Лекция 5. Машинные модели для расчета свойств электронной структуры молекул.

Курс: Машинное обучение в структурной биологии

Головин А.В. [1]

[1]МГУ им М.В. Ломоносова, Факультет Биоинженерии и Биоинформатики

Москва, 2021

» Химическое разнообразие

* QM методы позволяют точно рассчитывать электронные свойства молекул и материалов.
* Статистическая физика позволяет оценить макроскопические свойства.

Рассчёты очень затратны для сканирования химического пространства, $\approx 10^{23}$

» Корреляции

- Модели строят для поиска в узком диапозоне свойств: аффиность, токсичность, стабильность
- Такие модели применимы в рамках обучающей выборки
- Причина это отсутствие ”физики”при построении модели

» Терминология

- ”... we refer to the combinations of QM and SM approaches with ML as QML models.”
- ”QML refers to the idea of applying modern statistical learning theory to predict electronic and atomistic properties and processes in molecules and materials. ”
- ”.. the goals and reaches of QML models should not be confused with quantum ML algorithms executed on quantum computers.”

» Цели QML

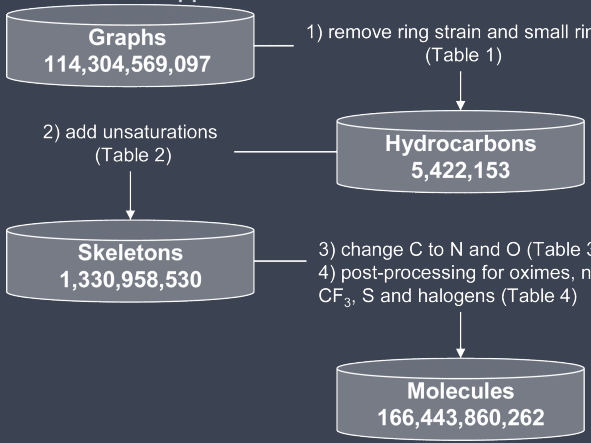
- Создание наборов данных для успешной экстраполяции на всё химическое разнообразие
- Поиск эффективного представления молекул и материалов для ML
- Реализация моделей способных к качественным предсказаниям для всего химического разнообразия

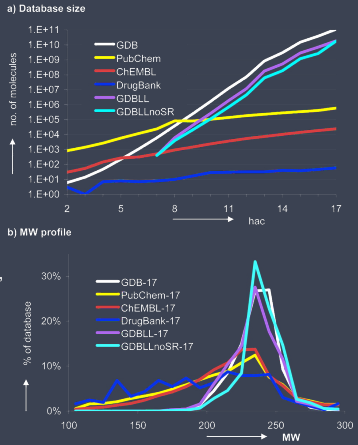
» Набор данных GDB

- Область применения: drug like соединения
- Генерация комбинаций для графов из 11 узлов, генерация стереоизомеров, 13.9 млн. соединений
- Построение 3D структуры и фильтрация

» GDB-17

166 млн. соединений



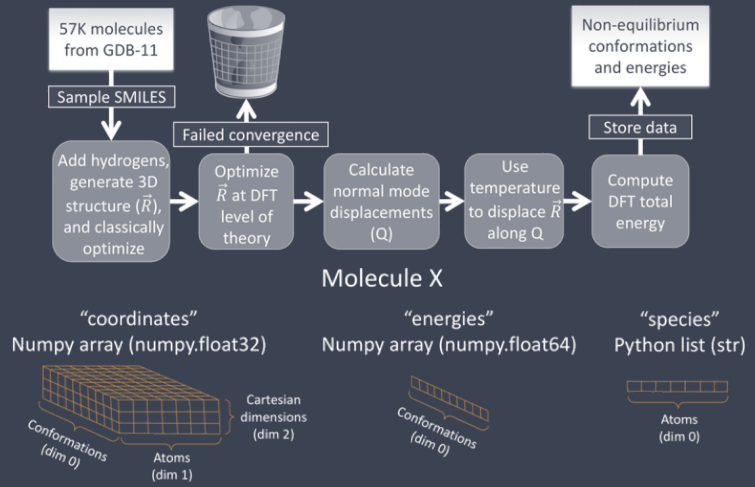


10.1021/ci300415d

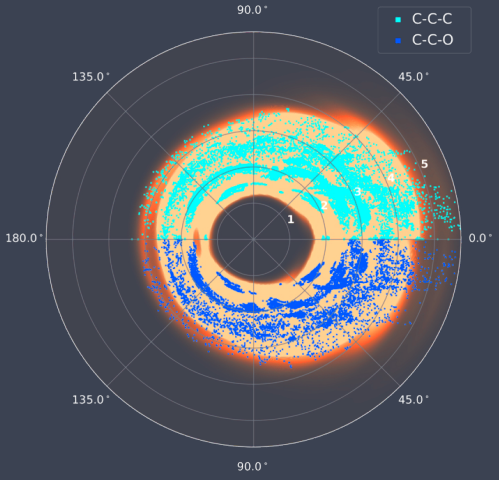
» QM7, QM9, MD17 и ANI1

- QM7, 7K: Для выборки соединений были рассчитаны эллектронные свойства: HOMO, LUMO, E, E_i
- QM9,134K сединений: равновесные геометрии, дипольные моменты, поляризуемость, термохимия и т.д.
- MD17: 10 соединений и 50K-1M конформаций для каждого из AIMD при 500K
- ANI1, 57K соединений и 20M конформаций

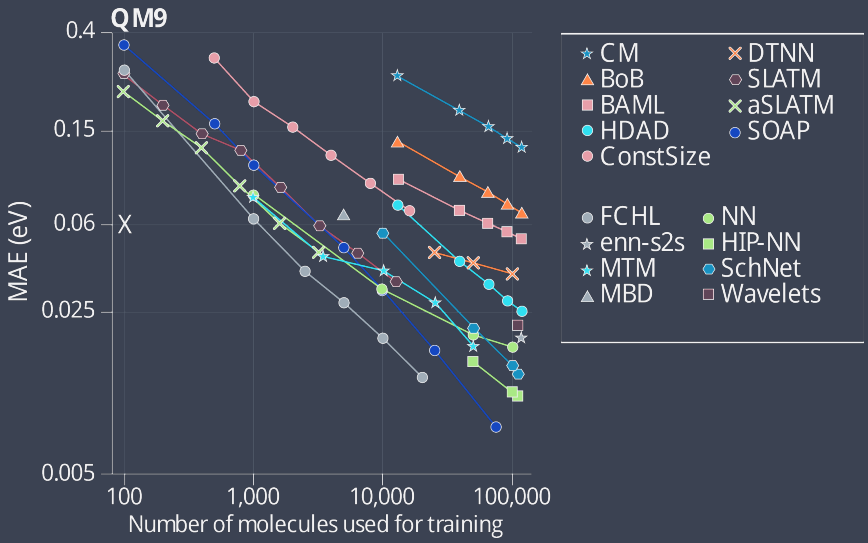
» ANI-1


57K molecules from GDB-11
Sample SMILES
Add hydrogens, generate 3D structure ($\vec{R}$), and classically optimize
Optimize $\vec{R}$ at DFT level of theory
Failed convergence
Calculate normal mode displacements (Q)
Use temperature to displace $\vec{R}$ along Q
Compute DFT total energy
Store data
Non-equilibrium conformations and energies
Molecule X
"coordinates"
Numpy array (numpy.float32)
Conformations (dim 0)
Atoms (dim 1)
Cartesian dimensions (dim 2)
"energies"
Numpy array (numpy.float64)
Conformations (dim 0)
"species"
Python list (str)
Atoms (dim 0)

» ANI-1, сравнение с равновесными наборами



» Размер набора данных



» Представление молекул

- не существует «универсального» представления, которая удовлетворяет всем желаемым свойствам одновременно
- Предложен ряд представлений, каждое из которых удовлетворяет только часть общих требований.
- Все существующие модели в основном основаны на позициях атомов и заряде ядра
- kernel machines и нейронные сети используют для старта молекулярное представление, нейронные сети также могут генерировать представление как часть процесса обучения.

» Атомцентрированные функции симетрии

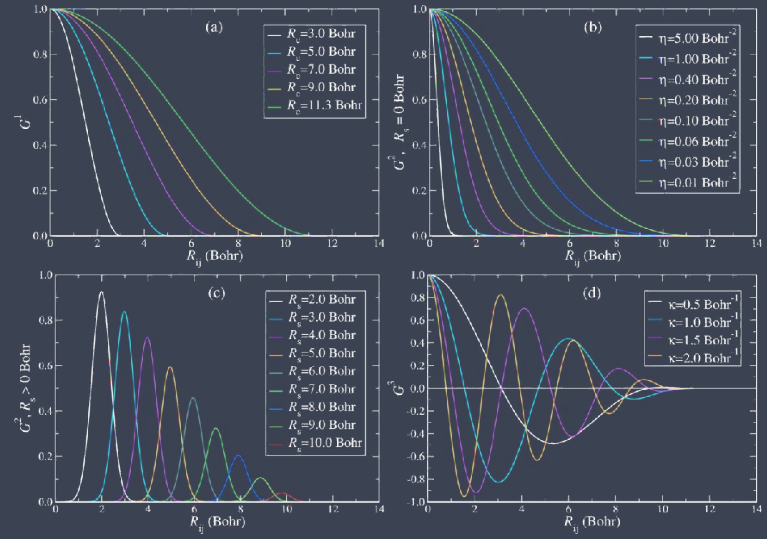
- Произведения функций радиальной и угловой симметрии используется для представления молекулярной среды.
- Преимущество - компактное и физически осмысленное представление
- Ограничение - что сложность быстро растет с увеличением количества типов атомов.

$$G_i^1 = \sum_i f f_c(Rij)$$

$$G_i^2 = \sum_i e^{n(R_{ij}-R)s)^2} \cdot f_c(R_{ij})$$

$$G_i^3 = \sum_i cos(kR_{ij} \cdot f_c(R_ij)$$

» Атомцентрированные функции симетрии



» Кулоновская матрица

- Кулоновская матрица: матрица обратных расстояний, которая представляет межъядерное кулоновское отталкивание между атомами.
- Очень эффективное, глобальное и элегантное представление, но не удовлетворяет симетрии перестановки для эквивалентных атомов и может привести к разрывам функции.

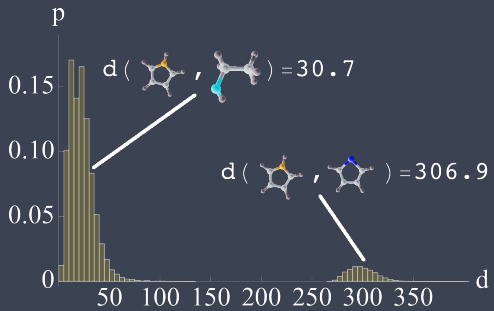
$$\hat{V}_C = \frac{1}{2} \sum_{I \neq J} \frac{Z_I Z_J}{|\mathbf{R}_I - \mathbf{R}_J|}$$

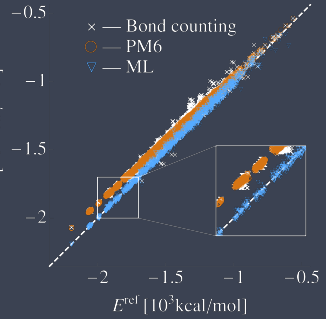
10.1126/sciadv.1603015

» Кулоновская матрица

$$d(\mathbf{M}, \mathbf{M}') = d(\epsilon, \epsilon') = \sqrt{\sum_I |\epsilon_I - \epsilon'_I|^2}$$

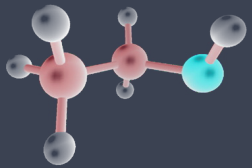
, где ϵ собственные значения **M**





» Bag of bonds

* Мешок связей : в этом случае представление кулоновской матрицы векторизуется, улучшая его меру сходства и эффективность.
* Симметрия перестановок и разрывы по-прежнему представляют собой проблему.



	O	C	C	H	H	H	H	H	H
O	O	OC	OC	OH	OH	OH	OH	OH	OH
C	OC	C	CC	CH	CH	CH	CH	CH	CH
C	OC	CC	C	CH	CH	CH	CH	CH	CH
H	OH	CH	CH	H	HH	HH	HH	HH	HH
H	OH	CH	CH	HH	H	HH	HH	HH	HH
H	OH	CH	CH	HH	HH	H	HH	HH	HH
H	OH	CH	CH	HH	HH	HH	H	HH	HH
H	OH	CH	CH	HH	HH	HH	HH	H	HH
H	OH	CH	CH	HH	HH	HH	HH	HH	H

O
C
C
H
H
H
H
H
H

OC
OC
OH
OH
OH
OH
OH
OH
CC

CH
CH
CH
CH
CH
CH
CH
CH
CH
CH

HH
HH
HH
HH
HH
HH
HH
HH
HH
HH
HH
HH
HH
HH
HH

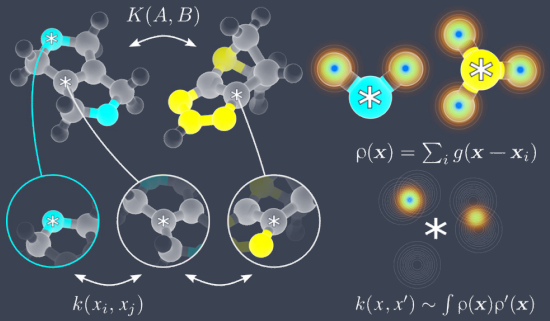
0
O-bag
0
C-bag
0
H-bag
0
OC-bag
0
OH-bag
0
CC-bag
0
CH-bag
0
HH-bag
0

» Многотельное представление

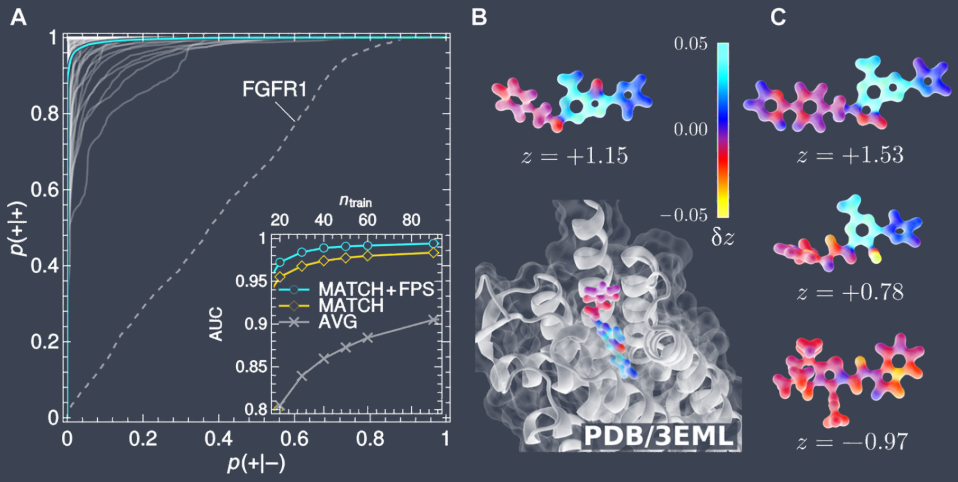
- Многотельная репрезентация: локализованное расширение мешка связей с явной обработкой различных шкал расстояний, включая термины с тремя и больше телами.
- Решает симетрию перестановок , однако, может стать неэффективным при использовании более чем трехтельных членов.

» Smooth overlap of atomic positions

- плавное перекрытие позиций атомов: в этом случае молекулы представлены в виде суперпозиции радиальных функций Гаусса и углового момента.
- Строго относится к вращению и перестановочные симметрии и избегает разрывов.
- Может стать очень дорогим из-за точного предсказания и для нескольких типов атомов.

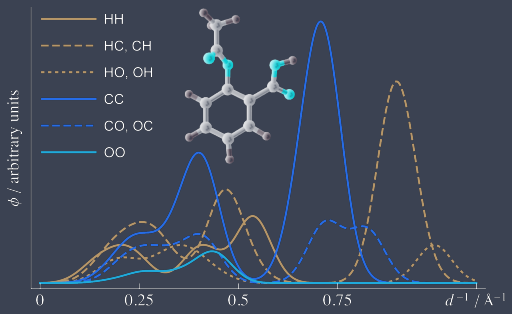


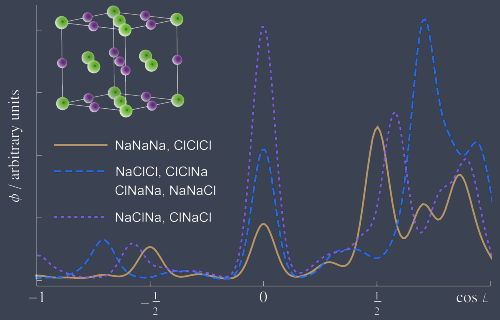
» Smooth overlap of atomic positions



» Многотельное тензорное представление

* Многотельное тензорное представление: тензорное представление для молекул и твердых тел.
* Обычно применим ко многим различным системам. Для больших систем оценка может стать дорогостоящей.





10.1126/sciadv.1701816

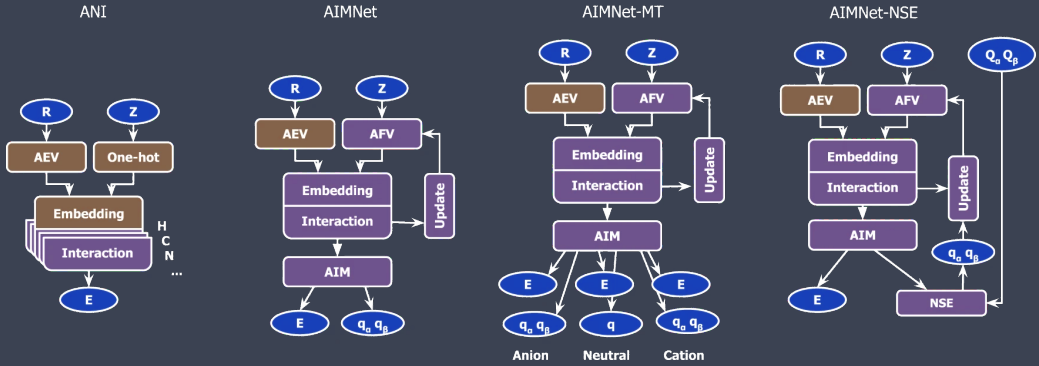
» wavelet scattering transform

- Вейвлет преобразование: многомасштабное представление молекулярных свойств на основе вейвлетов (волнообразные колебания с амплитудой, уменьшающейся до нуля вдали от начала координат).
- Естественно отражает многомасштабный характер молекулярных свойств, не навязывая локализацию этих свойств. Необходимы знания для построения соответствующего вейвлет-базиса

arxiv.org/pdf/1805.00571.pdf

- Еще продолжается дискуссия о преимуществах и ограничениях разных представлений для разных приложений.
- Модели машинного обучения на основе kernels требуют явной формулировки для представления, используя одну шкалу на ядро
- Глубокие нейронные сети, такие как глубокая тензорная нейронная сеть (DTNN), могут привести к неявному многомасштабному представлению через процесс обучения.

» Архитектуры NN



» Размер выборки, для моделей в FF

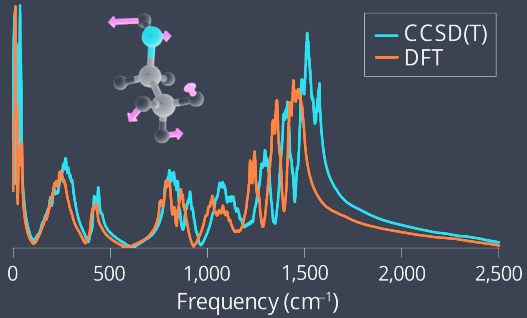
Для нескольких сотен молекулярных конформаций из траектории МД, sGDML (symmetric gradient domain machine learning) модель может давать глобальные силовые поля для небольших молекул (≤25 атомов), при точности сопоставимой с достижимой на уровне теории CCSD(T) .

» Маштабируемость

Была продемонстрирована применимость модели в больших системах после обучения на небольших системах. На основе AEV было показано, что можно получить многообещающие результаты для значений энергий, сил, сдвигов ЯМР и другие QM свойств для различных систем различного размера, до сотен тяжелых атомов

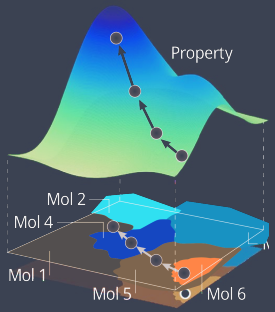
» Новые insights-1

Впервые удалось вычислить точные термодинамические и спектроскопические свойства для молекул размером с аспирин без ущерба для точности расчета атомных сил и с временной шкалой, доступной в ММ моделировании.



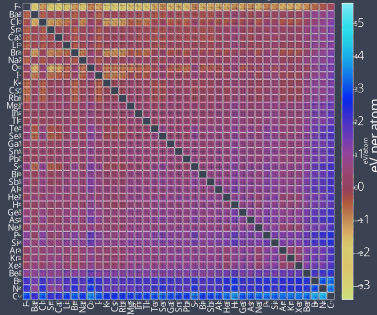
» Новые insights-2

Оптимизация новых соединений была реализована через использование генеративных моделей, в сжатом варианте CCS с уменьшенной размерностью, в результате предсказаны кандидаты на светопоглощающие материалы.



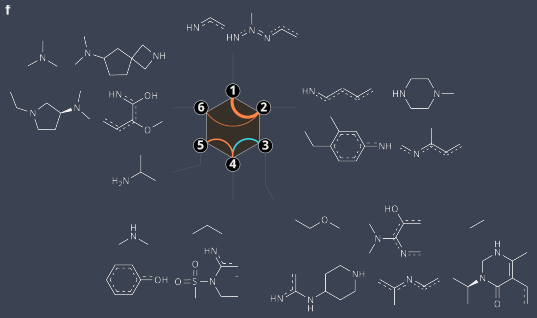
» Новые insights-3

QML, очевидно, также применим к материалам. Рассчет энергии образования 2 млн кристаллов эльпасолита привел к идентификации почти 100 потенциально стабильных новых кристаллов
подробный анализ степеней окисления привел к открытию экзотического кристалла, NFAl 2 Ca 6, в котором Al несет отрицательную степень окисления.



» Новые insights-4

Ли и сотрудники применили ML к экспериментальным данным в чтобы понять лиганд - белковое связывание.
Теория случайных матриц была использована для идентификации химические группы и особенностей, которые сильно влияют на связывание и те, которые нет.



» Multiscale QML models

Конечная цель - реализовать глобальные и универсальное исследование химического разнообразия, при сочетании с QM, SM и ML.

» Towards molecular design with QML

- Больше наблюдаемых значений. Использование QML для оценки наблюдаемых статистической механикой згачений, например, для предсказания энергетических профилей редких событий
- Экспериментальный дезайн. QML формирует основу для программного обеспечения, которое может работать как вспомогательное решение в экспериментальном дизайне
- Дизайн реакции с использованием QML. Компьютер основанное планирование реакций и их поиск имеет давнюю историю в области химии, начиная с 1960-х годов