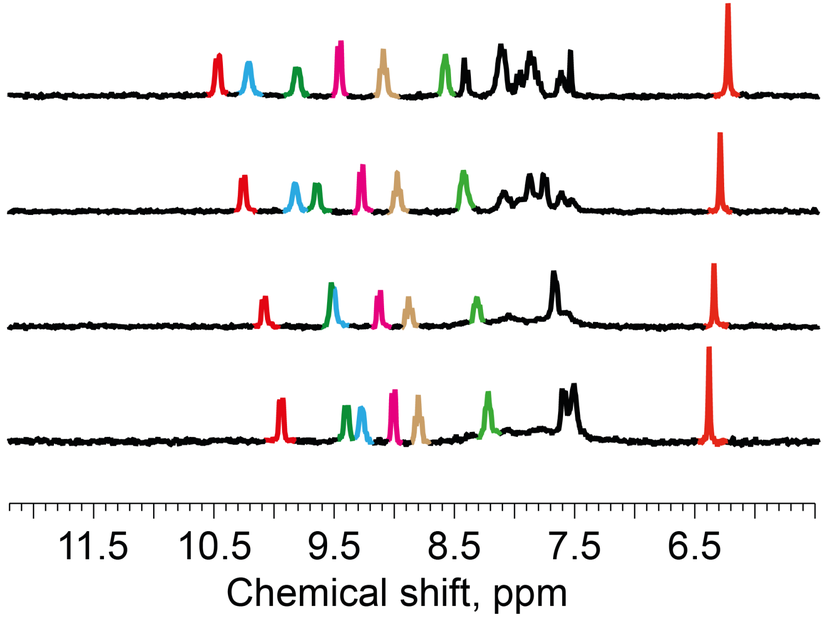
Spec2Mol

1) Марковский процесс принятия решения на основе Монте Карло поиска по дереву
2) Трансформерная архитектура




Mol2spec
(Легко)

Spec2mol
(Сложно)



# Источники данных

База данных OdanChem (https://www.odanchem.org)

- Содержит сведения о более чем шести миллионах спектров ЯМР $^{1}H$, $^{13}C$, $^{31}P$, $^{19}F$ и др.
- Данные получены парсингом научных статей в области органической химии

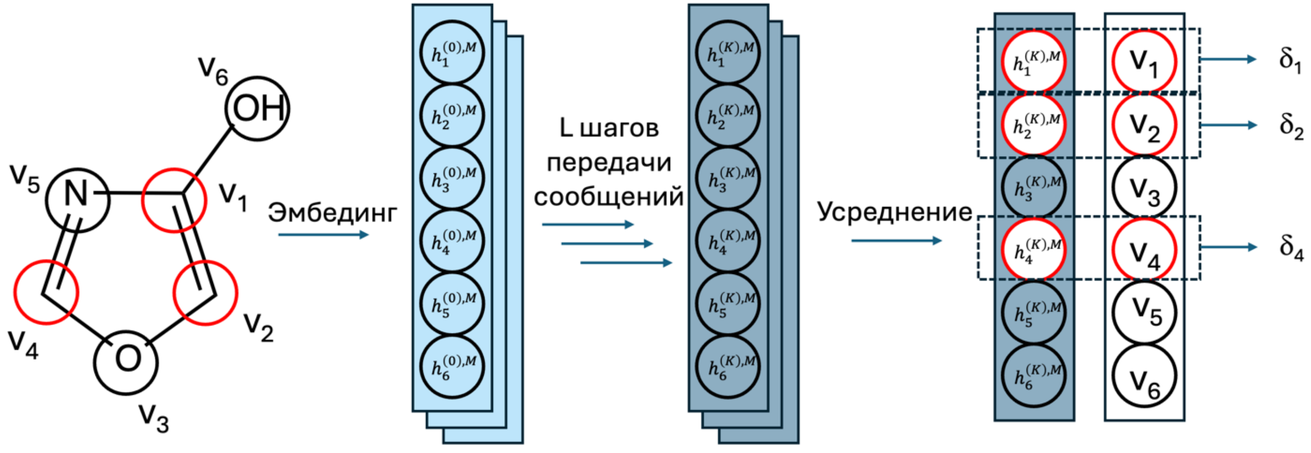
В ходе работы над ВКР была проведена дополнительная очистка данных

Текущий датасет включает спектры ЯМР $^{13}C$ для 1363100 органических соединений.

Датасет постоянно обновляется.

# Разметка данных

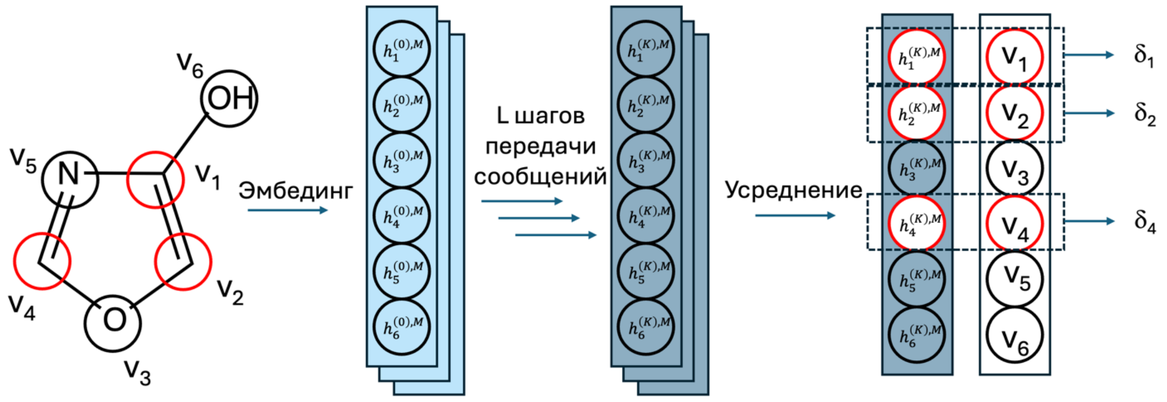
Для обучения моделей, предсказывающих химические сдвиги в спектрах 13C ЯМР для выбранного молекулярного графа, необходим размеченный датасет, в котором каждому сигналу в спектре соответствует конкретное ядро 13C в молекуле либо комбинация таких сигналов в случае ядер, связанных друг с другом операциями симметрии. В исходной базе спектров OdanChem собрано большое число спектров 13C ЯМР для множества соединений, но в большинстве случаев – без явно указанного соотнесения сигналов с конкретными ядрами $^{13}C$.




Адаптировано из J. Han с соавт., Phys. Chem. Chem. Phys., 2022, 24, 26878.

1. Предсказание спектра: с использованием графовой нейронной сети для каждой молекулы получали теоретический спектр $^{13}C$ ЯМР;
2. Упорядочивание химических сдвигов: предсказанный спектр сортировали по увеличению химических сдвигов;
3. Сопоставление с литературными данными: аналогичным образом сортировали экспериментальный спектр $^{13}C$ ЯМР, приведенный в литературе;
4. Назначение соответствий: каждому сигналу из экспериментального спектра приписывали значение из соответствующего предсказанного спектра, основываясь на минимальной разнице химических сдвигов.

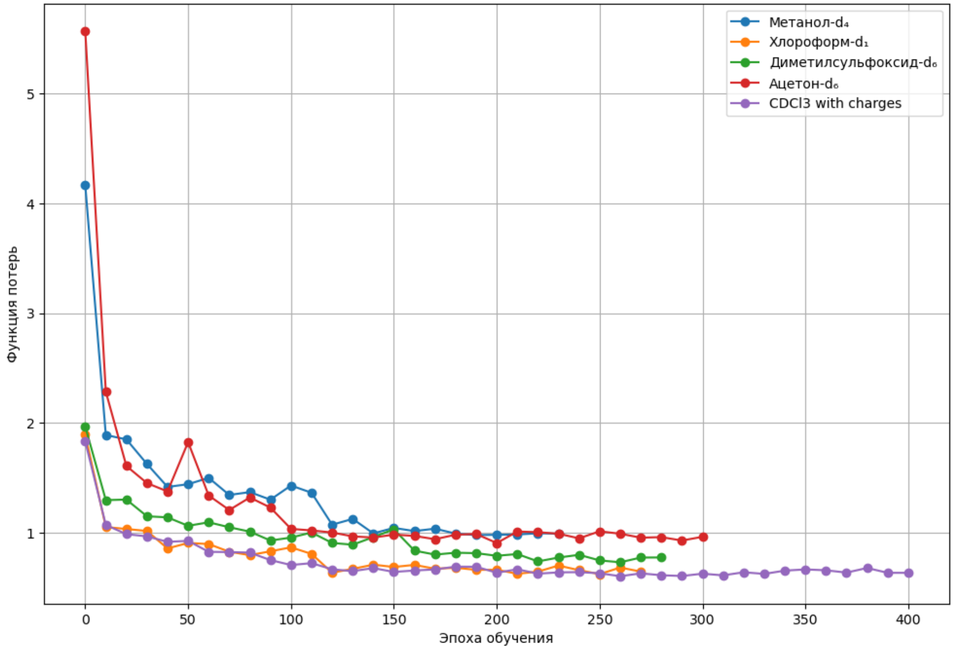
# mol2spec – обучение на собранных данных

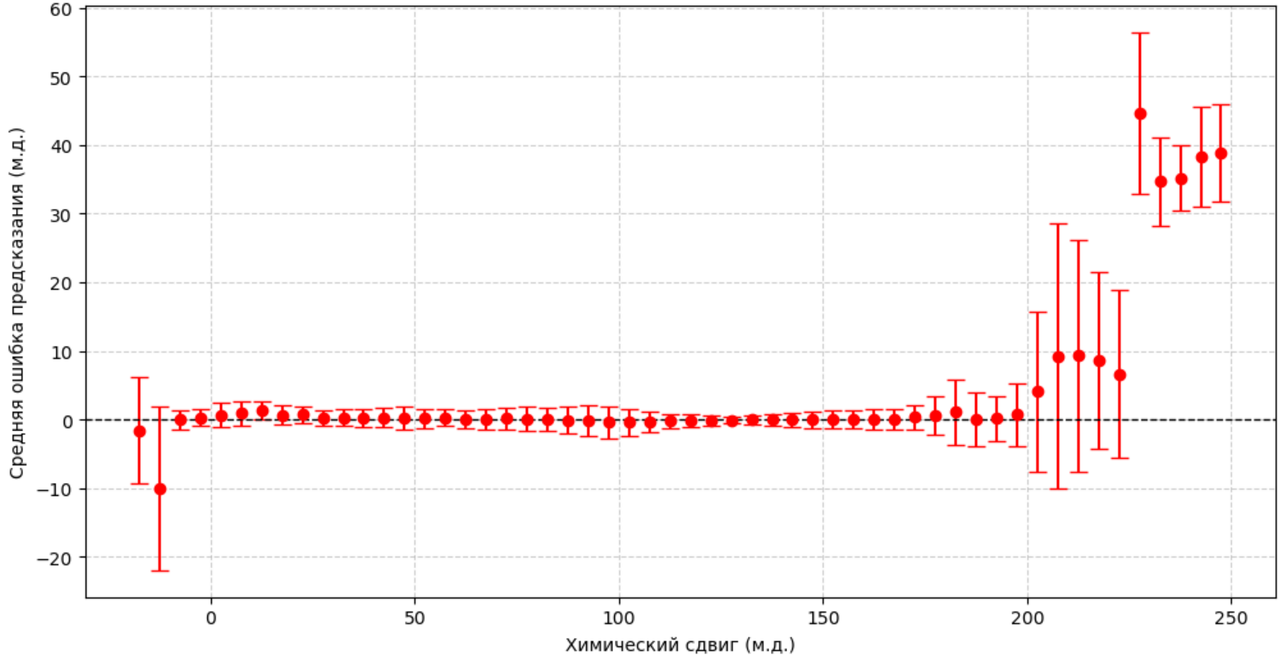



MAE (м.д.)

Было: 1.340 (PCCP, 2022, 24, 26878)

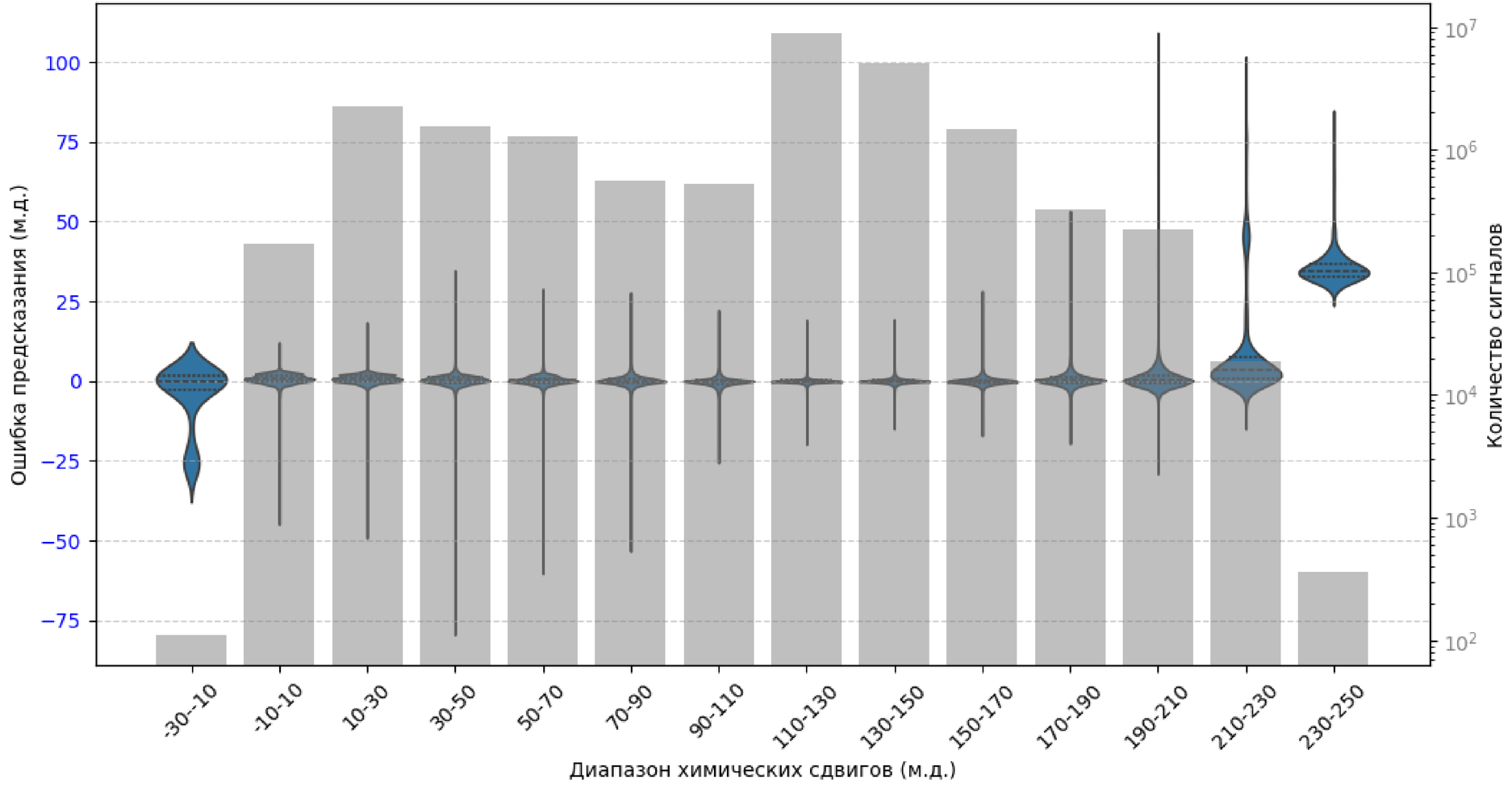
Стало: 0.455 (данное исследование)



# mol2spec – ошибка предсказания

Хорошо видно, что ошибка предсказания велика только для диапазонов химических сдвигов, для которых в базе очень мало данных



# spec2mol

Общая проблема – правда ли в спектре ЯМР закодирована вся информация о строении молекулы?

В связи с этим определение структуры молекулы проводится на основе:

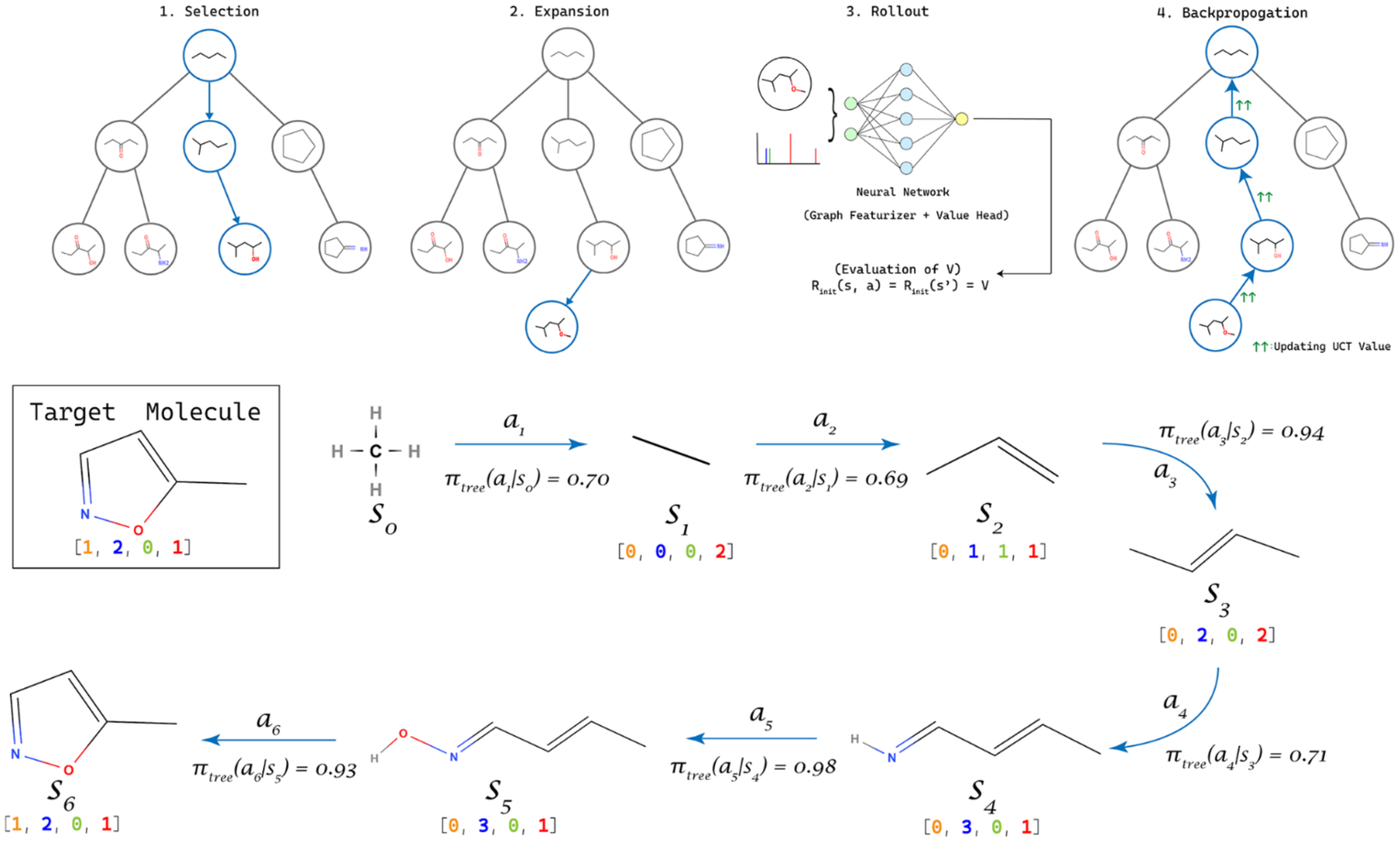
- ЯМР $^{1}H$
- ЯМР $^{13}C$
- Спектры в инфракрасном диапазоне
- Молекулярная масса
- Брутто-формула
- Присутствие конкретных молекулярных фрагментов

В литературе хорошие результаты были получены с использованием:

- Марковский процесс принятия решения (Монте-Карло движение по дереву)
- Трансформерные архитектуры

# spec2mol

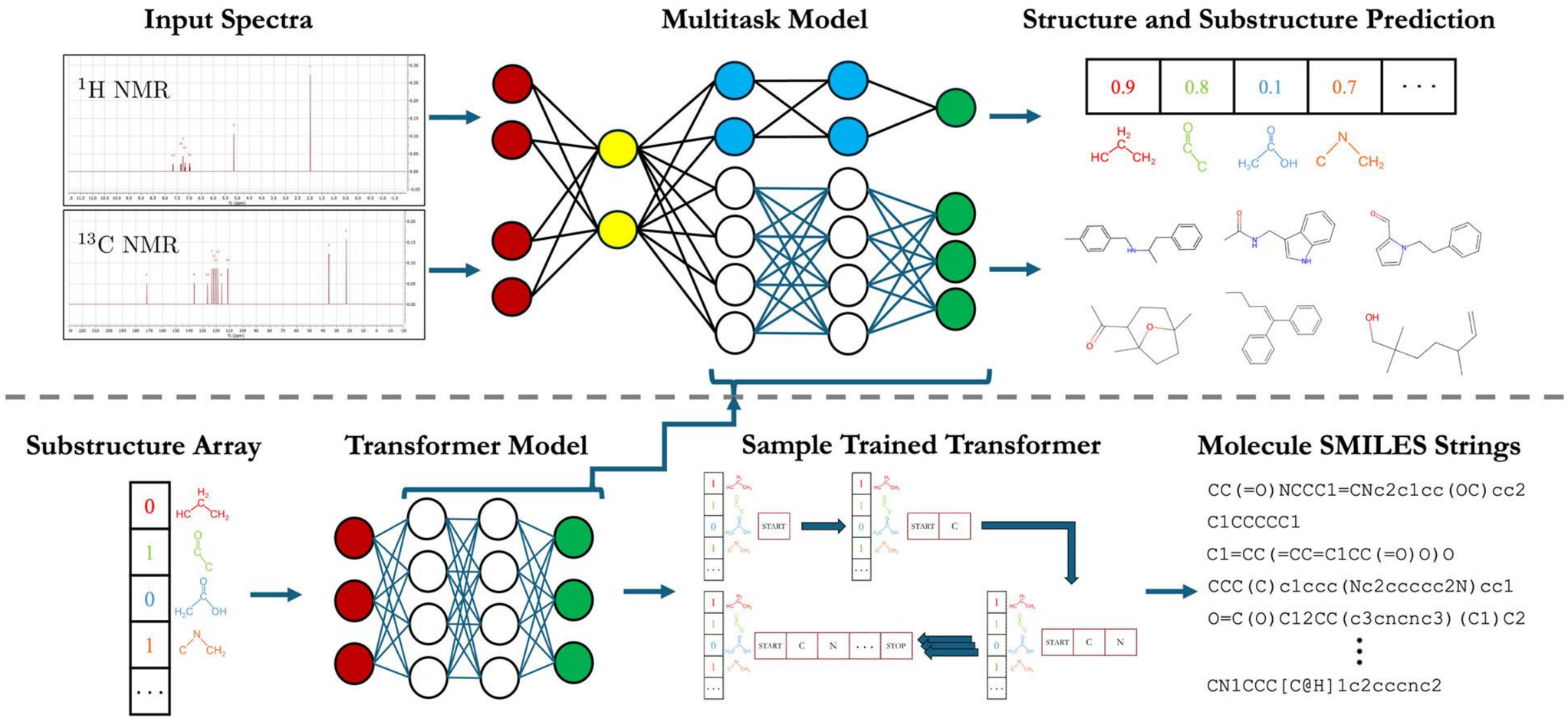
- Марковский процесс принятия решения (Монте-Карло движение по дереву)




Sridharan с соавт., J. Phys. Chem. Lett., 2022, 13, 4924
Kanakala с соавт, Digital Discovery 2024, 3, 2417

# spec2mol

- Трансформерные архитектуры

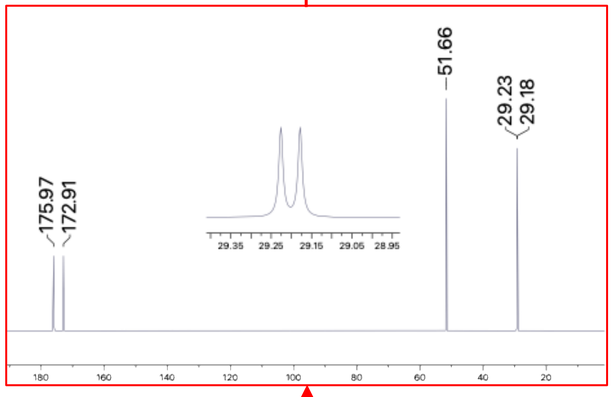


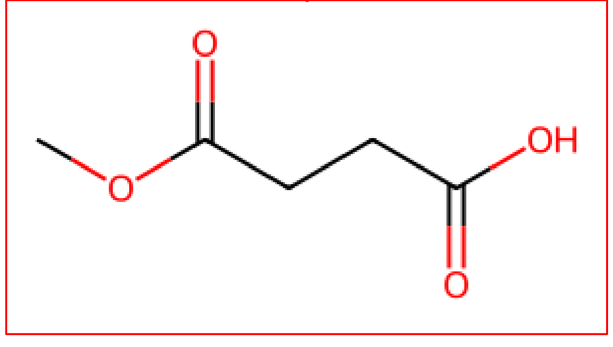

Hu с соавт., ACS Cent. Sci. 2024, 10, 2162

# Наша архитектура

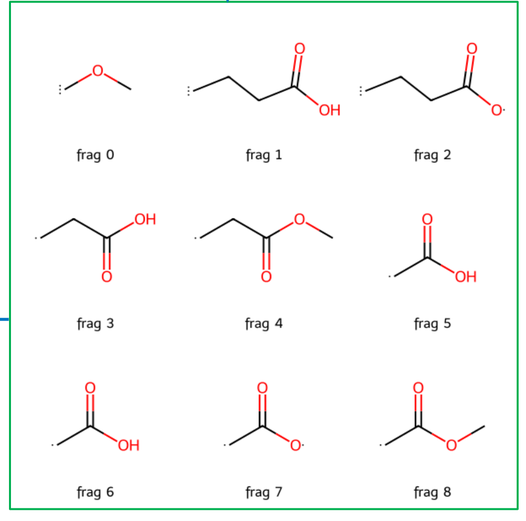
Representation in dataset:

[175.97, 172.91, 51.66, 29.23, 29.18]




frags2mol Encoder-decoder transformer




spec2frags Encoder-decoder transformer

`Original spectrum:` `[175.97, 172.91, 51.66, 29.23, 29.18]`

`Shifted by 50 ppm:` `[225.97, 222.91, 101.66, 79.23, 79.18]`

`Input vector:` **`[EOS, 79, 79, 102, 223, 226, BOS]`**

**Библиотека фрагментов:**

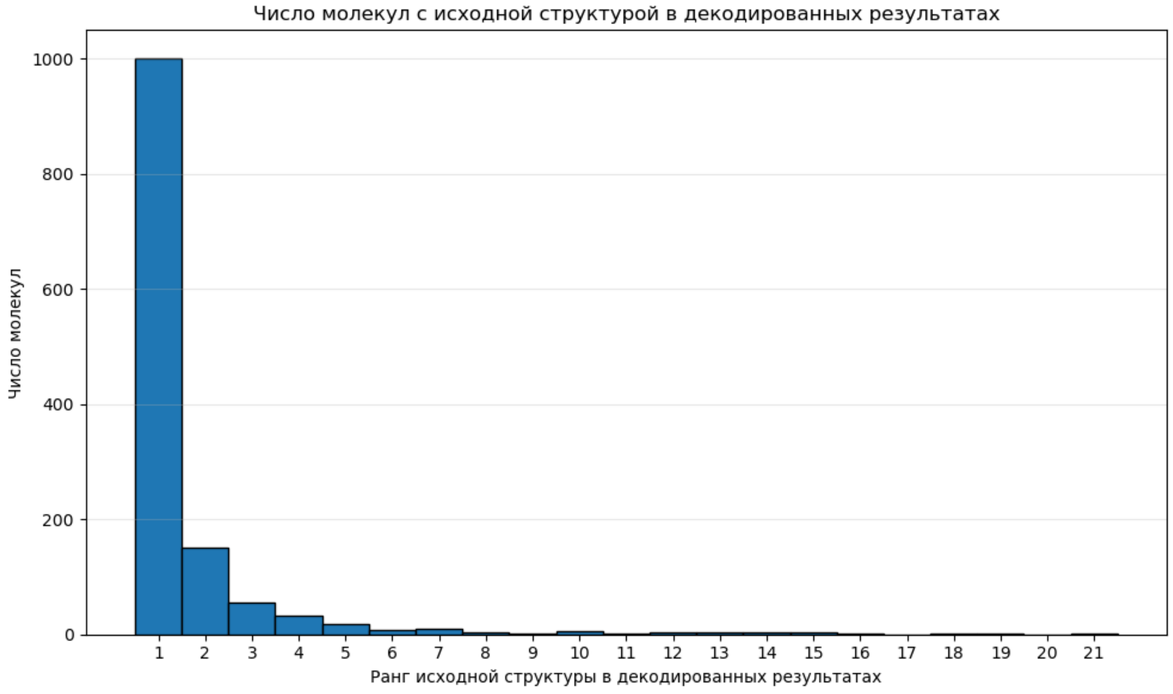
- Исходный датасет: 1363100 молекул
- Все соединения в датасете были разбиты на фрагменты радиуса 2
- Фрагменты молекул, встречаемые менее 300 раз в датасете, были удалены из словаря и корпуса.
- Записи в корпусе, принявшие после удаления фрагментов вид пустого вектора (т.е. все фрагменты в составе молекул оказались «редкими»), были удалены
- Итоговый размер датасета: 1362728 молекул

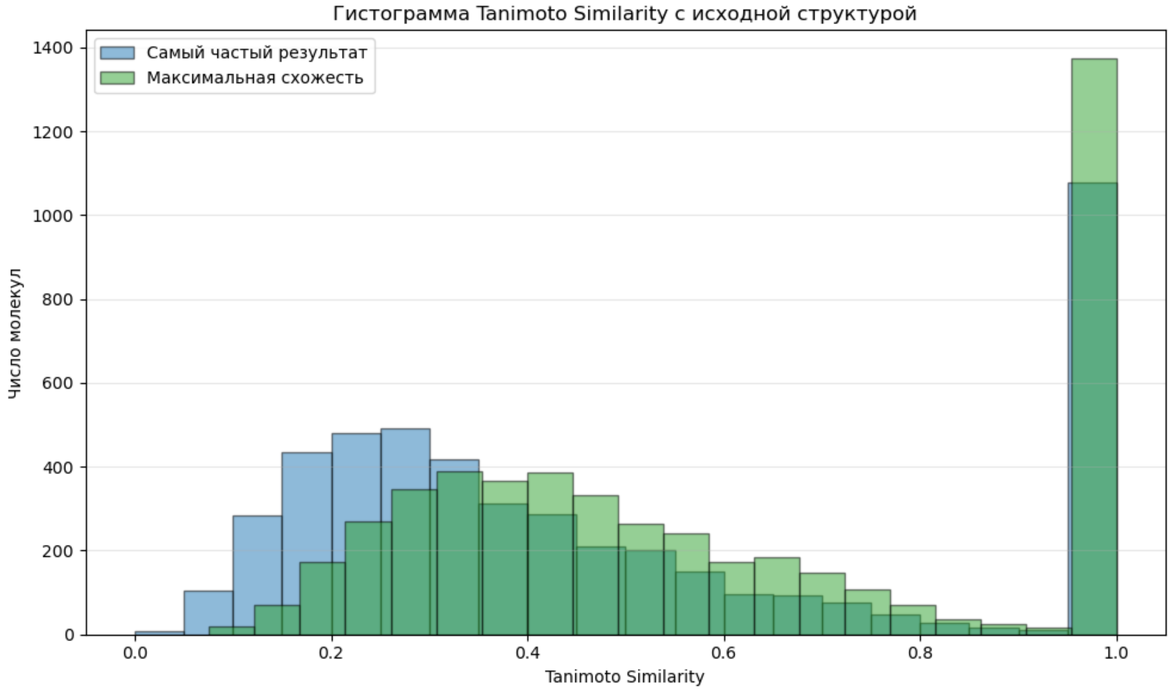
# Результаты работы модели

Уменьшенный валидационный датасет в 5000 структур, 10 генераций фрагментов, 4 генерации структуры

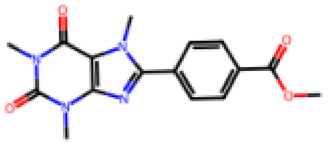
В 26% случаев целевая молекула была в результатах генерации

Почти всегда – в виде самого популярного результата

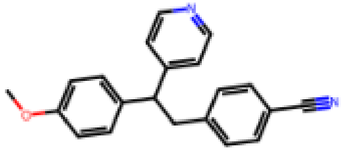


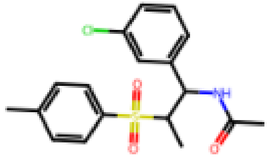
# Примеры корректных предсказаний



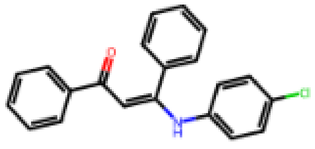
id=754161



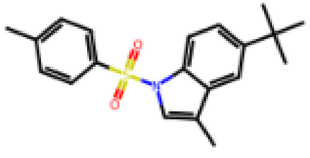
id=1242547



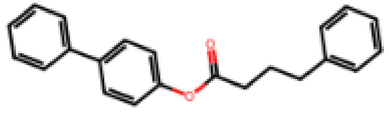
id=1060010



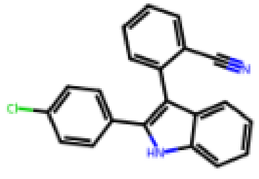
id=86986



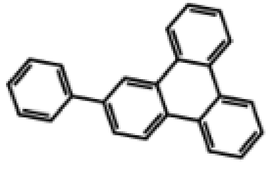
id=222549



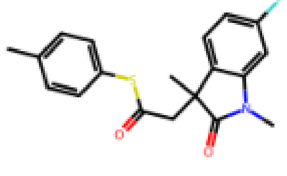
id=1136039



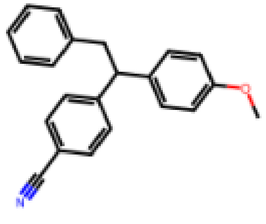
id=422357



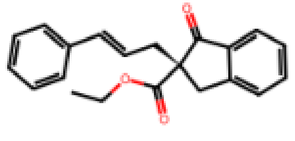
id=165782



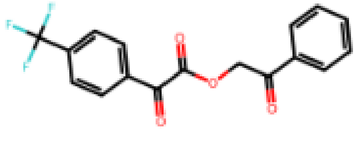
id=1242226



id=1242561

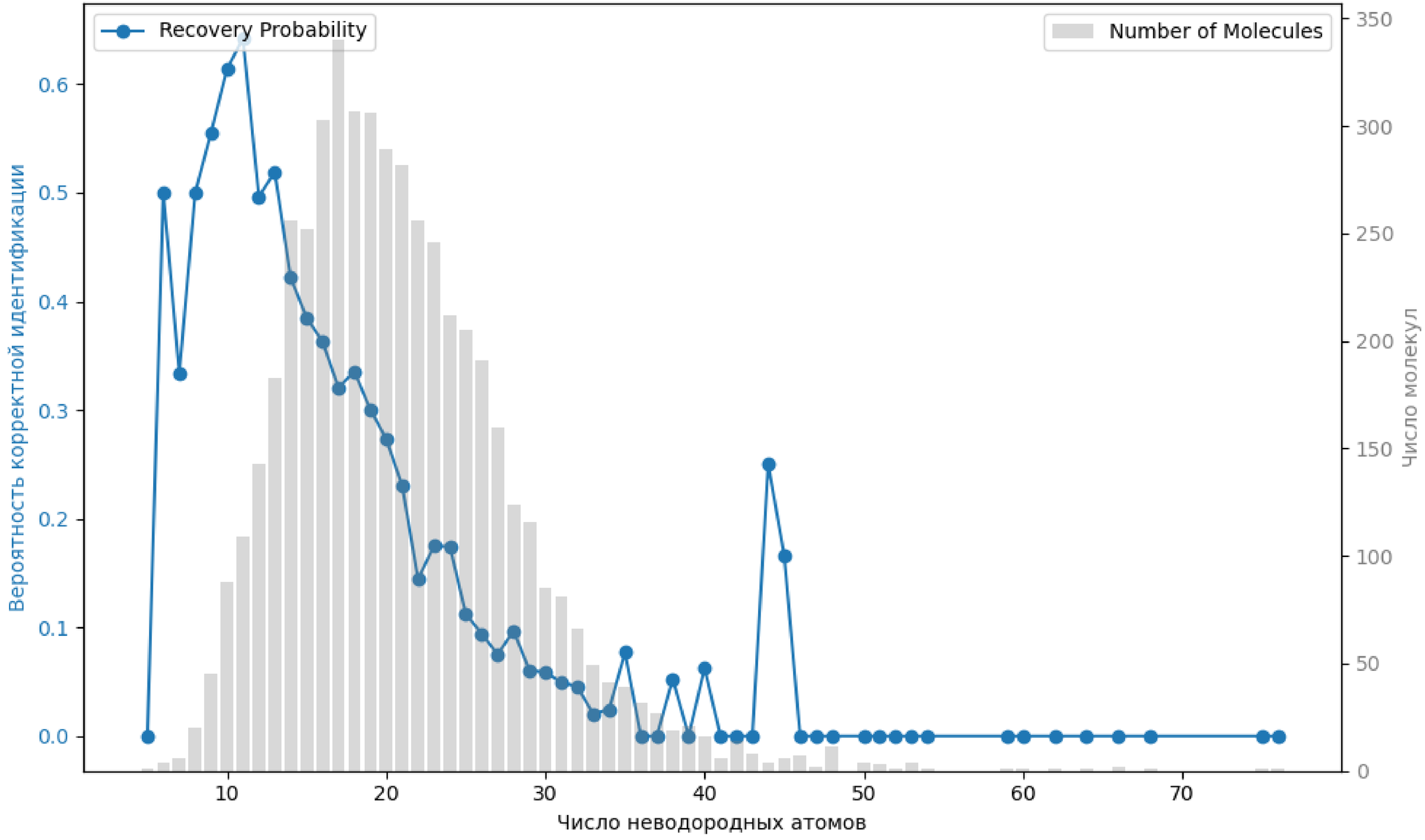


id=14058



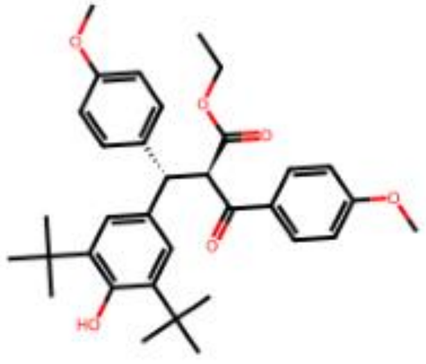
id=141298

# Ограничения на размер молекулы

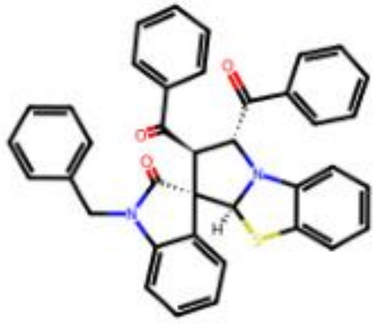


# Примеры корректных предсказаний

Molecules with 34–47 Non-H Atoms (correctly predicted)

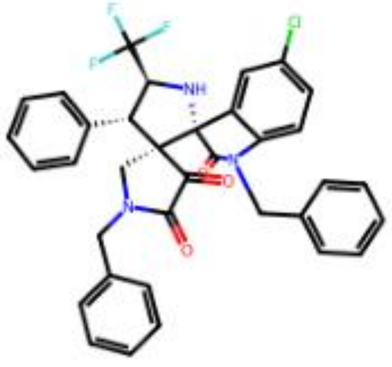


id=790269



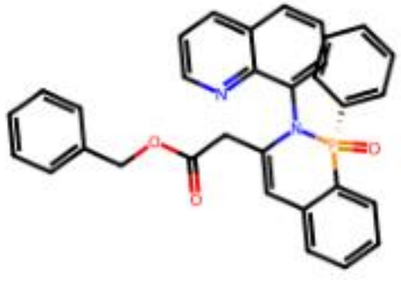
id=950046

id=1100271

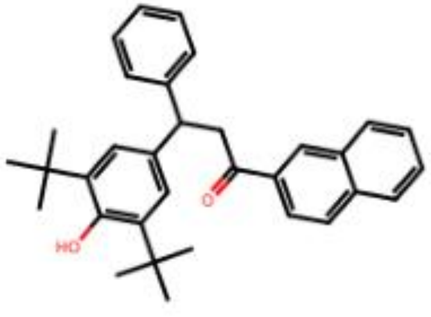


id=892325

id=899746



id=16553



id=858066

id=504543

# Выводы

1. Обучение графовой нейронной сети передачи сообщений на расширенном датасете позволило улучшить точность предсказания химических сдвигов в спектрах $^{13}C$ ЯМР. Основной вклад в ошибку предсказания связан с недостаточной представленностью редких классов соединений, крайне редко изучаемых в лабораторной практике.
2. Использование двухстадийной трансформерной архитектуры, основанной на промежуточной генерации молекулярных фрагментов, продемонстрировал высокую предсказательную способность, позволяя корректно предсказывать строение исходного соединения исключительно на основе химических сдвигов его сигналов в спектрах ЯМР $^{13}C$ в более чем четверти случаев.
3. В случае отсутствия целевой молекулы в результатах генерации, средний коэффициент сходства по Танимото результатов генерации составлял более 0.5, что говорит о том, что предсказаное строение молекулы было очень близко к целевому.

# Дальнейшее развитие

1. Разработка подхода для токенизации молекулярных фрагментов вместо FP2 фингерпринтов.
2. Добавление данных спектроскопии $^{1}H$ ЯМР (как отдельно, там и в комбинации со спектрами $^{13}C$)
3. Внедрение модели в работу сервиса OdanChem.

# Апробация исследования

**Представленные доклады на конференциях:**

- The 9th European Conference on Molecular Magnetism, ECMM2024, устный доклад «Exploring Molecular Magnets with Paramagnetic NMR Spectroscopy», Краков, Польша, 14 – 18 июля 2024
- 3rd Asian Conference on Molecular Magnetism (ACMM III), устный доклад "Paramagnetic NMR Spectroscopy for Molecular Magnets", Пусан, Южная Корея, 1 – 4 сентября 2024

Оба доклада имели отношение к интерпретации спектров ЯМР относительно узкого класса органических соединений, в состав которых входит парамагнитных ион металла, в том числе – с использованием методов машинного обучения.

**Планируемое участие в конференциях в 2024 году:**

- 21st European Magnetic Resonance Congress (EUROMAR 2025), Оулу, Финляндия, 6 – 10 июля 2025 (отправлены тезисы на устный доклад по результатам, полученным при работе над ВКР, информация о принятии тезисов в виде устного доклада или постера – после 17го апреля).
- 9th International Conference on Molecular Magnetism (ICMM), Бордо, Франция, 27 – 31 октября 2025 (планируется подача тезисов на устный доклад, дедлайн подачи до 14го апреля).

**Оценка проекта внешними экспертами:**

Результаты, полученные в ходе работы над ВКР, были использованы при подготовки заявки «Разработка системы автоматической интерпретации результатов ЯМР спектроскопии малых органических молекул» на проект конкурса «Старт-Искусственный интеллект-1» (очередь VIII) Фонда Содействия Инновациям. Было получено финансирование в объеме 4 млн. рублей.

# Список литературы

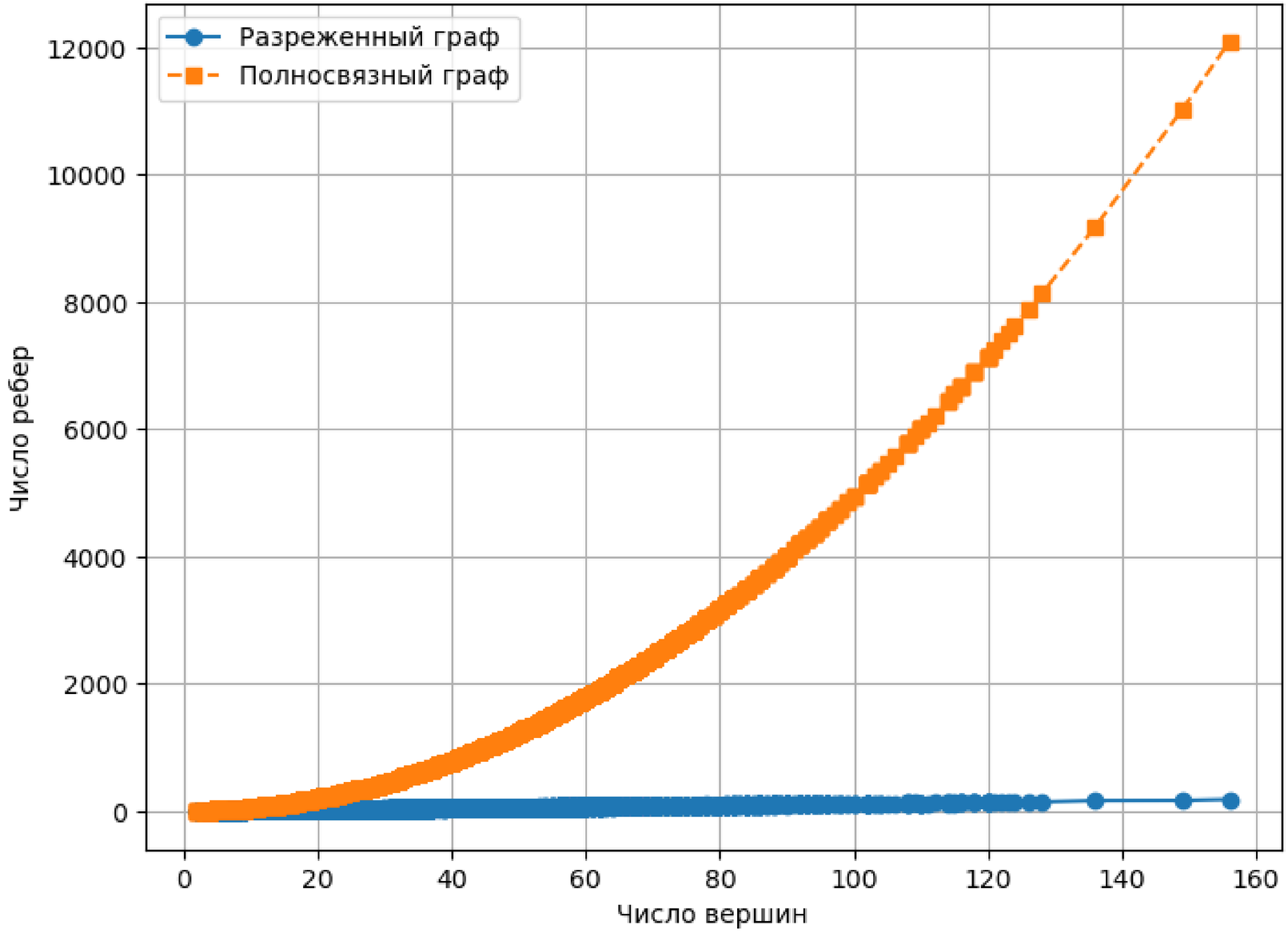
В данный момент список литературы включает 48 источников, в том числе:


1. Y. Guan, S. V. Shree Sowndarya, L. C. Gallegos, P. C. St. John, R. S. Paton, Chem. Sci. 2021, 12, 12012–12026.
2. Z. Yang, M. Chakraborty, A. D. White, Chem. Sci. 2021, 12, 10802–10809.
3. **Y. Kwon, D. Lee, Y.-S. Choi, M. Kang, S. Kang, J. Chem. Inf. Model. 2020, 60, 2024–2030.**
4. B. Sridharan, M. Goel, U. Deva Priyakumar, Chemical Communications 2022, 58, 5316–5331.
5. J. Zhang, K. Terayama, M. Sumita, K. Yoshizoe, K. Ito, J. Kikuchi, K. Tsuda, Science and Technology of Advanced Materials 2020, 21, 552–561.
6. **B. Sridharan, S. Mehta, Y. Pathak, U. D. Priyakumar, J. Phys. Chem. Lett. 2022, 13, 4924–4933.**
7. **S. Devata, B. Sridharan, S. Mehta, Y. Pathak, S. Laghuvarapu, G. Varma, U. Deva Priyakumar, Digital Discovery 2024, 3, 818-829.**
8. 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.
9. **F. Hu, M. S. Chen, G. M. Rotskoff, M. W. Kanan, T. E. Markland, *ACS Cent. Sci.* 2024, *10*, 2162–2170.**

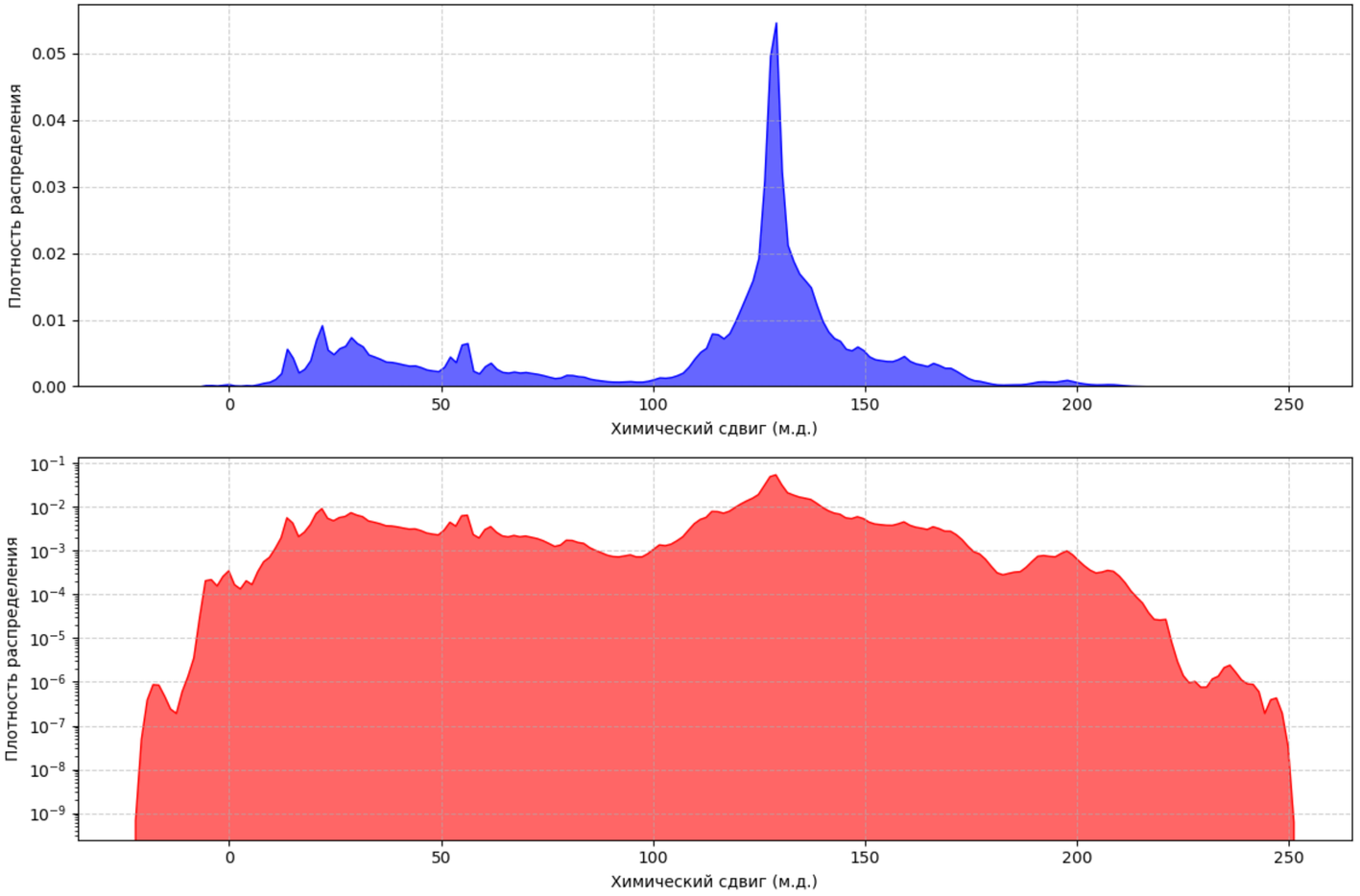
МФТИ

# Спасибо за внимание!

# mol2spec – сложность молекулярного графа



# Химические сдвиги в датасете




Ядерная оценка плотности (KDE) распределения химических сдвигов в датасете: верхний график представлен в линейной шкале, нижний — в логарифмической. Графики отражают частоту встречаемости сигналов с различными значениями химического сдвига.

# Выявленные тренды на основе анализа литературы

Переход от случайного поиска возможных структур, основанного на методе Монте Карло, к подходам, учитывающим «химическую логику» за счет использования практик, используемых для обработки естественного языка (например, BART).

Увеличение количества спектральных данных, используемых для обучения и валидации модели, однако до сих пор большая часть таких данных – синтетическая (получена путем квантовохимических расчетов)

Гипотеза на основе обнаруженных трендов: адаптация известных подходов генеративного искусственного интеллекта к области химии позволит на основе большого массива спектральных данных, приведенных в научных публикациях, построить генеративную модель, определяющую строение ранее не известного химического соединения на основе его спектров ЯМР.

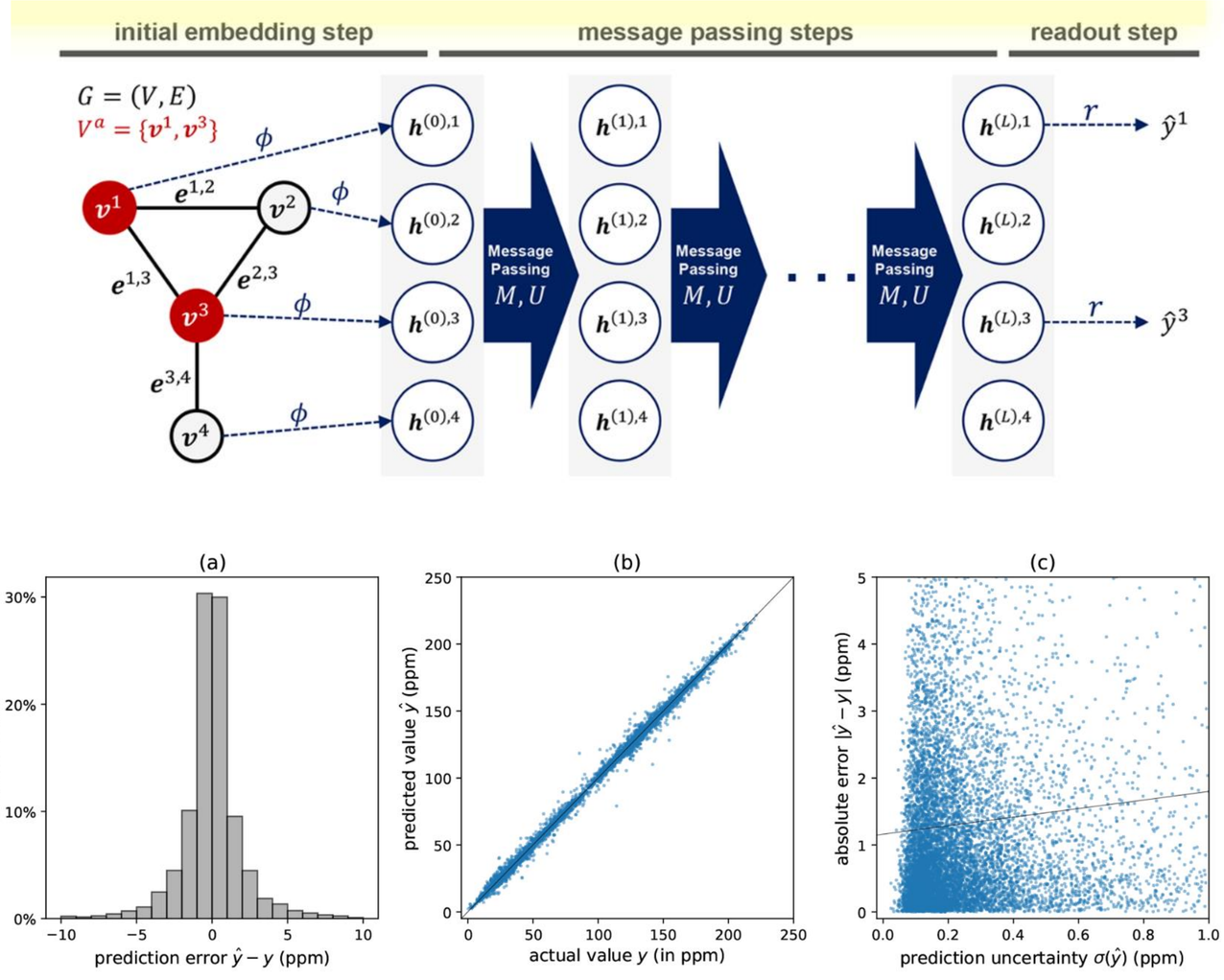
# GNN для Mol2Spec

Kwon, Y., Lee, D., Choi, Y.-S., Kang, M. & Kang, S. Neural Message Passing for NMR Chemical Shift Prediction. J. Chem. Inf. Model. 60, 2024–2030 (2020).

Хороший алгоритм, который на вход принимает строение соединения и на выходе выдает спектры ЯМР. Очень хорошая сходимость для спектров $^{13}C$, чуть худшая – для протонов (ожидаемо). Валидировали по большой базе, работает для широкого класса соединений.

**Плюсы:** готовые веса модели есть на github, можно использовать в нашей модели для валидации данных.

**Минусы:** протонные спектры предсказываются не идеально, нет информации о мультиплетности сигналов и значении констант спин-спинового взаимодействия.



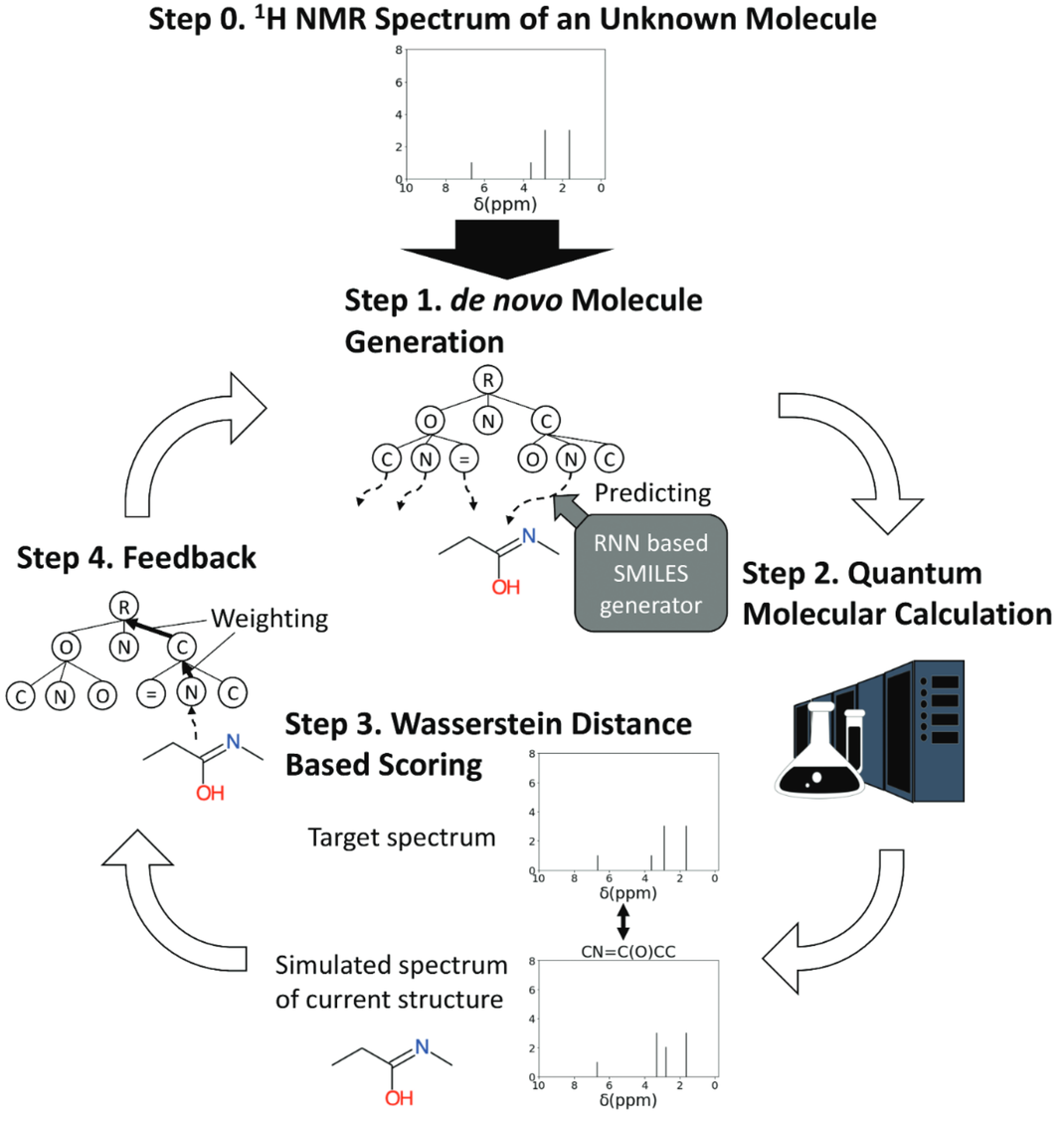
# Генерация с использованием DFT

Zhang, J. *et al.* NMR-TS: de novo molecule identification from NMR spectra. *Science and Technology of Advanced Materials,* **21**, 552–561 (2020).

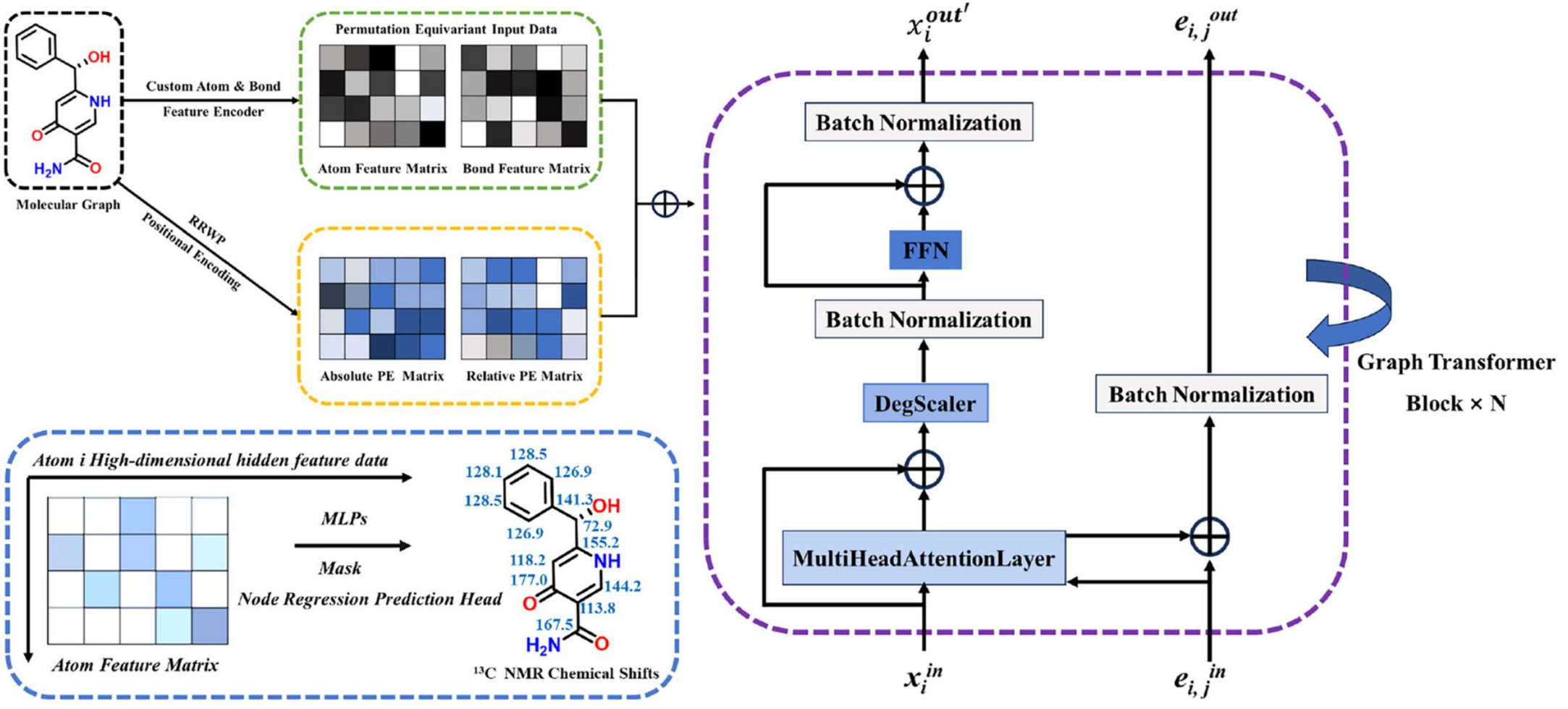
**Краткое описание:** на основе комбинация метода Монте Карло с рекуррентной нейронной сетью генерировали массив химических структур. Дальше использовали DFT для расчета спектров 1H ЯМР, из результатов скоринга корректировали веса генерирующей сети и повторяли цикл до достижения сходимости входного и рассчитанного спектра.

**Плюсы:** предложен рабочий подход для *de novo* генерации структур, работает даже для протонных спектров (правда, тут есть ряд сомнений – не учитываются константы спин-спинового взаимодействия, например, и их зависимость от использованной рабочей частоты спектрометра).

**Минусы:** протестировали только на девяти молекулах, и сработало на шести! Очень долгая процедура генерации, использование квантовохимических расчетов не только в ходе подготовки данных для обучения, но и непосредственно в цикле генерации – невозможно представить использование в реальной жизни.



Трансформеры для mol2spec




Chen, H.; Liang, T.; Tan, K.; Wu, A.; Lu, X. GT-NMR: A Novel Graph Transformer-Based Approach for Accurate Prediction of NMR Chemical Shifts. J Cheminform 2024, 16 (1), 132.

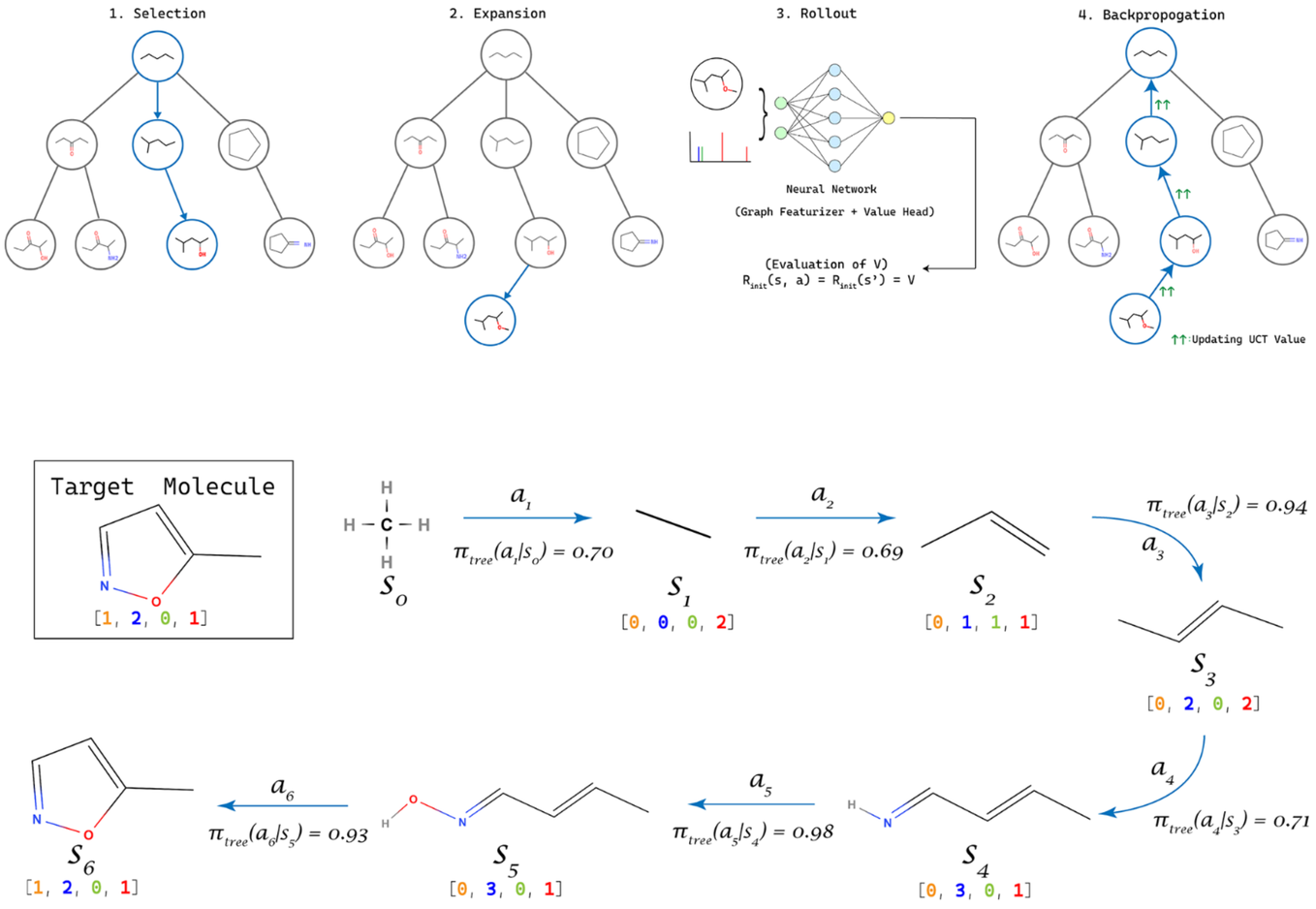
# Поиск Монте-Карло по дереву

Sridharan, B., Mehta, S., Pathak, Y. & Priyakumar, U. D. Deep Reinforcement Learning for Molecula Inverse Problem of Nuclear Magnetic Resonance Spectra to Molecular Structure. *J. Phys. Chem Lett.* **13**, 4924–4933 (2022).

**Краткое описание:** обратную спектральную задачу описали как Марковский процесс принятия решения, использовали комбинацию метода Монте Карло и графовой нейронной сети для итеративной генерации структуры соединения, соответствующего исходным спектрам.

**Плюсы:** неплохо работало на больших молекулах. Тестирование на относительно большой базе данных

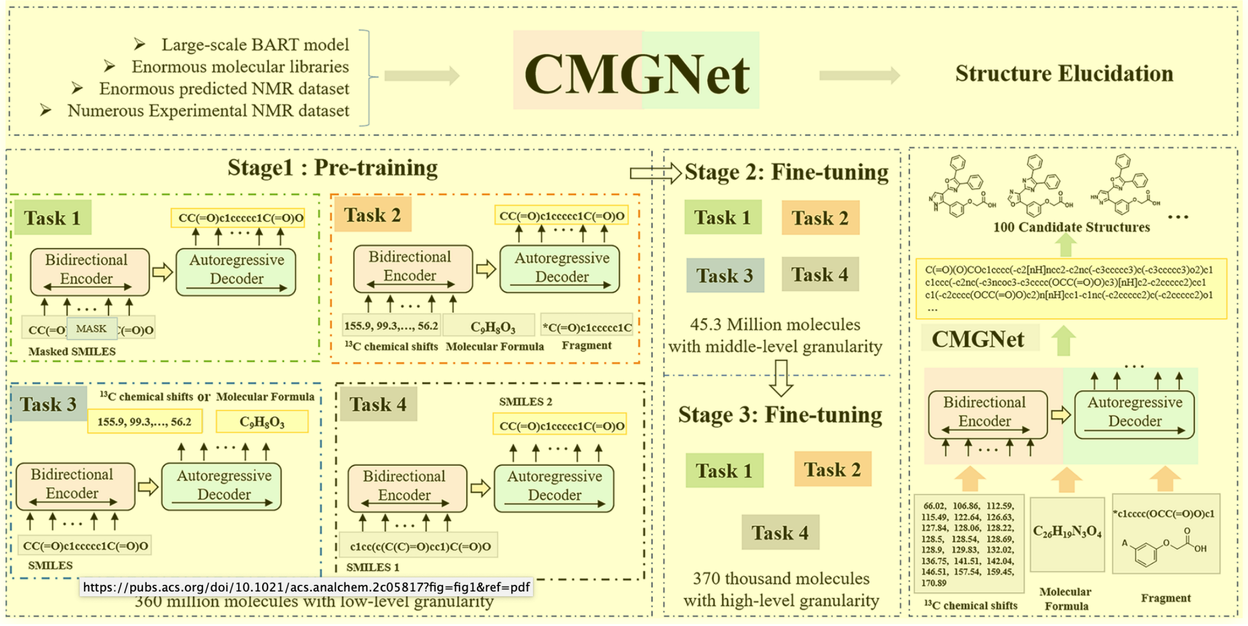
**Минусы:** по-прежнему процесс предсказания занимает время, сопоставимое с временем, которое потребуется специалисту чтобы решить задачу традиционными методами.



# Трансформеры для генерации

Yao, L. *et al.* Conditional Molecular Generation Net Enables Automated Structure Elucidation Based on 13C NMR Spectra and Prior Knowledge. *Anal. Chem.* **95**, 5393–5401 (2023).

**Краткое описание:** на основе базы спектров $^{13}C$ ЯМР была обучена модель CMGNet, позволяющая на основе брутто формулы, химических сдвигов $^{13}C$ 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 DFT), на финальном этапе дообучения использовали небольшую библиотеку экспериментальных спектров на 370 тысяч соединений.




**Плюсы:** предложен рабочий подход для de novo генерации структур

**Минусы:** модель обучали в основном на синтетических данных, **экспериментальных спектров было очень мало**. Слишком много соединений-лидеров, из которых затем нужно будет еще выбирать подходящее. Необходимость знания брутто-формулы и строения фрагментов, без этой информации метрики падают в два раза.