

Московский государственный университет
имени М.В. Ломоносова
Химический факультет
Кафедра физической химии
Лаборатория строения конденсированных систем

ХАЙБРАХМАНОВ АРТУР ИЛЬНУРОВИЧ

**ПРИМЕНЕНИЕ МЕТОДОВ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА ДЛЯ
РЕШЕНИЯ ЗАДАЧ РЕНТГЕНОДИФРАКЦИОННЫХ ИССЛЕДОВАНИЙ
КРИСТАЛЛОВ**

Дипломная работа

Научный руководитель:
к. х. н., с. н. с. Дмитриенко А. О.
д. х. н., профессор Лысенко К. С.

Москва, 2025

Оглавление

Введение	4
1 Обзор литературы	5
1.1 Рентгеновская кристаллография	5
1.2 Проблема фаз	6
1.3 Решение фазовой проблемы	8
1.3.1 Методы проб и ошибок	8
1.3.2 Прямые методы	9
1.3.3 Метод Паттерсона	13
1.3.4 Метод обращенного заряда (charge-flipping)	18
1.3.5 Метод Vive La Différence (VLD)	23
1.3.6 Методы искусственного интеллекта	26
1.4 Заключение	29
2 Методика решения	31
2.1 Подход	31
2.2 Рентгенодифракционные данные	32
2.3 Модели машинного обучения	36
3 Результаты и обсуждение	40
3.1 Автоматизированный вычислительный конвейер	40
3.2 Модели для структурных факторов (разрешение 1.5 Å)	40
3.2.1 Результаты обучения	40
3.2.2 Определение структур	41
3.3 Модели для структурных факторов (разрешение 1.2 Å)	42
3.3.1 Результаты обучения	42
3.3.2 Определение структур	43
3.4 Модели для нормализованных структурных факторов	44
3.4.1 Результаты обучения	44
3.4.2 Определение структур	45
3.5 Интерпретируемость моделей	45
3.6 Уточнение структуры 6-бром-1Н-индол-2,3-диона	50
4 Результаты и выводы	53

Список литературы	54
Приложение	58

Введение

Проблема фаз — определение фаз структурных факторов — давно является одной из главных задач рентгеновской кристаллографии. На сегодняшний день существуют такие решения, как прямые методы, методы Паттерсона и обращенного заряда (charge-flipping), однако они, как правило, ограничены атомным разрешением рентгенодифракционных данных. Данные методы подразумевают использование полного набора интенсивностей дифракционных максимумов и часто требуют выращивания высококачественных образцов кристаллов, что является сложной задачей для слабо рассеивающих кристаллов или больших молекул, таких как белки. Для определения структур макромолекулярных соединений разработаны методы молекулярного замещения и изоморфного замещения, требующие дополнительную информацию - знание о структуре кристаллов других белков с той же аминокислотной последовательностью или результаты рентгенодифракционных экспериментов с добавлением тяжелых атомов в исследуемую структуру.

Применение машинного обучения в области рентгеновской дифракции сейчас бурно развивается. Чаще всего авторы пытаются обойти решение проблемы фаз, предсказывая кристаллическую структуру из доступных экспериментальных данных. Однако на сегодняшний день уже показано, что с помощью методов искусственного интеллекта можно предсказывать фазы дифракционных отражений центросимметричных кристаллов, что позволяет решить фазовую проблему с атомным разрешением всего 2 \AA , используя всего 10–20% данных, требуемых традиционными методами. Таким образом, методы машинного обучения способны преодолеть проблему фаз, снимая ограничение на высокое разрешение рентгенодифракционных данных.

Целью нашей работы стала разработка подходящих средств решения проблемы фаз с помощью методов искусственного интеллекта и анализ их эффективности решения проблемы фаз для нецентросимметричных структур.

1 Обзор литературы

1.1 Рентгеновская кристаллография

Рентгеновская кристаллография является важнейшим инструментом определения трехмерных структур кристаллов. Этот метод позволяет получить бесценные сведения об атомных и молекулярных структурах кристаллов, что крайне важно для понимания свойств и функций материалов в различных областях, включая химию, биологию и материаловедение. Рентгеноструктурный анализ основан на упругом рассеянии монохроматического рентгеновского излучения на трехмерной регулярной решетке атомов твердого вещества, что приводит к интерференции рентгеновских лучей.

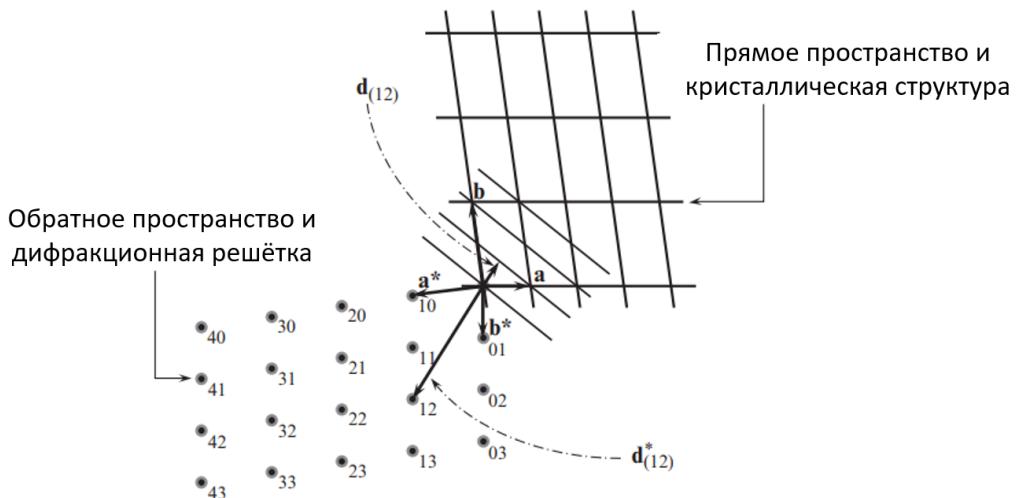


Рисунок 1. Связь прямого и обратного пространства [1]

Дифракция в кристалле описывается как отражения от семейств параллельных кристаллографических плоскостей элементарной ячейки кристалла (рис. 1). Каждая точка дифракционной картины задается набором из индексов Миллера h, k, l . Угол отражения (рассеяния) Θ для дифракционного максимума с межплоскостным расстоянием d_{hkl} , полученного при рассеянии рентгеновского излучения с длиной волны λ , определяется из закона Вульфа–Брэгга (уравнение 1). Также вводится обратное пространство, где каждая точка задаёт семейства кристаллографических плоскостей структуры, а значит, и отражения. Вместо индексов Миллера можно ввести вектор обратного пространства, задающий отражение $\vec{H} = (h, k, l)$. Таким образом, дифракционная картина кристалла является трансформацией упорядоченной атомной структуры в обратное пространство.

$$2d_{hkl} \sin \Theta = \lambda, \quad (1)$$

Любой дифракционный максимум также имеет математическое описание его интенсивности, известное как структурный фактор $F(\vec{H})$, зависящий от расположения $r_j = (x, y, z)$ и коэффициентов рассеяния f_j всех атомов в элементарной ячейке кристалла:

$$F(\vec{H}) = |F(\vec{H})| \exp(i\phi) = \sum_{j=1}^N F_j(\vec{H}) = \sum_{j=1}^N f_j \exp(2\pi i(\vec{H}, \vec{r}_j)), \quad (2)$$

где $(\vec{H}, \vec{r}_j) = hx_j + ky_j + lz_j$, F_j — структурный фактор каждого из N симметрично независимых атомов в кристаллической ячейке, в котором закодирована информация об амплитуде f_j и фазе $\phi_j = 2\pi(\vec{H}, \vec{r}_j)$ рассеянной этим атомом волны.

1.2 Проблема фаз

Структурный фактор, как следует из определения (уравнение 2), является комплексной величиной, которая описывает вклад дифракции всех атомов кристаллической решетки в отражение. Стоит отметить, что оно было получено в следующем приближении:

- Элементарная ячейка поделена на N атомов в точках \vec{r}_j .
- Каждый из атомов рассеивает волну с амплитудой f_j и фазой $\phi_j = 2\pi(\vec{H}, \vec{r}_j)$.

Перейдем к более общему описанию — заменим атомы на маленькие параллелепипеды, внутри которых находятся электроны и рассеивают излучение. Тогда предыдущее приближение превращается в следующее:

- Элементарная ячейка поделена на N маленьких параллелепипедов, объем каждого ΔV и позиция \vec{r}_j .
- Каждый из параллелепипедов рассеивает волну с амплитудой $\rho(\vec{r}_j) \Delta V$ и фазой $\phi_j = 2\pi(\vec{H} \cdot \vec{r}_j)$, где $\rho(\vec{r}_j)$ — электронная плотность внутри параллелепипеда.

Тогда выражение 2 можно записать следующим образом, устремив объемы параллелепипедов к нулю:

$$F(h, k, l) = \sum_{j=1}^N \rho(\vec{r}_j) \Delta V_j \exp(2\pi i(\vec{H} \cdot \vec{r}_j)) = \int_V \rho(\vec{r}) \exp(2\pi i(\vec{H} \cdot \vec{r})) dV, \quad (3)$$

где интегрирование ведется по всему объему элементарной ячейки.

Можно заметить, что уравнение 3 является обратным преобразованием Фурье, переводящее электронную плотность в структурные факторы. Тогда можно записать прямое преобразование Фурье:

$$\rho(\vec{r}) = \int F(\vec{H}) \exp(-2\pi i(\vec{H} \cdot \vec{r})) dV^* = \frac{1}{V} \sum_{\vec{H}} F(\vec{H}) \exp(-2\pi i(\vec{H} \cdot \vec{r}_j)), \quad (4)$$

где интегрирование ведётся по обратной ячейке.

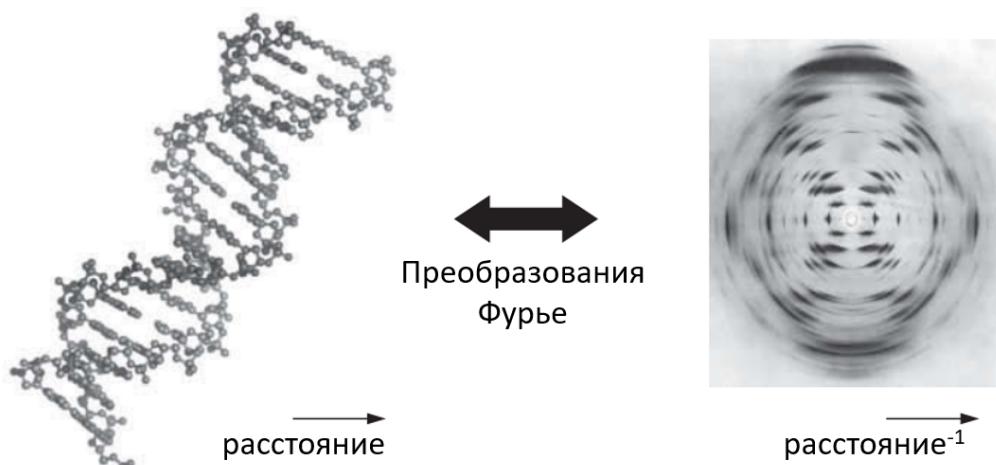


Рисунок 2. Схема связи электронной плотности и структурных факторов [1]

Если амплитуда и фазы всех дифрагированных лучей были бы зарегистрированы, можно было бы рассчитать распределение электронной плотности в кристалле, применив преобразование Фурье к дифракционной картине (рис. 2). Однако в ходе рентгенодифракционного эксперимента регистрируется лишь интенсивность излучения, информация о фазах теряется, и для получения кристаллической структуры требуется решить так называемую "проблему фаз" или "фазовую проблему".

Таким образом, в идеальном случае, при сохранении полной информации о фазах в ходе эксперимента, структура кристалла могла бы быть непосредственно восстановлена из измеренных амплитуд с помощью преобразования Фурье. Поскольку же фазовая составляющая оказывается неизвестной при рентгеновском рассеянии, расчет электронной плотности становится принципиально затрудненным. Эта утрата фазовой информации, известная как фазовая проблема, представляет собой центральную проблему рентгеновской кристаллографии, и в последующих разделах будут рассмотрены основные методы её решения.

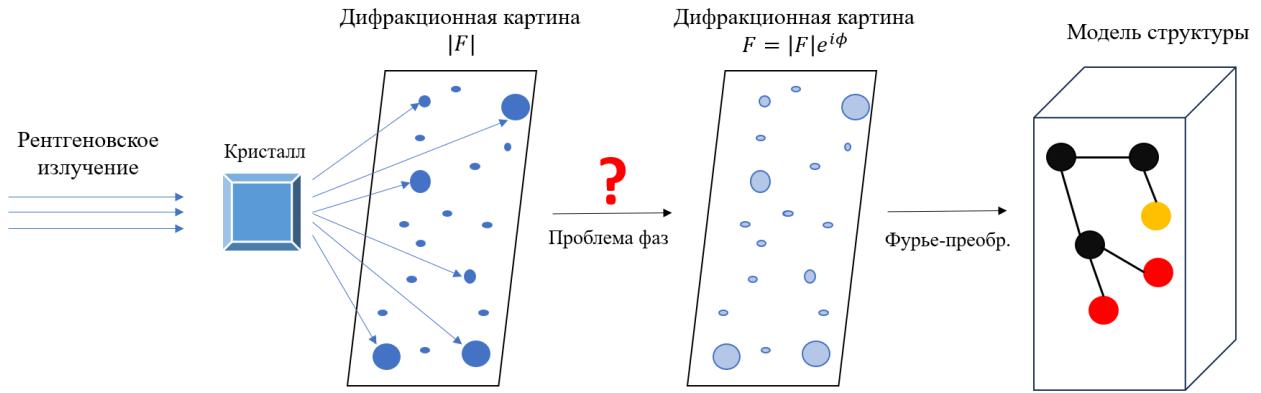


Рисунок 3. Схема определения кристаллической структуры

1.3 Решение фазовой проблемы

1.3.1 Методы проб и ошибок

В 1913 году физиком Уильямом Генри Брэггом и его сыном Уильямом Лоренсом Брэггом, были впервые описаны структуры таких веществ, как NaCl, KCl, KBr и KI, с помощью дифракции рентгеновских лучей на кристаллах [2]. Для расшифровки структур тогда использовали так называемые методы проб и ошибок (trial-and-error), которые основаны на достаточно простой идеи. Вместо того, чтобы анализировать дифракционную картину, ученые предполагали структуру из известной информации о веществе. Методы проб и ошибок могут использовать следующую информацию для формирования модели структуры: возможные пространственные группы симметрии, число формульных единиц в элементарной ячейке Z , симметрия молекулы, структуры подобных кристаллов и соединений, а также физические свойства кристалла. После этого рассчитывали структурные факторы и интенсивности с помощью обратного преобразования Фурье (уравнение 3) и сравнивали полученный набор отражений с экспериментальным. Если наборы данных соответствуют друг другу, то предполагаемую структуру можно считать похожей на истинную и уточнить её, изменяя параметры, пока не будет достаточное совпадение с наблюдаемой дифракционной картиной. В ином случае предположение отвергалось и проверялась другая структура.

Несмотря на кажущуюся простоту, методы проб и ошибок актуальны по сей день. Так, методы молекулярного и изоморфного замещения являются одними из основных способов определения структур белковых соединений, поскольку их кристаллы слабо рассеивают и полученного набора рентгенодифракционных максимумов недостаточно для решения методами, эффективными для низкомолекулярных соединений [3], [4].

Основаны они на дополнительном знании о структурах белков с той же аминокислотной последовательностью (молекулярное замещение) или результатах дифракционных экспериментов структуры того же белка с тяжелыми атомами (изоморфное замещение). Информация используется для построения начальной структуры, которая затем подбирается и уточняется.

Также нельзя не упомянуть важный результат эволюции методов проб и ошибок — программное обеспечение FOX (Free Objects for Xtallography) [5]. В ней имплементирована глобальная оптимизация структуры с помощью алгоритма Монте–Карло. Бесплатная программа с открытым доступом FOX сейчас широко используется в порошковой рентгеновской или нейтронной дифракции для определения структур кристаллов [6], [7].

1.3.2 Прямые методы

В работе [8] впервые было предложено решение проблемы фаз для центросимметричной группы $P\bar{1}$, которое требует только знание амплитуд структурных факторов и химического состава кристалла. Данный подход, основанный на вероятностном подходе и предположении, что в распределение амплитуд структурного фактора уже заложена информация о фазах, в дальнейшем был распространен на все центросимметричные группы, а затем адаптирован для нецентросимметричных структур. В дальнейшем этот метод и подобные были названы прямыми, их объединяют статистические взаимосвязи между двумя, тремя и четырьмя сильными отражениями — структурные инварианты и полуинварианты [9].

Для того, чтобы избавиться от явной зависимости структурного фактора от угла рассеяния, авторы заменяют реальный кристалл с электронной плотностью $\rho(\vec{r})$ на идеальный, элементарная ячейка которого состоит из дискретных неподвижных точечных атомов, которые расположены в максимумах электронной плотности. Тогда структурный фактор $F(\vec{H}) = |F(\vec{H})| \exp(i\phi(\vec{H}))$ следует заменить на нормализованный структурный фактор $E(\vec{H}) = |E(\vec{H})| \exp(i\phi(\vec{H}))$ [10]:

$$|E|^2 = \frac{|F|^2}{\langle |F|^2 \rangle} \quad (5)$$

В уравнении 5 параметр $\langle |F|^2 \rangle$ есть математическое ожидание (среднее) квадрата амплитуды структурного фактора, для расчета которого нужна *a priori* информация. Существует множество способов вычисления нормализованных структурных факторов, которые исходят из количества доступных данных, приведём некоторые из них [11]:

1. Если отсутствует структурная информация, то координаты атомов следует принять случайными. Пусть ϵ — некоторый параметр, зависящий от группы симметрии, тогда:

$$\langle |F|^2 \rangle = \epsilon \sum_{j=1}^N f_j^2 \quad (6)$$

2. Если известны M групп по M_i атомов в каждой, конфигурация которых известна, но неизвестны ориентация и позиция самих групп. Пусть некоторое количество межатомных расстояний $r_{j_1 j_2}$ известно, тогда для отражения \vec{H} :

$$\langle |F|^2 \rangle = \epsilon \left[\sum_{j=1}^N f_j^2 + \sum_{i=1}^M \sum_{j_1 \neq j_2=1}^{M_i} f_{j_1} f_{j_2} \frac{\sin(2\pi |\vec{H}| r_{j_1 j_2})}{2\pi |\vec{H}| r_{j_1 j_2}} \right] \quad (7)$$

3. Если известны M групп по M_i атомов в каждой, конфигурация и ориентация которых известны, но неизвестна позиция самих групп. Пусть некоторое количество межатомных расстояний $r_{j_1 j_2}$ зафиксированы, тогда для отражения \vec{H} :

$$\langle |F|^2 \rangle = \epsilon \left[\sum_{j=1}^N f_j^2 + \sum_{i=1}^M \sum_{j_1 \neq j_2=1}^{M_i} f_{j_1} f_{j_2} \exp(2\pi i(\vec{H}, \vec{r}_{j_1 j_2})) \right] \quad (8)$$

4. Если известны M групп атомов и их позиции, тогда:

$$\langle |F|^2 \rangle = |F_M|^2 + \epsilon \sum_{i=1}^Q f_i^2, \quad (9)$$

где F_M — структурный фактор известной подструктуры,, Q — количество неизвестных атомов.

Однако чтобы рассчитать нормализованные структурные факторы по уравнению 5 требуется учесть, что величины должны быть в абсолютной шкале, а наблюдаемые амплитуды являются относительными величинами. Также в формулах для $\langle |F|^2 \rangle$ выше никак не учтено тепловое движение атомов. Оба обстоятельства можно преодолеть с использованием графика Вилсона [12], согласно которому наблюдаемые данные разделяются на несколько промежутков по переменной $s = (\frac{\sin\theta}{\lambda})^2$, в каждом из которых вычисляется средняя интенсивность $\langle I_{obs} \rangle = \langle |F_{obs}|^2 \rangle$. Для каждого промежутка можно вычислить $K \langle I \rangle$, где K — параметр, который необходим для перевода интенсивности рентгеновского излучения в абсолютные величины:

$$K \langle I \rangle = \langle |F_{obs}|^2 \rangle \exp(-2Bs^2), \quad (10)$$

где B — термический параметр, $\langle |F_{obs}|^2 \rangle$ вычисляется по уравнениям 6, 7, 8, 9.

Чтобы найти параметры B и K , уравнение 10 логарифмируют, строят линейный график по уравнению 11 в координатах $(s^2, \ln \frac{\langle I \rangle}{\langle |F_{obs}|^2 \rangle})$ и получают параметры B и K после аппроксимации.

$$\ln \frac{\langle I \rangle}{\langle |F_{obs}|^2 \rangle} = -\ln K - 2Bs^2 \quad (11)$$

Таким образом, после построения графика Вилсона и нахождения нужных параметров, нормализованные структурные факторы вычисляются по итоговой формуле:

$$|E|^2 = \frac{KI_{obs}}{\langle |F_{obs}|^2 \rangle \exp(-2Bs^2)} \quad (12)$$

Также известны [11] плотности распределения полученных нормализованных структурных факторов, которые различаются в зависимости от симметрии кристаллической структуры (уравнения 13, 14), их вид представлен на рис. 4. Нетрудно показать, что первый момент $|E|$ независимо от структуры равен 1, а математическое ожидание величины $(E^2 - 1)^2$, которое можно рассчитать по уравнению 15, для центросимметричных кристаллов больше (2.0), чем для нецентросимметричных (1.0), что можно использовать как критерий центросимметричности структуры.

$$\text{Центросимметричная: } P(|E|) = \sqrt{\frac{2}{\pi}} \exp\left(-\frac{E^2}{2}\right) \quad (13)$$

$$\text{Нецентросимметричная: } P(|E|) = 2|E| \exp(-|E|^2) \quad (14)$$

$$\langle (E^2 - 1)^2 \rangle = \int_0^\infty P(E)(E^2 - 1)^2 dE \quad (15)$$

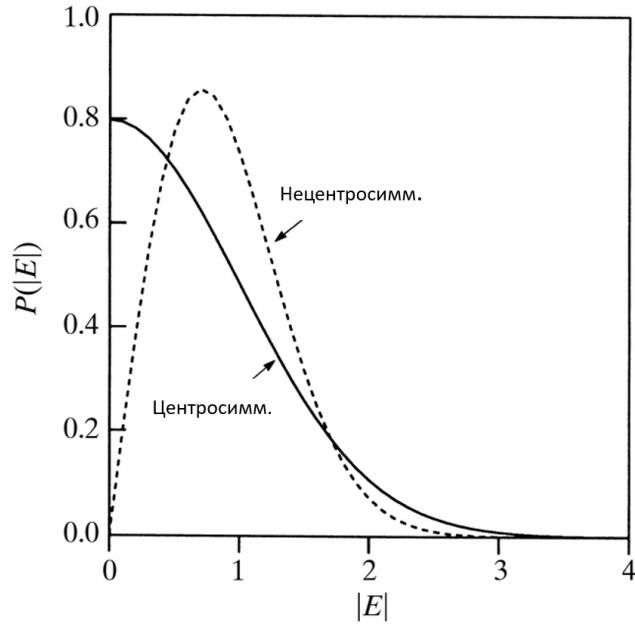


Рисунок 4. Плотность распределения $|E|$ для центросимметричных и нецентросимметричных кристаллов [11]

Прямые методы также основаны на структурных инвариантах — комбинации фаз рентгеновских отражений, сумма значений которых не меняется при сдвиге начала координат. Приведём некоторых из них. В работах [13], [14] была показана следующая взаимосвязь между высокоинтенсивными отражениями:

$$s(h + h', k + k', l + l') \approx s(h, k, l)s(h', k', l') \quad (16)$$

$$\phi(h + h', k + k', l + l') \approx \phi(h, k, l) + \phi(h', k', l') \quad (17)$$

Из закона Фриделя следует, что:

$$\phi(-h - h', -k - k', -l - l') = -\phi(h + h', k + k', l + l') \quad (18)$$

Тогда, сложив уравнения 18 и 17 мы получим выражение, которое получило название триплетное отношение:

$$\phi(h, k, l) + \phi(h', k', l') + \phi(-h - h', -k - k', -l - l') \approx 0 \quad (19)$$

Из данного уравнения и закона Фриделя так же нетрудно получить выражение для квартета отражений:

$$\phi(h, k, l) + \phi(h', k', l') + \phi(h'', k'', l'') + \phi(-h - h' - h'', -k - k' - k'', -l - l' - l'') \approx 0 \quad (20)$$

Также нельзя не упомянуть важное равенство, которое является ключевым для прямых методов, а также нашло применение в других способах решения фазовой проблемы — формулу тангенсов, которая позволяет рассчитать фазу отражения [15]:

$$\tan \phi(\vec{H}) = \frac{\sum_{\vec{K}} |E(\vec{K})E(\vec{H} - \vec{K})| \sin(\phi(\vec{K}) + \phi(\vec{H} - \vec{K}))}{\sum_{\vec{K}} |E(\vec{K})E(\vec{H} - \vec{K})| \cos(\phi(\vec{K}) + \phi(\vec{H} - \vec{K}))} \quad (21)$$

На сегодняшний день прямые методы являются одним из наиболее используемых и популярных подходов для решения проблемы фаз в рентгеновской кристаллографии, незаменимыми для исследования малых и средних молекул. Основываясь на анализе статистических соотношений между нормализованными структурными факторами, эти методы позволяют оценить вероятные значения фаз с помощью структурных инвариантов, что делает возможным восстановление электронной плотности без дополнительных экспериментальных данных. Подробности расчётов и имплементации метода достаточно громоздки и достойны отдельного изучения [10].

Прямые методы непрерывно совершенствуются: современные алгоритмы интегрируют классическую формулу тангенсов для оценки фаз с передовыми техниками модификации электронной плотности [16]. Такой гибридный подход позволяет не только получить более надёжные фазы благодаря доказанной статистической основе формулы тангенсов, но и существенно повысить качество и скорость сходимости решения за счёт итеративного улучшения электронной плотности в прямом пространстве. В результате объединения этих подходов удается достичь значительного повышения эффективности и надёжности решения проблемы фаз даже для сложных кристаллических структур.

1.3.3 Метод Паттерсона

Функция Паттерсона $P(\vec{u})$ представляет собой автокорреляционную функцию (свёртку функции с собой) электронной плотности [1]:

$$P(\vec{u}) = \int_V \rho(\vec{r})\rho(\vec{r} + \vec{u})d\vec{r}, \quad (22)$$

где \vec{u} — вектор координат ячейки Паттерсона, которая совпадает по размерам с

обычной, V — объем кристаллической решетки в прямом пространстве.

Из свойств автокорреляционной функции прямо следует, что функция Паттерсона $P(\vec{u})$ принимает ненулевые значения тогда и только тогда, когда электронная плотность принимает ненулевые значения в точке $\vec{r} = (x, y, z)$ и точке $\vec{r} + \vec{u} = (x+u, y+v, z+w)$, то есть в этих точках расположены атомы. Пики функции Паттерсона достигаются в таких точках, которые отвечают межатомным векторам кристалла.

Из экспериментальных данных функция Паттерсона рассчитывается следующим образом:

$$P(\vec{u}) = \frac{1}{V} \sum_{\vec{H}} |F|^2(\vec{H}) \exp(-2\pi i(\vec{H}, \vec{u})) \quad (23)$$

Также можно выделить следующие свойства функции Паттерсона:

- Функция Паттерсона всегда чётная, поскольку для каждой пары атомов существует пара межатомных векторов: $P(\vec{u}) = P(-\vec{u})$.
- Карта Паттерсона (графическое представление функции) обладает той же симметрией, что и группа Лауэ кристалла.
- Карта Паттерсона всегда имеет большой пик в начале координат — явно следует из определения.
- Максимумы функции широкие и размазанные из-за перекрывания электронной плотности атомов.

Проанализируем количество максимумов функции Паттерсона [17]. Пусть в элементарной ячейке N атомов с атомными факторами рассеяния f_j . Из определения структурного фактора (уравнение 2) можно получить следующее выражение для $|F|^2$:

$$|F(\vec{H})|^2 = F(\vec{H})F^*(\vec{H}) = \left[\sum_{j=1}^N f_j \exp(2\pi i(\vec{H}, \vec{r}_j)) \right] \left[\sum_{j=1}^N f_j \exp(-2\pi i(\vec{H}, \vec{r}_j)) \right] \quad (24)$$

$$|F(\vec{H})|^2 = \sum_{j=1}^N f_j^2 + \sum_{j=1}^N \sum_{i=1, i \neq j}^N f_i f_j \exp(2\pi i(\vec{H}, \vec{r}_j - \vec{r}_i)) \quad (25)$$

Таким образом, объединив уравнения 25 и 23, получаем, что функция Паттерсона является суммой N^2 межатомных взаимодействий, из которых N в начале координат с весами f_j^2 и $N(N - 1)$ попарных взаимодействий, пропорциональных $f_i f_j$. Таким образом, в карте Паттерсона $N(N - 1)$ максимумов, которые зависят от межатомных

векторов и типов атомов в ячейке. Нетрудно показать, что интенсивность этих пиков пропорциональна атомным номерам соответствующих атомов Z_i (уравнение 26, m — фактор мультиплетности). Поскольку функция Паттерсона является четной, для описания уникальных попарных взаимодействий достаточно $N(N-1)/2$ значений, поэтому часто используемым для представления данных является верхнетреугольная матрица.

$$P(\vec{u}_{ij}) = \frac{m Z_i Z_j}{\sum_{j=1}^N f_j^2} \quad (26)$$

Поскольку N максимумов карты Паттерсона, отвечающим свёртке электронной плотности каждого атома с самим собой, являются не очень информативными, от них можно избавиться, изменив структурные факторы перед расчётом функции Паттерсона по уравнению 27. Аналогично, если положения каких-то атомов заранее точно известны, пики, отвечающим их межатомным векторам, можно убрать благодаря модификации $|F(\vec{H})|^2$ в уравнении 28, где r_1, r_2 — позиции известных атомов 1 и 2, σ_1, σ_2 — их коэффициенты теплового движения.

$$|F_{\text{mod}}(\vec{H})|^2 = |F(\vec{H})|^2 - \sum_{j=1}^N f_j^2 \quad (27)$$

$$|F_{\text{mod}}(\vec{H})|^2 = |F(\vec{H})|^2 - \sum_{j=1}^N f_j^2 \sigma_j^2 - 2f_1 f_2 \sigma_1 \sigma_2 \cos(2\pi(\vec{H}, \vec{r}_1 - \vec{r}_2)) \quad (28)$$

Для борьбы с шириной пиков, которые могут сильно повлиять на корректность определения фаз дифракционных отражений, используют процедуру сужение карты Паттерсона. Для этого вместо структурных факторов и их интенсивностей $|F|^2$ используют нормализованные структурные факторы $|E|^2$. Поскольку интенсивность нормализованных структурных факторов не так сильно падает с увеличением угла отражения θ благодаря поправочному множителю, полученный набор получен как бы от атомов меньшего размера, в результате чего максимумы Паттерсона также будут более узкие.

Карту Паттерсона можно представить суммой нескольких копий исходной структуры с разными весами [18], которые различаются тем, какой атом находится в начале паттерсоновских координат (u, v, w) (рис. 5). Для простых структур низкомолекулярных соединений возможно расшифровать карту, сопоставив каждому максимуму межатомный вектор. Однако количество пиков функции Паттерсона для структуры из N атомов

растёт по квадратичному закону, что делает невозможным прямую интерпретацию без дополнительных операций над картой.

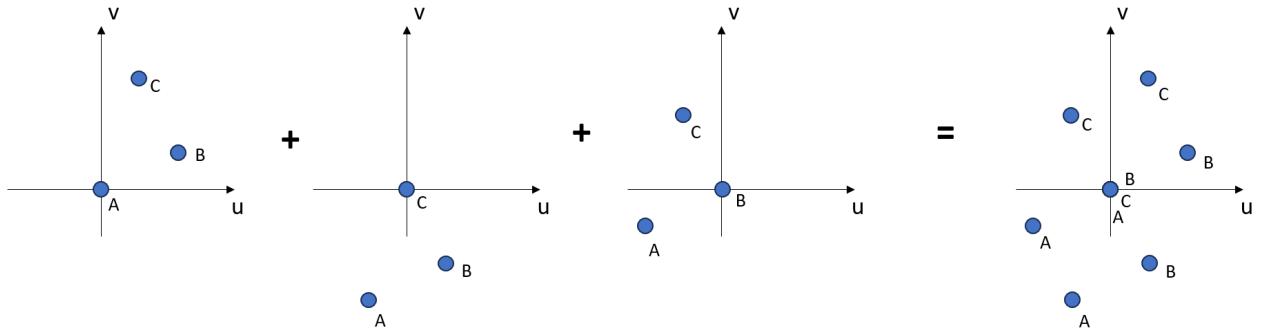


Рисунок 5. Схема карты Паттерсона для простейшей структуры из 3 атомов

С развитием компьютерных технологий во второй половине 20 века произошел расцвет методов, основанных на функции Паттерсона. Так, особенно полезным для решения фазовой проблемы оказался метод суперпозиции карт Паттерсона [19]. Было показано, что суперпозиция исходной карты Паттерсона со смещенной на какой-то трансляционный вектор может привести к меньшему набору межатомных векторов (рис. 6), поскольку точки, соответствующие образу исходной структуры, будут всегда повторяться. Суперпозиция представляет собой, например, пересечение или сумму наборов векторов (см. далее). Через множество применений данной процедуры можно получить исходную структуру.

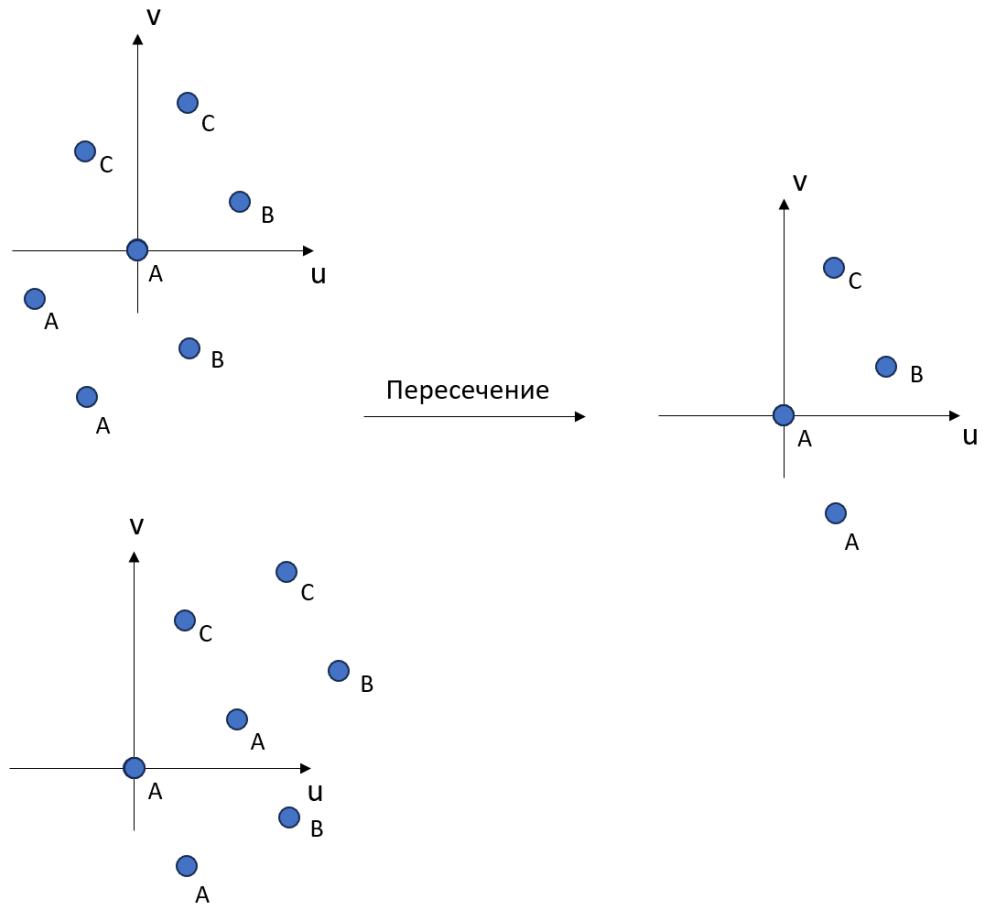


Рисунок 6. Суперпозиция (пересечение) исходной и смещенной на вектор \overrightarrow{AB} карт Паттерсона

Есть три основных подхода, которые используются для выделения структуры молекулы из наложенных карт Паттерсона — с помощью функций суммы, произведения и минимума суперпозиции [17]. Пусть есть N карт Паттерсона, каждая из которых смещена на вектор \vec{u}_i . Функция суммы (уравнение 29) основывается на том, что она будет иметь наибольшие значения в точках пересечения всех карт, что будет отвечать расположениям атомов.

$$S(\vec{r}) = \sum_{i=1}^N P(\vec{r} + \vec{u}_i) = \sum_{\vec{H}} \left[|F(\vec{H})|^2 \exp(2\pi i(\vec{H}, \vec{r})) \left(\sum_{j=1}^N \exp(2\pi i(\vec{H}, \vec{u}_i)) \right) \right] \quad (29)$$

Функция произведения суперпозиции (уравнение 30) является более сильной по сравнению с суммой в том смысле, что все точки, в которых нулевое значение хотя бы у одной из карт, будут обнулены.

$$\Pr(\vec{r}) = \prod_{i=1}^N P(\vec{r} + \vec{u}_i) \quad (30)$$

Если при наложении двух карт в точке ненулевая плотность, то суммарное значение функции Паттерсона будет больше, чем от одного изображения структуры. Если взять минимум от накладывающихся карт, то при удачной суперпозиции останется лишь плотность от одной копии структуры [20]. Так, в SHELXS-96 [21] имплементирован вариант суперпозиции карт Паттерсона с функцией минимума, которая берётся от двух копий утонченной функции Паттерсона, смещенных на векторы \vec{u} и $-\vec{u}$.

1.3.4 Метод обращенного заряда (charge-flipping)

В работе [22] был предложен простой, но эффективный алгоритм — метод обращенного заряда (charge-flipping), использующий свойства электронной плотности, а именно её неотрицательность. Данный алгоритм основан на итеративном приближении фаз дифракционных максимумов, которые в начале инициализируются случайными, к настоящим фазам, что позволяет решить фазовую проблему и определить кристаллическую структуру по дифракционным данным.

Пусть в эксперименте был зарегистрирован набор дифракционных отражений, которые характеризуются амплитудами $F_{obs}(\vec{H})$. Незарегистрированные отражения в других точках обратного пространства принимаются за нулевые. Алгоритм начинается с инициализации фаз зарегистрированных отражений, которые выбираются случайным образом так, чтобы выполнялся закон Фриделя: $\phi(-\vec{H}) = -\phi(\vec{H})$. Получаем набор структурных факторов $F = F_{obs}(\vec{H}) \exp(i\phi(\vec{H}))$.

Запишем основные шаги алгоритма:

1. Для набора структурных факторов $F = F_{obs}(\vec{H})$ рассчитаем распределение электронной плотности $\rho(\vec{r})$, применив преобразование Фурье.
2. Модифицируем электронную плотность $\rho(\vec{r})$ с помощью обращения заряда:

$$\rho(\vec{r}) \geq \delta : g = \rho, \quad (31)$$

$$\rho(\vec{r}) < \delta : g = -\rho, \quad (32)$$

где $\delta > 0$ — пороговое значение, параметр алгоритма.

3. Переходим в обратное пространство, применив обратное преобразование Фурье к модифицированной электронной плотности $g(\vec{r})$. Получаем набор структурных факторов $G(\vec{H}) = |G(\vec{H})| \exp(i\psi(\vec{H}))$.
4. Последнее преобразование называется ремодуляция. Собираем новый набор структурных факторов F следующим образом: $F = F_{obs}(\vec{H}) \exp(i\psi(\vec{H}))$ — амплитуды равняются экспериментальным, фазы берутся из набора G , полученного в ходе обращения заряда. Вернуться к шагу 1.

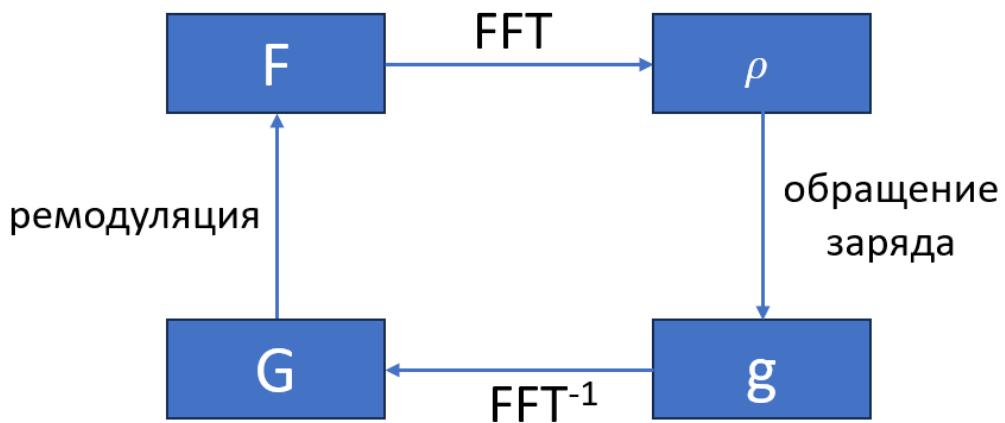


Рисунок 7. Схема итеративного цикла алгоритма charge-flipping

В итерационный процесс (рис. 7) не заложено условие окончания алгоритма, для мониторинга процесса можно следить за метриками, такими как фактор расходимости R . Главными достоинствами метода являются его простота и отсутствие необходимости в дополнительных данных — метод позволил получить структуры множества соединений *ab initio* при наличии лишь экспериментальных дифракционных данных высокого разрешения, задавая только пороговое значение δ для электронной плотности.

В следующей работе [23] авторы модифицировали четвертый шаг метода, явно используя отражения низкой интенсивности. Перед запуском алгоритма дифракционные максимумы сортируются по наблюдаемой амплитуде и размечаются на две группы — сильные и слабые отражения, к которым будут применяться разные преобразования в ходе вычислений. Для сильных всё остается без изменений — фазу берут из шага 2, амплитуда структурного фактора равняется экспериментальной. Фаза слабых отражений дополнительно смещается на $\Delta\phi$, а их амплитуда не изменяется на наблюдаемую. Это значит, что экспериментальные данные слабых отражений не используются в алгоритме, кроме начальной разметки на группы.

В модифицированном алгоритме charge-flipping появилось два дополнительных параметра: сдвиг фазы $\Delta\phi$ и доля отражений, которые можно считать слабыми, принятая за 20%. Авторы показали, что оптимальное значение $\Delta\phi = \frac{\pi}{2}$ — поскольку при сдвиге на такую величину волна слабого отражения становится ортогональной изначальной. Такое возмущение волн распространения низкоинтенсивных максимумов позволило на порядок увеличить долю успешных решений структур по сравнению с начальным алгоритмом.

Описанная модификация происходит в обратном пространстве, и в работе [24] авторы описывают подход к улучшению алгоритма с помощью изменения процедуры обработки функции электронной плотности. Charge-flipping в его классическом варианте можно рассматривать как алгоритм локального возмущения низких плотностей, то есть он никак не влияет на области с высокими значениями электронной плотности. Если в начале выполнения расчета будут получены некорректные высокие значения электронной плотности, то их практически невозможно исправить тривиальным обращением знака. В качестве аналога использования слабых отражений предлагается следующее улучшение в прямом пространстве (уравнения 33, 34). Суть операции в следующем: измененная электронная плотность g^{n+1} в $(n + 1)$ -й цикл алгоритма будет вне интервала, сформированными ρ^{n-1}, ρ^n , причем она будет ближе к ρ^n . Процедура получила название 'flip-mem' и значительно улучшает классический алгоритм.

$$\rho^n(\vec{r}) < \delta : g^{n+1}(\vec{r}) = -\rho^n(\vec{r}), \quad (33)$$

$$\rho^n(\vec{r}) \geq \delta : g^{n+1} = \rho^n + \beta(\rho^n - \rho^{n-1}), \quad (34)$$

где обычно $\beta \in [0.5, 1.0]$.

В этой же работе отмечено, что использование нормализованных структурных факторов $E(\vec{H})$ (уравнение 5) вместо стандартных амплитуд позволяет увеличить скорость сходимости для больших структур. Кроме того, все модификации, которые были ранее упомянуты, подходят и для варианта с нормализованными амплитудами. Использование $E(\vec{H})$ уменьшает на порядок число итераций до сходимости по сравнению с $F(\vec{H})$ для всех вариантов алгоритма.

Метод обращенного заряда позволяет в некоторых случаях решать макромолекулярные структуры *ab initio* [25]. Так, для таких структур, как лизоцим (2385 атомов, $d_{min} = 1.1\text{\AA}$), алкогольдегидрогеназа (5866 атомов, $d_{min} = 1.0\text{\AA}$), апамин (385 атомов, $d_{min} = 0.95\text{\AA}$), фазы, рассчитанные алгоритмом, позволили корректно

определить структуру. Авторы использовали метод в варианте с нормализованными структурными факторами $E(\vec{H})$, использованием слабых отражений (доля слабых отражений $\omega = 0.1$). Также авторы продемонстрировали, что charge-flipping является эффективным инструментом для нахождения фаз для рентгенодифракционных данных низкого разрешения сложных структур с тяжелыми атомами, а также с значимым аномальным рассеянием.

В работе [26] авторы представили свой вариант алгоритма обращенного заряда с использованием нормализованных структурных факторов и новым способом возмущения, который применяет знание о формуле тангенсов, связывающей три отражения:

$$\tan \phi_{tf}(\vec{H}) = \frac{\sum_{\vec{K}} |E(\vec{K})E(\vec{H})E(\vec{H} - \vec{K})| \sin(\phi(\vec{K}) + \phi(\vec{H} - \vec{K}))}{\sum_{\vec{K}} |E(\vec{K})E(\vec{H})E(\vec{H} - \vec{K})| \cos(\phi(\vec{K}) + \phi(\vec{H} - \vec{K}))} \quad (35)$$

Итоговый алгоритм (после случайной расстановки фаз для зарегистрированных отражений) выглядит следующим образом:

1. Обнулить 50% структурных факторов $E(\vec{H})$ с самыми низкими значениями амплитуд. Модуль остальных структурных факторов равен наблюдаемым.
2. Рассчитать электронную плотность $\rho(\vec{r})$ с помощью обратного преобразования Фурье.
3. Определить граничное значение δ так, что 60% значений $\rho(\vec{r})$ лежат ниже порога.
4. Преобразовать плотность:

$$\rho < \delta : g = -\rho \quad (36)$$

$$\rho \geq \delta : g = \delta + (\rho - \delta)^{1/2} \quad (37)$$

5. Рассчитать набор структурных факторов G , применив преобразование Фурье к электронной плотности $g(\vec{r})$
6. Добавить к фазам с наибольшими значениями $E(\vec{H})$ долю разницы между фазами, полученной обратным зарядом ϕ_{cf} и рассчитанной по формуле тангенсов ϕ_{tf} :

$$\phi(\vec{H}) = \phi_{\text{cf}}(\vec{H}) + \alpha(\vec{H})(\phi_{\text{tf}}(\vec{H}) - \phi_{\text{cf}}(\vec{H})), \quad (38)$$

где $\alpha(\vec{H})$ — параметр, рассчитываемый алгоритмом для каждого отражения, который отражает достоверность рассчитанных в ходе выполнения программы фаз. Вернуться к шагу 1.

Авторы продемонстрировали, что предложенный вариант метода позволяет решать за несколько минут структуры, которые ранее требовали сотни тысяч итераций. Использование формулы тангенсов для высокоинтенсивных отражений позволило повысить устойчивость и эффективность алгоритма, особенно при работе с низким разрешением данных (более $d_{\min} = 1\text{\AA}$), где классический вариант алгоритма чаще всего терпит неудачу.

В работе [27] был проведен сравнительный анализ метода обращенного заряда с другими стандартными инструментами рутинного определения кристаллографических структур низкомолекулярных соединений. Для этого автор использовал charge-flipping в реализации программы SUPERFLIP [28] и традиционные прямые методы — SHELXS, SHELX86, SHELXD, SIR2004 [29]. Тестирование проводилось на наборе данных из 518 структур, включая органические, металлоганические и неорганические соединения. Метод обращенного заряда показал эффективность, сравнимую с прямыми методами, в среднем процент успешного решения структур составляет более 92%. На рис. 8 представлены значения среднего R-фактора, достигаемого в ходе уточнения после успешного решения структур каждым методом. Также в статье показано, что charge-flipping является самым быстрым методом из рассматриваемых при той же эффективности решения. Таким образом, рассматриваемый алгоритм является полностью пригодным для определения низкомолекулярных структур в качестве рутинного метода.

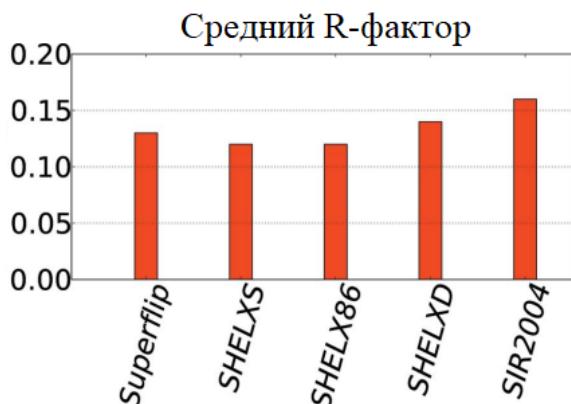


Рисунок 8. Средний R-фактор успешных решений различных методов решения

1.3.5 Метод Vive La Différence (VLD)

Следующий рассматриваемый метод под названием Vive La Différence (VLD) является эффективным и универсальным методом для определения кристаллических структур, основанным на преобразовании электронной плотности и разностном Фурье–синтезе исходя из вероятностных и статических зависимостей между фазами и амплитудами [30]. Как и charge-flipping, метод является итеративным, но более комплексным, чем метод обращенного заряда. Перед стартом расчёта фаз всем наблюдаемым амплитудам нормализованных структурных факторов E_{obs} присваиваются случайные фазы ϕ_m и рассчитывается электронная плотность кристаллической ячейки ρ_m . Здесь и далее величины, определяемые в ходе выполнения расчёта, назовём модельными. Алгоритм состоит из следующих шагов (схема представлена на рис. 9):

1. Модифицируем электронную плотность $\rho(\vec{r})$ следующим образом: обнуляем функцию во всех точках, кроме 2.5% с наибольшим значением плотности. По полученной измененной электронной плотности $g(\vec{r})$ рассчитываем модельные нормализованные структурные факторы E_m с помощью Фурье–преобразования.
2. Вычисляем разностную электронную плотность — синтез Фурье со следующими коэффициентами для каждого отражения:

$$\Delta E = (mE_{obs} - E_m) \exp(i\phi_m) \quad (39)$$

где m — коэффициент корреляции между модельными и реальными фазами.

3. Модифицируем полученную разностную плотность: обнуляем функцию во всех точках, кроме 4% с наибольшим значением по модулю. С помощью Фурье–преобразования рассчитываем разностные структурные факторы с амплитудами и фазами E_{diff}, ϕ_{diff} .
4. Используем полученные из разностной электронной плотности структурные факторы для расчёта коэффициента Фурье–синтеза (шаг 2), рассчитываем новую разностную плотность. Повторяем настоящий и предыдущий шаги β раз, итоговые параметры структурных факторов E'_{diff}, ϕ'_{diff} .
5. С помощью формулы тангенсов, которая была получена из вероятностного анализа и распределения фон Мизеса, можно рассчитать новые фазы ϕ_{calc} :

$$\tan(\theta) = \frac{E_m \sin \phi_m + \omega_{\text{diff}} E_{\text{diff}} \sin \phi_{\text{diff}}}{E_m \cos \phi_m + \omega_{\text{diff}} E_{\text{diff}} \cos \phi_{\text{diff}}} \quad (40)$$

где ω_m — параметр сходства модельной и реальной структур, отражающий среднюю корреляцию между моделью и структурой.

6. Экспериментальные E_{obs} и рассчитанные на предыдущем шаге фазы ϕ_{calc} используются для расчета электронной плотности, которая претерпевает γ циклов модификации, как в шаге 1.
7. Рассчитывается метрика $RESID$ — средняя ошибка между наблюдаемыми и рассчитанными амплитудами:

$$RESID = \frac{\sum_{\vec{H}} |E_{\text{obs}} - E_m|}{\sum_{\vec{H}} E_{\text{obs}}} \quad (41)$$

Если значение меньше 0.3 — можно считать, что решение достигнуто и цикл останавливается. Иначе — возвращение к шагу 1.

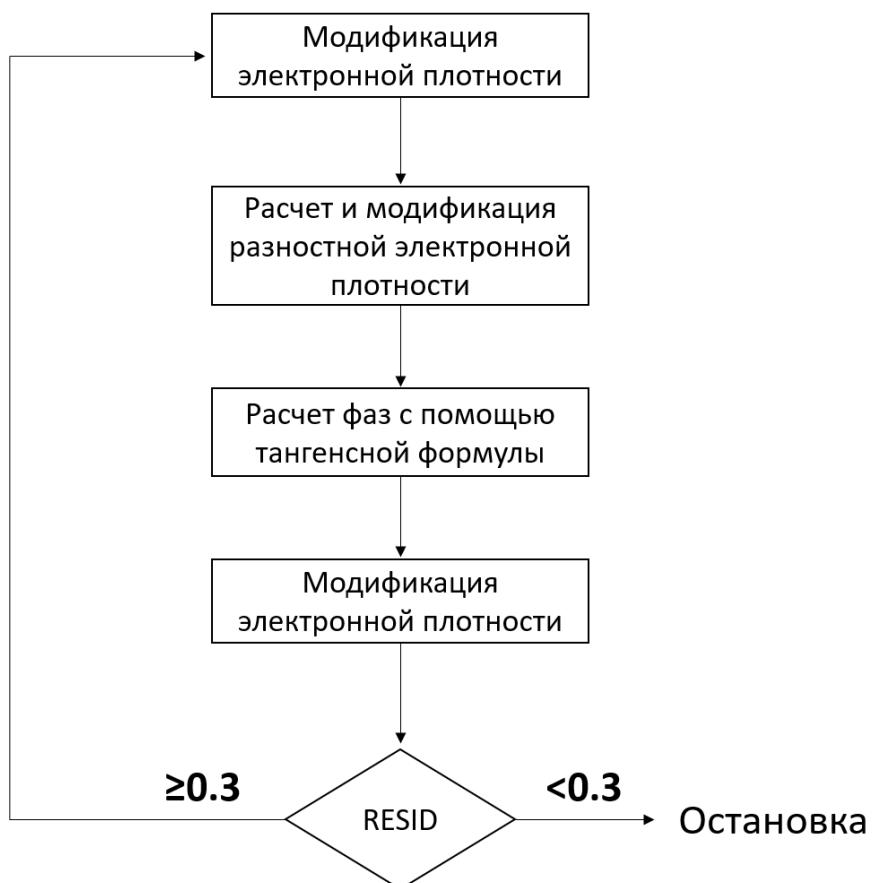


Рисунок 9. Схема итеративного цикла алгоритма VLD

Подробности процедур выбора настраиваемых параметров алгоритма и расчета критических параметров, как ω_m и m , подробно описаны в оригинальной работе. Авторы протестирували предложенный метод на 33 низкомолекулярных структурах, разных по пространственной группе симметрии и наличию тяжелых атомов, из которых VLD успешно решил 30 структур за не более, чем 300 циклов. При дополнительных запусках с другими случайными фазами при инициализации модели привели к решению всего набора данных. Таким образом, показано, что метод является достаточно стабильным и обладает быстрой сходимостью (менее, чем 1 минута). Важно отметить, что в отличие от метода обращенного заряда, метод решает структуру в правильной группе симметрии, а не P1.

В следующей работе [31] авторы развивают идеи первой работы и описывают существенные усовершенствования оригинального подхода к решению кристаллографической фазовой проблемы. Ключевым усовершенствованием является внедрение процедуры RELAX, которая позволяет автоматически переместить правильно ориентированную, но смещённую модель в корректное начальное положение в пространственной группе. RELAX основана на наблюдении, что прямые методы определяет молекулярные фрагменты, которые корректно ориентированы, но неправильно расположены. Использование процедуры существенно повысило процент успешно решённых структур, особенно в случае макромолекулярных соединений. Также в статье представлено значительное количество улучшений, связанные с улучшенной оценкой параметров качества модели, оптимизацией этапов модификаций электронной плотности, а также дополнительные параметры качества решений для управления остановкой VLD. Тестирование на малых, средних молекулах и белках (разрешение до $d_{min} = 1.2\text{\AA}$) показало, что доработанный алгоритм решает структуры быстрее в среднем в 3–6 раз, определяет структуры белков за время, сравнимое с прямыми методами (SIR2011), а добавление RELAX позволило находить решения для крупных молекул именно благодаря этой процедуре.

В публикации [32] представлено множество дополнительных вариантов алгоритма VLD (4 протокола), которые ориентированы на решение проблемы фаз для среднемолекулярных и белковых соединений. Вариации метода различаются подходами к контролю метрик сходимости и внутренних параметров, а также способами комбинирования модели, разностной и обычной электронных плотностей (количество преобразований Фурье, использование или игнорирование экспериментальных данных, а также отказ от формулы тангенсов). В ходе оценки эффективности показано, что новые подходы уступают по успешности и скорости решения варианту с процедурой RELAX,

но могут быть в дальнейшем оптимизированы для *ab initio* решения макромолекулярных соединений.

1.3.6 Методы искусственного интеллекта

Применение методов машинного обучения в рентгенодифракционных исследованиях — область, демонстрирующая стремительное развитие. Традиционно анализ дифракционных данных и решение обратной задачи в кристаллографии базировались на строго детерминированных алгоритмах, использующих априорные физические и химические знания о структуре вещества. Однако с ростом объёмов экспериментальных данных, увеличением доступной вычислительной мощности и успехами в смежных областях — таких как обработка изображений, анализ временных рядов и предсказание структур белков — возник интерес к использованию ИИ как инструмента для автоматизации и улучшения интерпретации дифракционных данных.

Первая публикация с решением фазовой проблемы методами глубокого обучения была опубликована почти год назад [33]. Для предсказания были выбраны центросимметричные структуры (группы симметрии $P2_1/c, C2/c, Pbc\alpha, Pnma, Pb\bar{c}n, C2/m$), фазы дифракционных максимумов которых принимают два возможных значения — 0 и π [34]. Для обучения авторы сгенерировали 49 млн. синтетических структур, подавляющее большинство которых — органические и небольшая доля металлоганических. Нейронная сеть PhAI представляет собой бинарных классификатор из блоков трехмерных свёрток и многослойных перцептронов (рис. 10). Также в архитектуре сети представлен рециклинг фаз — данные прогоняются несколько раз через модель, чтобы добиться лучшего качества определения фаз отражений. Нейронная сеть использует только амплитуды структурных факторов, и достигает сопоставимого качества решения с методом обращенного заряда. Нельзя не упомянуть, что модель способна решить фазовую проблему при разрешении рентгенодифракционных данных всего 2.0 Å, то есть используя лишь 10–20% объема данных, необходимых классическим методам. Авторы продемонстрировали высокую точность на экспериментальных данных, а также способность к предсказанию фаз для отсутствующих фаз (расширение фаз, phase extension). Таким образом, впервые был продемонстрирован потенциал машинного обучения для определения фаз. Недостатком модели является то, что она заточена под центросимметричные структуры. Исключительный интерес же представляют нецентросимметричные кристаллы, к которым нельзя будет применить подход с бинарным классификатором, описанным в

работе.

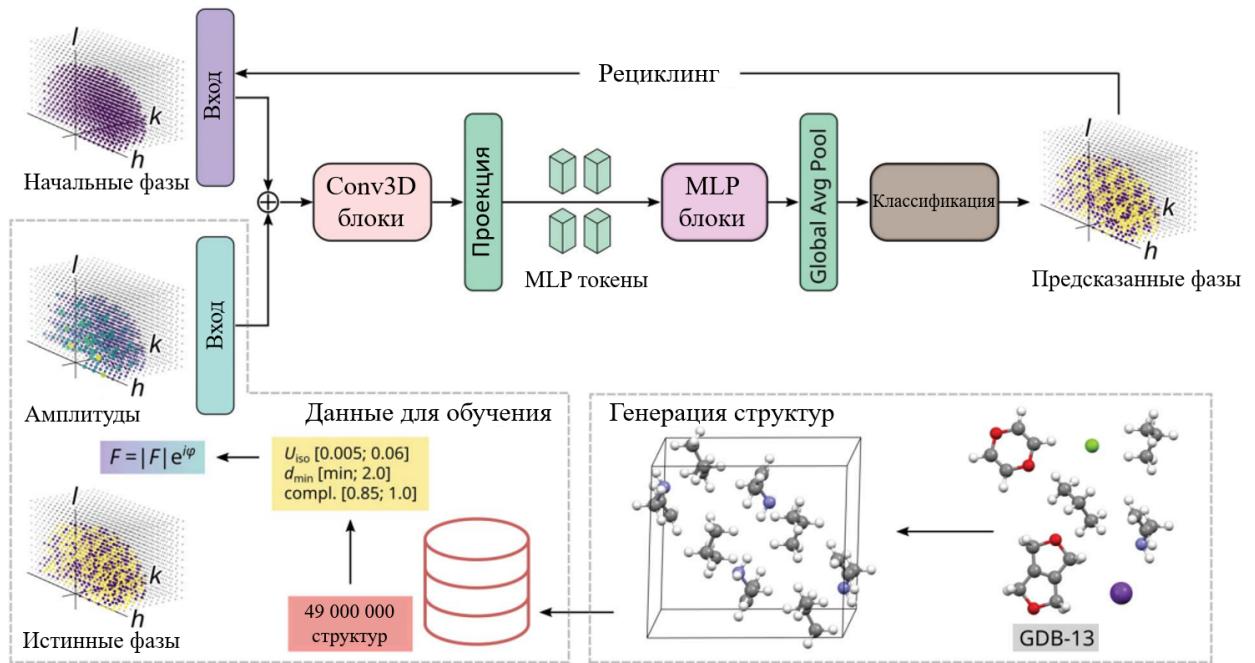


Рисунок 10. Схема нейронной сети PhAI для решения проблемы фаз [33]

Вдохновившись описанным исследованием, авторы [35] предлагают адаптацию его идеи для нецентросимметричных структур. В работе предложен метод phase seeding, который позволяет переформулировать задачу из регрессионной с определением непрерывных величин от 0 до 2π в задачу мультиклассовой классификации. Сначала весь диапазон фаз дискретизируется в ограниченный набор промежутков, в каждом из которых фазы отражений заменяют на одинаковое значение (рис. 11). Затем фаза каждого отражения случайным образом относится к одному из подмножеств. Авторы продемонстрировали на различных структурах, что для решения структур детерминированными методами достаточно будет предсказать с помощью методов искусственного интеллекта одно из дискретных значений фаз для каждого дифракционного максимума. Также предложено для решения фазовой проблемы методами ИИ использовать отражения с наибольшими значениями амплитуды нормализованного структурного фактора E .

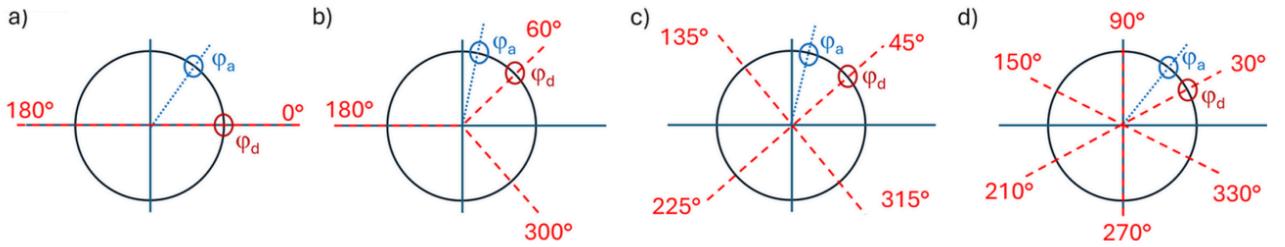


Рисунок 11. Дискретизация значений фаз: на 2 промежутка (а), на 3 промежутка (б), на 4 промежутка (в), на 6 промежутков (г). Истинное значение фазы — ϕ_a (синий круг), дискретизованное — ϕ_d (красный круг) [35]

Также были совершены попытки предсказать структуры белков напрямую из экспериментальных данных, минуя этап с определением фазовой информации. Так, была предложена модель RecCrysFormer, позволяющая напрямую предсказывать электронную плотность макромолекулярных соединений на основе карт Паттерсона и известной плотности отдельных белковых фрагментов [36]. Архитектура сети представляет собой трехмерные сверточные слои, переводящие карту Паттерсона и плотность в пространство признаков, которые затем подаются на вход в трансформер, выходные данные которого обрабатываются с помощью слоёв многослойного перцептрона и трехмерных свёрток для получения карт плотности всего белка. Примечательно, что в слое трансформера реализовано одностороннее внимание, поэтому только токены функции Паттерсона "смотрят" на токены фрагментов соединения. Авторы смогли добиться решения 93% тестовых структур, хотя и признают, что выбранные ими соединения меньше реальных белков.

Были найдены решения фазовой проблемы с помощью методов глубокого обучения в области физики, а именно в рамках метода когерентной безлинзовой микроскопии [37]. В обзорной статье выделены 3 подхода (рис. 12) — DL-pre-processing, в котором обученная модель повышает разрешение дифракционных данных, после чего проблема фаз решается классическими детерминированными методами; DL-in-processing, в рамках которого из экспериментальных интенсивности с помощью нейронной сети рассчитывают фазы напрямую; DL-post-processing, в котором уточняются зашумленные приближенные фазы, полученные из исходных интенсивностей. Несмотря на то, что для рентгеновской кристаллографии нельзя напрямую использовать уже готовые методы микроскопии, поскольку данные сильно различаются разрешением, эти идеи можно адаптировать. Так, существующее решение PhAI [33] можно отнести к DL-in-processing.

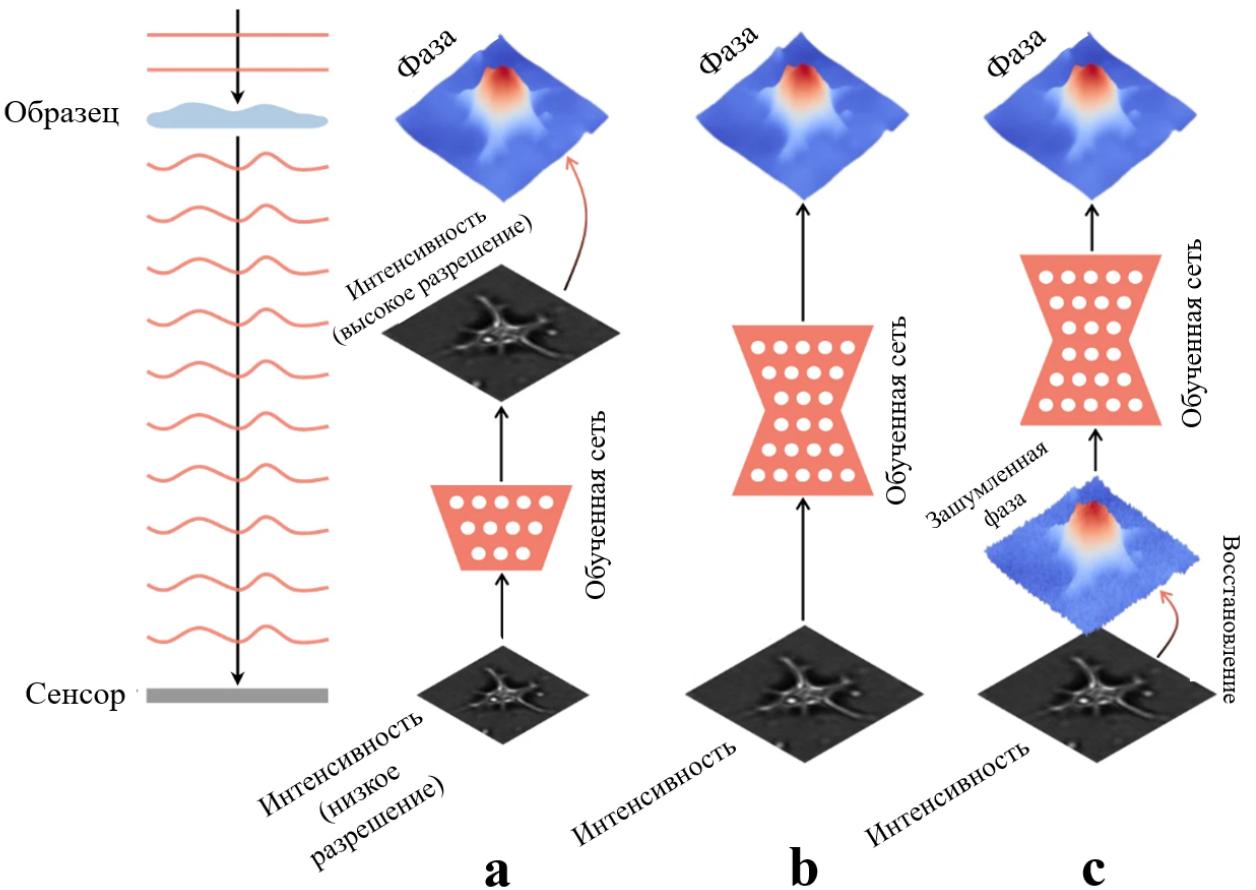


Рисунок 12. Подходы к восстановлению фаз в микроскопии: DL-pre-processing (a), DL-in-processing (b), DL-post-processing (c) [37]

1.4 Заключение

Проблема фаз является фундаментальной задачей рентгенодифракционных исследований кристаллов, без решения которой невозможно определить структуру кристалла. На сегодняшний день существует множество методов, позволяющих преодолеть отсутствие информации о фазах дифракционных максимумов, но они имеют ряд ограничений, связанных с требованием к достаточно высокому разрешению данных. Данное требование может не выполняться в рамках белковой кристаллографии, где фазовая проблема исключительно редко решается рутинными методами. Для определения структур макромолекулярных соединений требуется дополнительная информация — знание о структуре белка с той же аминокислотной последовательностью или результаты рентгенодифракционных экспериментов той же структуры с добавлением тяжелых атомов. Поэтому решение проблемы фаз является особенно актуальной задачей для исследователей белковых структур.

Однако инструменты на основе методов глубокого обучения способны преодолеть

ограничения традиционных подходов, поскольку уже продемонстрирован их потенциал к определению кристаллических структур на основе ограниченных рентгенодифракционных данных с низким разрешением. По мере развития исследований в этой области, вероятно, алгоритмы искусственного интеллекта станут незаменимым инструментом в области кристаллографии, облегчая решение сложных структур.

Таким образом, целью данной работы является разработка и исследование применимости методов искусственного интеллекта для решения проблемы фаз.

В рамках данной цели были выделены следующие задачи:

1. Разработать программное обеспечение, позволяющее создать набор синтетических рентгенодифракционных данных для обучения нейронных сетей;
2. Разработать программное обеспечение, с помощью которого можно проводить воспроизводимые численные эксперименты по обучению и тестированию моделей;
3. Разработать и обучить подходящие модели глубокого обучения;
4. Произвести анализ полученных моделей на эффективность решения фазовой проблемы для реальных структур.

2 Методика решения

2.1 Подход

В работе было предложено предсказывать амплитуды структурных факторов дифракционных максимумов, которые нельзя получить из эксперимента, по известным из того же эксперимента. После предсказания достаточного количества отражений, разрешения данных должно хватить для определения фаз и расчета электронной плотности одним из рутинных *ab initio* методов, в качестве которого был выбран метод, реализованный в комплексе программ SHELXTL PLUS [29] (схема представлена на рис. 13). Разрешение — минимальное межплоскостное расстояние из набора дифракционных максимумов структуры [1].

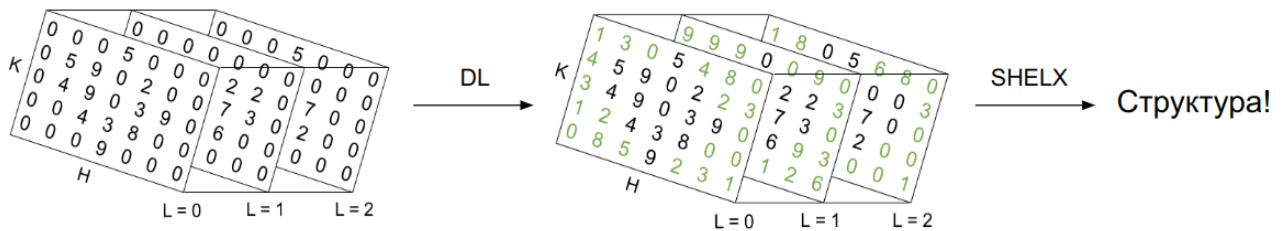


Рисунок 13. Схема решения проблемы фаз с помощью методов глубокого обучения (DL).
Дифракционные картины (тензоры отражений \mathbf{F}) представлены в виде
параллелепипедов

Дифракционные отражения являются точками обратного пространства, каждое из них можно однозначно описать индексами Миллера (h, k, l). Тогда дифракционную картину можно описать трехмерным тензором $\mathbf{F} \in \mathbb{R}^{H \times K \times L}$, в котором записаны амплитуды каждого отражения: $\mathbf{F}[h, k, l] = F(h, k, l)$. В точках, где не зарегистрированы дифракционные максимумы, в тензоре записаны нули. Таким образом, задача сводится к восстановлению трехмерного тензора. Также значения в каждом тензоре были отнормированы в диапазон 0–1. Получение результата (inference) моделей глубокого обучения должен выглядеть следующим образом: на вход подается тензор с рентгенодифракционными экспериментальными данными, на выходе должен быть тензор с амплитудами дополнительных отражений. В ходе обучения планируется научить модель восстанавливать тензор отражений по данным органических молекул.

В работе также была проведена обработка после получения результата моделью (постпроцессинг), не входящая в обучение и включающая в себя учёт систематических погасаний — "зануления" некоторых значений структурных факторов, что определяется симметрией структуры; в ходе обучения происходит явное восстановления части

тензора, которую не нужно предсказывать. Эффективность предсказания обученных моделей глубокого обучения проверялась на тестовой части синтетического датасета, а также тестовой части рентгенодифракционных данных моноклинных структур из CSD (Кембриджской Базы Структурных данных).

2.2 Рентгенодифракционные данные

Было разработано программное обеспечение, позволяющее генерировать случайные структуры и рассчитывать для них рентгенодифракционные данные (github.com/blackwood168/xrd_simulator). С помощью открытой библиотеки на языке Python CCTBX (Computational Crystallography Toolbox) [38] создаются кристаллические решетки, в которых случайным образом с учётом симметрии расставляются случайные атомы. Для получаемых синтетических структур реализованы расчёт дифракционной картины — индексов и структурных факторов отражений, вычисление порошковой дифрактограммы (реализована профильная функция Псевдо–Войдта [39] и осевая расходимость рентгеновского пучка, рис. 14), а также карты Паттерсона. Нужный расчет выбирает пользователь исходя из своей текущей задачи. Также генератор поддерживает последующий расчет данных рентгеновской порошковой дифракции. Созданное ПО может быть полезно для решения прикладных задач рентгенодифракционных исследований кристаллов с помощью методов машинного обучения.

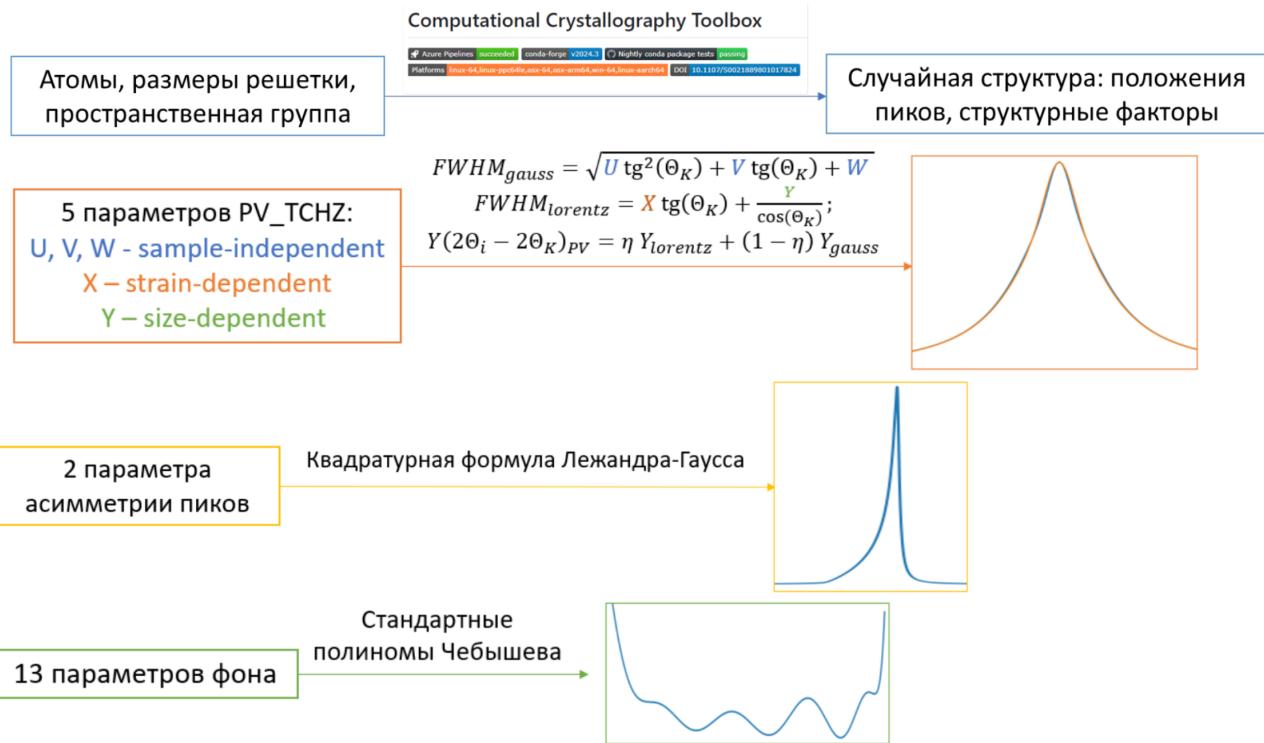


Рисунок 14. Схема генерации случайных дифрактограмм в разработанной программе генерации

Начальное обучение моделей глубокого обучения было решено проводить на синтетических структурах, которые были получены с помощью созданного генератора. Так как общее количество симметрично независимых отражений зависит от класса Лауэ, было решено сосредоточиться на моноклинных структурах, поскольку моноклинные группы симметрии являются одними из наиболее распространенных для белковых структур в базе данных белков (rcsb.org/stats/distribution-space-group). При генерации структур её основные параметры (группа симметрии, типы атомов, их количество) определяются случайным образом (случайное сэмплирование) из следующих значений:

- группы симметрии: P2₁, C2;
- атомы: C, N, O, Cl, Br;
- число симметрично независимых атомов в ячейке: 10–30;

При выполнении численных экспериментов со структурными факторами было решено работать с их амплитудами. Типичные распределения этих данных для сгенерированных структур представлены на рис. 15. Как можно заметить, распределение амплитуд более пологое, чем интенсивностей, поэтому именно амплитуды были выбраны для решения задачи.

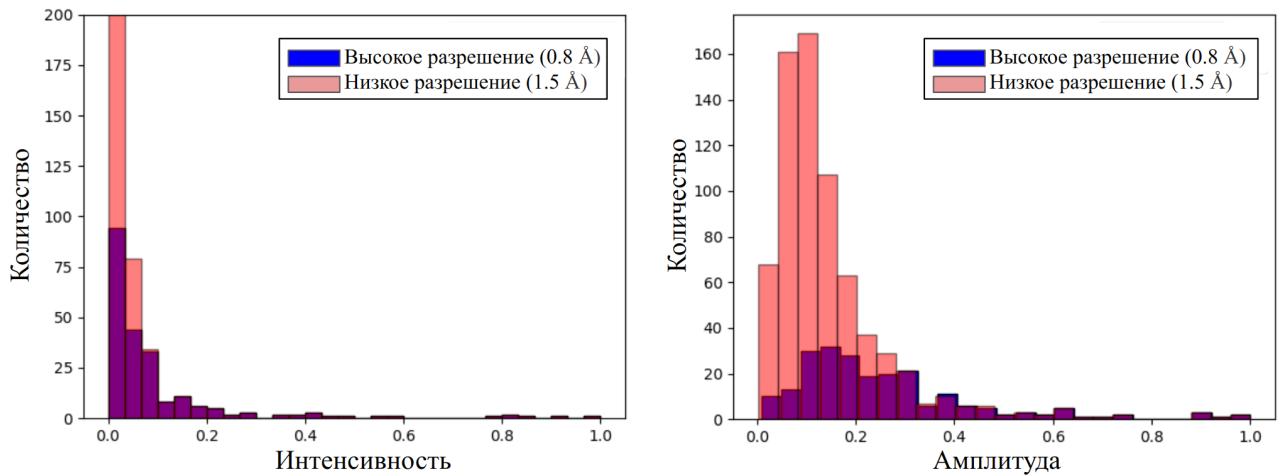


Рисунок 15. Типичные распределения интенсивностей (слева) и амплитуд (справа) дифракционной картины

600.000 полученных структур были разделены на наборы следующих размеров: 400.000, 100.000 и 100.000 для тренировочной, валидационной и тестовой выборок, соответственно. Из этих структур были собраны следующие наборы данных, с помощью которых будет проведено обучение моделей:

1. $D_1 = \{\mathbf{F}(1.5\text{\AA})_i, \mathbf{F}(0.8\text{\AA})_i\}_{i=1}^n$
2. $D_2 = \{\mathbf{F}(1.2\text{\AA})_i, \mathbf{F}(1.0\text{\AA})_i\}_{i=1}^n$
3. $D_3 = \{\mathbf{E}(1.2\text{\AA})_i, \mathbf{E}(1.0\text{\AA})_i\}_{i=1}^n$

Также в работе использовались реальные моноклинные молекулярные структуры малых молекул из Кембриджского Банка Структурных Данных [40], для которых были расчитаны дифракционные отражения. 10.000 структур использовались для дообучения моделей на реальных структурах, 2000 — были отложены для тестирования. На рис. 16 представлена зависимость среднего количества отражений от выбранного разрешения дифракционной картины. При повышении разрешения с 1.5 Å до 0.8 Å (набор D_1) требуется с помощью нейронной сети увеличить число максимумов более чем в 5 раз, поэтому был также собран набор D_2 , для которого потребуется расширить дифракционную картину в 1.7 раз.

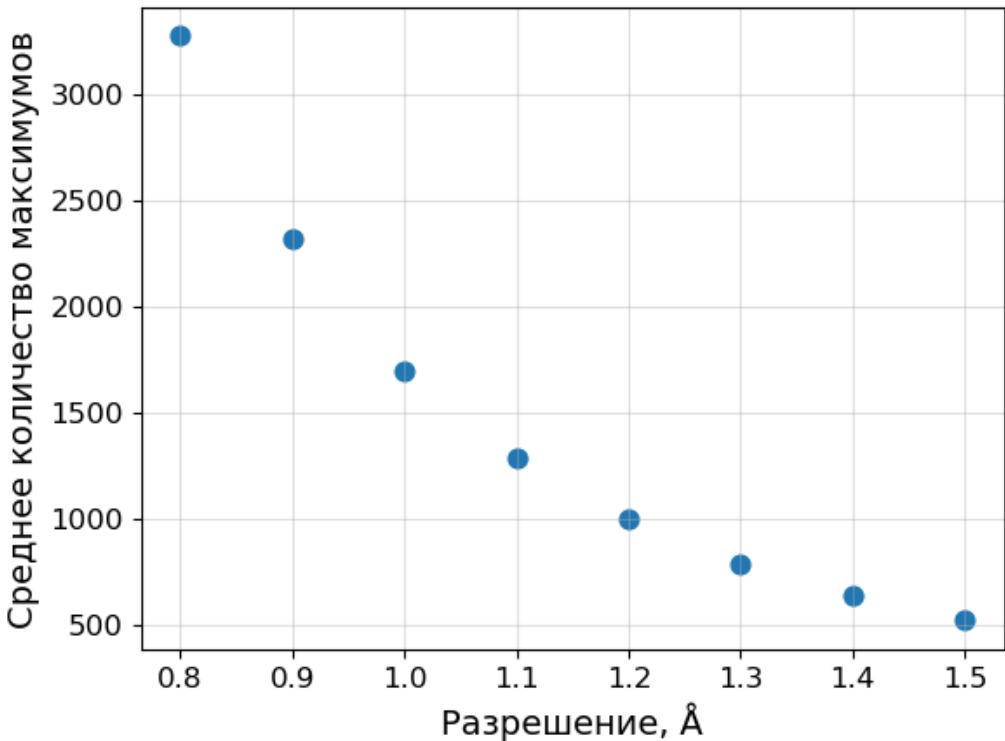


Рисунок 16. Зависимость среднего количества дифракционных отражений моноклинных ($P2_1$, $C2$) структур из CSD от разрешения

Рассчитанный набор амплитуд нормализованных структурных факторов D_3 использован из соображения, что нормализованные структурные факторы E лишены явной зависимости амплитуды от $\frac{\sin \theta}{\lambda}$, где Θ — угол отражения, λ — длина волны рентгеновского излучения. Это приводит к тому, что распределение $|E|$ менее смещено в сторону высокоинтенсивных максимумов по сравнению с $|F|$ (рис. 17), что делает данный набор данных более перспективным для решения с помощью машинного обучения, поскольку требуется предсказывать низкоинтенсивные отражения.

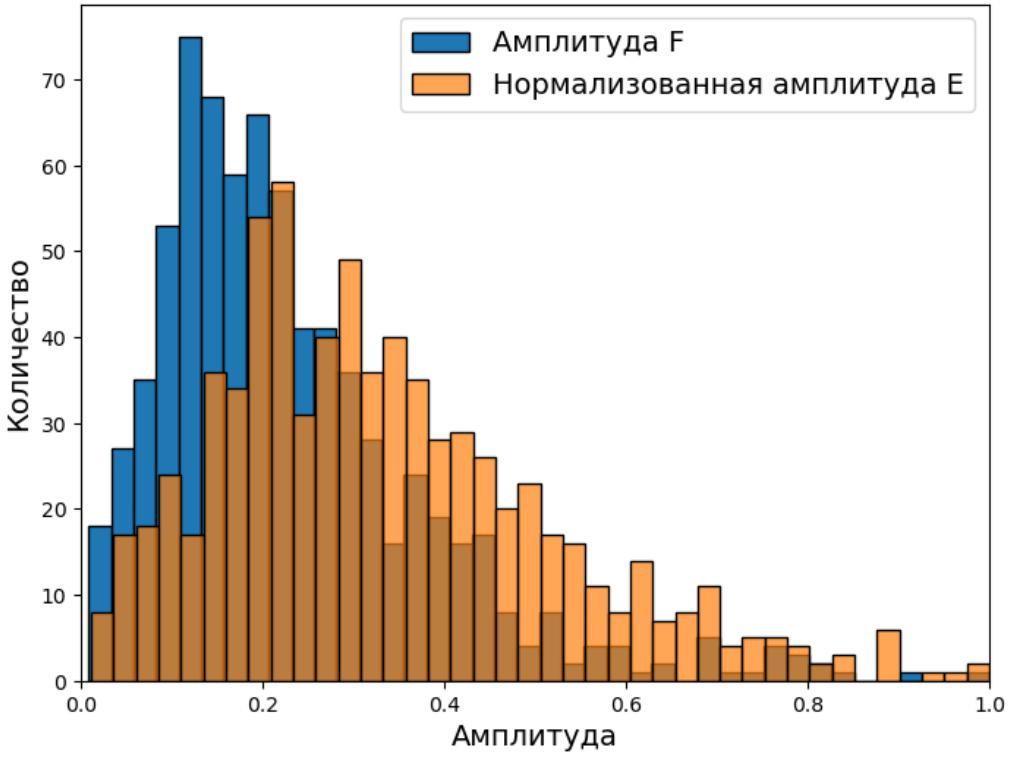


Рисунок 17. Типичные распределения амплитуд структурного фактора F и нормализованного структурного фактора E (разрешение 1.0 \AA)

Размер полученных тензоров составляет (26, 18, 23) для набора данных D_1 и (23, 16, 21) для наборов D_2 , D_3 — в тензоры таких размерностей помещаются все отражения для самой большой моноклинной структуры из синтетических данных (для разрешения 0.8 \AA и 1.0 \AA). Также значения в каждом тензоре были отнормированы в диапазон 0–1.

2.3 Модели машинного обучения

Обозначим модель машинного обучения как $g(\Theta, \mathbf{F})$, которая задаётся параметрами Θ и принимает на вход тензор рентгенодифракционных отражений \mathbf{F} . Модель рассчитывает дополненный тензор отражений $\mathbf{F}_{\text{high}}^{\text{g}} = g(\Theta, \mathbf{F}_{\text{low}})$, который должен быть максимально близок к реальной дифракционной картине высокого разрешения \mathbf{F}_{high} . Задача является регрессионной, процесс обучения на наборе данных, состоящего из пар $\{\mathbf{F}_{\text{low},i}, \mathbf{F}_{\text{high},i}\}_{i=1}^n$ стремится оптимизировать параметры Θ , минимизируя целевую функцию, в качестве которой выбрана среднеквадратичная ошибка (MSE):

$$\Theta^* = \arg \min_{\Theta} \left[\text{MSE}(\Theta) := \frac{1}{n} \sum_{i=1}^n \|g(\Theta, \mathbf{F}_{\text{low}}) - \mathbf{F}_{\text{high}}\|_F^2 \right] \quad (42)$$

Как уже было отмечено, в качестве функции потерь была выбрана среднеквадратичная ошибка, минимизация которой должна приводить к восстановлению тензора

рентгеновских отражений. В качестве метрики для оценки эффективности моделей также был использован фактор расходимости R:

$$R = \frac{\sum_{h,k,l} ||F_{obs} - |F_{calc}|||}{\sum_{h,k,l} |F_{obs}|}, \quad (43)$$

где $|F_{obs}|$ — экспериментальные структурные факторы, $|F_{calc}|$ — рассчитанные по модели структурные факторы. R-фактор является общепринятым стандартом в кристаллографическом сообществе для оценки качества структурных моделей. Нулевое значение R-фактора отвечает идеальному соответствуанию между данными модельной структуры и экспериментальными данными. В качестве экспериментальных структурных факторов в работе были использованы структурные факторы, рассчитанные по структуре, а F_{calc} — результат вычислений с помощью нейронных сетей.

Также в качестве метрики сравнения изображений рассматривался индекс структурного сходства SSIM [41], который зарекомендовал себя как хорошая метрика для оценки восстановления и увеличения разрешения изображений, однако эксперименты с ним в качестве добавки к функции потери привели к более низкому качеству восстановления модели с точки зрения R-фактора. Данный результат можно объяснить отсутствием локальной связности наших данных.

В качестве базовой модели (baseline) была обучена модель UNet [42], адаптированная для трехмерных данных (рис. 18). Данная модель выбрана, потому что она хорошо себя проявляет в простейших задачах увеличения разрешения изображения за счет генерации недостающих пикселей на входных данных низкого разрешения (Super Resolution).

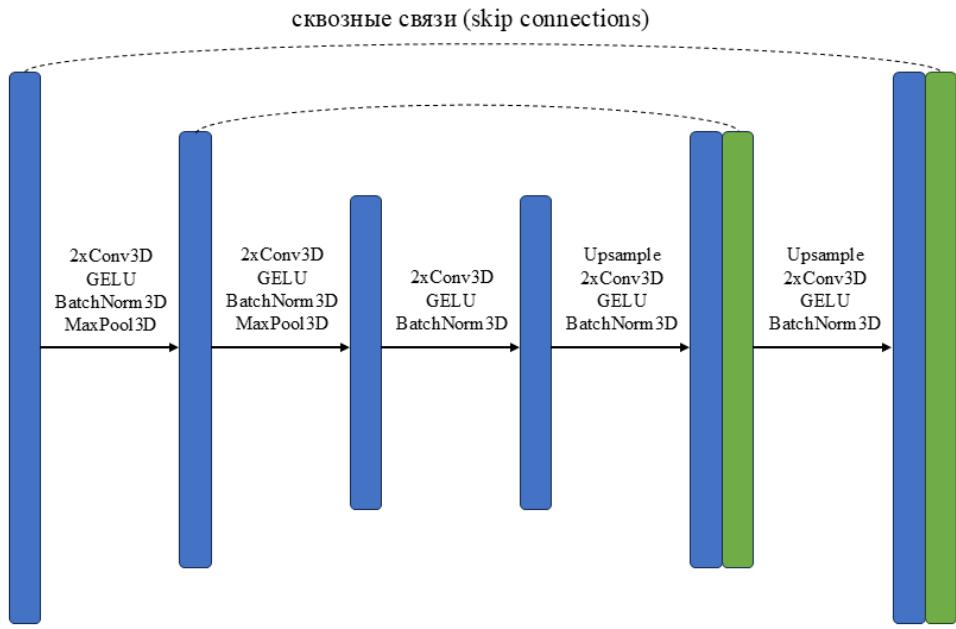


Рисунок 18. Схемы архитектуры модели UNet

Также была разработана и обучена модель на основе UNet с улучшенными слоями, содержащими Фурье-преобразование FFT_UNet (рис. 19). Данный подход был продемонстрирован [43] при работе с дифракционными данными и он является многообещающим и для нашей задачи, поскольку при переходе в прямое пространство наши данные являются локально связанными.

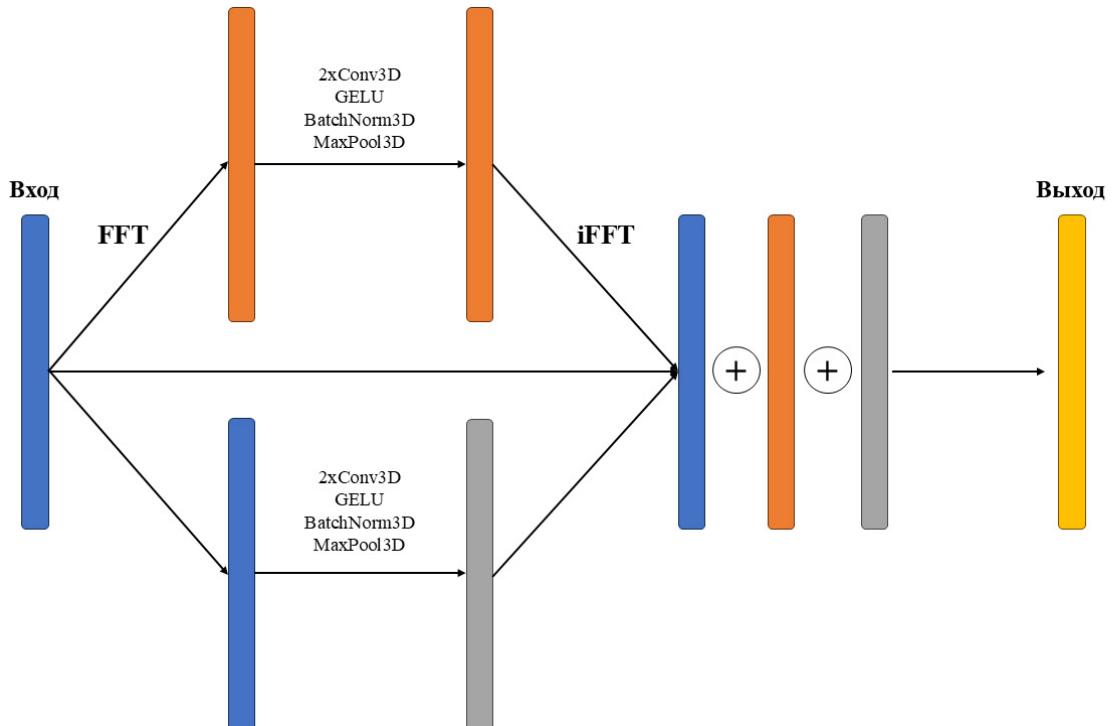


Рисунок 19. Схема слоёв с преобразованием Фурье

Поскольку тензоры отражений не являются локально связанными, актуально

использование механизма внимания. Он позволит модели находить связи между дальними отражениями. Так, был разработан трансформер XRD_Transformer для нашей задачи (рис. 20).

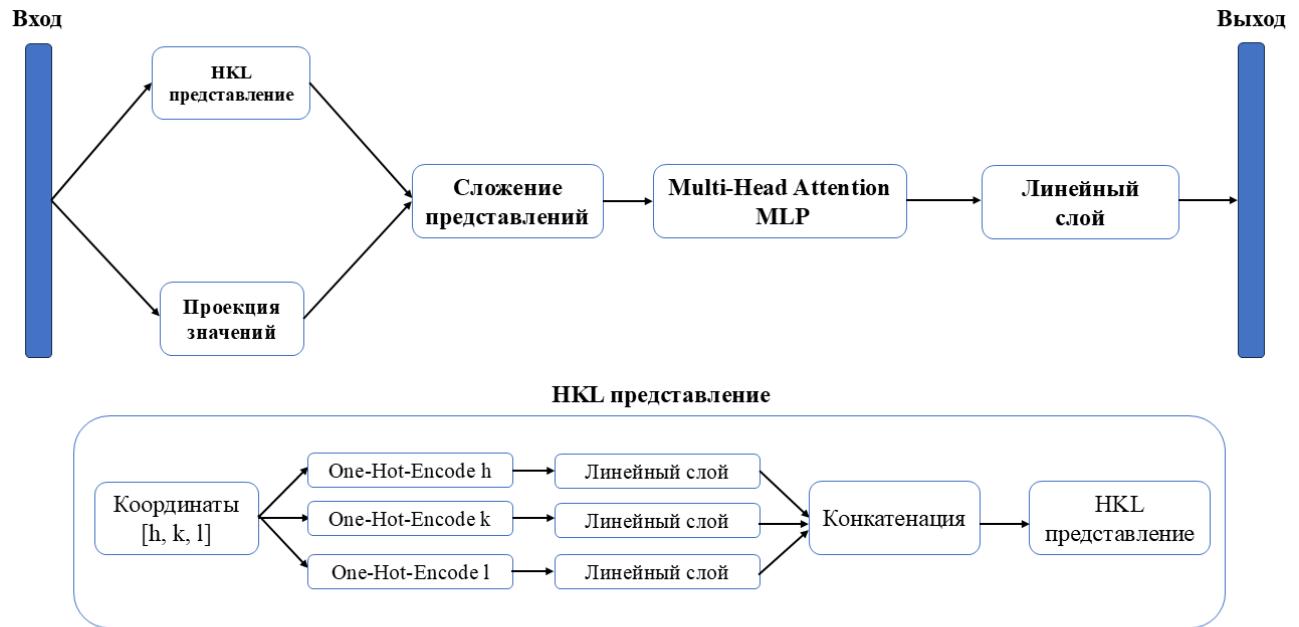


Рисунок 20. Схема архитектуры XRD_Transformer

В модели формируется единое векторное представление (embedding) из индексов Миллера и проекции значений амплитуды отражений, затем он проходит через 5 слоев трансформера, которые состоят из слоев многоголового внимания (Multi-Head Attention) и блоков многослойного перцептрона (MLP). Затем после нормализации (LayerNorm) и обратного проецирования в исходное пространство получается восстановленный тензор дифракционной картины.

Так как нам известны все возможные значения индексов Миллера (h, k, l) для структур выбранных размеров с заданным разрешением, HKL представления (рис. 20) формируется следующим образом: индексы кодируются с помощью унитарного кодирования (One-Hot Encoding), после чего каждый вектор с помощью обучаемого линейного слоя проецируется в вектор размерности $\frac{embed_dim}{3}$ после конкатенации размерность полученного векторного представления составляет $embed_dim$. Также в модели реализована возможность получения представления через полносвязный слой, который проецирует позицию (h, k, l) сразу в вектор, однако она не использовалась при обучении модели, так как первый способ является более физическим для нашей задачи, поскольку мы используем только симметрически независимые отражения.

3 Результаты и обсуждение

3.1 Автоматизированный вычислительный конвейер

В рамках данной работы на первом этапе был разработан автоматизированный конвейер (пайплайн), позволяющий проводить воспроизводимые численные эксперименты по решению проблемы фаз с помощью предложенного подхода (github.com/blackwood168/xrd_phase_ml, рис. 21). В нем реализовано обучение и тестирование моделей, а также получение результатов (inference) на реальных массивах данных и структурах кристаллических соединений. В репозитории присутствуют маленькие наборы данных из сгенерированных и реальных структур малых органических молекул, также там представлены веса обученных в работе моделей. Воспроизводимость обучения обеспечивает фиксирование начальных значений генераторов случайных состояний.



Рисунок 21. Схема разработанного автоматизированного контейнера, использующего модели глубокого обучения (DL) для предсказывания амплитуд структурных факторов

3.2 Модели для структурных факторов (разрешение 1.5 Å)

3.2.1 Результаты обучения

Было проведено обучение на синтетических данных (набор D_1) и последующее дообучение на рентгенодифракционных данных реальных структур из Кембриджской Базы Структурных Данных (CSD). В ходе обучения модель повышала разрешение с 1.5 до 0.8 Å. Сравнение R-фактора для моделей до и после дообучения представлено в таблице 1. После дообучение на реальных данных точность предсказания моделей увеличивается минимум на треть, однако теряется точность на синтетических данных, кроме UNet с Фурье-преобразованием, который лишь прибавляет в точности на сгенерированных структурах после дообучения на реальных. Лучшую точность имеет модель UNet_FFT ($R=0.336$), от которой немного отстает трансформер. В сводной таблице 2 представлены значения метрик на реальных данных финальных моделей. Таким образом, UNet_FFT занимает меньше видеопамяти, работает быстрее и достигает лучшей метрики на тестовых реальных данных.

Таблица 1. Результаты обучения и эффективность дообучения моделей, повышающих разрешение с 1.5 до 0.8 Å (сint. — синтетические тестовые данные)

Модель	Метрика	Синт.	CSD
UNet	До, R	0.477	0.590
	После, R	0.632	0.393
	$\Delta, \%$	-32.6	33.4
FFT_UNet	До, R	0.726	0.646
	После, R	0.619	0.336
	$\Delta, \%$	14.7	48.0
XRD_Transformer	До, R	0.346	0.581
	После, R	0.615	0.358
	$\Delta, \%$	-77.9	38.4

Таблица 2. Значения метрик на тестовом реальном наборе данных моделей, повышающих разрешение с 1.5 до 0.8 Å после дообучения

Metric \ Model	UNet	FFT_UNet	XRD_Transformer
MSE $\cdot 10^{-3}$	1,39	1,20	1,31
R	0,393	0,336	0,358

3.2.2 Определение структур

Модели демонстрируют достаточно большие значения R-фактора на реальных моноклинных структурах из Кембриджского Банка Структурных данных, качество восстановления тензора отражений даже после дообучения недостаточно для решения структуры с помощью программы SHELXT. На рис. 22 представлены характерные сечения тензора отражений для реальной структуры, где R-фактор восстановленного тензора с помощью FFT_UNet составляет 0.343, но SHELXT, относящийся к дуальным методам решения проблемы фаз, не смог определить кристаллическую структуру по таким данным.

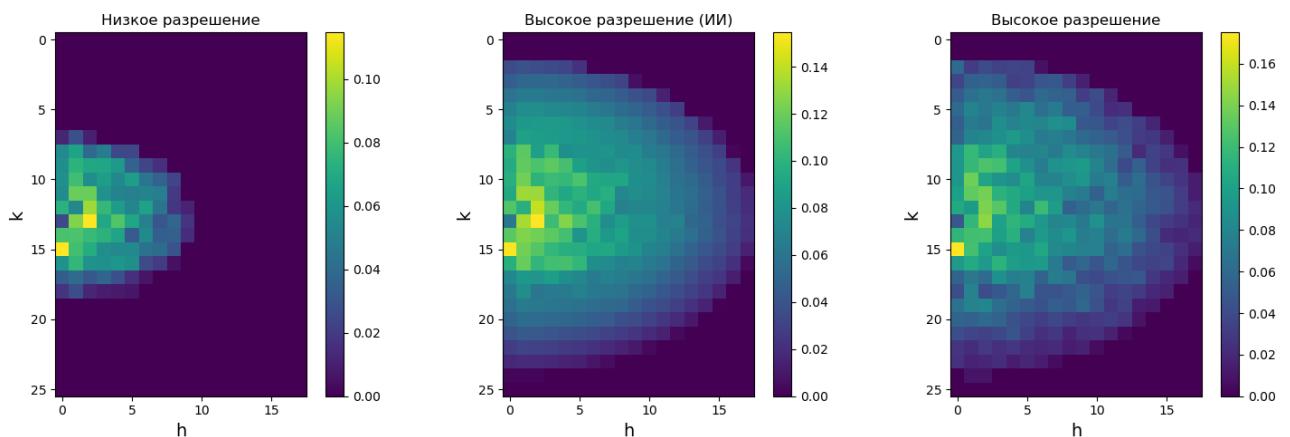


Рисунок 22. Типичное восстановление дифракционной картины реальной структуры, $R = 0.343$ (с 1.5 до 0.8 Å, усреднено по индексу 1)

Можно заметить, что изображение восстановленного тензора отражений по сравнению с истинным размытое, что можно объяснить результатом минимизации среднеквадратичной ошибки. Поскольку дифракционная картина не является локально связанный, как изображения, соседние пиксели могут сильно различаться. Модели машинного обучения не могут точно предсказать значение амплитуды структурного фактора в каждом пикселе, но справляются с вычислением среднего значения. Стоит отметить, что несмотря на низкую точность решения поставленной задачи регрессии, модель достаточно точно определяет границы дифракционной картины, то есть какие отражения будут ненулевые в тензоре отражений более высокого разрешения.

3.3 Модели для структурных факторов (разрешение 1.2 Å)

3.3.1 Результаты обучения

Было проведено обучение на синтетических данных (набор D_2) и последующее дообучение на реальных моделей, повышающих разрешение дифракционной картины с 1.2 до 1.0 Å. Сравнение R-фактора для моделей до и после дообучения представлено в таблице 3. После дообучение на настоящих органических структурах из CSD точность предсказания дифракционных максимумов для структур увеличивается больше всего для UNet с Фурье–преобразованием — более чем на треть. Лучшую точность показывает UNet_FFT ($R=0.336$), от которой немного отстает UNet. На этот раз трансформер показывает худший результат на реальных структурах. В сводной таблице 4 представлены значения метрик на реальных данных финальных моделей. Таким образом, UNet_FFT снова становится наиболее точной моделью машинного обучения для увеличения разрешения дифракционной картины.

Таблица 3. Результаты обучения и эффективность дообучения моделей, повышающих разрешение с 1.2 до 1.0 Å (синт. — синтетические тестовые данные)

Модель	Метрика	Синт.	CSD
UNet	До, R	0.119	0.255
	После, R	0.109	0.207
	Δ , %	8.4	18.8
FFT_UNet	До, R	0.107	0.234
	После, R	0.319	0.152
	Δ , %	-198.1	35.0
XRD_Transformer	До, R	0.236	0.337
	После, R	0.104	0.267
	Δ , %	55.9	26.2

Таблица 4. Значения метрик на тестовом реальном наборе данных моделей, повышающих разрешение с 1.2 до 1.0 Å после дообучения

Метрика \ Модель	UNet	FFT_UNet	XRD_Transformer
MSE·10 ⁻³	0.684	0.489	0.199
R	0.207	0.152	0.267

3.3.2 Определение структур

Хотя факторы расходимости принимают низкие значения для данных, посчитанных при сравнении дифракционной картины, восстановленой нейронной сетью FFT_UNet, и данных, рассчитанных по реальным структурам (рис. 23), качества данных для определения структуры с помощью программного обеспечения SHELXT все еще не хватает. Так, лишь для 83% низкомолекулярных органических структур из CSD, выбранных в качестве тестовой выборки, метод нашёл какое-то подходящее решение, средний R-фактор этих моделей равен 0.49. Структуры, чьё решение имеет R-фактор не более 0.25, составляют 10% всей тестовой выборки. Для сравнения — если попытаться решить структуры с помощью рентгенодифракционных данных, рассчитанных по структуре, с разрешением 1.0 Å, SHELXT выдаёт решение для 97% структур со средним R-фактором равным 0.13. Результаты представлены в таблице 5.

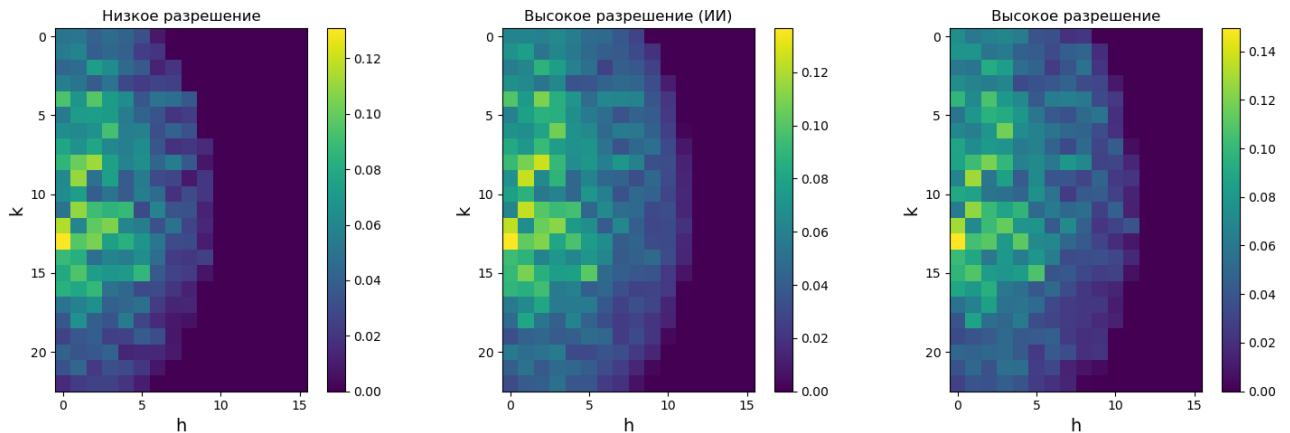


Рисунок 23. Типичное восстановление дифракционной картины реальной структуры, $R = 0.133$ (с 1.2 до 1.0 Å, усреднено по индексу 1)

Таблица 5. Данные по эффективности определения решения по рентгенодифракционным данным с разрешением 1.0\AA , полученных с помощью ИИ, и рассчитанным по структуре

Показатель \ Данные	Восстановленные (ИИ)	Рассчитанные
Наличие решений, %	83	97
Наличие решений с $R \leq 0.25$, %	10	94
Средний R	0.49	0.13

3.4 Модели для нормализованных структурных факторов

3.4.1 Результаты обучения

Было проведено обучение на нормализованных структурных факторах синтетических (набор D_3) и последующее дообучение на реальных структурах моделей, повышающих разрешение дифракционной картины с 1.2 до 1.0\AA . Сравнение фактора расходимости R для моделей до и после дообучения представлено в таблице 6. После дообучения на низкомолекулярных органических структурах из CSD эффективность предсказания дифракционных максимумов выросла только для модели UNet с Фурье–преобразованием на 21.5%. Лучшую точность в очередной раз показывает UNet_FFT ($R=0.213$). В сводной таблице 7 представлены значения метрик на реальных данных финальных моделей. Таким образом, UNet_FFT является наиболее точной моделью машинного обучения для увеличения разрешения дифракционной картины во всех рассмотренных случаях. В дальнейшем для проблемы фаз будет использоваться именно эта модель.

Таблица 6. Результаты обучения и эффективность дообучения моделей, повышающих разрешение с 1.2 до 1.0\AA (сint. — синтетические тестовые данные)

Модель	Метрика	Сint.	CSD
UNet	До, R	0.423	0.401
	После, R	0.401	0.397
	Δ , %	5.2	1.0
FFT_UNet	До, R	0.219	0.279
	После, R	0.285	0.213
	Δ , %	-30.1	21.5
XRD_Transformer	До, R	0.342	0.400
	После, R	0.307	0.384
	Δ , %	10.2	4.0

Таблица 7. Значения метрик на тестовом реальном наборе данных моделей, повышающих разрешение с 1.2 до 1.0\AA после дообучения

Метрика \ Модель	UNet	FFT_UNet	XRD_Transformer
MSE $\cdot 10^{-3}$	8.909	2.482	7.543
R	0.397	0.213	0.384

3.4.2 Определение структур

Несмотря на ожидания большей эффективности предсказывания нормализованных амплитуд по сравнению с обычными, дополненные моделью FFT_UNet наборы рентгенодифракционных отражений (рис. 24) оказываются сильно расходящимися с истинными значениями реальных структур (табл. 8). Так, средний фактор расходимости модельных структур, полученных SHELXT, составил 0.62, что на 10% выше, чем при предсказании ненормализованных амплитуд $|F|$. Низкое качество численного восстановления нормализованных факторов требует дополнительного исследования, но можно попробовать это объяснить неверным подбором архитектуры моделей машинного обучения или требованием дополнительной нормализации тензоров отражений.

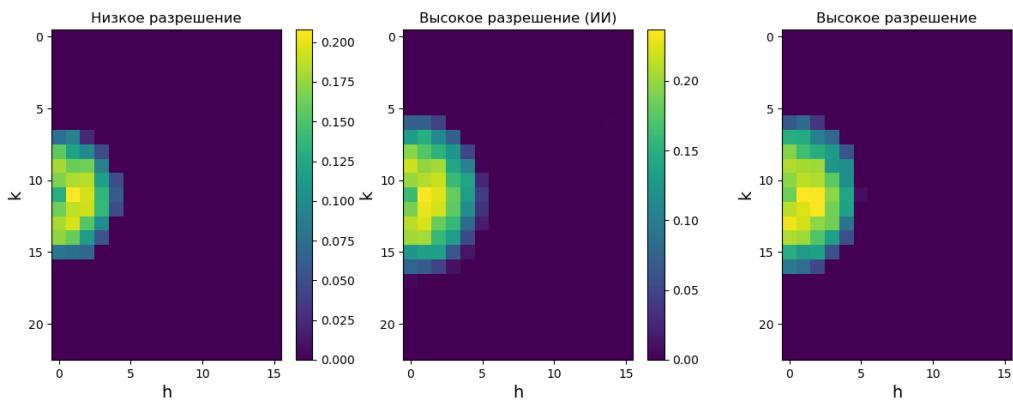


Рисунок 24. Типичное восстановление дифракционной картины реальной структуры, $R = 0.202$ (с 1.2 до 1.0 Å, усреднено по индексу l)

Таблица 8. Данные по эффективности определения решения по рентгенодифракционным данным с разрешением 1.0 Å, полученных с помощью ИИ, и рассчитанным по структуре

Показатель \ Данные	Восстановленные (ИИ)	Рассчитанные
Наличие решений, %	42	97
Наличие решений с $R \leq 0.25$, %	7	94
Средний R	0.62	0.13

3.5 Интерпретируемость моделей

Для более глубокого понимания трансформера был проведен анализ внутренней работы модели. Особый интерес представляют карты внимания (рис. 25), которые визуализируют, как модель распределяет свое внимание при обработке дифракционных данных тестовой реальной кристаллической структуры. Механизм внимания в трансформере позволяет модели определять, какие части входных данных наиболее

важны для принятия решения в каждый момент времени. В первых слоях трансформера внимание распределено равномерно и рассеянно по всей структуре данных. Это соответствует этапу сбора общего контекста, когда модель пытается получить целостное представление о кристаллической структуре. В последующих слоях внимание становится более сфокусированным и локализованным, что указывает на то, что модель научилась выделять специфические взаимосвязи между различными частями данных. Важно отметить способность модели устанавливать связи между отражениями, находящимися на значительном расстоянии друг от друга в обратном пространстве. Это критически важно для решения проблемы фаз, так как часто ключевые взаимосвязи существуют между отражениями, которые не являются ближайшими соседями. Кроме того, наблюдаются характерные диагональные паттерны внимания, которые коррелируют с известными систематическими погасаниями в кристаллографии. Можно предположить, что модель глубокого обучения без явного указания из обучающей выборки выучила механизм погасаний.

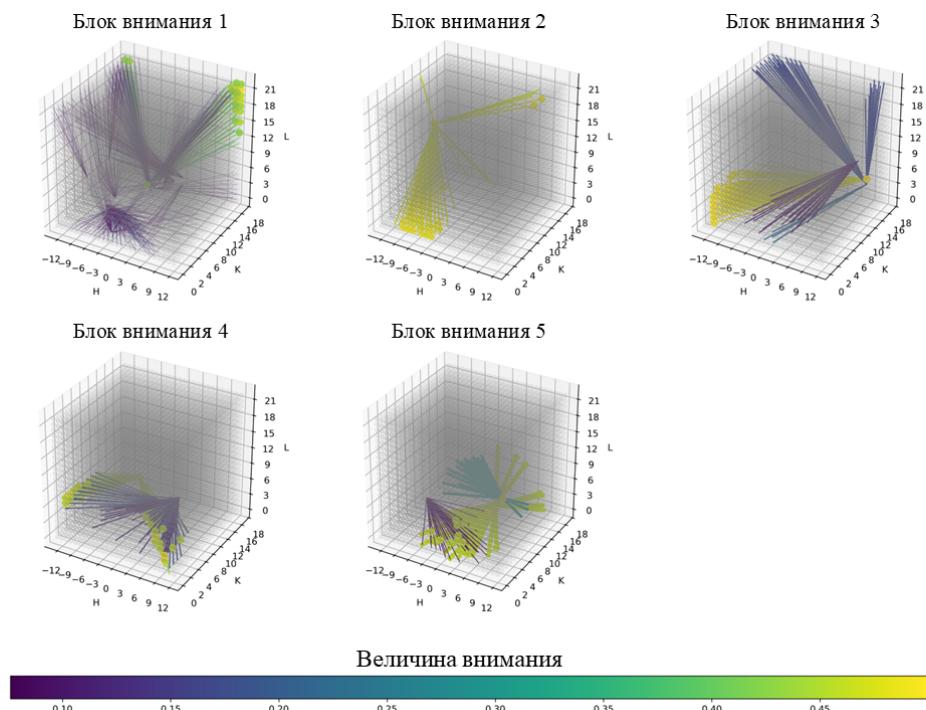


Рисунок 25. Связи внимания в блоках трансформера

Для количественного подтверждения этой гипотезы, было проанализировано распределение значений внимания для наиболее сильных связей в каждом блоке трансформера (рис. 26). Особый интерес представляет сравнение двух типов связей: между точками, соответствующим систематическим погасаниям и обычными отражениями. Можно заметить, что во время сбора общего контекста в первом блоке

распределение значений внимания для связей с погасшими отражениями сопоставимо с таковым для обычных отражений. Это логично, поскольку на этом этапе модель еще не дифференцирует типы отражений, а просто собирает общую информацию о структуре. Однако уже во втором блоке наблюдается значительно уменьшение связей внимания, соединяющих погасания. Это указывает на то, что модель начинает осознавать, что эти отражения несут меньше полезной информации, требуемой для предсказывания дальних отражений. В третьем же блоке модель игнорирует точки обратного пространства, соответствующие систематическим погасаниям, и в выходном тензоре на соответствующих местах стоят нули.

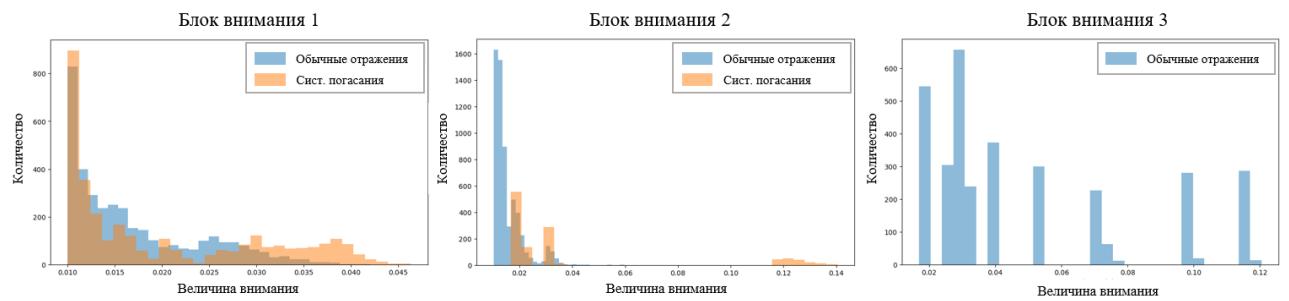


Рисунок 26. Распределение величин внимания в первых блоках трансформера
(сист. — систематические)

Такое поведение модели демонстрирует, что она не просто запомнила шаблоны из обучающей выборки, а научилась автоматически определять и игнорировать отражения, соответствующие систематическим погасаниям (если такие присутствуют). Трансформер действительно успешно выучил кристаллографическую закономерность, значит, данная архитектура может являться ключевой для дальнейших исследований применения методов глубокого обучения для решений кристаллографических задач.

Для понимания внутренней работы моделей на основе UNet, был проведен анализ первого блока обеих архитектур с помощью метода GradCAM (рис. 27) [44]. Этот метод позволяет визуализировать, какие области входных данных имеют наибольшее влияние для принятия решений моделью.

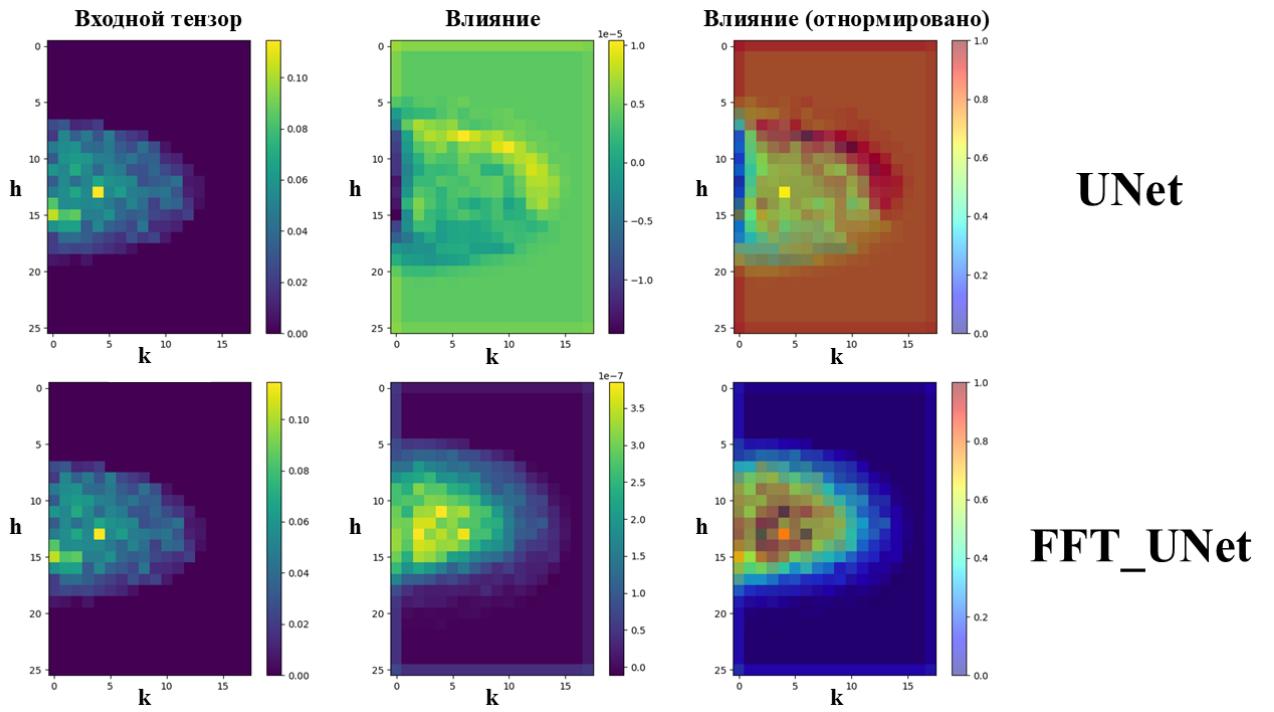


Рисунок 27. Тепловые карты влияния GradCAM (усреднено по индексу 1)

Анализ тепловой карты влияния для модели UNet показал интересную особенность: модель концентрируется во всем пространстве за пределами изначальной дифракционной картины низкого разрешения. Это может указывать на то, что модель UNet не полностью учитывает физические ограничения задачи и пытается извлечь информацию из областей, где она физически не может существовать.

В отличие от UNet, паттерн внимания для FFT_UNet выглядит более физически обоснованным. Значения влияния распространяются преимущественно на область входной дифракционной картины и немного выходят за её границы. Это поведение более физически обосновано, так как модель ищет информацию вблизи границ дифракционной картины, где могут находиться важные детали структуры.

Это различие в поведении моделей может объяснять, почему FFT_UNet показывает лучшие результаты в восстановлении дифракционной картины. Её способность более точно определять области, где может находиться полезная информация, и игнорировать физически нереалистичные области, делает её более эффективной в поиске правильного решения.

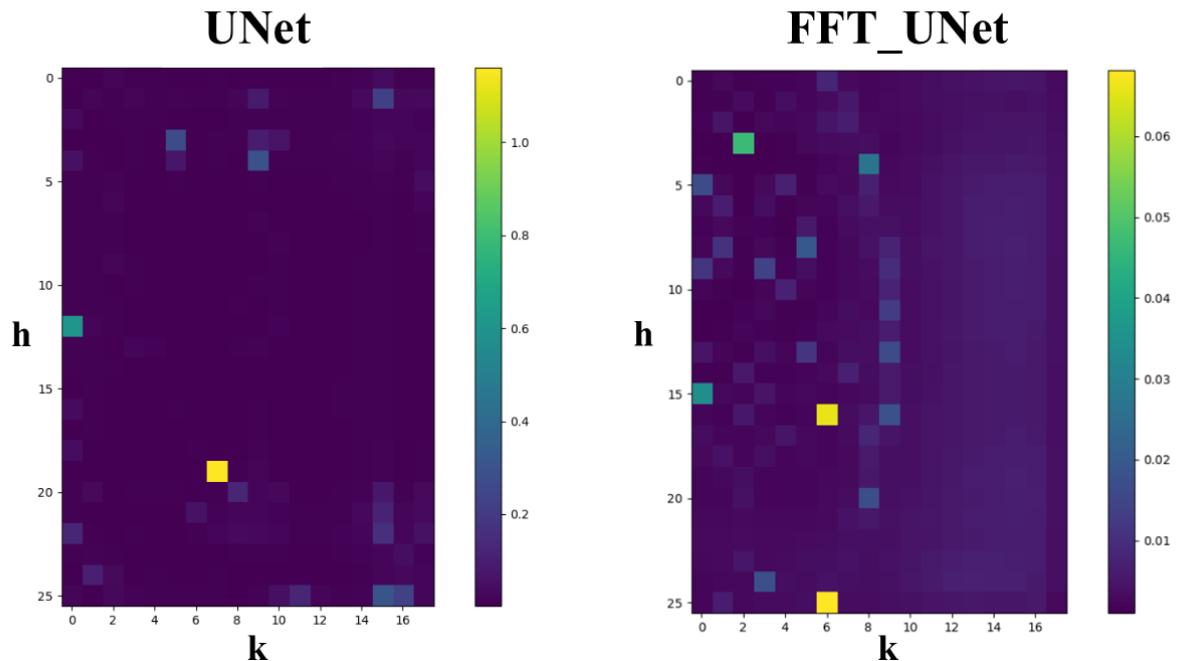


Рисунок 28. Тепловые карты чувствительности, усреднено по индексу 1

Для оценки устойчивости моделей к шуму в экспериментальных данных был проведен анализ карт чувствительности (рис. 28). Карты чувствительности показывают, как сильно меняется выход модели при добавлении случайного шума к входным данным, что позволяет оценить устойчивость модели в различных областях обратного пространства. Базовая модель UNet демонстрирует неустойчивое поведение – выход модели может измениться при добавлении шума на очень большие значения. Модель с Фурье-преобразованием FFT_UNet стабильно работает при наличии шума в данных. Высокая рабочесть в большей части пространства указывает на то, что модель научилась извлекать надежные признаки из данных и не переобучилась на конкретные значения амплитуд структурных факторов.

Анализ результатов показывает, что нейронная сеть успешно справляется с определением общих характеристик дифракционной картины. Она точно предсказывает границы дифракционной картины и корректно восстанавливает средние значения амплитуд структурных факторов по небольшим областям. Однако изменения амплитуд между соседними точками обратного пространства недостаточно четко выражены, что приводит к потере важных деталей в распределении амплитуд.

Это ограничение становится особенно критичным, если учесть фундаментальную особенность дифракционных отражений: каждое отражение содержит информацию о всей кристаллической структуре, и для успешного решения проблемы фаз необходимо

чрезвычайно точное определение амплитуды в каждой точке. Хотя модели глубокого обучения выявляют некоторые кристаллографические закономерности, они не достигают необходимой точности в численном определении значений для каждой точки.

3.6 Уточнение структуры 6-бром-1Н-индол-2,3-диона

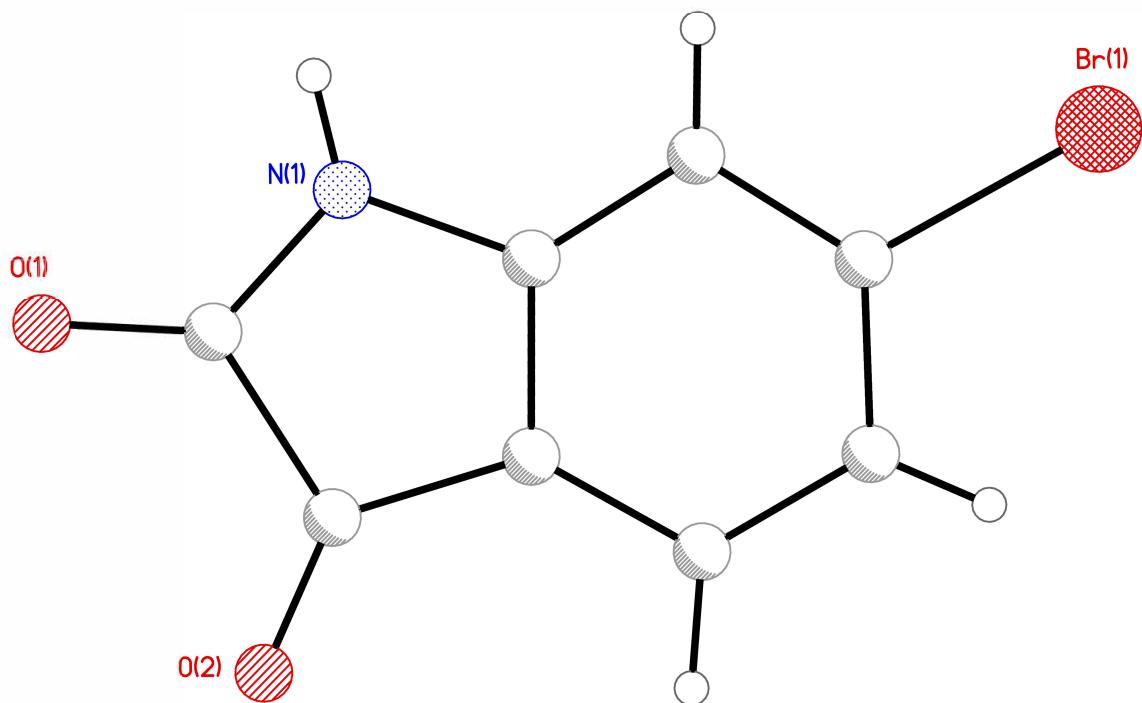


Рисунок 29. Молекула 6-бром-1Н-индол-2,3-диона

Для проверки эффективности предсказывания рентгенодифракционных данных моделями машинного обучения и оценки схожести структурных факторов с истинными значениями было проведено уточнение структуры 6-бром-1Н-индол-2,3-диона (рис. 29) по восстановленным методами искусственного интеллекта отражениями. Данное соединение кристаллизуется в пространственной группе $P2_1$ с параметрами элементарной ячейки $a = 5.6456\text{\AA}$, $b = 5.0594\text{\AA}$, $c = 12.7207\text{\AA}$, $\beta = 91.254^\circ$ (рис. 30). Структура является моноклинной с небольшим количеством атомов, поэтому она идеально подходит для исследования эффективности обученных нейронных сетей.

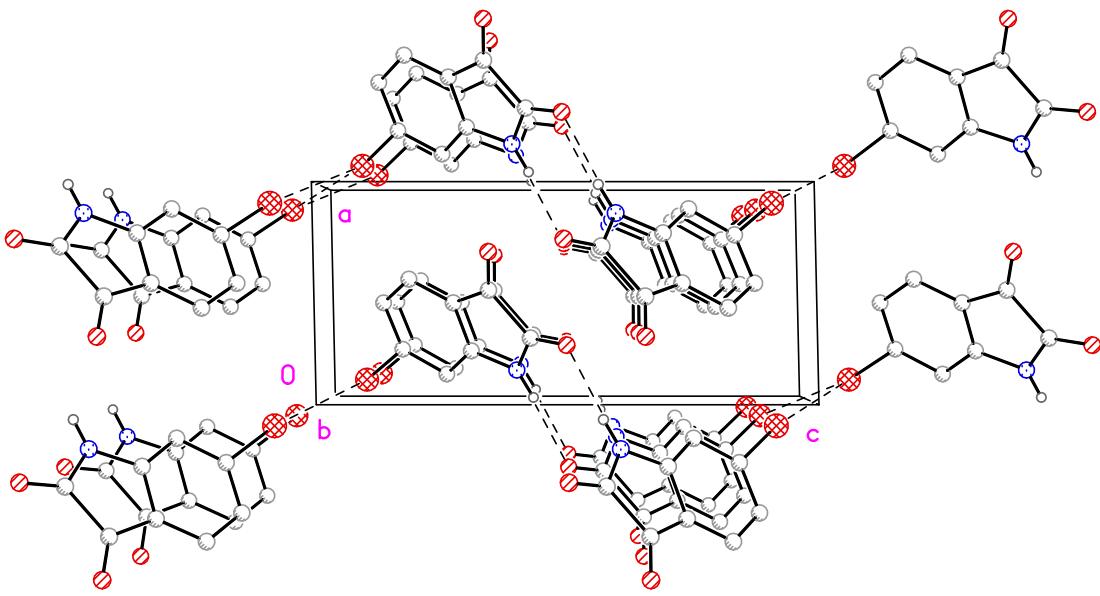


Рисунок 30. Фрагмент кристаллической упаковки 6-бром-1Н-индол-2,3-диона. Молекула показана без атомов водорода

Поскольку подход с повышением разрешения дифракционных данных с 1.5 Å до 0.8 Å имеет слишком низкое качество численного предсказания, сравнение производится только для лучших моделей, предсказывающих отражения с межплоскостными расстояниями от 1.2 Å до 1.0 Å, а именно FFT_UNet для структурных факторов F и нормализованных амплитуд E .

Набор дифракционных максимумов 6-бром-1Н-индол-2,3-диона был обрезан до 1.2 Å и подан на вход нейронных сетей. Для модели, обученных на нормализованных структурных факторах, была произведена предварительная нормализация данных, выходные же данные были обратно переведены в ненормализованные $|F|$. После этого новые данные, дополненные моделями ИИ до разрешения дифракционных данных 1.0 Å, были использованы для уточнения кристаллической структуры. Координаты атомов были зафиксированы, уточнялся только шкальный фактор, который отвечает за приведение расчетных по модели и экспериментальных значений модулей структурных амплитуд к одной шкале. Процедура уточнения проводилась методом наименьших квадратов с

помощью программного обеспечения SHELXT [29].

После уточнения структуры 6-бром-1Н-индол-2,3-диона по данным, полученных с помощью модели FFT_UNet, предсказывающей амплитуды структурных факторов $|F|$, фактор расходимости расчетных и экспериментальных данных R составляет 0.096; после уточнения по выходным данным обученной на нормализованных амплитудах $|E|$ модели — 0.102. Полученный результат может говорить о том, что модели машинного обучения FFT_UNet имеют достаточно высокое качество восстановления дифракционной картины.

Таким образом, методы глубокого обучения демонстрируют значительный потенциал в распознавании кристаллографических закономерностей в обратном пространстве, но сталкиваются с фундаментальным ограничением при решении проблемы фаз. Это ограничение связано с необходимостью чрезвычайно точного численного восстановления амплитуд отражений, что требует более точного подхода к определению значений в каждой точке. Это наблюдение указывает на необходимость разработки новых архитектур или подходов, которые могли бы сочетать способность к распознаванию паттернов с более точным численным восстановлением.

4 Результаты и выводы

1. Разработано программное обеспечение по генерации синтетических рентгенодифракционных данных, которое может быть использовано для решения прикладных задач рентгеновской дифракции с помощью ИИ.
2. Разработан единый автоматизированный конвейер, позволяющий проводить воспроизводимые численные эксперименты по обучению и тестированию моделей машинного обучения, повышающих разрешение рентгенодифракционных данных.
3. Лучшей разработанной моделью оказалась FFT_UNet, предсказывающая ненормализованные амплитуды, фактор расходности при уточнении структуры 6-бром-1Н-индол-2,3-диона составил 9.6%.
4. Проведена интерпретация моделей, подтверждающая, что модели глубокого обучения выявляют кристаллографические закономерности и имеют потенциал решения прикладных задач рентгеновской дифракции.
5. У моделей машинного обучения пока не достает качества численного восстановления рентгеновских отражений для решения проблемы фаз в случае нецентросимметричных структур.

Список литературы

- [1] Girolami G. X-ray Crystallography. — Mill Valley : University Science Books, 2016.
- [2] Hauptman H. The phase problem of X-ray crystallography // Reports on Progress in Physics. — 1951. — no. 54. — P. 1427–1454.
- [3] Lattman E. and Loll P. Protein Crystallography: A Concise Guide. Protein Crystallography. — Johns Hopkins University Press, 2008. —
- [4] Margiolaki I. and Wright J. P. Powder crystallography on macromolecules // Acta Crystallogr A Found Crystallogr. — 2008. — Vol. 64, no. 1. — P. 169–180. — Publisher: International Union of Crystallography (IUCr).
- [5] Favre-Nicolin V. and Černý R. FOX: Modular Approach to Crystal Structure Determination from Powder Diffraction // European Powder Diffraction EPDIC 8.— Trans Tech Publications Ltd. — 2004. — Jan. — Vol. 443 of Materials Science Forum. — P. 35–38.
- [6] Černý R. and Favre-Nicolin V. FOX: A friendly tool to solve nonmolecular structures from powder diffraction // Powder Diffraction. — 2005. — Vol. 20, no. 4. — P. 359–365.
- [7] Černý R., Favre-Nicolin V., Rohlíček J., and Hušák M. FOX, Current State and Possibilities // Crystals. — 2017. — Vol. 7, no. 10. —
- [8] Hauptman H. and Karle J. Solution of the phase problem for space group $P\bar{1}$ // Acta Crystallographica. — 1954. — Vol. 7, no. 4. — P. 369–374.
- [9] Hauptman H. The Direct Methods of X-ray Crystallography // Science. — 1986. — Vol. 233, no. 4760. — P. 178–183.
- [10] Giacovazzo C. Direct Phasing in Crystallography: Fundamentals and Applications. — Oxford University Press, 1998.
- [11] Giacovazzo C. International Tables for Crystallography. — 2 ed. — International Union of Crystallography, 2010. — Vol. B: Reciprocal space.
- [12] Wilson A. J. C. Determination of Absolute from Relative X-Ray Intensity Data // Nature. — 1942. — Vol. 150, no. 3796. — P. 151–152.
- [13] Sayre D. The squaring method: a new method for phase determination // Acta Crystallographica. — 1952. — Vol. 5, no. 1. — P. 60–65.

- [14] Cochran W. A relation between the signs of structure factors // *Acta Crystallographica*. — 1952. — Vol. 5, no. 1. — P. 65–67.
- [15] Karle J. and Karle I. L. The symbolic addition procedure for phase determination for centrosymmetric and non-centrosymmetric crystals // *Acta Crystallographica*. — 1966. — Vol. 21, no. 6. — P. 849–859.
- [16] Burla M. C., Giacovazzo C., and Polidori G. A robust tangent procedure // *Journal of Applied Crystallography*. — 2013. — Vol. 46, no. 6. — P. 1592–1602.
- [17] Rossmann M. G. and Arnold E. Patterson and molecular-replacement techniques // *International Tables for Crystallography Volume B: Reciprocal space* / ed. by Shmueli U. — Dordrecht : Springer Netherlands, 2001. — P. 235–263.
- [18] Buerger M. J. Solution Functions for Solving Superposed Patterson Syntheses // *Proceedings of the National Academy of Sciences*. — 1953. — Vol. 39, no. 7. — P. 674–678.
- [19] Hendrixson T. L. and Jacobson R. A. Locating symmetry elements in Patterson superposition maps // *Zeitschrift für Kristallographie - Crystalline Materials*. — 1997. — Vol. 212, no. 8. — P. 577–585.
- [20] Pavelčík F., Kuchta L., and Sivý J. Patterson-oriented automatic structure determination. Utilizing Patterson peaks // *Acta Crystallographica Section A*. — 1992. — Vol. 48, no. 6. — P. 791–796.
- [21] Sheldrick G. M. Patterson superposition and ab initio phasing // *Macromolecular Crystallography Part A*. — Academic Press, 1997. — Vol. 276 of *Methods in Enzymology*. — P. 628–641.
- [22] Oszlányi G. and Süto A. Ab initio structure solution by charge flipping // *Acta Crystallogr A*. — 2004. — Vol. 60, no. Pt 2. — P. 134–141.
- [23] Oszlányi G. and Süto A. it Ab initio structure solution by charge flipping. II. Use of weak reflections // *Acta Crystallographica Section A*. — 2005. — Vol. 61, no. 1. — P. 147–152.
- [24] Oszlányi G. and Süto A. The charge flipping algorithm // *Acta Crystallographica Section A*. — 2008. — Vol. 64, no. 1. — P. 123–134.
- [25] Dumas C. and van der Lee A. Macromolecular structure solution by charge flipping // *Acta Crystallographica Section D*. — 2008. — Aug. — Vol. 64, no. 8. — P. 864–873.

- [26] Coelho A. A. A charge-flipping algorithm incorporating the tangent formula for solving difficult structures // *Acta Crystallographica Section A*. — 2007. — Vol. 63, no. 5. — P. 400–406.
- [27] van der Lee A. Charge flipping for routine structure solution // *Journal of Applied Crystallography*. — Vol. 46, no. 5. — P. 1306–1315.
- [28] Palatinus L. and Chapuis G. It SUPERFLIP – a computer program for the solution of crystal structures by charge flipping in arbitrary dimensions // *Journal of Applied Crystallography*. — 2007. — Vol. 40, no. 4. — P. 786–790.
- [29] Sheldrick G. M. *SHELXT*— Integrated space-group and crystal-structure determination // *Acta Crystallogr A Found Adv*. — 2015. — Vol. 71, no. 1. — P. 3–8.
- [30] Burla M. C., Giacovazzo C., and Polidori G. From a random to the correct structure: the VLD algorithm // *Journal of Applied Crystallography*. — 2010. — Vol. 43, no. 4. — P. 825–836.
- [31] Burla M. C., Carrozzini B., Cascarano G. L., Giacovazzo C., and Polidori G. Advances in the VLD algorithm // *Journal of Applied Crystallography*. — 2011. — Vol. 44, no. 6. — P. 1143–1151.
- [32] Burla M. C., Giacovazzo C., and Polidori G. Phasing medium-size structures and proteins by the VLD algorithm // *Journal of Applied Crystallography*. — 2011. — Vol. 44, no. 1. — P. 193–199.
- [33] 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.
- [34] Cowtan K. Phase problem in X-ray crystallography, and its solution // *eLS*. — 2003.
- [35] Carrozzini B., De Caro L., Giannini C., Altomare A., and Caliandro R. The phase-seeding method for solving non-centrosymmetric crystal structures: a challenge for artificial intelligence // *Acta Crystallographica Section A*. — 2025. — Vol. 81, no. 3. — P. 188–201.
- [36] Pan T., Dramko E., Miller M. D., Jr G. N. P., and Kyriolidis A. RecCrysFormer: Refined Protein Structural Prediction from 3D Patterson Maps via Recycling Training Runs. — 2025.

- [37] Wang K., Song L., Wang C., Ren Z., Zhao G., Jiazhen D., Di J., Barbastathis G., Zhou R., Zhao J., and Lam E. On the use of deep learning for phase recovery // Light: Science & Applications. — 2024. — Vol. 13.
- [38] 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 // J Appl Crystallogr. — 2002. — Vol. 35, no. 1. — P. 126–136.
- [39] David W. I. F. Powder diffraction peak shapes. Parameterization of the pseudo-Voigt as a Voigt function // Journal of Applied Crystallography. — 1986. — Vol. 19, no. 1. — P. 63–64.
- [40] Groom C. R. and Allen F. H. The Cambridge Structural Database in Retrospect and Prospect // Angew Chem Int Ed. — 2014. — Vol. 53, no. 3. — P. 662–671.
- [41] Zeng K. and Wang Z. 3D-SSIM for video quality assessment // 19th IEEE International Conference on Image Processing. — 2012. — P. 621–624.
- [42] Ronneberger O., Fischer P., and Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. — 2015.
- [43] Yang Y., Lian Q., Zhang X., Zhang D., and Zhang H. HIONet: Deep priors based deep unfolded network for phase retrieval // Digital Signal Processing. — 2023. — Vol. 132. — P. 103797.
- [44] Selvaraju R. R., Cogswell M., Das A., Vedantam R., Parikh D., and Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization // International Journal of Computer Vision. — 2020. — Vol. 128, no. 2. — P. 336–359.

Приложение

Приложение 1. Общий вид скрипта для генерации структурных данных

```
1 SPACE_GROUPS = ['P21', 'C2']
2 ELEMENTS = ["C", "N", "O", "Cl", "Br"]
3 N_ATOMS_LIMS = (10, 30)
4 ATOM_VOLUME_START_WIDTH = (14, 8)
5 d_high = 1.0 # High resolution limit
6 d_low = 1.2 # Low resolution limit
7 n_atoms = sample_gen(range(*N_ATOMS_LIMS))
8
9 # Initialize structure generator
10 str_generator = GenBuilder(
11     classname=core.CctbxStr.generate_packing,
12     sg=sample_gen(SPACE_GROUPS),
13     atoms=sample_gen(ELEMENTS, size=n_atoms),
14     atom_volume=distr_gen(sts.uniform(*ATOM_VOLUME_START_WIDTH)),
15     seed=utils.distr_gen(sts.randint(1, 2**32-1))
16 )
17 def runner(pattern):
18     """Process a single crystal structure pattern.
19     Args:
20         pattern: Crystal structure pattern object
21     Returns:
22         dict: Contains structure parameters and intensity data
23     """
24     ### Здесь разный код в зависимости от задачи ####
25
26 # Processing parameters
27 CHANKS = 50
28 CHANK_SIZE = 10000
29 CORES = 12
30 # Process structures in chunks
31 for i in range(CHANKS):
32     chank = it.islice(str_generator, CHANK_SIZE)
33     pool = mp.Pool(CORES)
34     with pool as p: results = p.map(runner, chank)
35     directory = f'...'
36     if not os.path.exists(directory): os.makedirs(directory)
37     # Save results
38     np.savez_compressed(f'{directory}/{i}',
39                         db=np.array(results)
40     )
41     pool.close()
```

Приложение 2. Функция runner для расчета интенсивностей случайных структур

```
1 def runner(pattern):
2     """Process a single crystal structure pattern.
3
4     Args:
5         pattern: Crystal structure pattern object
6
7     Returns:
8         dict: Contains structure parameters and intensity data
9     """
10    params = pattern.report_params()
11    structure = pattern.structure
12
13    # Calculate structure factors and intensities
14    a_high = structure.structure_factors(d_min=d_high).f_calc().sort()
15    I_high = a_high.as_intensity_array().data().as_numpy_array()
16    ind_high = np.array(list(a_high.indices()))
17
18    a_low = structure.structure_factors(d_min=d_low).f_calc().sort()
19    ind_low = np.array(list(a_low.indices()))
20
21    return {
22        'structure_params': params,
23        'ind_low': ind_low,
24        'I_high': I_high,
25        'ind_high': ind_high
26    }
```

Приложение 3. Функция runner для расчета нормализованных структурных факторов случайных структур

```
1 def runner(pattern):
2     """Process a single crystal structure pattern.
3
4     Args:
5         pattern: Crystal structure pattern object
6
7     Returns:
8         dict: Contains structure parameters and intensity data
9     """
10    params = pattern.report_params()
11    structure = pattern.structure
12
13    # Calculate structure factors and intensities
14    a_high = structure.structure_factors(d_min=d_high).f_calc().sort()
15    elements_to_parse = ['N', 'O', 'C', 'Cl', 'Br']
16    asu_content = {}
17    for scatterer in structure.scatterers():
18        for elem in elements_to_parse:
19            if scatterer.label == elem or scatterer.label.startswith(elem):
20                asu_content[elem] = asu_content.get(elem, 0) + 1
21    I_high = structure.structure_factors(d_min=d_high).f_calc().sort().as_intensity_array()
22    e_high = a_high.as_amplitude_array().normalised_amplitudes(asu_content).array().data()
23    e_high = e_high.astype(np.float32)
24    ind_high = np.array(list(a_high.indices()))
25
26    a_low = structure.structure_factors(d_min=d_low).f_calc().sort()
27    e_low = a_low.as_amplitude_array().normalised_amplitudes(asu_content).array().data()
28    e_low = e_low.astype(np.float32)
29    ind_low = np.array(list(a_low.indices()))
30
31    return {
32        'structure_params': params,
33        'ind_low': ind_low,
34        'I_high': I_high,
35        'ind_high': ind_high,
36        'e_high': e_high,
37        'e_low': e_low
38    }
```

Приложение 4. Функция runner для расчета карт Паттерсона случайных структур

```
1 def runner(pattern):
2     """Process a single crystal structure pattern.
3
4     Args:
5         pattern: Crystal structure pattern object
6
7     Returns:
8         dict: Contains Patterson maps, structure parameters and intensity data
9     """
10    params = pattern.report_params()
11    structure = pattern.structure
12
13    # Calculate structure factors
14    I_high = structure.structure_factors(d_min=d_high).f_calc().sort().as_intensity_array()
15    ind_high = np.array(list(I_high.indices()))
16    I_low = structure.structure_factors(d_min=d_low).f_calc().sort().as_intensity_array()
17    ind_low = np.array(list(I_low.indices()))
18
19    patt_low = pu.calculate_patterson_fft(I_low.data().as_numpy_array(),
20                                         miller_indices = ind_low, map_shape = (12,12,12))
21    patt_high = pu.calculate_patterson_fft(I_high.data().as_numpy_array(),
22                                         miller_indices = ind_high, map_shape = (24,24,24))
23    assert patt_low.min() == 0 and patt_high.min() == 0
24    assert patt_low.max() == 1 and patt_high.max() == 1
25
26    # Prepare intensity and index data
27    I_high = I_high.data().as_numpy_array()
28    I_low = I_low.data().as_numpy_array()
29
30    return {
31        'patt_low': patt_low,
32        'patt_high': patt_high,
33        'structure_params': params,
34        'ind_low': ind_low,
35        'ind_high': ind_high
36    }
```

Приложение 5. Общий вид скрипта для генерации случайных порошковых дифрактограмм

```
1 BKG_MAX_ORDER = 13
2 BKG_MIN_ORDER = 2
3 bkg_generator = ...
4 SPACE_GROUPS = ["P-1", "P21/c", "C2/c", "Pbca", "I41"]
5 ELEMENTS = ["C", "N", "O", "Cl"]
6 N_ATOMS_LIMS = (3, 30)
7 ATOM_VOLUME_START_WIDTH = (14, 8)
8 n_atoms = sample_gen(range(*N_ATOMS_LIMS))
9 str_generator = ...
10 GAUSS_STEPS = 0.01
11 symmetric_profile_generator = ...
12 profile_generator = ...
13 phase_generator = ...
14 GRID = np.linspace(3.0, 90.0, 4351)
15 CUKA1 = [[1.540596, 1]]
16 CUKA12 = [[1.540596, 2/3], [1.544493, 1/3]]
17 pattern_generator = GenBuilder(
18     classname= core.Pattern,
19     waves = sample_gen([CUKA1, CUKA12]),
20     phases = ([phase] for phase in phase_generator),
21     bkg = bkg_generator,
22     scales = distr_gen(sts.uniform(100,20000), size = 1),
23     bkg_range = (sorted(ii) for
24                   ii in utils.distr_gen(sts.uniform(500, 7000), size = 2)))
25 def runner(pattern):
26     return (pattern.report_params(), pattern.pattern(GRID))
27 CHANKS = 2
28 CHANK_SIZE = 50
29 CORES = 4
30 for i in range(CHANKS):
31     chank = it.islice(pattern_generator, CHANK_SIZE)
32     pool = mp.Pool(CORES)
33     with pool as p:
34         results = p.map(runner, chank)
35     y = pd.DataFrame( [ y for y,_ in results ] )
36     x = pd.DataFrame( [ x for _,x in results ] )
37     y.to_csv(f'test_y_{i}.csv')
38     x.to_csv(f'test_x_{i}.csv')
```

Приложение 6. Реализация асимметрии максимумов в рентгеновских порошковых
дифрактограммах

```
1 class AxialCorrection:
2     def __init__(self, profile, HL, SL, N_gauss_step):
3         self.L = 1
4         self.H = HL
5         self.S = SL
6         self.N_gauss_step = N_gauss_step
7         self.profile = profile
8     def h(self, phi, peak):
9         return self.L*np.sqrt(np.cos(phi*np.pi/180)**2/np.cos(peak*np.pi/180)**2 - 1)
10    def phi_min(self, peak):
11        a = np.cos(peak*np.pi/180) * np.sqrt( ((self.H+self.S)/self.L)**2 + 1 )
12        if a > 1 :
13            return 0
14        else:
15            return 180/np.pi*np.arccos( a )
16    def phi_infl(self, peak):
17        a = np.cos(peak*np.pi/180)*np.sqrt( ((self.H-self.S)/self.L)**2 + 1 )
18        if a > 1 :
19            return 0
20        else:
21            return 180/np.pi*np.arccos(a)
22    def W2(self, phis, peak):
23        result = np.zeros(len(phis))
24        cond1 = (self.phi_min(peak) <= phis) & (phis <= self.phi_infl(peak))
25        result[cond1] = self.H + self.S - self.h(phis[cond1], peak)
26        cond2 = (phis > self.phi_infl(peak)) & (phis <= peak)
27        result[cond2] = 2 * min(self.H, self.S)
28        return result
29    def calc(self, Th2, peak):
30        phmin = self.phi_min(peak)
31        dd = np.abs(peak - phmin) / self.N_gauss_step
32        N_gauss = np.ceil(dd).astype(int)
33        if (N_gauss == 1):
34            return self.profile.calc(Th2, peak)
35        xn, wn = np.polynomial.legendre.leggauss(N_gauss)
36        step = Th2[1] -Th2[0]
37        deltan = (peak+phmin)/2 + (peak-phmin)*xn/2
38        tmp_assy = np.zeros(len(Th2))
39        arr1 = wn*self.W2(deltan, peak)/self.h(deltan, peak)/np.cos(deltan*np.pi/180)
40        for dn in range(len(deltan)):
41            tmp_assy += arr1[dn] * self.profile.calc(Th2, deltan[dn])
42        tmp_assy = tmp_assy / np.sum(arr1)
43        return(tmp_assy)
```

Приложение 7. Реализация функции Псевдо-Войдта

```
1  class PV_TCHZ:
2      def __init__(self, parameters):
3          self.U = parameters[0] / 1083 # Follow GSAS conventions
4          self.V = parameters[1] / 1083 # Follow GSAS conventions
5          self.W = parameters[2] / 1083 # Follow GSAS conventions
6          self.X = parameters[3] / 100 # Follow GSAS conventions
7          self.Y = parameters[4] / 100 # Follow GSAS conventions
8          self.Z = parameters[5] / 100 # Follow GSAS conventions
9
10     def fwhmL(self, peak):
11         peak = peak / 180 * np.pi
12         return (self.X * np.tan(peak/2) + self.Y/np.cos(peak/2))
13
14     def fwhmG(self, peak):
15         peak = peak / 180 * np.pi
16         return np.sqrt(self.U * np.tan(peak/2) ** 2 +
17                         self.V * np.tan(peak/2) +
18                         self.W +
19                         self.Z / np.cos(peak/2) ** 2)
20
21     def lorenz(self, Th2, peak, l):
22         return (2 / np.pi / l) / (1 + 4 * (Th2 - peak)**2 / l**2)
23
24     def gauss(self, Th2, peak, g):
25         return (2 * (np.log(2)/np.pi) ** 0.5 / g) *
26             np.exp(-4 * np.log(2) * (Th2 - peak)**2 / g**2)
27
28     def n_for_tchz(self, l, g):
29         G = g ** 5 + 2.69269*g ** 4 * l + 2.42843 * g ** 3 * l ** 2 +
30             4.47163 * g ** 2 * l ** 3
31         G += 0.07842 * g * l ** 4 + l ** 5
32         G = l / (G ** 0.2)
33         n = 1.36603 * G - 0.47719 * G ** 2 + 0.11116 * G ** 3
34         return n
35
36     def tchz(self, Th2, peak, l, g, n):
37         return n* self.lorenz(Th2, peak, l) + (1 - n)* self.gauss(Th2, peak, g)
38
39     def calc(self, Th2, peak):
40         wl = self.fwhmL(peak)
41         wg = self.fwhmG(peak)
42         n = self.n_for_tchz(wl, wg)
43         return self.tchz(Th2, peak, wl, wg, n)
```