# Решение проблемы фаз с помощью методов глубокого обучения

Хайбрахманов Артур Ильнурович
Колпинский Сергей Викторович, Дмитриенко Артем Олегович

## Введение

Порошковая рентгеновская дифракция – метод исследования, который используется для качественного и количественного фазового анализа, определения кристаллографических параметров, а также получения кристаллической структуры. Данный метод основан на упругом рассеянии монохроматического рентгеновского излучения на трехмерной регулярной решетке атомов твердого вещества, что приводит к интерференции рентгеновских лучей.

Дифракцию в кристалле можно описать как отражение (рефлекс) от кристаллографических плоскостей кристаллической решетки. Семейство таких параллельных плоскостей полностью задается набором из трех целых чисел (h,k,l), которые называют индексами Миллера. Тогда угол отражения $\theta$ определяется по закону Вульфа-Брэгга: $2d_{hkl}sin\theta = \lambda$. Здесь $d_{hkl}$ – межплоскостное расстояние, $\lambda$ – длина волны. Вводят так называемое обратное пространство (рис. 1), базисные вектора которого по модулю обратны базисным векторам прямого пространства, а индексы Миллера являются координатами всех векторов. Точки обратной решетки задают семейства кристаллографических плоскостей прямой решетки, а значит и дифракционные отражения. Таким образом, дифракционная картина кристалла является трансформацией упорядоченной атомной структуры в обратное пространство.

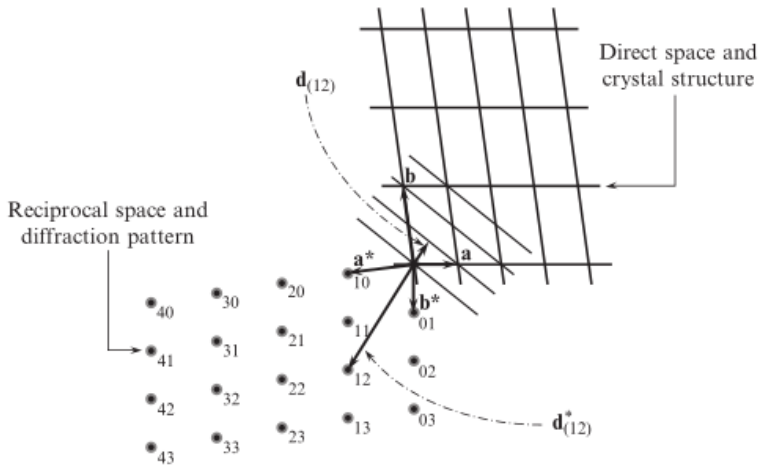


Рисунок 1. Пример перехода кристаллографических плоскостей прямой решетки в обратную. Соответствующиие индексы Миллера показаны около точек обратного пространства [1]

Чтобы учесть вклад атомов кристаллической решетки в отражение вводят структурный фактор $F(hkl) = \sum_{j=1}^{N} f_j exp[2\pi i(hx_j + ky_j + lz_j)]$, где $f_j$ – атомный фактор, отражающий вклад атома, $(x_j, y_j, z_j)$ – координаты атома. Структурный фактор является комплексной величиной. В ходе эксперимента регистрируется интенсивность отражения как безразмерная величина, отражающая число детектированных рентгеновских лучей, которая равна квадрату амплитуды структурного фактора $|F(hkl)|^2$.

Трехмерное распределение атомов в решетке может быть получено только после перехода дифракционной картины из обратного в прямое пространство с помощью Фурье-преобразования: $\rho(x, y, z) = V^* \sum_{hkl} F(hkl) exp[(-2\pi i(hx + ky + lz)]$ ($V^*$ – объем элементарной обратной ячейки) [2]. Как видно из формулы, для расчета электронной плотности требуется структурный фактор, но из эксперимента можно определить только его амплитуду. Данная проблема получила название проблема фаз. Имея достаточный набор отражений из эксперимента, с помощью различных рутинных методов можно решить фазовую проблему и определить структуру вещества. Имея недостаточное количество экспериментальных отражений, при использовании рутинных методов и последующего расчета электронной плотности получаются диффузные электронные облака, не соответствующие атомам, из чего невозможно определить кристаллическую структуру.

Получение трехмерной структуры биологических макромолекул является важной задачей для понимания механизма их функций и активности [3]. Получение структуры методом порошковой дифракции осложнено перекрыванием отражений, особенно на дальних углах. Белки же обычно образуют большие, но плохие кристаллы, что приводит к большому числу и более широкому профилю линии рефлексов [2]. Низкое разрешение дальних отражений не позволяет получить требуемый набор данных для "решения"таких больших структур. Для уменьшения перекрывания линий используют синхротронный источник излучения, собирают несколько наборов данных, варьируя, например, температуру или предпочтительную ориентацию образца. Зная аминокислотную последовательность белка, можно получить его структуру, если уже известна белковая структура с той же последовательностью, методом молекулярного замещения. Если же такая структура недоступна, то кристаллографическая проблема фаз может быть решена методом изоморфного замещения, в рамках которого собирают дополнительные наборы данных порошковой дифракции, добавляя тяжелые атомы в структуру [4].

В литературе результатов решения фазовой проблемы с помощью нейронных сетей не обнаружено. Значит, baseline – преодоление экспериментатором проблематики нерутинными методами с последующим определением белковой структуры, требующими особого внимание на каждый случай.

Для решения проблемы фаз предлагается предсказывать интенсивности дифракционных рефлексов по имеющимся (более подробно в разделе методология). Это может быть реализовано с помощью вариационного автоэнкодера [5], зарекомендовавшего

себя в генерации и восстановлении изображений. Так как мы не уверены, что наши данные обладают локальной связанностью, планируется проверить архитектуры автоэнкодеров со сверточными, а также линейными слоев. Также может оказаться перспективным использование архитектур на основе визуальных трансформеров [6]. Однако в наших данных нет связанной последовательности для восстановления, неясно, как можно приспособить трансформеры для данной задачи.

Таким образом, решение фазовой задачи белковой кристаллографии является актуальной задачей, нерешаемой рутинными методами. Создание инструментов на основе методов глубокого обучения для преодоления данной проблемы является целью работы.

Применение методов машинного обучение в обработке рентгенодифракционных экспериментов бурно развивается. Полученные результаты в этой области были опубликованы как в главных кристаллографических журналах (IUCR Journal), так и в высокорейтинговых журналах, рассчитанных на более широкую аудиторию (Nature Communications, npj Computational Materials). Планируемый вклад в данную тему с ориентацией на решение реальной кристаллографической задачи обладает необходимой научной новизной.

## Данные

Для генерации наборов данных планируется модифицировать для данной задачи и использовать собственное программное обеспечение (`codeberg.org/dmitrienka/pxrd_simulator`), в котором с помощью библиотеки CCTBX (Computational Crystallography Toolbox [7]) создаются случайные достоверные структуры и рассчитываются структурные факторы. Для создания набора данных для обучения планируется использовать наиболее распространенные для молекулярных кристаллов пространственные группы (P-1, $P2_1$, $P2_1/c$, C2/c, $P2_12_12_1$, Pbca), типы атомов (C, N, O, F) и число симметрийно независимых атомов 20–30. Параметры элементарной ячейки для случайных структур планируется выбрать из нормального распределения с центрами 10.05Å, 12.19Å и 15.11Å для параметров $a$, $b$ и $c$ соответственно и стандартным отклонением 4.6; все углы — из нормального распределения с центром 90 градусов и стандартным отклонением 15 градусов. Такие значения отражают распределение параметров элементарных ячеек реальных молекулярных кристаллов в Кембриджском Банке Структурных Данных [8]. В ходе работы планируется собрать набор литературных данных порошковой дифракции белков.

## Методология

Предлагается предсказывать интенсивности дифракционных отражений белков, которые нельзя получить из экспериментальных данных из-за низкого разрешения, по известным из того же эксперимента. После предсказания достаточного количество

рефлексов, набора данных должно хватить для определения фаз одним из рутинных методов.

Так как отражения являются точками обратного пространства, каждое из них можно однозначно описать индексами Миллера (h,k,l). Тогда дифракционную картину можно описать трехмерным тензором, в котором записаны интенсивности каждого отражения. Таким образом, задача сводится к восстановлению трехмерного тензора. Inference моделей глубокого обучения должен выглядеть следующим образом: на вход подается тензор с рентгенодифракционными экспериментальными данными, на выходе должен быть тензор с дополнительными интенсивностями. В ходе обучения планируется научить модель восстанавливать тензор отражений по данным малых органических молекул, для этого интенсивности, соответствующие дальним отражениям, будут обращены в нуль.

В качестве модели предлагается использовать вариационный автоэнкодер. Планируется протестировать архитектуры декодера и энкодера из трехмерных сверточных и полносвязных линейных слоев из-за неуверенности в локальной связанности данных. Рассматривается также использование архитектур типа трансформеры, механизм внимания которых может быть полезен для данной задачи.

Эффективность предсказания обученных моделей глубоко обучения будет проверяться на тестовой части синтетического датасета, а также собранном наборе экспериментальных данных.

## Метрики

В качестве метрики предлагается использовать среднюю квадратичную ошибку. В качестве функции потерь планируется использовать среднеквадратичную ошибку для всех возможных в данной задаче архитектур, при обучении VAE также в функцию будет добавлена дивергенция Кульбака-Лейблера.

## План

- Модифицировать имеющееся программное обеспечение по генерации синтетических рентгенодифракционных данных
- Сгенерировать набор данных, разделить его на обучающую, валидационную и тестовую выборки
- Обучить вариационные автоэнкодеры на основе сверточных и полносвязных линейных слоев
- Попробовать использование трансформера в рамках данной задачи и при возможности обучить его
- Собрать экспериментальный набор данных по порошковой дифракции белков из литературы

- Провести тестирование моделей на синтетических и реальных данных

## Журнал

Journal of Applied Crystallography

# Список литературы


[1] Pecharsky V. K. and Zavalij P. Y. Fundamentals of Powder Diffraction and Structural Characterization of Materials. — Springer US, 2009.

[2] Girolami G. S. X-ray Crystallography. — University Science Books, 2016.

[3] Spiliopoulou M., Valmas A., Trandafillidis D., and Kosinas C. Applications of X-ray Powder Diffraction in Protein Crystallography and Drug Screening // Crystals. — 2020. — Vol. 10, no. 2.

[4] Margiolaki I. and Wright J. P. Powder crystallography on macromolecules // Acta Crystallographica Section A. — 2008. — Vol. 64, no. 1. — P. 169–180.

[5] Kingma D. P. and Welling M. Auto-Encoding Variational Bayes. — 2022. — 1312.6114.

[6] Wu B., Xu C., Dai X., Wan A., Zhang P., Yan Z., Tomizuka M., Gonzalez J., Keutzer K., and Vajda P. Visual Transformers: Token-based Image Representation and Processing for Computer Vision. — 2020. — 2006.03677.

[7] Grosse-Kunstleve R. W., Sauter N. K., Moriarty N. W., and Adams P. D. The *Computational Crystallography Toolbox*: crystallographic algorithms in a reusable software framework // Journal of Applied Crystallography. — 2002. — Vol. 35, no. 1. — P. 126–136.

[8] Groom C. R. and Allen F. H. The Cambridge Structural Database in Retrospect and Prospect // Angewandte Chemie International Edition. — 2014. — Vol. 3. — P. 662–671.