Всероссийский конкурс исследовательских и проектных работ
школьников “Высший пилотаж”

Создание генеративно-состязательной нейронной сети для генерации металло-органических каркасов со свойствами адсорбции.

Исследовательская работа
Направление “Computer science”

Автор: Грузинцев Егор Алексеевич,
учащийся 11 класса,
МБОУ лицей “Технический” г. Самара

2023 г.

Оглавление

Введение

Машинное обучение, нейронные сети и искусственный интеллект уже внесли огромный вклад в современное общество. Только за последний год возник целый рынок вакансий в области промт-инжиниринга, профессии, связанной с составлением запросов к искусственному интеллекту. Midjourney [1], Dall-E [2] и Stable Diffusion [3] стали горячей темой для обсуждения в среде художников, а ChatGPT от OpenAI [4] произвёл революцию в, кажется, всех областях, связанных с работой с текстом, начиная от разработки и заканчивая поисковыми запросами.

Сложно представить себе дальнейшее развитие цивилизации без использования этих технологий повсеместно, от кухни до космоса. И, естественно, нам бы хотелось видеть, как новообретенные вычислительные и интеллектуальные мощности идут на пользу не только отдельным людям, но и человечеству в целом, решая глобальные задачи вроде энергетического кризиса и глобальных климатических изменений.

Одним из ведущих факторов, с которым приходится справляться нашему поколению, является так называемый углеродный след, который является следствием функционирования цивилизации. Его влияние на климат и биологическое состояние планеты активно изучают экологи, химики и физики, однако нет сомнений в том, что регуляция, улавливание и хранение различных газов становится важной темой для обсуждения в перспективе ближайших лет. Более того, эта тема обсуждается в научных кругах уже около десятка лет.

За этот срок был сделан значительный прогресс в предсказании поведения газов при взаимодействии с различными материалами, предсказаны и синтезированы сотни и тысячи различных микропористых соединений, было найдено химическое решение проблемы, которым стали металлоорганические каркасы (МОК) [5]. Но вместе с этим возникла и новая задача. Не всякий МОК подходит для улавливания газов, и определение подходящих становится большой нагрузкой на исследователей.

Именно в этот момент мы предлагаем подключить технологии нейросетей для предсказания и генерации структур уже подходящих под хранение газов.

Таким образом, цель нашего проекта:

- Разработать новый способ для генерации МОК со свойствами адсорбции.

В процессе достижения этой цели мы сформулировали следующие задачи:

- Найти общедоступные кристаллографические базы данных, содержащие в себе информацию об металлоорганических каркасах.
- Определить метод представления полученных данных.
- Разработать архитектуру генеративной нейронной сети.
- Обработать данные и собрать в пригодный для нейронной сети датасет.
- Обучить генеративную нейронную сеть.
- Визуализировать полученные результаты.

Основная часть

Металлоорганические каркасы

Водород является перспективным энергоносителем, который можно использовать в топливных элементах для выработки электроэнергии с высокой эффективностью и нулевыми выбросами парниковых газов. Однако хранение и транспортировка водорода может быть затруднена из-за его низкой плотности и низкой температуры кипения. Одним из потенциальных решений является хранение водорода в металлоорганических каркасах (МОК), которые представляют собой пористые материалы, состоящие из ионов или кластеров металлов, соединенных органическими лигандами. МОК имеют высокую площадь поверхности, регулируемый размер пор и высокую химическую стабильность, что делает их подходящими для различных применений по хранению и разделению газа. В последние годы МОК изучаются в качестве среды для хранения водорода и улавливания CO2, предлагая потенциал для одновременного решения двух важных энергетических и экологических проблем.
За последнее время использование искусственного интеллекта (ИИ) и машинного обучения набирает обороты в различных областях науки, включая материаловедение и химическую инженерию. Эти методы применяются для оптимизации дизайна и синтеза МОК, прогнозирования их свойств и характеристик, а также для понимания их поведения на молекулярном уровне. Используя возможности искусственного интеллекта и машинного обучения, исследователи могут ускорить открытие и оптимизацию МОК для хранения водорода и адсорбции CO2, а также потенциально обнаружить новые идеи и тенденции, которые было бы трудно выявить только на основе экспериментальных или теоретических исследований.

Прогнозирование адсорбционных свойств металлоорганических каркасов (МОК) может быть сложной задачей из-за сложной структуры и поведения этих материалов. Некоторые из трудностей, которые могут возникнуть при прогнозировании адсорбционных свойств МОК, включают:

1. Структурная сложность: МОК состоят из ионов металлов или кластеров, соединенных органическими лигандами, которые могут быть расположены различными способами, образуя различные структуры и топологии. Конкретное расположение металлических и органических компонентов в МОК может сильно влиять на ее адсорбционные свойства, поэтому трудно предсказать поведение МОК, основываясь только на ее химическом составе.

2. Размер и форма пор: МОК имеют высокую площадь поверхности и регулируемый размер пор благодаря пористой природе их структуры. Размер и форма пор в МОК могут влиять на адсорбцию газов и других молекул, и могут варьироваться в зависимости от типа металла и органических лигандов, использованных для синтеза МОК.
3. Химическая стабильность: МОК обычно стабильны в широком диапазоне условий, но они могут быть чувствительны к определенным факторам, таким как влажность, температура и давление, которые могут влиять на их адсорбционные свойства.
4. Динамическое поведение: МОК могут претерпевать структурные изменения, такие как расширение или сжатие, при адсорбции или десорбции газов, что может повлиять на их адсорбционные свойства. Такое динамическое поведение может быть трудно предсказать, основываясь только на статической структурной информации.
5. Ограниченность экспериментальных данных: синтез и характеризация МОК - сложный и трудоемкий процесс, который может ограничить количество экспериментальных данных, доступных для моделирования и прогнозирования.

Несмотря на эти проблемы, исследователи активно работают над созданием вычислительных методов и моделей, которые могут точно предсказать адсорбционные свойства МОК и помочь направить их разработку и синтез для конкретных применений.

МОК, или металлоорганические каркасы, - это класс материалов, которые имеют широкий спектр потенциальных применений, включая хранение газов, катализ и доставку лекарств. Однако одной из основных проблем при изучении МОК является доступность, качество и количество данных об этих материалах.

Одной из проблем является ограниченная доступность экспериментальных данных по МОК. Эти материалы относительно новые и не так хорошо изучены. Означает, что имеется ограниченное количество данных о структуре, свойствах и эффективности МОК. Это может затруднить исследователям полное понимание потенциала этих материалов и выявление новых и улучшенных конструкций.

Другой проблемой является качество имеющихся данных. МОК - это сложные материалы, которые могут иметь широкий спектр свойств, в зависимости от их состава и структуры. Однако имеющиеся данные об этих материалах часто неполны или противоречивы, что может затруднить исследователям точное прогнозирование их свойств и характеристик.

Наконец, проблемой является и количество доступных данных. Количество МОК, которые были синтезированы и охарактеризованы, все еще относительно невелико, что затрудняет обучение моделей для прогнозирования свойств и поведения новых МОК. Кроме того, синтез и экспериментальное определение характеристик МОК требует больших затрат и времени, что приводит к нехватке данных.

Эти проблемы с доступностью, качеством и количеством данных затрудняют исследователям полное понимание потенциала МОК и разработку новых и улучшенных конструкций. Чтобы преодолеть эти проблемы, исследователи обращаются к вычислительным методам, таким как генеративные модели машинного обучения, которые помогают генерировать новые конструкции МОК и предсказывать их свойства на основе ограниченных доступных данных.

Генеративные модели машинного обучения уже используются различными компаниями в области химии, медицины, энергохранилищ и экологии. Эти компании используют технологию для улучшения своих исследований и разработок, а также для улучшения своих продуктов и услуг.

В области химии такие компании, как Atomwise [6], используют генеративные модели машинного обучения, чтобы помочь в открытии новых молекул для использования в лекарствах и других химических продуктах. Платформа ИИ компании может анализировать огромные объемы данных о существующих молекулах и предсказывать, какие из них имеют наибольший потенциал для использования в новых лекарствах.

В области медицины такие компании, как Insilico Medicine [7], используют генеративные модели машинного обучения для разработки новых лекарств и методов лечения. Компания использует генеративные модели на основе искусственного интеллекта (ИИ) для анализа больших объемов биологических данных и прогнозирования того, какие соединения с наибольшей вероятностью будут эффективны при лечении того или иного заболевания.

В области хранения энергии компании используют генеративные модели машинного обучения для повышения эффективности и производительности своих продуктов. Одним из примеров является компания Form Energy [8], которая использует генеративные модели на основе ИИ для разработки и оптимизации передовых аккумуляторных систем для использования в системах хранения энергии в масштабах энергосистемы. Платформа ИИ компании может анализировать большие объемы данных о существующих аккумуляторных системах и предсказывать, какие

конструкции будут наиболее эффективны в различных сценариях хранения энергии. Это позволяет им разрабатывать более эффективные и экономичные решения для хранения энергии, которые могут помочь в интеграции возобновляемых источников энергии в энергосистему. Другие компании, такие как FermionX [9], используют аналогичные подходы, применяя ИИ для оптимизации конструкции литий-серных батарей, которые потенциально могут стать более устойчивой и эффективной альтернативой традиционным литий-ионным батареям.

В целом, использование генеративных моделей машинного обучения становится все более распространенным в самых разных отраслях, причем компании используют эту технологию для улучшения своих исследований и разработок, а также для совершенствования своих продуктов и услуг. Ожидается, что в ближайшие годы эта тенденция сохранится, поскольку все больше компаний обращаются к генеративным моделям на базе ИИ, чтобы получить конкурентное преимущество.

Существует два основных подхода к использованию ИИ для поиска новых молекул: генерация новых молекул с заданными свойствами и поиск молекул с желаемыми свойствами среди уже предсказанных.

Первый подход - генерирование новых молекул с заданными свойствами - предполагает обучение генеративных моделей машинного обучения для создания новых молекул, обладающих определенными характеристиками, например, высокой степенью связывания с конкретным целевым белком или высокими показателями адсорбции. Сеть обучается на наборе данных существующих молекул и способна генерировать новые молекулы, похожие на те, что есть в наборе данных, но ранее не встречавшиеся. Этот подход часто используется, например, в процессе поиска лекарств, где целью является поиск новых молекул, которые могут связываться с целевым белком и потенциально лечить заболевание.

Второй подход - поиск молекул с желаемыми свойствами среди уже предсказанных - предполагает обучение модели машинного обучения для выявления молекул из набора данных, обладающих определенными желаемыми свойствами. Например, модель машинного обучения может быть обучена определять молекулы, которые обладают высокой стабильностью и низкой токсичностью. Этот подход часто используется в процессе открытия материалов и поиска лекарств, где целью является поиск новых молекул, обладающих определенными свойствами, полезными для конкретного применения.

Оба подхода имеют свои преимущества и недостатки. Первый подход, генерирующий новые молекулы, позволяет обнаружить большее разнообразие соединений, но поиск молекулы, отвечающей всем желаемым критериям, может занять больше времени. Второй подход - выявление молекул с желаемыми свойствами - быстрее и эффективнее, но он ограничен разнообразием набора данных.

Объединив возможности ИИ и ГНС (генеративные нейронные сети) , мы сможем генерировать основанные на данных идеи и модели, которые можно использовать для оптимизации синтеза кристаллических материалов и открытия новых материалов с желаемыми свойствами.

Методы репрезентации данных

Данные о составе и структуре МОК обычно хранятся в различных форматах кристаллографических файлов и системах, в зависимости от типа данных и программного обеспечения, используемого для их сбора и анализа.

Одним из часто используемых форматов является формат CIF (Crystallographic Information File) [10], который является стандартным форматом для хранения кристаллографических данных. Файлы CIF содержат информацию о кристаллической структуре, включая положения атомов, длины и углы связей, операции симметрии и другие структурные детали. Они также содержат информацию о составе МОК, включая типы и количество присутствующих атомов. Файлы CIF могут быть прочитаны и обработаны широким спектром кристаллографического программного обеспечения, что делает их широко распространенным форматом для обмена данными.

```
data_functionalizedCrystal
_audit_creation_method   'MofGen! by Chris Wilmer'
_symmetry_space_group_name_H-M 'P1'
_symmetry_Int_Tables_number    1
_symmetry_cell_setting   triclinic
loop_
_symmetry_equiv_pos_as_xyz
  x,y,z
_cell_length_a     12.759393
_cell_length_b     12.759401
_cell_length_c     12.759399
_cell_angle_alpha  89.983359
_cell_angle_beta   89.967969
_cell_angle_gamma  90.019837
loop_
_atom_site_label
_atom_site_type_symbol
_atom_site_fract_x
_atom_site_fract_y
```

```
_atom_site_fract_z
Zn1   Zn    -0.594017   -1.606252   -0.418161
Zn2   Zn    -0.418336   -1.606145   -0.593880
Zn3   Zn    -0.418352   -1.430554   -0.418220
```

Таблица 1. Пример содержимого CIF-файла

Другим широко используемым форматом является формат PDB (Protein Data Bank) [11], который представляет собой стандартный формат для хранения макромолекулярных структур. Файлы PDB содержат информацию о трехмерной структуре белков, нуклеиновых кислот и других биомолекул. Файлы PDB также могут использоваться для хранения информации о структуре МОК, включая положения атомов, длины и углы связей и другие структурные детали. Файлы PDB можно читать и обрабатывать с помощью широкого спектра программного обеспечения, включая инструменты визуализации и анализа.

```
HEADER COMPLEX (ANTIBODY-ANTIGEN)   27-AUG-90   1FDL           1FDL  2
COMPND IG*G1 FAB FRAGMENT (ANTI-LYSOZYME ANTIBODY D1.3, KAPPA) 1FDL  3
COMPND 2 - LYSOZYME (E.C.3.2.1.17) COMPLEX                     1FDL  4
SOURCE MOUSE (MUS $MUSCULUS) FROM BALB/$C STRAIN AND           1FDL  5
SOURCE 2 HEN (GALLUS $GALLUS) EGG WHITE                        1FDL  6
AUTHOR T.O.FISCHMANN,R.J.POLJAK                                1FDL  7
REVDAT 1   15-OCT-91 1FDL    0                                 1FDL  8
JRNL  AUTH   T.O.FISCHMANN,G.A.BENTLEY,T.N.BHAT,G.BOULOT,      1FDL  9
JRNL  AUTH 2 R.A.MARIUZZA,S.E.V.PHILLIPS,D.TELLO,R.J.POLJAK    1FDL  10
JRNL  TITL   CRYSTALLOGRAPHIC REFINEMENT OF THE                1FDL  11
JRNL  TITL 2 THREE-DIMENSIONAL STRUCTURE OF THE                1FDL  12
JRNL  TITL 3 FAB*D1.3-*LYSOZYME COMPLEX AT 2.5-*ANGSTROMS      1FDL  13
JRNL  TITL 4 RESOLUTION                                        1FDL  14
JRNL  REF    J.BIOL.CHEM. V. 266 12915 1991                    1FDL  15
JRNL  REFN   ASTM JBCHA3  US ISSN 0021-9258 071                1FDL  16
```

```
ATOM 1658  OE2 GLU L 213 -2.454  63.348 -21.174  1.00 60.81      1FDL1750
ATOM 1659  N   CYS L 214  2.669  65.274 -17.328  1.00 67.92      1FDL1751
ATOM 1660  CA  CYS L 214  3.648  65.325 -16.253  1.00 76.32      1FDL1752
ATOM 1661  C   CYS L 214  3.046  65.575 -14.891  1.00 82.45      1FDL1753
ATOM 1662  O   CYS L 214  1.879  65.984 -14.818  1.00 84.00      1FDL1754
ATOM 1663  CB  CYS L 214  4.667  66.394 -16.551  1.00 73.55      1FDL1755
ATOM 1664  SG  CYS L 214  3.737  67.825 -17.102  1.00 74.12      1FDL1756
ATOM 1665  OXT CYS L 214  3.745  65.292 -13.910  1.00 87.87      1FDL1757
TER  1666      CYS L 214                                        1FDL1758
```

Таблица 2. Выдержки из PDB-файла

Существует также множество различных программных пакетов и баз данных, доступных для хранения, управления и анализа данных МОК. Некоторые из наиболее популярных программных пакетов включают Кембриджскую структурную базу данных (CSD) [12], база данных NIST [13] и проект Materials Project [14]. Эти программные пакеты и базы данных предоставляют широкий спектр инструментов для поиска, визуализации и анализа данных МОК, облегчая исследователям доступ к этим данным и работу с ними.

В дополнение к структурным данным, многие из этих программных пакетов и баз данных также включают информацию об адсорбционных свойствах МОК. Адсорбционные данные относятся к способности МОК адсорбировать или поглощать определенные газы или жидкости, что является важным свойством для многих приложений, таких как хранение, разделение и очистка газа. Эти данные обычно включают такую информацию, как тип адсорбированного газа или жидкости, количество адсорбированного газа и условия, при которых происходила адсорбция.

Например, Кембриджская структурная база данных (CSD) и база данных неорганических кристаллических структур (ICSD) включают информацию об адсорбционных свойствах МОК. CSD включает информацию об адсорбции таких газов, как водород, метан и углекислый газ, а ICSD - информацию об адсорбции различных газов и жидкостей.

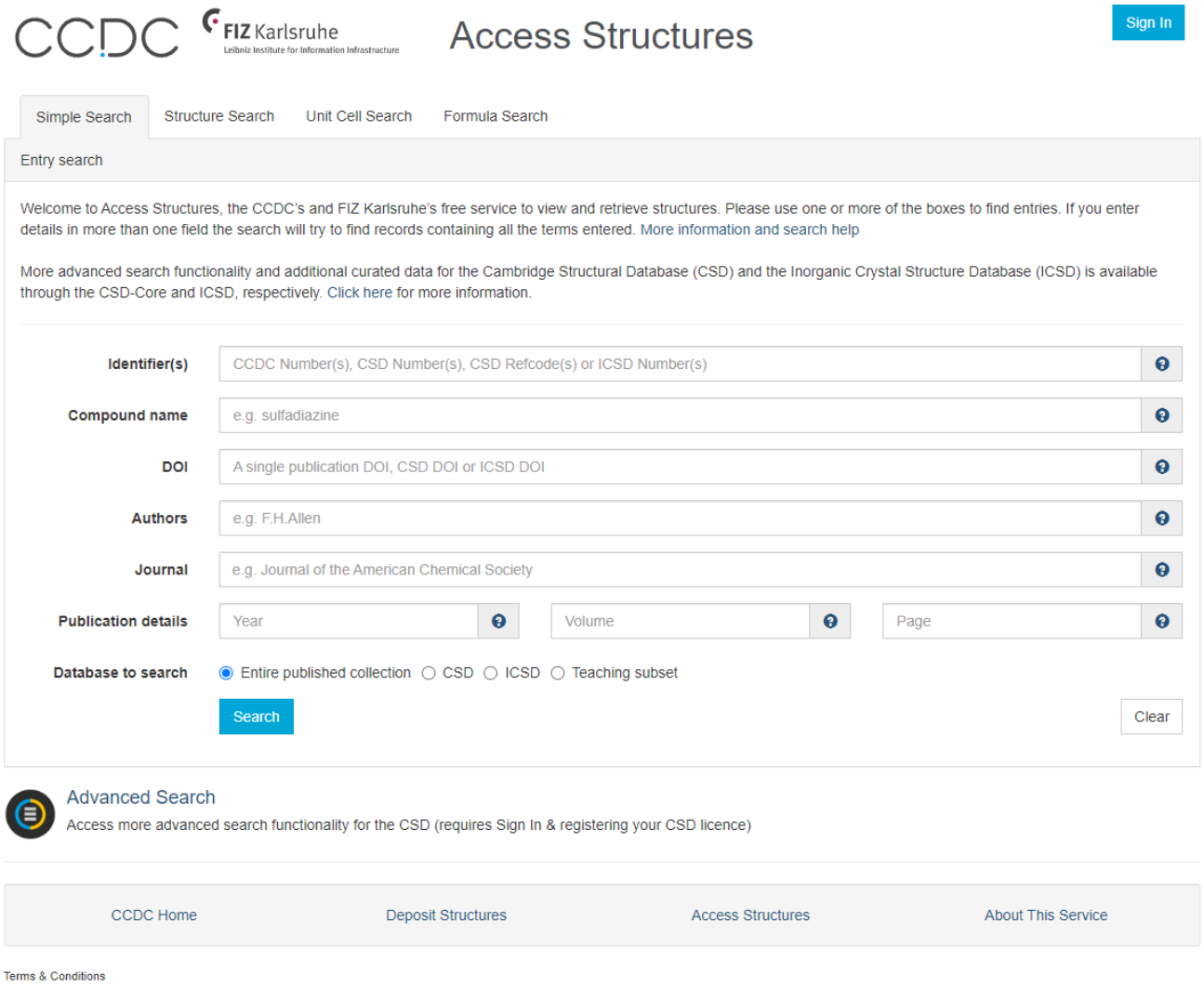


Рисунок 1. Кембриджская структурная база данных

Проект Materials Project также содержит информацию об адсорбционных свойствах МОК. У них есть база данных рассчитанных изотерм адсорбции различных газов, таких как CO2, H2, CH4, N2 и других, на МОК, которые можно искать и фильтровать на основе различных параметров.

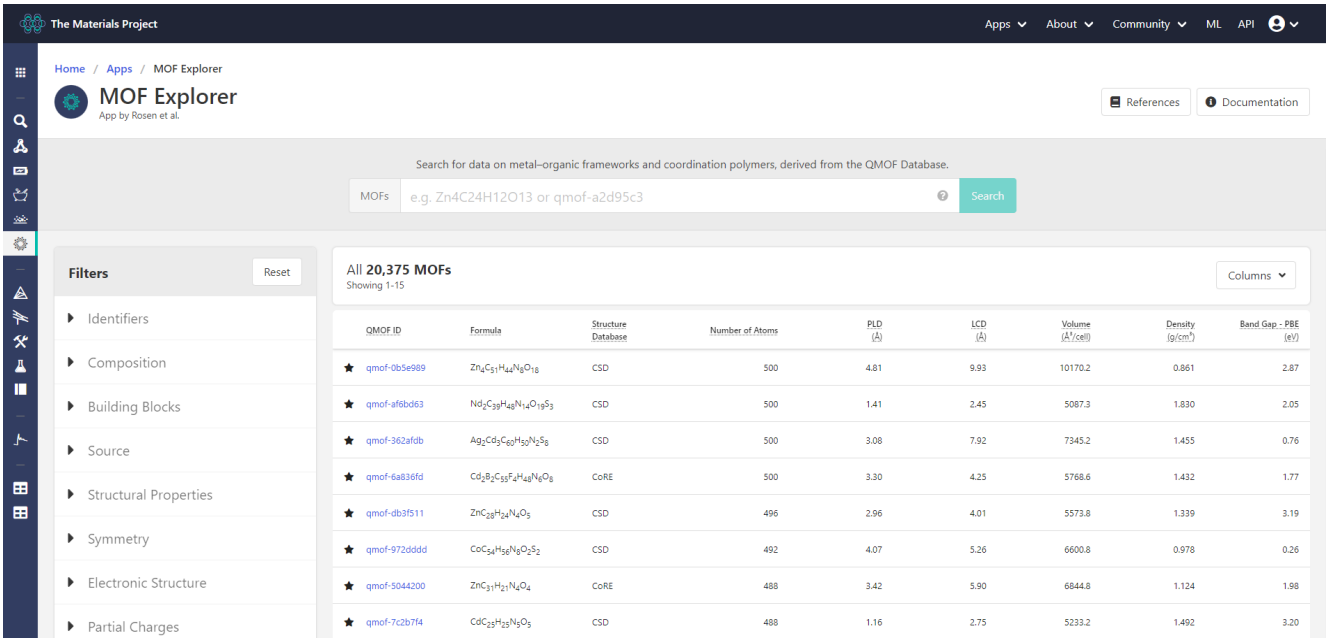


Рисунок 2. База данных МОК materialsproject

Еще одним важным источником данных о МОК является база данных NIST (Национального института стандартов и технологий). База данных NIST представляет собой всеобъемлющую коллекцию экспериментальных данных по широкому спектру материалов, включая МОК. База данных NIST содержит огромное количество информации о физических и химических свойствах МОК, включая структурные данные, термические данные и данные по адсорбции.

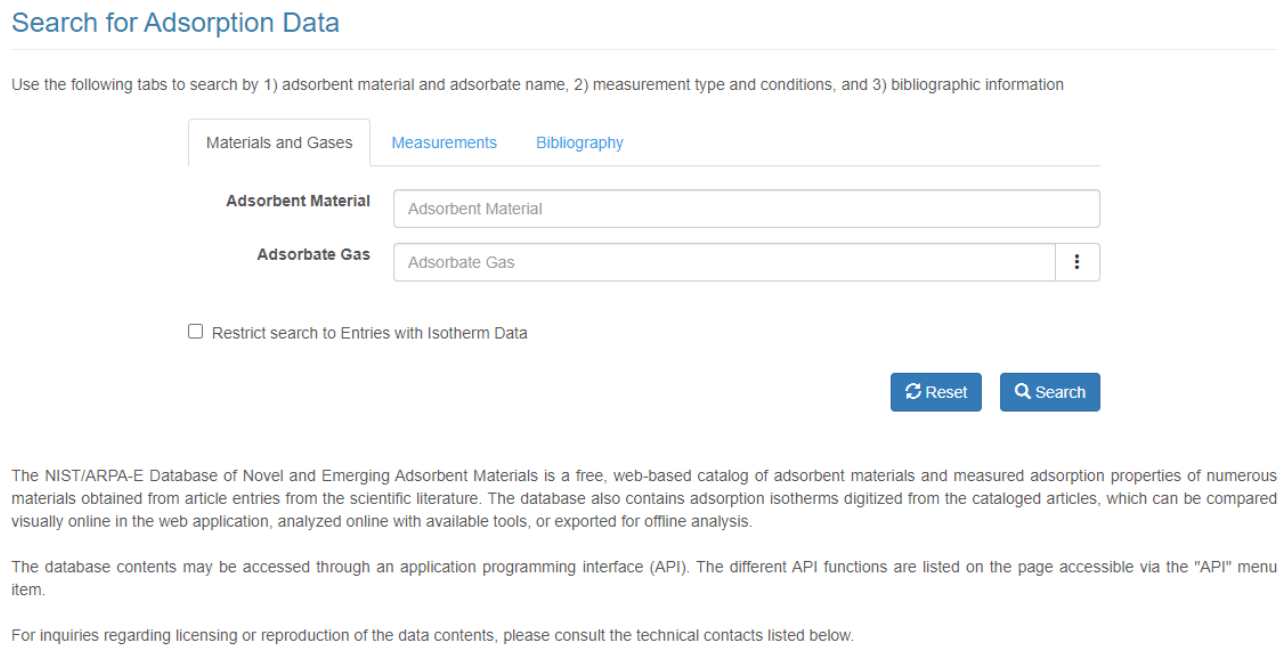


Рисунок 3. База данных NIST

В дополнение к базам данных и программным пакетам, упомянутым ранее, существует также база данных, специально ориентированная на МОК, под названием MOFXDB [15] Северо-Западного университета.

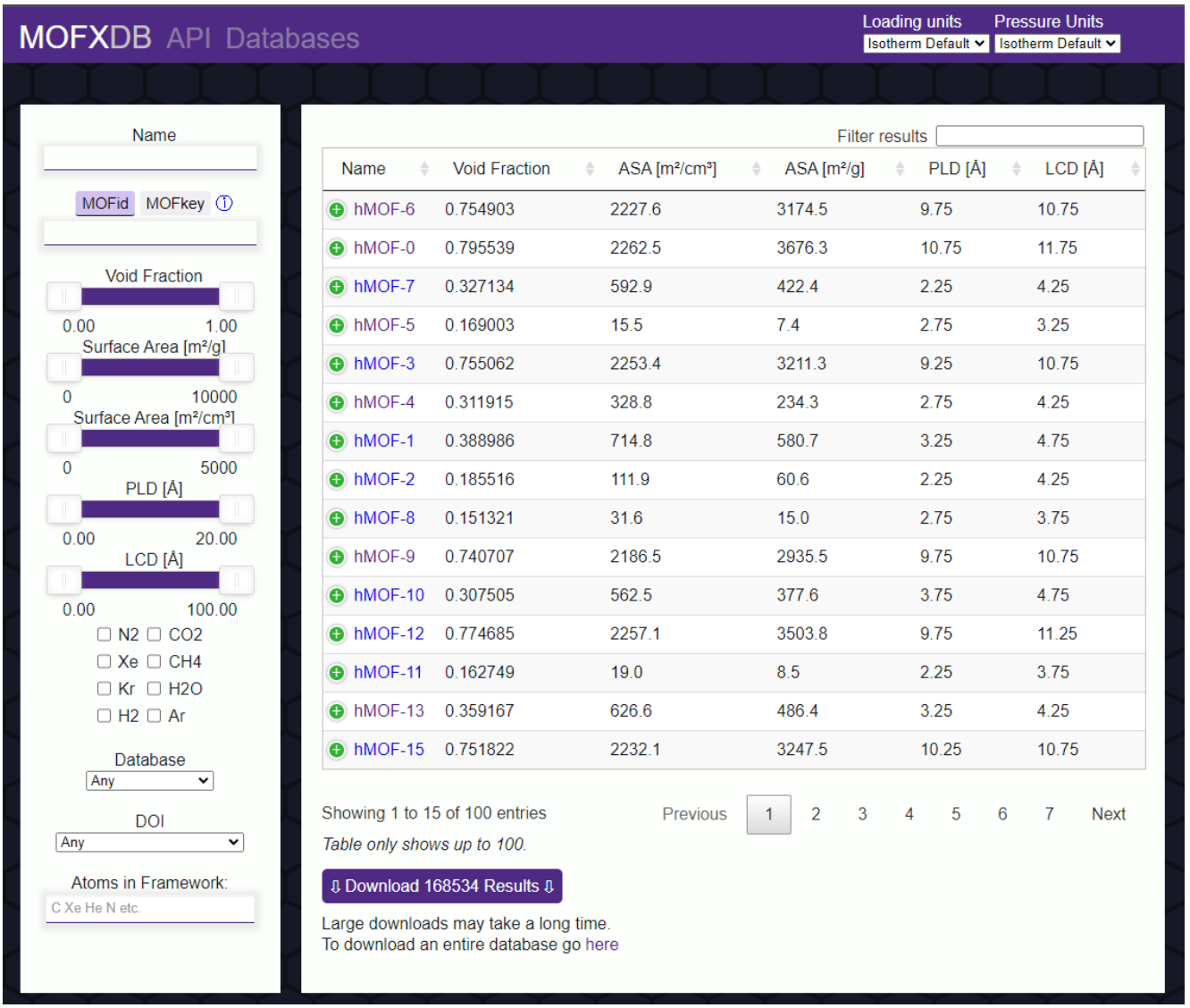


Рисунок 4. База данных Северо-Западного Университета

Каждая из баз данных имела собственный API, позволяющий получить CIF файл для конкретного МОК. Помимо CIF файла при обработке json запроса мы получаем информацию об адсорбции. В результате объединения всех доступных баз данных нам удалось создать датасет, содержащий в себе 42000 значений.

В этот момент возникает первое препятствие для написания нейросети. Дело в том, что водород, углекислый газ и прочие субстанции запасаются на поверхности МОК, занимая энергетически выгодные позиции около лигандов (органические соединители между атомами \ кластерами

металла) и металлических кластеров. Однако CIF-файл содержит только информацию об атомах, составляющих сам кристалл и ничего не говорит о пространстве вокруг. Это, а также тот факт, что позиции атомов, их количество и углы связей сложно использовать как материал для обучения нейросети, толкнуло нас на необходимость создания пространственного представления химических соединений.

Пространственное представление соединений часто используется для предсказания поведения соединений при наличии сторонних взаимодействий. Молекулярная механика описывает возможные варианты и пороговые значения реакций соединения и разложения. В задачи, решаемые этим подходом входит, например, способ сжигания топлива в самолетах и ракетоносителях с образованием минимального количества сажи.

Программы расчета свойств веществ из первопринципов (*ab initio*) моделируют пространство внутри кристалла для предсказания стабильности соединений и поиска подходящих материалов для использования в качестве составного элемента литий-ионных батарей, а также для поиска альтернативных способов хранения электроэнергии, без использования лития. К таким программам можно отнести VASP [16], Wien2k [17], CRYSTAL [18] и другие.

Особняком стоит упомянуть программу USPEX (УСПЕХ) [19]. USPEX - это вычислительная программа для предсказания кристаллической структуры, разработанная Артемом Р. Огановым и его исследовательской группой в Университете Стоуни Брук. Она использует эволюционный алгоритм, расчеты ab initio и ограничения симметрии для предсказания структуры материалов, включая МОК. Она была широко принята исследователями в этой области и используется для предсказания структур новых МОК и других материалов.

Однако прямое использование этих подходов сопряжено со значительным количеством трудностей, главная из которых – огромное время обработки данных. Оценка свойств даже одного соединения может занять не просто часы, а дни, недели, а в некоторых случаях и месяцы.

В нашем случае выходом стало использование молекулярно-динамического описания соединений при помощи функции Леннарда-Джонса 6-12 [20]. Функция Леннарда-Джонса это двухчастичная функция распределения энергии между частицами, зависящая от расстояния между частицами и типов атомов.

$$E = 4\epsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right] \quad \text{(I)}$$

Функция Леннарда-Джонса, в силу своей простоты, часто используется для моделирования адсорбции молекул водорода внутри различных структур, от МОК и цеолитов до углеродных нанотрубок.

Взаимодействие между парой химических элементов определяется двумя параметрами, ε и σ, которые можно получить из проверенных открытых источников [21]. В итоге для пространственной репрезентации химического соединения мы разбили пространство ячейки на 2^18 ячеек, для каждой из которых был рассчитан суммарный уровень энергии.

Таким образом мы получили трёхмерный энергетический ландшафт соединения, используя стандартные подходы метода молекулярной динамики и химического моделирования адсорбции. Поскольку ε и σ имеют, соответственно, размерности энергии и расстояния, мы получили распределения энергетически выгодных вакантных позиций в соединении, а также области, занимаемые атомами. Это распределение позволяет нам и воссоздать вещество по его энергетическому ландшафту.

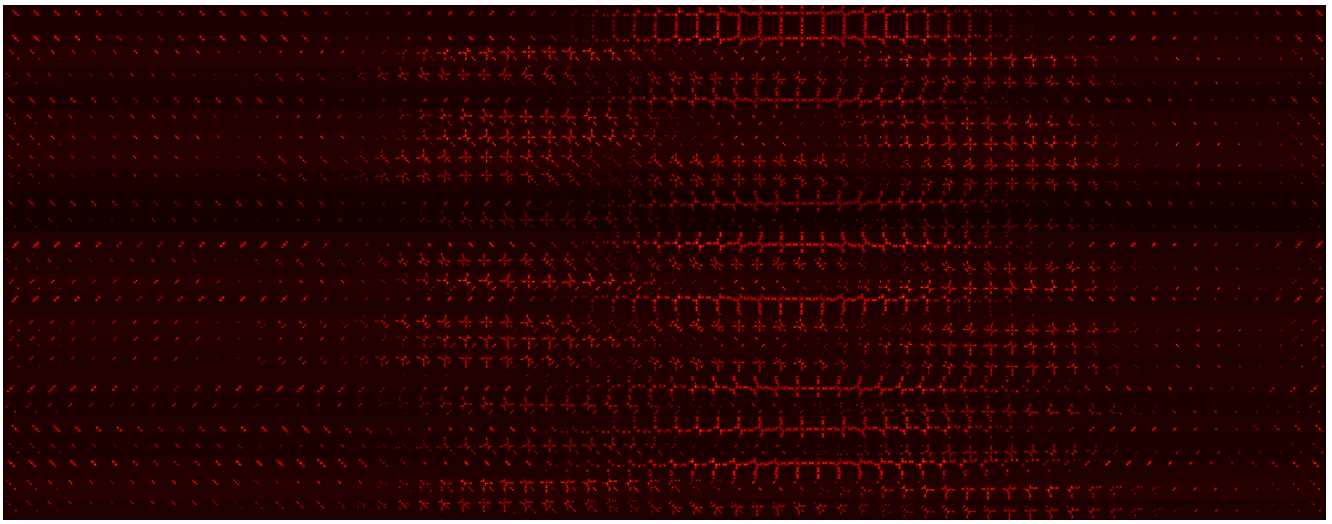


Рисунок 5. Энергетический ландшафт 25 соединений

Адсорбция

Адсорбция газа обычно описывается в статьях с использованием различных параметров и измерений. Наиболее распространенными параметрами, используемыми для описания адсорбции газа на МОК, являются адсорбционная емкость и энергия адсорбции [22].

Адсорбционная емкость - это мера количества газа, который может быть адсорбирован МОК, и обычно указывается в единицах массы или объема газа на массу или объем МОК. Этот параметр предоставляет информацию о способности материала адсорбировать конкретный газ.

Энергия адсорбции является мерой силы взаимодействия между газом и МОК и обычно указывается в единицах энергии на молекулу газа. Этот параметр дает информацию о стабильности адсорбированного газа и легкости десорбции.

Другим параметром, который часто используется для описания адсорбции газа на МОК, является изотерма. Изотерма - это график количества адсорбированного газа в зависимости от давления газа при постоянной температуре. Различные формы изотерм могут дать информацию о типе адсорбции, происходящей в МОК (например, физисорбция или хемосорбция).

Помимо этих параметров, статьи также часто включают информацию об условиях, при которых проводились измерения адсорбции, таких как температура и давление, а также тип газа, который был адсорбирован.

Поскольку нашей целью была адсорбция водорода, нам потребовалось извлечь параметры адсорбции из имеющихся источников и оцифровать их для нашего датасета. База данных NIST [13] содержит в себе набор точек изотерм адсорбции различных газов, полученных из натурных экспериментов. База данных MOFXDB [15] содержит данные об адсорбции водорода и азота при двух различных температурах.

Существует несколько математических функций, которые обычно используются для описания изотерм адсорбции газов на МОК. Эти функции обычно используются для подгонки экспериментальных данных и выбираются в зависимости от типа адсорбции, которая происходит в МОК. Наиболее распространенными функциями, используемыми для описания изотерм адсорбции, являются:

Изотерма Ленгмюра [23] – это широко используемая функция, которая описывает адсорбцию газов на поверхности. Она основана на предположении, что адсорбционные участки на поверхности эквивалентны и что адсорбция является обратимым процессом. Изотерма Ленгмюра представлена следующим уравнением:

$$q = Q_{max} \frac{P}{1+bP} \qquad \text{(II)}$$

Где q - количество адсорбированного газа, P - давление газа, Q max - максимальная адсорбционная емкость, b - постоянная Ленгмюра, которая представляет собой энергию адсорбции.

Изотерма БЭТ (Брунауэр-Эмметт-Теллер) [24] - это математическая функция, которая описывает адсорбцию газов на пористой поверхности. Она основана на предположении, что адсорбция происходит на многослойной поверхности, и учитывает площадь поверхности MOF. Изотерма БЭТ представлена следующим уравнением:

$$q = \frac{CP}{1-P} \quad \text{(III)}$$

Где q - количество адсорбированного газа, P - давление газа, а C - постоянная величина.

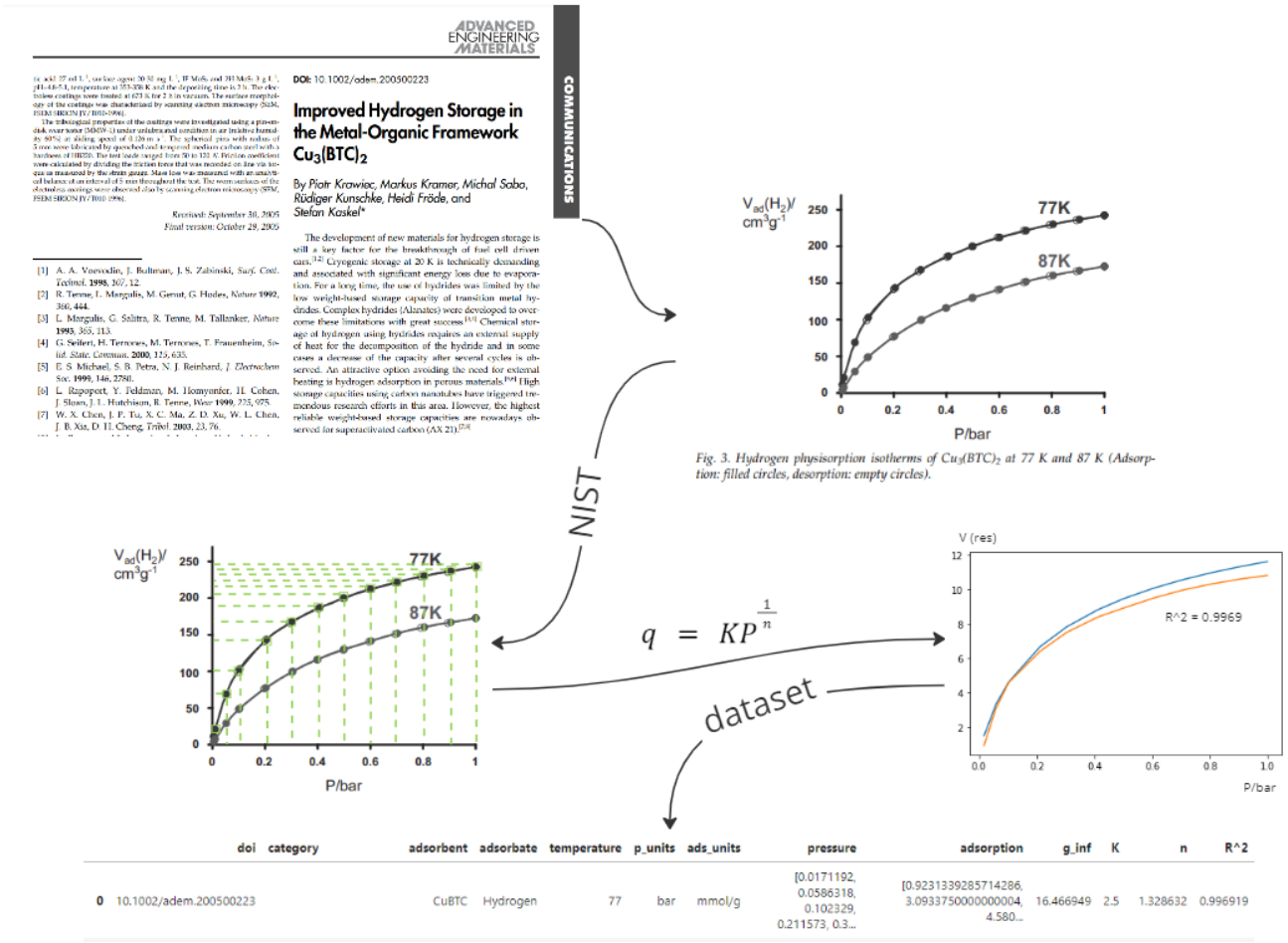
Другой математической функцией, которая обычно используется для описания изотерм адсорбции, является изотерма Фрейндлиха [25]. Изотерма Фрейндлиха - это простое эмпирическое уравнение, которое описывает адсорбцию газов на пористой поверхности. Она основана на предположении, что адсорбция происходит на многослойной поверхности, и учитывает площадь поверхности МОК. Изотерма Фрейндлиха представлена следующим уравнением:

$$q = KP^{\frac{1}{n}} \quad \text{(IV)}$$

Где q - количество адсорбированного газа, P - давление газа, K - константа, связанная с адсорбционной способностью, а n - константа Фрейндлиха, которая представляет собой интенсивность адсорбции.

Модель Фрейндлиха полезна для описания адсорбции газов на пористых поверхностях, она часто используется, когда процесс адсорбции не является линейным, и полезна для оценки процесса адсорбции на поверхности состоящей из разных видов атомов. Она также часто используется для описания адсорбции газов на неидеальных поверхностях и может быть применена к МОК, когда модель Ленгмюра или BET не дает хороших результатов.

Таким образом, мы выбрали для описания адсорбции модель Фрейндлиха. В данном случае адсорбционные свойства каждого из элементов датасета стало возможно описать при помощи всего двух констант: K и n.



	doi	category	adsorbent	adsorbate	temperature	p_units	ads_units	pressure	adsorption	g_inf	K	n	R^2
0	10.1002/adem.200500223		CuBTC	Hydrogen	77	bar	mmol/g	[0.0171192, 0.0586318, 0.102329, 0.211573, 0.3...	[0.9231339285714286, 3.0933750000000004, 4.580...	16.466949	2.5	1.328632	0.996919

Рисунок 6. Пайплайн сохранения данных адсорбции в датасете

Для определения этих параметров мы воспользовались оцифрованными значениями величины адсорбции и применили метод наименьших квадратов для вычисления коэффициентов.

Реализация генеративной модели машинного обучения

Выбор генеративного алгоритма

Среди множества генеративных алгоритмов было принято решение использовать GAN (Generative adversarial network) [26] модель машинного обучения без учителя. Почему нами был выбран именно такой тип моделей? Генеративно-состязательные сети достаточно сложны в обучении, но при этом выдают наилучшие результаты генерации. Это обусловлено архитектурой этой модели.

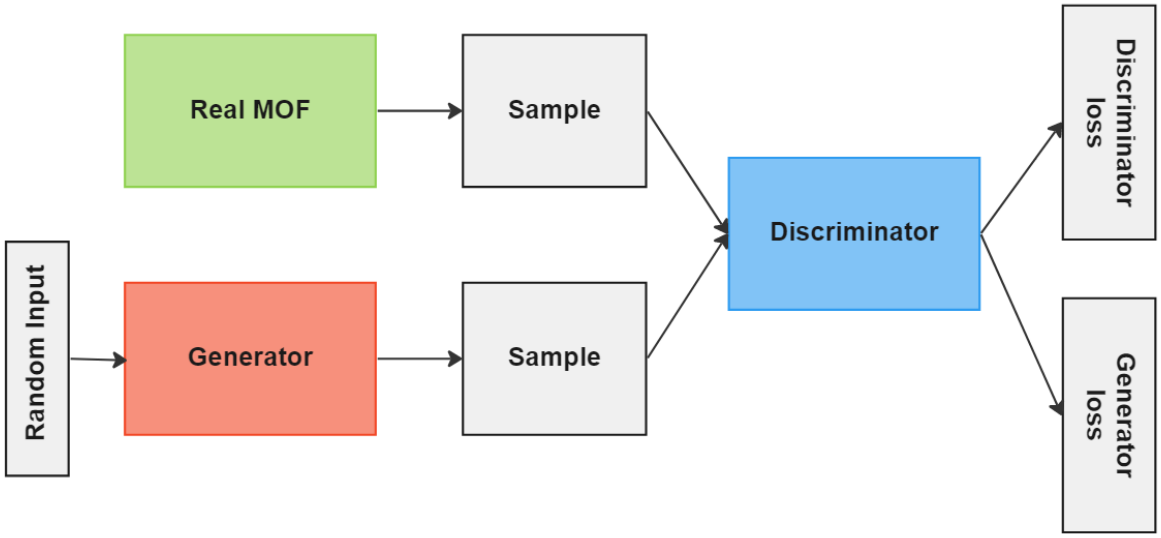


Рисунок 7. Архитектура GAN

GAN состоит из двух нейронных сетей: генератор и дискриминатор.

- Генератор учится генерировать правдоподобные данные.
- Дискриминатор учится отличать генерируемые данные от реальных.

Генератор, словно фальшивомонетчик старается обмануть своими данными дискриминатор, а дискриминатор в свою очередь, словно полиция наказывает генератор за неправдоподобные результаты. Выход генератора подключен ко входу дискриминатора. Через обратное распространение классификация дискриминатора предоставляет сигнал, который генератор использует для обновления своих весов.

GAN можно использовать с любыми дискретными данными. Предположительно, мы бы могли просто использовать значение координат из CIF файла, что показало бы их взаиморасположение, но тогда бы нейронная сеть не знала бы об их взаимодействии друг с другом. А также представление данных в CIF файле не унифицировано,а матрица содержащая в

себе энергетический ландшафт унифицирована, что также сказалось на результате. Для получения качественных результатов требуется универсальное представление данных.

Конкретно в данном исследовании проводится работа с трехмерными матрицами размерностью 64x64x64, содержащими в себе энергетический ландшафт.

Трехмерные матрицы данных и свертки

Обработка трехмерных матриц не является классической задачей в обучении нейронных сетей. Самая явная проблема - размер входного слоя. При большой размерности входного слоя, а также для выявления новых параметров традиционно используют свертки [27]. Их, например часто используют в работе с изображениями.

Свертка - линейная функция, которую применяют для выявления новых параметров в данных, а также для изменения их размерности. То есть используют для перевода данных в латентное пространство.

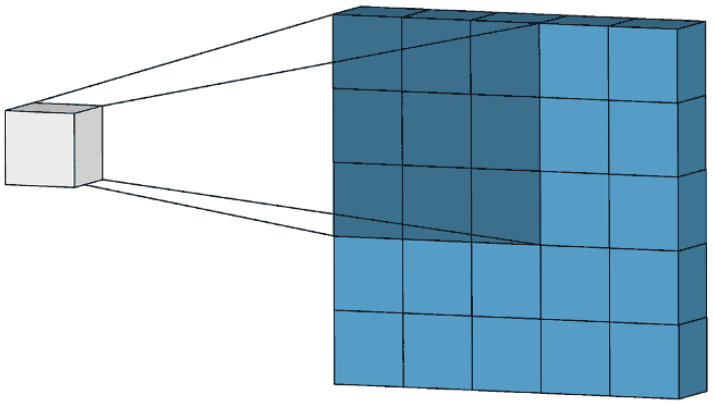


Рисунок 8. Визуальное представление работы свертки

В свертках есть несколько настраиваемых гипер-параметров: размер ядра, шаг ядра, увеличение размерности (Приложение №1) и смещение полученных значений (Приложение №2).

Гипер-параметр	Значение
kernel_size (размер ядра)	(4x4x4)

straid (шаг ядра)	2
padding (увеличение размерности)	1
bias (смещение полученных значений)	False

Таблица №3. Параметры сверток.

Специфика поставленной задачи состоит в том, что происходит работа с трехмерными объектами. Предположительно, мы бы могли работать подобно свертке цветных изображений. Так как в цветных изображениях у нас три разных канала, то мы разбиваем их по слоям и применяем свертку к каждому из слоев.

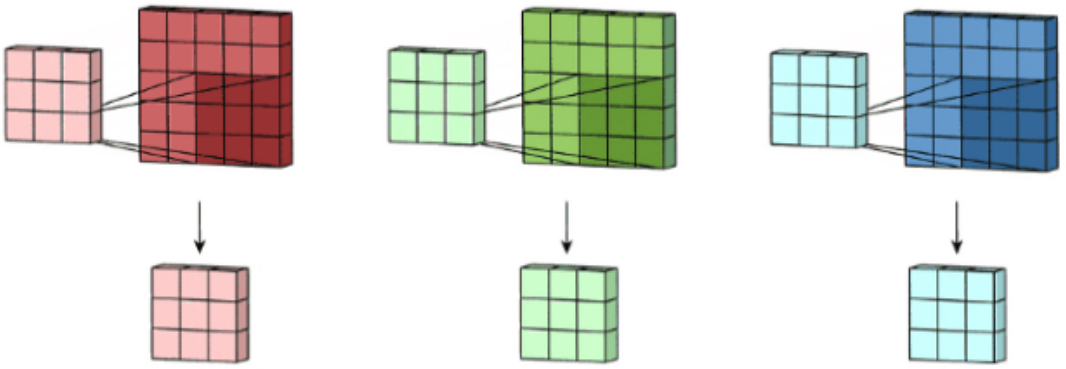


Рисунок 9. Визуализация работы сверток в цветных изображениях.

После чего происходит суммирование каждой из результирующих матриц.

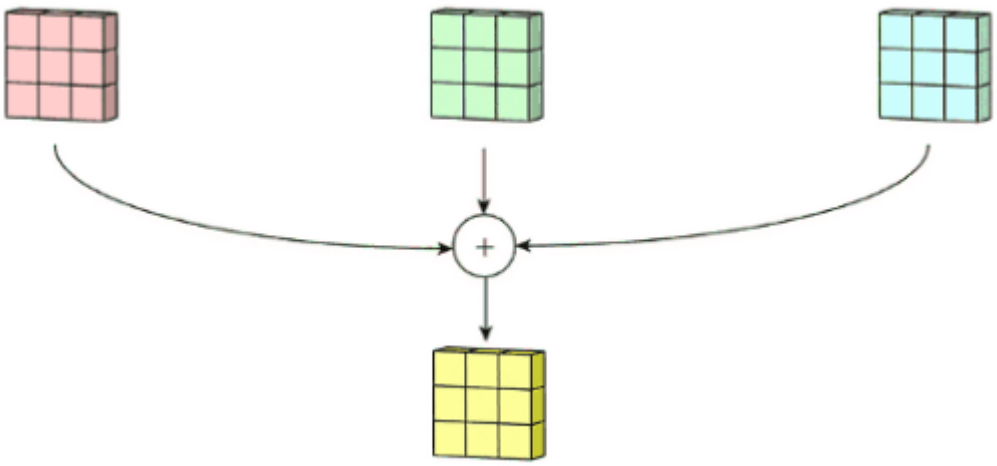


Рисунок 10. Суммирование результирующих матриц.

Однако, в таком случае нам бы пришлось разбивать наш трехмерный массив на 64 отдельных матрицы из-за чего произошла потеря информации, так как нам интересно их общее взаимодействие без привязки к слоям. При суммировании без привязки к слоям приводит к потере пространственной информации о распределении атомов вдоль одной из осей. В

зависимости от МОКа основное скопление молекул может находиться на разных слоях. Более того МОКи имеют трехмерную периодичность, из-за чего невозможно редуцировать до меньшей размерности без потери общности.

Единственным возможным решением было использовать трехмерные свертки и работать с трехмерными объектами во всей модели.

Архитектура GAN

Для этого была написана генеративно-состязательная модель, архитектура которой была основана на исследовании [28]. В этой статье была впервые показана модель 3DGAN для взаимодействия с трехмерным пространством.

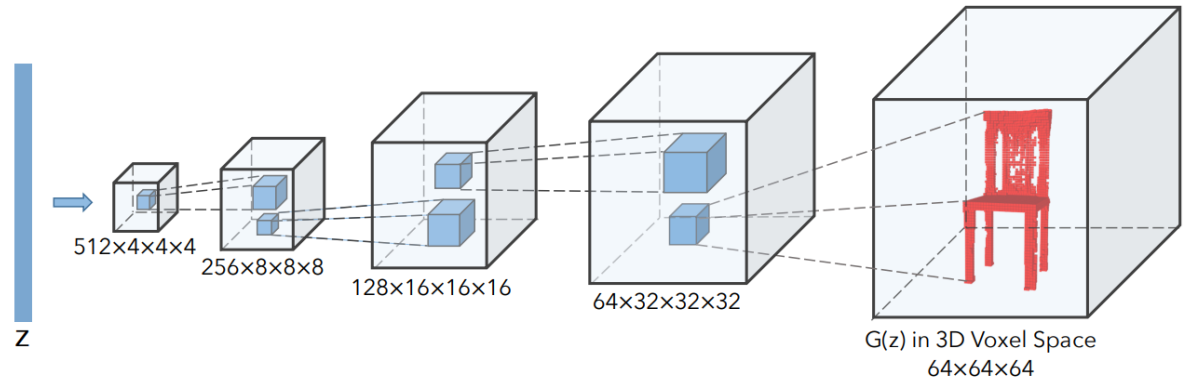


Рисунок 11. Визуализация работы модели 3DGAN

Модель нужно было адаптировать под задачу генерации энергетического ландшафта многих атомов взаимосвязанных друг с другом.

Реализация нейронной сети была произведена с помощью библиотеки pytorch [29], так как благодаря pytorch мы можем получать промежуточные результаты обучения, что позволит более тщательно контролировать процесс обучения, а также иметь возможность простого взаимодействия с обученными моделями.

Архитектуры генератора и дискриминатора очень схожи и обратны друг другу, как отражение в зеркале. В дискриминаторе были использованы 4 слоя свертки с гиперпараметрами, указанными выше. После каждого из слоев была применена пакетная нормализация (Приложение №3) и функция активации ReLU [30] (Приложение №4).

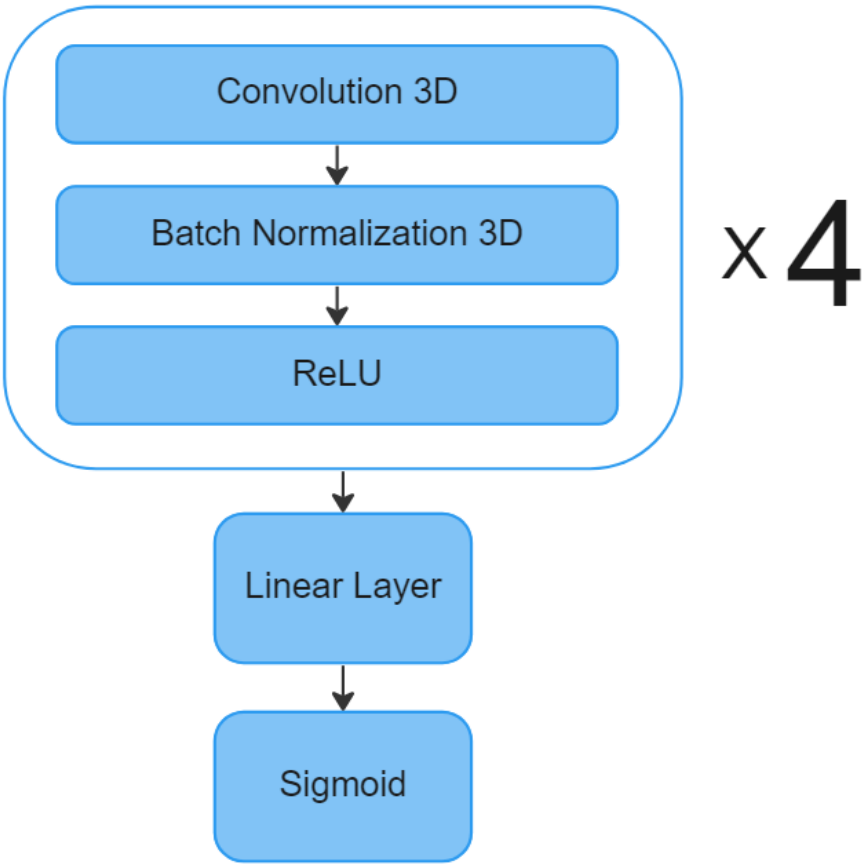


Рисунок 12. Архитектура дискриминатора.

Генератор же получает первоначально латентное пространство (рандомно сгенерированный вектор) размерностью 200, делает линейное преобразование для восстановления размерности. После чего идут 4 слоя обратной свертки, после каждого из которых мы снова используем нормализацию и используем функцию активации ReLU.

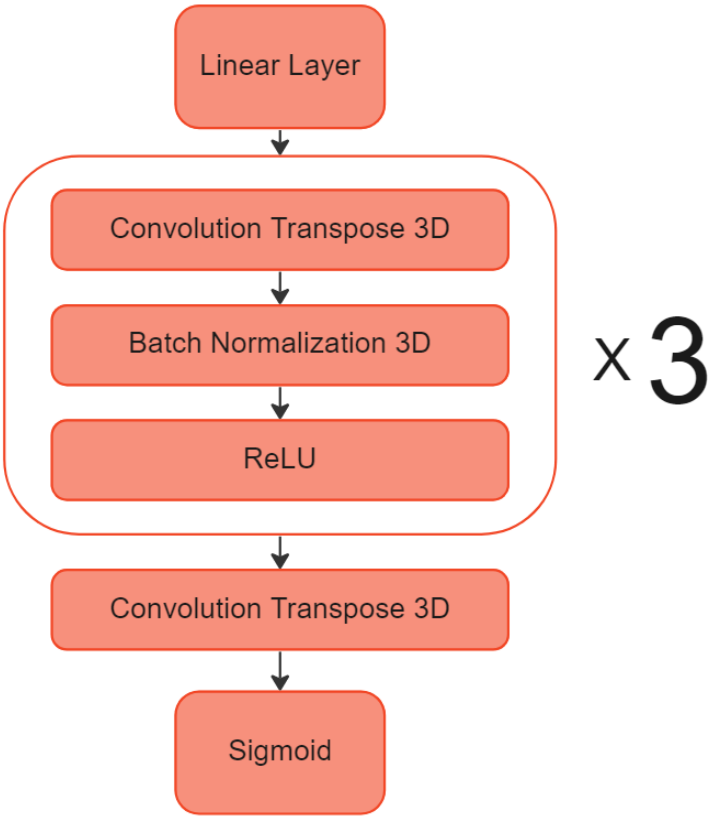


Рисунок 13. Архитектура генератора.

Нормализация и функция активации

Пакетная нормализация играет важную роль в глубоких нейронных сетях, так как повышает стабильность и ускоряет обучение нейронных сетей, так как в слой активации мы будем подавать признаки одной размерности. Математическое представление пакетной нормализации представлено в Приложении №3.

Активация с помощью функции ReLU была выбрана для того чтобы сеть была легче. Саму функцию можно записать с помощью формулы

$$f(z) = max(0, z) \quad \text{(V)}$$

что позволит получать градиент равный 0 при отрицательных значениях и 1 при положительных.

В данном случае нам пришлось инвертировать значения энергетического ландшафта, так как в интересующих точках будет отрицательные значения энергии. После инверсии эти точки будут иметь самые высокие положительные значения, как максимально пригодные для адсорбции. Можно предположить, что мы теряем часть данных ,“выравнивая впадины” на местах уже присутствующих атомов, так как размеры пиков становятся отрицательными, а функция

активации их зануляет. На самом деле мы можем пренебречь ими, так как эти позиции будут в любом случае не выгодны для атомов. Точки повышенной энергии вблизи атомов представляют интерес в динамике и поэтому в программах моделирования существуют специальные вероятностные функции размещения. В нашем случае мы смотрим на конечный результат, поэтому можем игнорировать эти точки вовсе.
Обобщенное описание модели можно посмотреть в Приложении №5

Функция потерь и оптимизатор

При классическом обучении GAN в качестве функции потерь используется BCE (Binary Cross Entropy) loss [31], описанный формулой:

$$L = E_x[\log(D(x)] + E_z[\log(1 - D(G(z))] \qquad \text{(VI)}$$

Основываясь на практическом опыте реализации GAN моделей, то BCE loss не является эффективным решением при обучении глубоких нейронных сетей. Поэтому в качестве функции потерь для нашей модели был выбран MSE (Mean Squared Error) loss [32], описанный формулой:

$$\min_D V_{LSGAN}(D) = \frac{1}{2}\mathbb{E}_{x\sim p_{data^{(x)}}}[(D(x) - b)^2] + \frac{1}{2}\mathbb{E}_{z\sim p_z(z)}[(D(G(z)) - a)^2] \qquad \text{(VII)}$$

$$\min_G V_{LSGAN}(G) = \frac{1}{2}\mathbb{E}_{z\sim p_z(z)}[(D(G(z)) - c)^2] \qquad \text{(VIII)}$$

MSE loss позволяет эффективнее обучать дискриминатор, что будет улучшать точность генератора.

В качестве оптимизатора был выбран Adam [33]. Этот оптимизатор лучше всего обрабатывает разряженные градиенты в зашумленных задачах, что позволяет ему быть одним из самых эффективных алгоритмов оптимизации.

Оперативная память и “сборка мусора”

При обучении возникла проблема с невозможностью хранения всего датасета в ОЗУ обучающего устройства, так как вес всего датасета составлял порядка 290 ГБ информации. Для того, чтобы обработать весь датасет, он был разбит на 420 одинаковых по размерности датасетов, содержащие в себе по 100 значений энергетического ландшафта. Процесс обучения был изменен с добавлением последовательной обработки каждого из датасетов, что позволило не хранить в ОЗУ весь объем информации.

Оценка работы нейросети

Благодаря составленной архитектуре мы смогли обучать нейронную сеть без заметного переобучения 48 эпох, что вы можете видеть на графике:

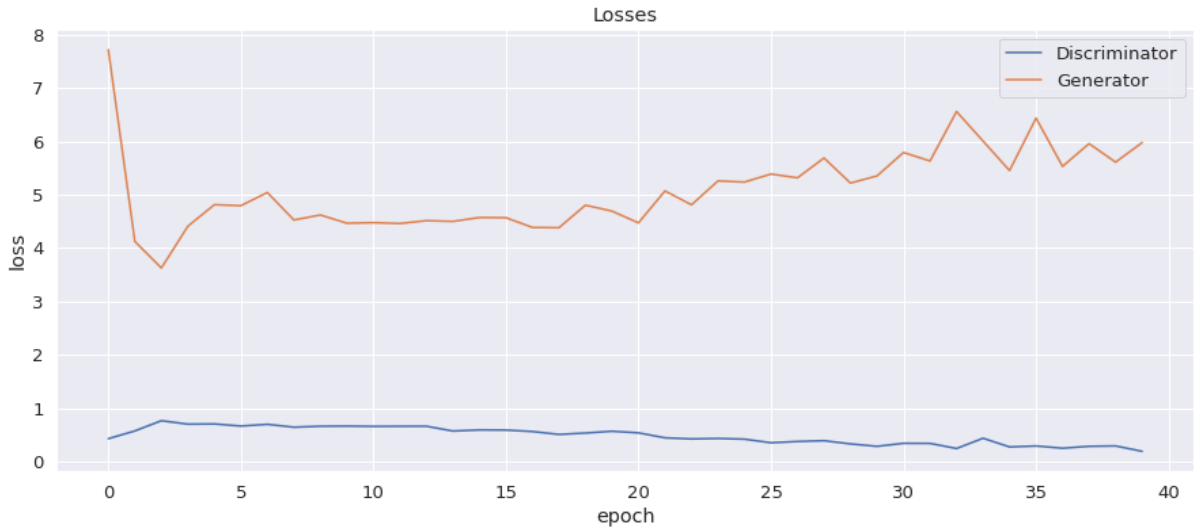


Рисунок 14. График изменения значений функций потерь.

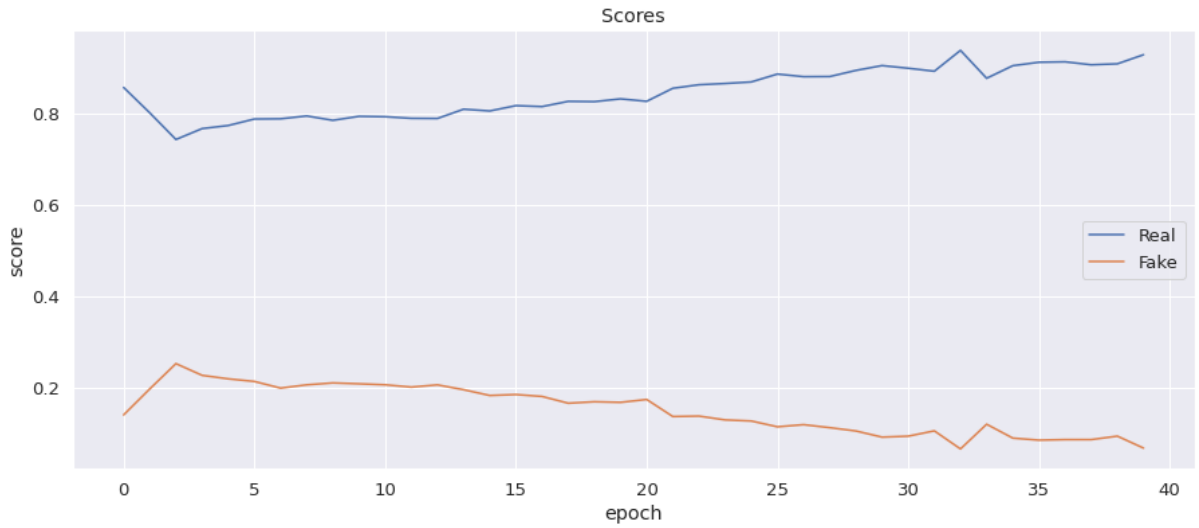


Рисунок 15. График изменения параметра score.

Для достижения наибольшей точности были протестированы множество вариантов гиперпараметров, вот несколько из них:

Количество эпох	30
Размер латентного пространства	200
Размерность батчей	16
Коэффициент скорости обучения	0.001

Количество эпох	48
Размер латентного пространства	200
Размерность батчей	2
Коэффициент скорости обучения	0.0002

Количество эпох	60
Размер латентного пространства	100
Размерность батчей	10
Коэффициент скорости обучения	0.002

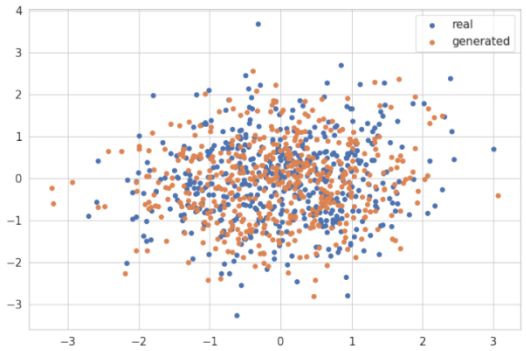
Рисунок 16. Несколько гиперпараметров.

Самые лучшие результаты показали:

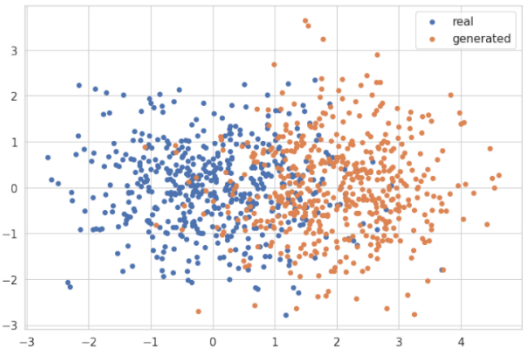
Количество эпох	48
Размер латентного пространства	200
Размерность батчей	2
Коэффициент скорости обучения	0.0002

Рисунок 17. Лучшие гиперпараментры.

Точность полученных результатов была оценена помимо визуальной оценки с помощью метода Leave-one out 1-NN classifier accuracy [34]. Суть метода состоит в том, чтобы представить одинаковые количество сгенерированных данных и реальных на одной плоскости.



Хорошо



Плохо

Рисунок 18. Визуализация работы метода Leave-one out 1-NN classifier accuracy

На смешанных значениях обучить KNN [35] классификатор с k = 1. Классификатор определяет параметры всей выборки и если сгенерированные признаки не отличаются от реальных, то он не сможет их отличить.

Используемый нами метод дал оценку в 58% (идеал: 50%). Это высокие значения, которые свидетельствуют о том, что полученная нейронная сеть создает МОК, который почти что не отличается по параметрам от уже существующих. Следовательно сгенерированные нами МОК имеют свойства адсорбции.

Визуализация и обратная задача

Визуализация данных

Визуализация - важный аспект науки о данных, поскольку она позволяет специалистам по анализу данных эффективно доносить до других понимание и результаты своего анализа. Она помогает создать более интуитивное понимание данных и взаимосвязей внутри них. Хорошо продуманная визуализация может выявить закономерности и тенденции, которые могут быть не сразу очевидны при просмотре необработанных данных, а также облегчить выявление выбросов и аномалий.

Визуализация также может использоваться для представления сложных наборов данных широкому кругу аудитории, включая неспециалистов. Это позволяет улучшить сотрудничество и принятие решений между различными командами и заинтересованными сторонами. Она также может использоваться для создания интерактивных информационных панелей, которые позволяют пользователям самостоятельно изучать данные и делать собственные открытия.

Кроме того, визуализация данных также является важным инструментом для исследовательского анализа данных, она позволяет специалисту по исследованию данных быстро выявить закономерности и взаимосвязи в данных и принять решение о том, как продолжить анализ. Она может помочь определить, какие переменные важны, а какие нет.

Традиционно химические соединения и молекулы визуализируют в виде шариков-атомов и палочек-связей. Иногда, в случае больших молекул, вроде белков, шарики убирают, оставляя только каркас соединения, для лучшей читаемости.

Однако, в случае представления соединения как энергетического ландшафта, такое представление будет нерациональным, поскольку каждая элементарная клетка пространства будет иметь свой цвет, рассчитанный из величины потенциальной энергии атомов, действующих на точку. Это всё равно что пытаться разглядеть внутреннее устройство огромного кубика Рубика, глядя на его внешние грани.

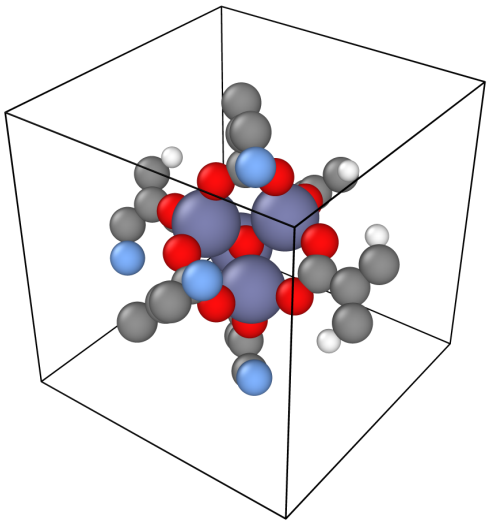


Рисунок 19. Представление примитивной ячейки МОК при помощи программы OVITO[36].

Именно поэтому традиционное представление не подходит для репрезентации ландшафта. Вместо него мы обратились к совершенно другой области знаний и воспользовались методом представления данных, полученных в результате компьютерной томографии.

При компьютерной томографии специалист делает огромное количество “фотографий-срезов” внутреннего органа человека, которые можно просмотреть как анимацию, так и разобрать на отдельные кадры. Также поступили и мы, создав 64 среза из кадров размером 64x64. Разместив их рядом друг с другом мы получили энергетический портрет соединения.

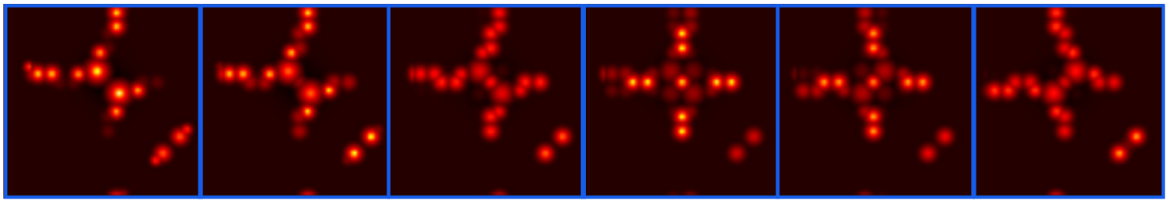


Рисунок 20. Энергетический ландшафт МОК на разных уровнях

Обратная задача

Последним этапом проекта стала задача получения состава и структуры сгенерированной молекулы из энергетического ландшафта. Сразу стоит упомянуть, что такую информацию можно сохранить в виде CIF-файла, но это не имеет большого практического смысла. Для извлечения свойств симметрии, топологии и прочих кристаллографических индексов потребовалась бы еще одна система машинного обучения, поэтому мы решили хранить только информацию о типе и положении атомов. Подобный упрощенный формат хранения информации о соединении хорошо укладывается в так называемый XYZ-файл.

Информации в полученном XYZ-файле хватает для того, чтобы при загрузке его в качестве входных данных в системы молекулярной динамики и молекулярной механики исследователь мог рассчитать параметры и изотермы адсорбции сгенерированного МОКа.

Для анализа типа и положения атома мы использовали алгоритмы кластеризации. Из-за зануления отрицательной инверсивной энергии функцией ReLU [30] мы не можем оценить типологию атома при помощи его максимумов и градиентов энергии, однако размеров кластера достаточно, чтобы оценить параметр σ и отличить один атом от другого.

В данном случае наш метод поиска состава соединения имеет свои минусы. В первую очередь проблема возникает в “шубе” из атомов водорода. Поскольку атомы водорода не создают очень большого энергетического поля, особенно на фоне наличия в системе дополнительных атомов молекулярного же водорода, их наличие и положение методом кластеризации определить довольно сложно. То же самое можно сказать про другие атомы с низкими значениями энергии связей. Это означает, в том числе, что наша схема не будет предполагать в составе соединения, например, благородных газов. К счастью, МОКов с таким составом – абсолютное меньшинство.

Проблемы подобного толка не уникальны для нашего подхода. Исследования соединений при помощи дифракционной рентгеноскопии порошка или монокристалла также дают нам надежную информацию только о крупных атомах в соединении.

Однако в случае анализа мы можем надеяться, что при оптимизации алгоритмов и увеличении точности прогнозов мы сможем предсказывать положения атомов точнее без значительного увеличения стоимости анализа в вычислительном смысле. В случае практического анализа увеличение точности прогноза может быть очень дорого, а иногда и вовсе невозможно.

Выводы

В ходе работы было достигнуто значительное достижение: создание генеративно-состязательной нейронной сети (GAN) для генерации металлоорганических каркасов (МОК) с адсорбционными свойствами.

Одним из ключевых достижений этого проекта стал сбор большой базы данных, содержащей информацию о МОК, включая сведения об их адсорбционных свойствах, энергетические ландшафты, формулы и файлы CIF. Эта база данных предоставляет богатую информацию для исследователей в области кристаллографии и окажет большую помощь в дальнейших исследованиях в этой области.

Кроме того, был определен единый метод представления МОК, позволяющий эффективно обрабатывать эти структуры с помощью алгоритмов машинного обучения. В рамках данного проекта также были разработаны скрипты и алгоритмы для разбора и обработки данных из API университетских баз данных.

Архитектура GAN, используемая в данном проекте, была разработана специально для работы в трехмерном пространстве с МОК. Для оптимизации обучения использовались методы сжатия данных и их стакинга. В результате нейронная сеть была успешно обучена и теперь способна обнаруживать совершенно новые металл-органические каркасы с интересующими нас адсорбционными свойствами.

Потенциальные возможности применения этих новых МОК велики и могут произвести революцию во многих областях науки и техники. Например, открытие новых МОК с улучшенными адсорбционными свойствами может значительно повысить безопасность транспортировки водорода.

В заключение следует отметить, что все задачи данного проекта были выполнены, и цель создания GAN для генерации металлоорганических каркасов с адсорбционными свойствами была успешно достигнута. Эта работа представляет собой значительное достижение в области генеративных нейронных сетей и, несомненно, приведет к дальнейшим важным открытиям в будущем.

В процессе написания работы нами использовались различные открытые источники, доступ к которым предоставляли через свои API различные организации и университеты. Для поддержания принципов открытости и доступности науки мы выложили весь код, а также все базы данных в публичном репозитории на Github: https://github.com/Gruz2520/3DGAN-for-MOF

Список использованной литературы

1. https://midjourney.com/
2. Hierarchical Text-Conditional Image Generation with CLIP Latents, Aditya Ramesh et al., https://arxiv.org/abs/2204.06125
3. https://stablediffusionweb.com/
4. Efficient Training of Language Models to Fill in the Middle, Mohammad Bavarian et al., https://arxiv.org/abs/2207.14255
5. Metal-organic framework functionalization and design strategies for advanced electrochemical energy storage devices, Avery E. Baumann et al., Communications Chemistry V. 2, #86 (2019)
6. Atomwise, https://www.atomwise.com/publications/
7. Insilico Medicine, https://insilico.com/publications
8. Form Energy, https://formenergy.com/
9. FermionX, https://www.fermionx.com/
10. Specification of the Crystallographic Information File (CIF), S. R. Hall, J. D. Westbrook, N. Spadaccini, I. D. Brown, H. J. Bernstein & B. McMahon, International Tables for Crystallography Volume G: Definition and exchange of crystallographic data pp 20–36
11. Westbrook J, Ito N, Nakamura H, Henrick K, Berman HM (April 2005). "PDBML: the representation of archival macromolecular structure data in XML". Bioinformatics. 21 (7): 988–992. doi:10.1093/bioinformatics/bti082. PMID 15509603
12. The Cambridge Structural Database C. R. Groom, I. J. Bruno, M. P. Lightfoot, S. C. Ward, Acta Cryst. (2016) B72, 171-179. DOI: 10.1107/S2052520616003954
13. Siderius, D.W., Shen, V.K., Johnson III, R.D. and van Zee, R.D., Eds., NIST/ARPA-E Database of Novel and Emerging Adsorbent Materials, National Institute of Standards and Technology, Gaithersburg MD, 20899, https://dx.doi.org/10.18434/T43882
14. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation, Anubhav Jain et al., APL Materials 1, 011002 (2013); https://doi.org/10.1063/1.4812323
15. MOFX-DB: An Online Database of Computational Adsorption Data for Nanoporous Materials N. Scott Bobbitt et al., Journal of Chemical & Engineering Data Article ASAP DOI: 10.1021/acs.jced.2c00583
16. G. Kresse and J. Hafner, Phys. Rev. B 47 , 558 (1993); ibid. 49 , 14 251 (1994)
17. WIEN2k: An APW+lo program for calculating the properties of solids. P. Blaha, K.Schwarz, F. Tran, R. Laskowski, G.K.H. Madsen and L.D. Marks, J. Chem. Phys. 152, 074101 (2020)
18. A. Erba et al., J/ Chem. Theory Comput., https://doi.org/10.1021/acs.jctc.2c00958

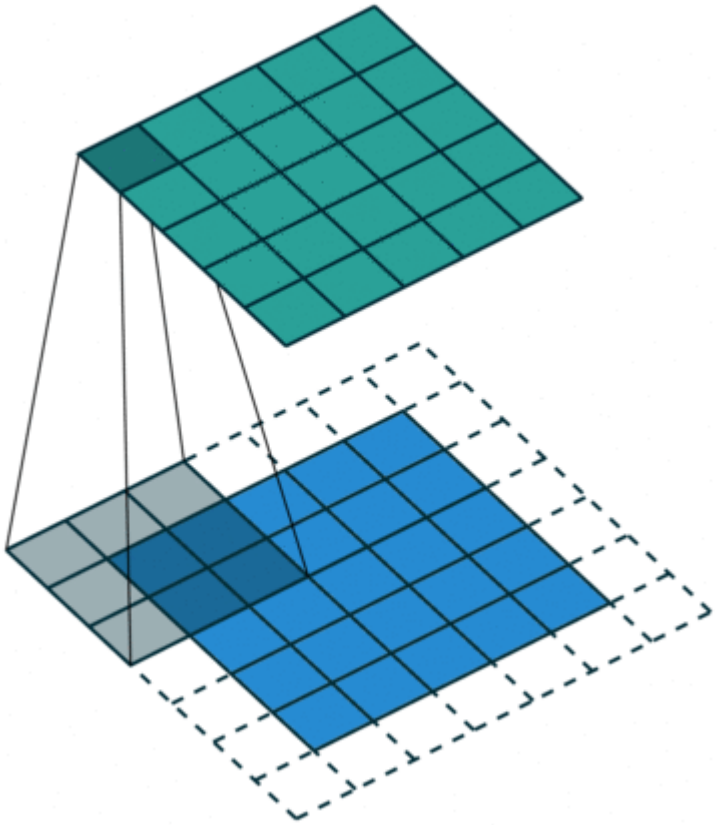
19. A.R. Oganov and C.W. Glass. Crystal structure prediction using ab initio evolutionary techniques: Principles and applications. The Journal of Chemical Physics, 124:244704, 2006.
20. Jones, J. E. (1924). "On the determination of molecular fields.—I. From the variation of the viscosity of a gas with temperature". Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character. 106 (738): 441–462. Bibcode:1924RSPSA.106..441J. doi:10.1098/rspa.1924.0081. ISSN 0950-1207.
21. LAMMPS - a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales, A. P. Thompson, H. M. Aktulga, R. Berger, D. S. Bolintineanu, W. M. Brown, P. S. Crozier, P. J. in 't Veld, A. Kohlmeyer, S. G. Moore, T. D. Nguyen, R. Shan, M. J. Stevens, J. Tranchida, C. Trott, S. J. Plimpton, Comp Phys Comm, 271 (2022) 10817.
22. Fundamentals of the Adsorption Theory, Partition and Adsorption of Organic Contaminants in Environmental Systems (pp.39-52), 2003 DOI:10.1002/0471264326.ch4
23. Langmuir, Irving (June 1918). "The Adsorption of Gases on Plane Surface of Glass, Mica and Platinum". Journal of the American Chemical Society. 40 (9): 1361–1402. doi:10.1021/ja02242a004.
24. Brunauer, Stephen; Emmett, P. H.; Teller, Edward (1938). "Adsorption of Gases in Multimolecular Layers". Journal of the American Chemical Society. 60 (2): 309–319. Bibcode:1938JAChS..60..309B. doi:10.1021/ja01269a023. ISSN 0002-7863
25. Freundlich, Herbert (1907). " Über die Adsorption in Lösungen." Zeitschrift für Physikalische Chemie – Stöchiometrie und Verwandschaftslehre. 57 (4), 385–470.
26. Goodfellow, Ian; Pouget-Abadie, Jean; Mirza, Mehdi; Xu, Bing; Warde-Farley, David; Ozair, Sherjil; Courville, Aaron; Bengio, Yoshua (2014). Generative Adversarial Nets, Proceedings of the International Conference on Neural Information Processing Systems (NIPS 2014). pp. 2672–2680.
27. Valueva, M.V.; Nagornov, N.N.; Lyakhov, P.A.; Valuev, G.V.; Chervyakov, N.I. (2020). "Application of the residue number system to reduce hardware costs of the convolutional neural network implementation". Mathematics and Computers in Simulation. Elsevier BV. 177: 232–243. doi:10.1016/j.matcom.2020.04.031. ISSN 0378-4754. S2CID 218955622. Convolutional neural networks are a promising tool for solving the problem of pattern recognition.
28. Learning a Probabilistic Latent Space of Object Shapes via 3D Generative-Adversarial Modeling, Jiajun Wu, Chengkai Zhang, Tianfan Xue, William T. Freeman, Joshua B. Tenenbaum, arXiv:1610.07584
29. https://github.com/pytorch/pytorch

30. Brownlee, Jason (8 January 2019). "A Gentle Introduction to the Rectified Linear Unit (ReLU)". Machine Learning Mastery. Retrieved 8 April 2021
31. Murphy, Kevin (2012). Machine Learning: A Probabilistic Perspective. MIT. ISBN 978-0262018029.
32. "Mean Squared Error (MSE)". www.probabilitycourse.com. Retrieved 2020-09-12.
33. https://pytorch.org/docs/stable/generated/torch.optim.Adam.html
34. Yang, Tao & Cao, Longbing & Zhang, Chengqi. (2010). A Novel Prototype Reduction Method for the K-Nearest Neighbor Algorithm with K ≥ 1. 89-100. 10.1007/978-3-642-13672-6_10.
35. Hart, Peter E. (1968). "The Condensed Nearest Neighbor Rule". IEEE Transactions on Information Theory. 18: 515–516.
36. A. Stukowski, Modelling Simul. Mater. Sci. Eng. 18, 015012 (2010)
37. G. E. Hinton , N. Srivastava, A. Krizhevsky, I. Sutskever and R. R. Salakhutdinov (2012) “Improving neural networks by preventing co-adaptation of feature detectors” ., https://arxiv.org/pdf/1207.0580.pdf

Приложения

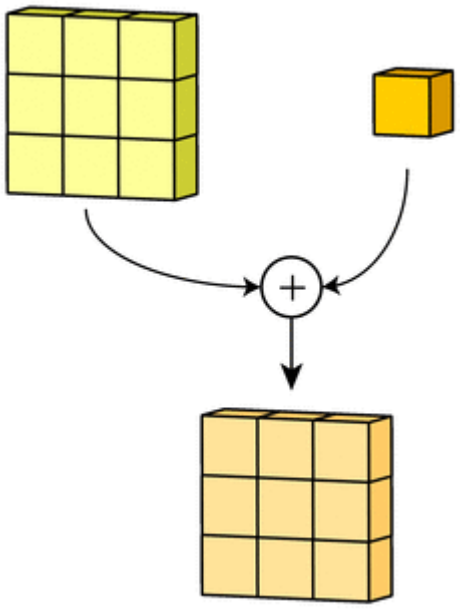
Приложение №1

Увеличение размерности (padding) - используется для сохранении данных при свертке. Чтобы не обрезать какое-либо количество данных мы добавляем вокруг матрицы дополнительные нулевые значения.



Приложение №2

Смещение полученных значений (bias) - погрешность данных, которые человек закладывает в модель.



Приложение №3

Алгоритм пакетной нормализации:

Вход: значения x из пакета $B = \{x_1, \cdots, x_m\}$; настраиваемые параметры γ, β; константа ϵ для вычислительной устойчивости

Выход: $\{y_i = BN_{\gamma,\beta}(x_i)\}$

$\mu_B = \frac{1}{m}\sum_{i=1}^{m} x_i$ – математическое ожидание пакета

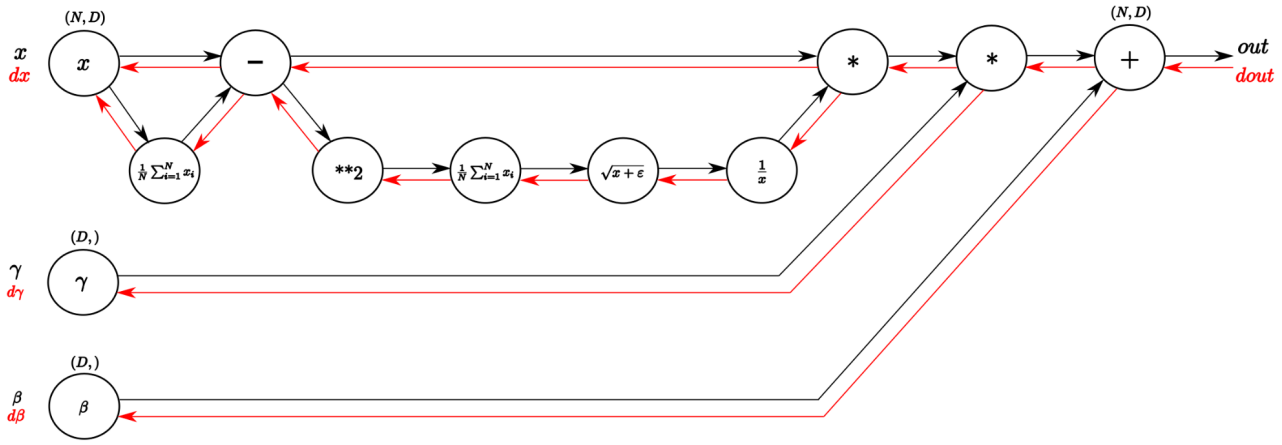
$\sigma_B^{\ 2} = \frac{1}{m}\sum_{i=1}^{m} (x_i - \mu_B)^2$ – дисперсия пакета

$\hat{x}_i = \frac{x_i - \mu_B}{\sqrt{\sigma_B^{\ 2} + \epsilon}}$ – нормализация

$y_i = \gamma \hat{x}_i + \beta \equiv BN_{\gamma,\beta}(x_i)$ – сжатие и сдвиг

Кроме того, использование пакетной нормализации обладает еще несколькими дополнительными полезными свойствами:

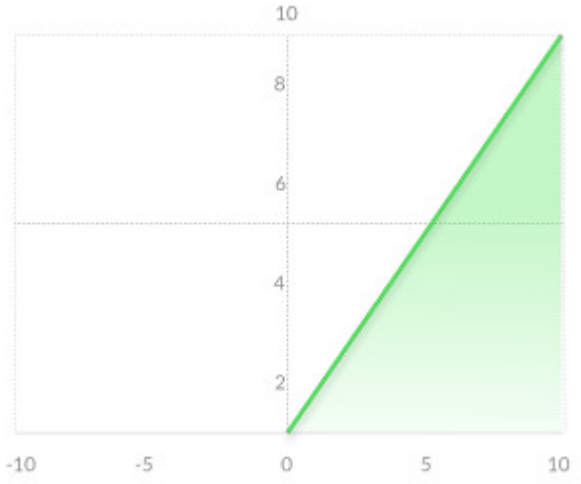
- достигается более быстрая сходимость моделей, несмотря на выполнение дополнительных вычислений;
- пакетная нормализация позволяет каждому слою сети обучаться более независимо от других слоев;
- становится возможным использование более высокого темпа обучения, так как пакетная нормализация гарантирует, что выходы узлов нейронной сети не будут иметь слишком больших или малых значений;
- пакетная нормализация в каком-то смысле также является механизмом регуляризации: данный метод привносит в выходы узлов скрытых слоев некоторый шум, аналогично методу dropout;
- модели становятся менее чувствительны к начальной инициализации весов.



Граф вычислений слоя пакетной нормализации алгоритмом обратного распространения ошибки. Слева-направо черными стрелками показана работа алгоритма в прямом направлении. А справа-налево красными стрелками — в обратном направлении, где вычисляется градиент функции потерь. Здесь N=m и D=d.

Приложение №4

ReLU (Rectified Linear Unit) - Самая популярная функция активации слоя, нужна для перевода нелинейных показателей в линейные для последующего нахождения производной от них.



Функция ReLU обладает несколькими преимущества перед сигмоидой и гиперболическим тангенсом:

- Очень быстро и просто считается производная. Для отрицательных значений — 0, для положительных — 1.
- Разреженность активации. В сетях с очень большим количеством нейронов использование сигмоидной функции или гиперболического тангенса в качестве активационный функции влечет активацию почти всех нейронов, что может сказаться на производительности обучения модели. Если же использовать ReLU, то количество включаемых нейронов станет меньше, в силу характеристик функции, и сама сеть станет легче.

Приложение №5

Обобщенное описание модели, сделанное с помощью summary из библиотеки torchsummary.

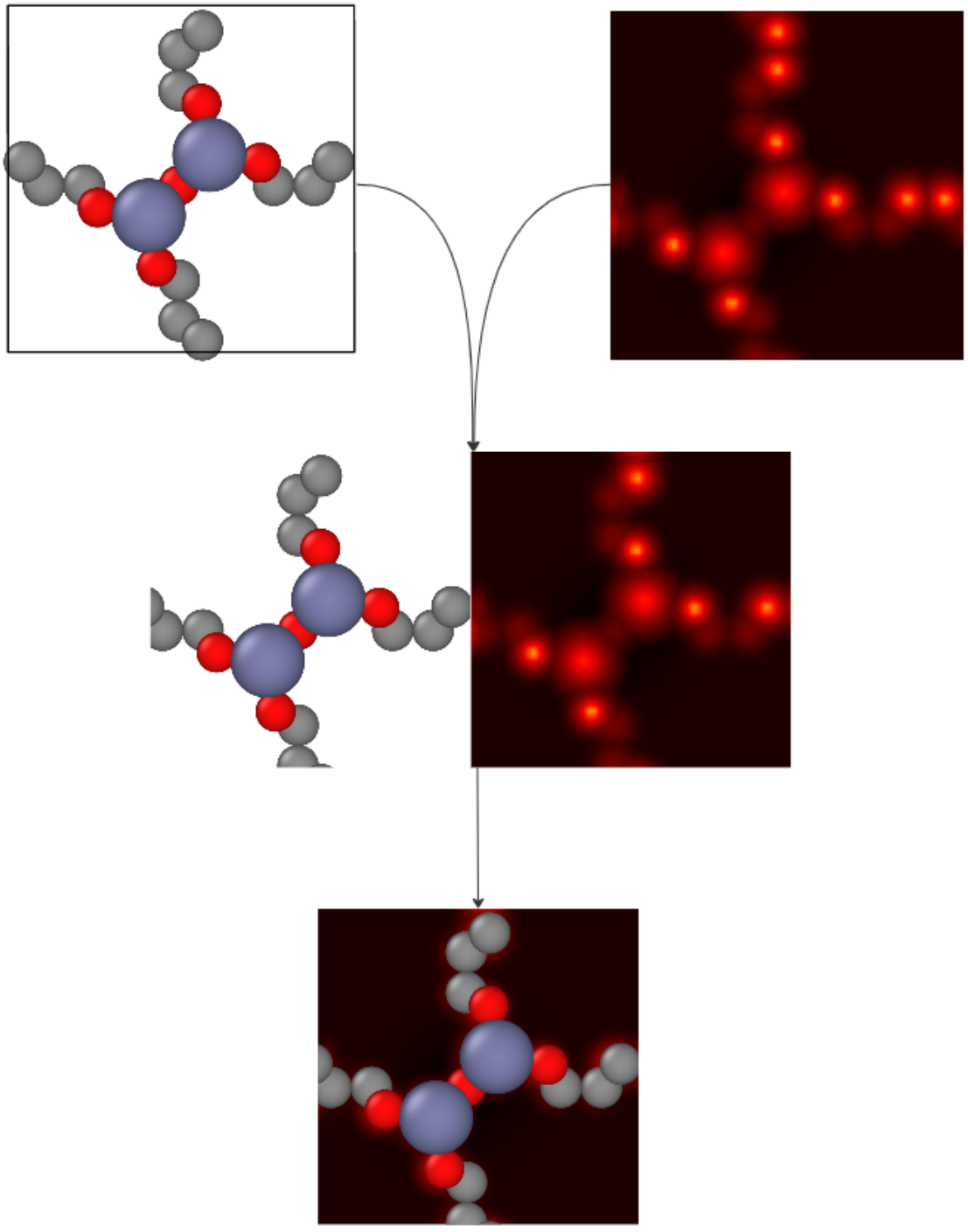
Генератор

```
----------------------------------------------------------------
        Layer (type)               Output Shape         Param #
================================================================
            Linear-1             [-1, 1, 32768]       6,586,368
   ConvTranspose3d-2         [-1, 256, 8, 8, 8]       8,388,608
       BatchNorm3d-3         [-1, 256, 8, 8, 8]             512
              ReLU-4         [-1, 256, 8, 8, 8]               0
   ConvTranspose3d-5      [-1, 128, 16, 16, 16]       2,097,152
       BatchNorm3d-6      [-1, 128, 16, 16, 16]             256
              ReLU-7      [-1, 128, 16, 16, 16]               0
   ConvTranspose3d-8       [-1, 64, 32, 32, 32]         524,288
       BatchNorm3d-9       [-1, 64, 32, 32, 32]             128
             ReLU-10       [-1, 64, 32, 32, 32]               0
  ConvTranspose3d-11        [-1, 1, 64, 64, 64]           4,096
          Sigmoid-12        [-1, 1, 64, 64, 64]               0
================================================================
Total params: 17,601,408
Trainable params: 17,601,408
Non-trainable params: 0
----------------------------------------------------------------
```

Дискриминатор

```
----------------------------------------------------------------
        Layer (type)               Output Shape         Param #
================================================================
            Conv3d-1       [-1, 64, 32, 32, 32]           4,096
       BatchNorm3d-2       [-1, 64, 32, 32, 32]             128
              ReLU-3       [-1, 64, 32, 32, 32]               0
            Conv3d-4      [-1, 128, 16, 16, 16]         524,288
       BatchNorm3d-5      [-1, 128, 16, 16, 16]             256
              ReLU-6      [-1, 128, 16, 16, 16]               0
            Conv3d-7         [-1, 256, 8, 8, 8]       2,097,152
       BatchNorm3d-8         [-1, 256, 8, 8, 8]             512
              ReLU-9         [-1, 256, 8, 8, 8]               0
           Conv3d-10         [-1, 512, 4, 4, 4]       8,388,608
      BatchNorm3d-11         [-1, 512, 4, 4, 4]           1,024
             ReLU-12         [-1, 512, 4, 4, 4]               0
           Linear-13                    [-1, 1]          32,769
          Sigmoid-14                    [-1, 1]               0
================================================================
Total params: 11,048,833
Trainable params: 11,048,833
Non-trainable params: 0
----------------------------------------------------------------
```

Приложение №6



Сопоставление атомарной модели молекулы CuBTC и её энергетического ландшафта