Химия

# Решение проблемы фаз с помощью методов глубокого обучения

Хайбрахманов Артур Ильнурович
Колпинский Сергей Викторович, Дмитриенко Артем Олегович

## Аннотация

Решение проблемы фаз является важной задачей рентгеноструктурного анализа, особенно актуальной для белковой кристаллографии ввиду отсутствия ab initio решений в этой области. Методы машинного обучения способны преодолеть данную задачу. В работе предпринята попытка решения путем повышения разрешения дифракционной картины с помощью сверточных нейронных сетей архитектуры UNet и вариационного автоэнкодера.

## Ключевые слова

Рентгеновская дифракция, проблема фаз, нейронные сети, машинное обучение, структурный фактор, рентгеноструктурный анализ

## Введение

Рентгеноструктурный анализ – дифракционный метод исследования кристаллической структуры вещества. 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 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, 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.

Чтобы учесть вклад атомов кристаллической решетки в отражение вводят структурный фактор:

$$F(h,k,l) = |F| \cdot e^{i\phi} = \sum_{j=1}^{N} f_j exp[2\pi i(hx_j + ky_j + lz_j)],$$

где $\phi$ — фаза отражения, $f_j$ – атомный фактор, отражающий вклад атома, $(x_j, y_j, z_j)$ – координаты атома. Структурный фактор является комплексной величиной. В ходе эксперимента регистрируется интенсивность отражения, отражающая число детектированных рентгеновских лучей, которая равна квадрату модуля структурного фактора $|F|^2$.

Трехмерное распределение атомов в решетке может быть получено только после перехода дифракционной картины из обратного в прямое пространство с помощью Фурье-преобразования:

$$\rho(x, y, z) = V^* \sum_{h,k,l} F(h, k, l) exp[(-2\pi i(hx + ky + lz)],$$

где $V^*$ – объем элементарной обратной ячейки [1]. 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, данную проблему можно решить с помощью прямых методов, полагающихся на атомность электронной плотности, или метода Charge flipping.

Получение трехмерной структуры биологических макромолекул является важной задачей для понимания механизма их функций и активности [2]. Дифракционные картины белков ограничены малыми углами отражения, так как их кристаллы из-за большого количества атомов в ячейке обладают меньшей кристалличностью, чем низкомолекулярные образцы [1]. Из-за низкого разрешения невозможно определить положения атомов в кристаллической ячейке белка рутинными ab initio методами, требуется дополнительная информация. Зная аминокислотную последовательность, можно получить структуру белка, если уже известна 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 проводят дополнительные рентгенодифракционные эксперименты, добавляя тяжелые атомы в структуру [3].

Применение машинного обучения в области рентгеновской дифракции лишь недавно зародилось и сейчас бурно развивается. Так, недавняя статья [4] является единственной публикацией по решению проблемы фаз с помощью методов глубокого обучения. В качестве объектов предсказания они выбрали центросимметричные структуры, для которых фазы отражений принимают два возможных значения — 0 и 1. Авторы презентовали нейронную сеть, представляющую собой бинарный классификатор из блоков трехмерных свёрток и многослойных перцептронов. Обучение проводилось на синтетических кристаллических молекулярных структурах. Также в работе реализована идея phase recycling — исходные данные прогоняются несколько раз через модель, что увеличивает точность классификации. Таким образом, впервые был продемонстрирован потенциал машинного обучения для решения проблемы фаз, но только для центросимметричных структур.

Но были найдены решения аналогичной проблемы с помощью методов глубокого обучения в области физики, а именно в рамках метода когерентной безлинзовой микроскопии. В обзорной статье [5] выделены 3 подхода — DL-post-processing, в котором уточняются "плохие фазы полученные из исходных интенсивностей; DL-in-processing, в рамках которого из интенсивностей с помощью нейронной сети рассчитывают фарзы; и DL-pre-processing, в котором обученная модель повышает разрешение микроскопической картины, и уже из полученного изображения фазы определяются классическими методами.

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.

## Данные

Для генерации данных для обучения использовалось собственное программное обеспечение (`github.com/blackwood168/xrd_simulator`), в котором с помощью библиотеки CCTBX (Computational Crystallography Toolbox [6]) создаются кристаллические решетки, в которых случайным образом расставлены атомы, и рассчитываются структурные факторы. Использовались наиболее распространенные для молекулярных кристаллов (набор №1) пространственные группы (P-1, $P2_1$, $P2_1/c$, C2/c, $P2_12_12_1$, Pbca), а также группы моноклинной ($P2_1$, C2) и орторомбической ($C222_1$, $P2_12_12$, $P2_12_12_1$) сингоний (набор №2, 3), соответствующие наиболее распространенным белковым структурам в базе данных белков (`rcsb.org/stats/distribution-space-group`); типы атомов — C, N, O, Cl; число симметрийно независимых атомов 10–30. Также в работе использовались данные из Кембриджского Банка Структурных Данных [7]. Высокое разрешение выбрано 0.8 Å, низкое — 1.5 Å.

В ходе выполнения работы были обучены нейросети, предсказывающие как интенсивность, так и модуль структурного фактора (амплитуду). Типичные распределения этих данных для сгенерированных структур представлено на рис. 1. Как можно заметить, распределение амплитуд больше похоже на стандартное. Дифракционные данные также были отнормированы в диапазон 0-1.

## Методология

В работе было предложено предсказывать интенсивности или амплитуды дифракционных отражений белков, которые нельзя получить из эксперимента, по известным из того же эксперимента. После предсказания достаточного количество рефлексов, разрешения должно хватить для определения фаз и расчета электронной плотности одним из рутинных ab initio методов.

Так как отражения являются точками обратного пространства, каждое из них можно однозначно описать индексами Миллера (h,k,l). Тогда дифракционную картину можно

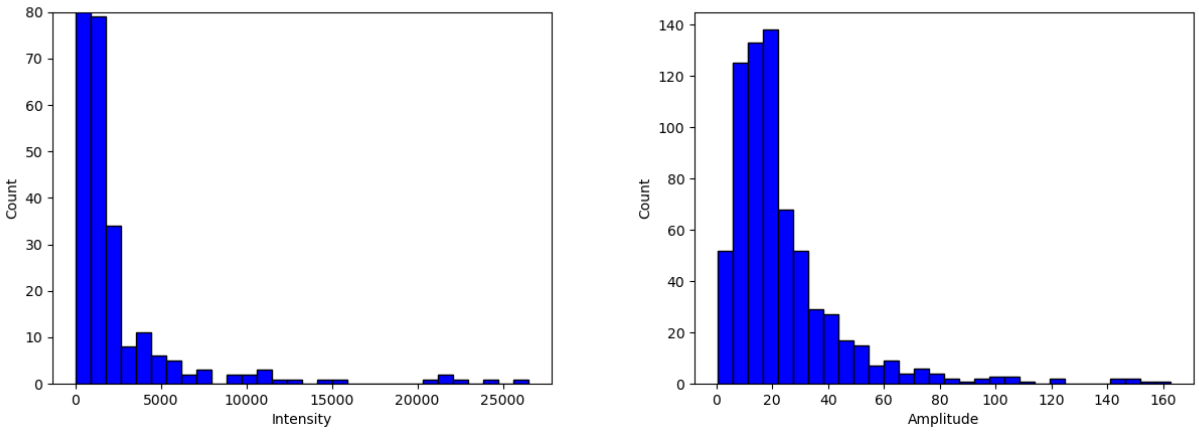


Рисунок 1. Типичные распределения интенсивностей (слева) и амплитуд (справа) дифракционной картины

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 будут обнулены интенсивности дальних отражений так, чтобы длина разрешения полученной дифракционной картины соответствовала типичному разрешению белковых соединений (1.5 Å).

В качестве моделей были предложены и проверены вариационный автоэнкодер и сверточная сеть UNet, адаптированные под трехмерные матрицы. Рассматривается также использование архитектур типа трансформеры, механизм внимания которых может быть полезен для данной задачи.

Эффективность предсказания обученных моделей глубокого обучения предлагается проверять на тестовой части синтетического датасета, а также собранных наборах экспериментальных данных.

В ходе выполнения работы также предложен постпроцессинг, включающий в себя учёт систематических погасаний — "зануления"некоторых значений интенсивностей, что определяется симметрией структуры; а также явного восстановления части тензора, которую не нужно предсказывать.

## Результаты

В рамках данной работы был разработан пайплайн, позволяющий проводить воспроизводимые эксперименты (`github.com/blackwood168/xrd_phase_ml`). В нем реализовано обучение и тестирование моделей, а также inference на реальных массивах данных и структурах кристаллических соединений.

### Наивный подход

По набору №1 структур, пространственные группы которых соответствуют наиболее распространенным для молекулярных кристаллов, были обучены UNet и вариационный автоэнкодер (рис. 2). Модели были обучены на датасете из 40 тысяч структур. На вход подавалась матрица интенсивностей дифракционных картин, архитектуры моделей были адаптированы под трехмерный случай. Значение среднеквадратичной ошибки на синтетической отложенной тестовой выборке составляет $16.1 \cdot 10^{-5}$ и $1.05 \cdot 10^{-5}$ для VAE и UNet, соответственно. Можно сделать вывод, что адаптированная архитектура UNet лучше справляется с восстановлением дифракционной картины.

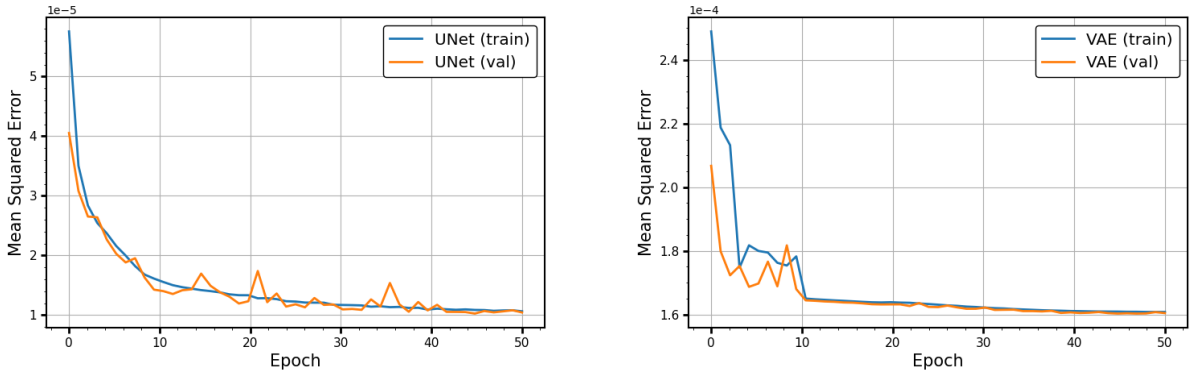


Рисунок 2. Кривые обучения для UNet (слева) и VAE (справа)

В ходе анализа данного подхода выяснилось, что он имеет ряд недостатков. Так, модели не выучивают паттерны систематических погасаний, от чего тензор интенсивностей не имеет нулевых значений. Также модели не сохраняют интенсивности ближних отражений, то есть центра тензора, по которому происходит восстановление в ходе inference.

Самым же важным недостатком является то, что для описания дифракционных картин требуются матрицы размером 33x36x29, но для большинства структур они будут сильно разрежены. Связано это с тем, что в дифракционные данные попадают только симметрично независимые отражения. У более симметричных структур будет меньше независимых отражений, и наоборот. Это приводит к тому, что для структур с высокой симметрией большая часть данных заполнена нулями. Поэтому имеет смысл обучать разные модели для каждой рассматриваемой кристаллической сингонии, чтобы избежать разреженных матриц.

### Моноклинная сингония

По набору №2 структур моноклинной сингонии были обучены разные модели UNet (рис. 3). Размер матриц для данной сингонии 26x18x23. Обучение происходило на датасете из 60 тысяч образцов и большом датасете на 200 тысяч. Также варьировался размер модели, в большой модели на 1 вертикальный блок больше.

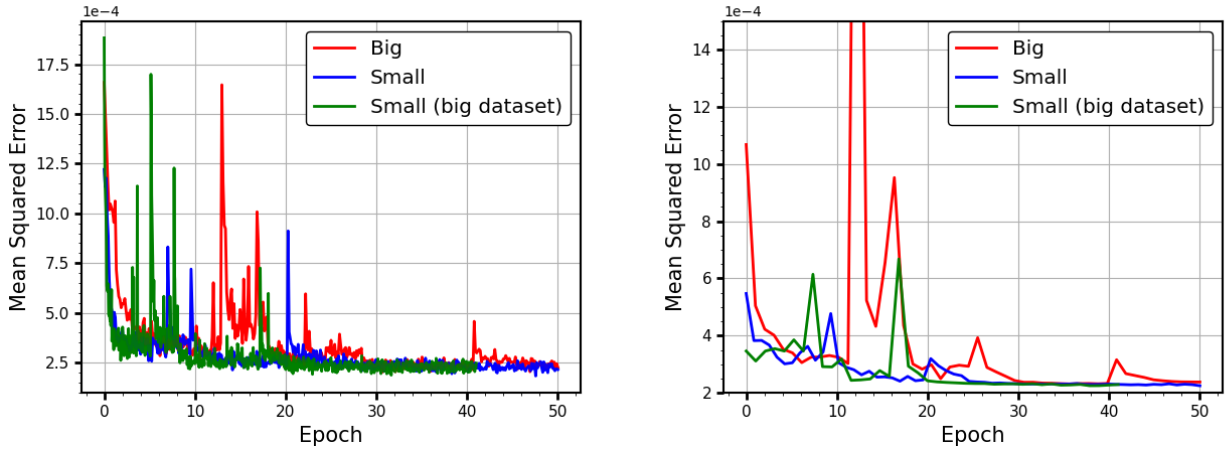


Рисунок 3. Тренировочная (слева) и валидационная (справа) кривые обучения

Таблица 1. Значения среднеквадратичной ошибки на отложенной тестовой выборке

| **Модель** | **MSE·$10^{-5}$** | **MSE-Postprocessing·$10^{-5}$** |
|---|---|---|
| Маленькая | 22.75 | 22.64 |
| Большая | 23.35 | 23.14 |
| Маленькая, большой датасет | 22.46 | 22.33 |

Значение MSE на отложенной тестовой выборке представлены в таблице 1, MSE-Postprocessing — значение среднеквадратичной ошибки, после применения постпроцессинга к восстановленным матрицам. Можно сделать вывод, что восстанавливающая способность моделей отличается незначительно. Наилучший же результат показывает маленькая модель, обученная на большом датасете (200 тысяч структур).

### Орторомбическая сингония

По набору №3 структур орторомбической сингонии была обучена модель с архитектурой UNet (рис. 4). Обучение происходило на датасете из 60 тысяч структур, размер матриц составил 17x22x29. На отложенной тестовой выборке значения ошибки составили 20.37·$10^{-5}$ и 20.26·$10^{-5}$ до и после постпроцессинга, соответственно.

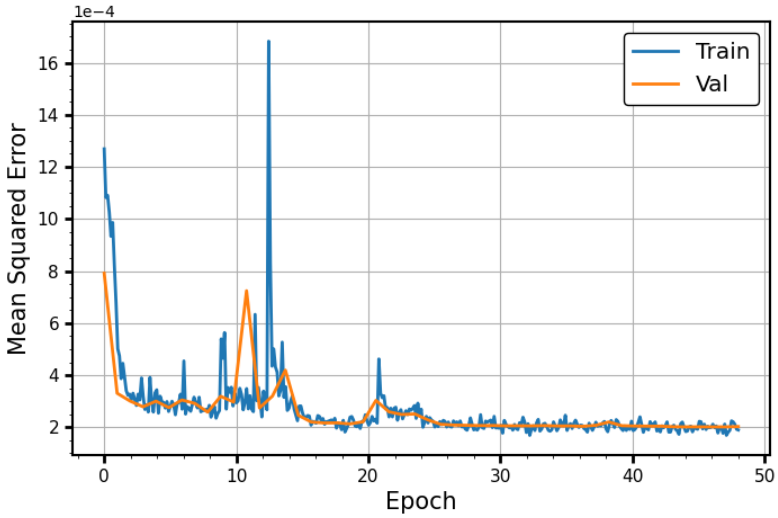


Рисунок 4. Кривые обучения

### Проверка на реальных данных

Было проведено тестирование на реальных молекулярных моноклинных структурах из Кембриджского Банка Структурных данных [7]. Среднее значение MeanAbsoluteError по результатам тестирования на 60 структурах составило 0.0013. Далее восстановленная нейронной сетью матрица подавалась на вход в кристаллографические программы для решения структур ShelXT и ShelXL. Однако ни одну структуру по восстановленным данным не получилось решить.

## Выводы

- Разработано программное обеспечение по генерации синтетических рентгенодифракционных данных (`github.com/blackwood168/xrd_simulator`)
- Разработан единый пайплайн (`github.com/blackwood168/xrd_phase_ml`), позволяющий проводить воспроизводимые эксперименты
- Обучены вариационный автоэнкодер и UNet на кристаллических структурах различной сингонии, лучшую эффективность продемонстрировала UNet
- Обучены сверточные нейронные сети с архитектурой UNet для структур орторомбической и моноклинной сингоний
- Проведено тестирование на реальных рентгенодифракционных данных на предмет возможности решения структуры по восстановленным данным

# Список литературы


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

[2] Gawas U. B., Mandrekar V. K., and Majik M. S. Structural analysis of proteins using X-ray diffraction technique // Advances in Biological Science Research. — 2019.

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

[4] Larsen A. S., Rekis T., and Madsen A. PhAI: A deep-learning approach to solve the crystallographic phase problem // Science. — 2024. — Vol. 385, no. 6708. — P. 522–528.

[5] Wang K., Song L., and Wang C. e. a. On the use of deep learning for phase recovery // Light Sci Appl. — 2024. — Vol. 13, no. 4.

[6] 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.

[7] 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.