Journal of Siberian Federal University. Chemistry 2 (2019 12) 188-200
~ ~ ~

УДК 54.022

# Possibilities of Neural Network Powder Diffraction Analysis Crystal Structure of Chemical Compounds

**Alexander N. Zaloga,**
**Vladimir V. Stanovov, Oksana E. Bezrukova,**
**Petr S. Dubinin and Igor S. Yakimov***
*Siberian Federal University*
*79 Svobodny, Krasnoyarsk, 660041, Russia*

Received 09.12.2018, received in revised form 04.04.2019, accepted 12.05.2019

*Some possibilities of using convolutional artificial neural networks (ANN) for powder diffraction structural analysis of crystalline substances have been investigated. First, ANNs are used to classify crystalline systems and space groups according to calculated full-profile diffractograms calculated from the crystal structures of the ICSD database (2017 year). The ICSD database contains 192004 structures, of which 80% was used for in-depth network training, and 20% for independent testing of recognition accuracy. The accuracy of classification by a network of crystalline systems was 87.9%, and that of space groups was 77.2%. Secondly, the ANN is used for a similar classification of structural models generated by the stochastic genetic algorithm in the search processes for triclinic crystal structures of test compound $K_4SnO_4$ according to their full-profile diffraction patterns. The classification criterion was the entry of one or several atoms into their crystallographic positions in the structure of a substance. Independent deep network training was performed on 120 thousand structural models of the $K_4PbO_4$ triclinic structure generated in several runs of the genetic algorithm. The accuracy of the classification of $K_4SnO_4$ structural models exceeded 50%. The results show that deeply trained convolutional ANNs can be effective for classifying crystal structures according to the structural characteristics of their powder diffraction patterns.*

*Keywords: crystal structure, powder diffraction, genetic algorithms, Rietveld method, artificial neural networks.*

Citation: Zaloga A.N., Stanovov V.V., Bezrukova O.E., Dubinin P.S., Yakimov I.S. Possibilities of neural network powder diffraction analysis crystal structure of chemical compounds, J. Sib. Fed. Univ. Chem., 2019, 12(2), 188-200. DOI: 10.17516/1998-2836-0118.


* Corresponding author E-mail address: i-s-yakimov@yandex.ru

# Возможности нейросетевого порошкового дифракционного анализа кристаллической структуры химических соединений

**А.Н. Залога, В.В. Становов,**
**О.Е. Безрукова, П.С. Дубинин, И.С. Якимов**
*Сибирский федеральный университет*
*Россия, 660041, Красноярск, пр. Свободный, 79*

*Исследованы некоторые возможности применения сверточных искусственных нейронных сетей (ИНС) для порошкового дифракционного структурного анализа кристаллических веществ. Во-первых, ИНС применены для классификации кристаллических систем и пространственных групп симметрии по расчетным полнопрофильным дифрактограммам, вычисленным из кристаллических структур базы данных ICSD 2017 г. База ICSD содержит 192004 структуры, из которых 80 % использовалось для глубокого обучения сети, а 20 % для независимого тестирования точности распознавания. Точность классификации сетью кристаллических систем составила 87,9 %, а пространственных групп – 77,2 %. Во-вторых, другая ИНС применена для классификации структурных моделей, сгенерированных стохастическим генетическим алгоритмом в процессах поиска кристаллических структур тестовых триклинных соединений $K_4SnO_4$ и $K_4SnO_4$, по их полнопрофильным дифрактограммам. Было сгенерировано около 150 тысяч структурных моделей каждой из этих структур. Глубокое обучение сети выполнялось на дифрактограммах структурных моделей $K_4PbO_4$. Обученная сеть была применена для классификации структурных моделей $K_4SnO_4$ по их дифрактограммам. Критерием классификации являлось попадание атомов в их кристаллографические позиции в структуре. Точность классификации адекватных позиций атомов в структурных моделях $K_4SnO_4$ превысила 50 %.*

*Ключевые слова: кристаллическая структура, порошковая дифракция, генетические алгоритмы, метод Ритвельда, искусственные нейронные сети.*

## Введение

Строение кристаллических веществ характеризуется периодичностью и симметрией пространственного расположения атомов и определяет многие физические и химические свойства материалов. Информация об атомной кристаллической структуре вещества включает координаты атомов в элементарной ячейке кристаллической решетки и параметры их тепловых колебаний, и накапливается в структурных базах данных [1, 2]. Основным методом для изучения структуры новых веществ, получаемых в поликристаллической форме, служит рентгеновская порошковая дифракция. Структурное исследование включает определение приближенной модели атомной кристаллической структуры и ее оптимизацию. Исходными данными для определения модели структуры служат химическая формула, параметры осей

***a***, ***b***, ***c*** кристаллической решетки, пространственная группа симметрии и полнопрофильная порошковая дифрактограмма вещества, которая может быть рассчитана из атомной кристаллической структуры [3]. Оптимизация модели кристаллической структуры осуществляется с помощью метода полнопрофильного анализа дифрактограмм Ритвельда [4]. Основным критерием оптимизации служит минимум профильного R-фактора – относительной невязки между расчетной и экспериментальной дифрактограммами. Проблемами дифракционного порошкового структурного анализа является поиск параметров решетки, пространственной группы и модели кристаллической структуры, подходящей для ее оптимизации методом Ритвельда.

Одним из наиболее перспективных направлений научных исследований в последние годы стало развитие методов и приложений искусственных нейронных сетей (ИНС), основанных на новых методиках их глубокого обучения [5]. Однако, для структурного анализа ИНС практически не применяются. Причиной является необходимость использования очень большого количества однородной структурной информации для глубокого обучения ИНС. Несколько ранних попыток применения ИНС представлены в [6 – 8]. В работе [6] была исследована возможность использования ИНС для прогнозирования каркасных кристаллических структур цеолитов на основе связывания структурных данных с данными порошковой рентгеновской дифракции (ПРД). В [7] предлагается применить ИНС к проблеме количественного анализа смоделированных бинарных и тройных смесей глинистых минералов по данным ПРД с использованием минимизации по алгоритму Левенберга-Марквардта. Интересный подход к определению параметров элементарной ячейки с помощью ИНС непосредственно из данных ПРД описан в [8]. Сеть была обучена с использованием имитированных данных ПРД и успешно применена для определения параметров элементарной ячейки двух веществ с использованием экспериментальных рентгенограмм, записанных на лабораторном дифрактометре.

Более эффективные подходы к структурному анализу начали недавно развиваться на основе глубокого обучения сверточных нейронных сетей [5]. В работе [9] предложен нейросетевой метод классификации двумерных изображений рассеяния рентгеновских лучей. Визуальный анализ изображений рассеяния рентгеновских лучей является мощным методом исследования физической структуры материалов в молекулярном масштабе. Изображения содержат визуальные образы, такие как кольца, пятна и ореолы, которые кодируют подробную информацию о размере, ориентации и упаковке атомов, молекул и наноразмерных областей. Современные рентгеновские детекторы могут генерировать от 50 000 до 1 000 000 изображений в день, таким образом, важно автоматизировать рабочий процесс обработки изображений. Для классификации изображений применены сверточные нейронные сети и сверточные автокодеры. Чтобы получить достаточное количество данных для глубокого обучения ИНС, использовано программное моделирование синтетических изображений рентгеновского рассеяния. В работе [10] предложен подобный инструмент для скрининга макромолекулярных рентгеноструктурных дифракционных двумерных изображений, полученных на рентгеновском лазерном источнике света с легкими электронами. Инструмент представляет собой сверточную нейронную сеть для детектирования Брэгговских рефлексов из шумовых данных с экспериментальными артефактами. Глубокое обучение основано на

данных редуцирования изображений, записанных современными детекторами, используемыми в экспериментах по рентгеновской кристаллографии. Для создания большего числа обучающих изображений применена их генерация, основанная на случайных перемещениях дискретизированных изображений на несколько пикселей. В работах [11, 12] предложен подход для автоматической классификации структуры по симметрии кристалла, основанный на глубоком обучении сверточных ИНС и классификации модельных двумерных монокристальных рентгенограмм. Идентификация симметрии решетки является первым шагом для характеристики структуры материалов и аналитики. Подход позволил правильно классифицировать набор данных, содержащих более 100 000 смоделированных кристаллических структур, включая сильно дефектные.

Применение сверточных нейронных сетей для классификации симметрии кристаллических веществ по порошковым дифрактограммам предложено в публикации [13]. В качестве входных данных для глубокого обучения нейронных сетей использовано 150 000 расчетных полнопрофильных дифрактограмм, вычисленных из структур, отобранных из базы данных ICSD (Inorganic Crystal Structure Database) с исключением наименее достоверных. Подход не использовал деконволюцию, дискретное положение или данные интенсивности пиков, но вместо этого дифрактограммы рассматривались как изображение. В результате была сконструирована архитектура сети, которая позволила определить кристаллическую систему, группу экстинкции и пространственную группу с вероятностью 94,99, 81,14, 83,83 %, соответственно. Обученная сеть затем была применена для идентификации симметрии нескольких новых неорганических соединений по экспериментальным порошковым дифрактограммам.

Из данного обзора можно сделать два вывода. Во всех новых работах использовано глубокое обучение сверточных нейронных сетей на основе большого количества модельных данных. ИНС пока не применяются для собственно дифракционного структурного анализа, т.е. для непосредственного определения атомно-кристаллической структуры веществ по дифрактограммам.

В данной работе сообщается о наших результатах исследования возможностей порошкового дифракционного структурного анализа на основе глубокого обучения ИНС. Во-первых, сконструирована сверточная сеть, аналогичная [13], и выполнены ее обучение и независимая оценка точности классификации кристаллических систем и пространственных групп симметрии по расчетным полнопрофильным дифрактограммам, вычисленным из кристаллических структур базы данных ICSD. Во-вторых, предложена и апробирована сверточная ИНС, обучаемая для селекции структурных моделей с адекватными атомными фрагментами с помощью дифрактограмм, генерируемых генетическим алгоритмом в процессах ab initio поиска кристаллических структур.

## 1. Нейросетевая классификация симметрии кристаллических структур

*Объекты и методы исследования*

Объектами исследования служат расчетные полнопрофильные дифрактограммы, вычисленные из кристаллических структур всей неорганической базы данных ICSD 2017 г., включа-

Таблица 1. Некоторые характеристики распределения структур в базе данных ICSD

Table 1. Some characteristics of the distribution of structures in the ICSD database

| Распределение структур по годам определения | | Распределение структур по объему ячейки | | Распределение структур по точности определения (по значениям R-фактора) | |
|---|---|---|---|---|---|
| годы | количество | объем ($A^3$) | количество | R-фактор | количество |
| <1900 | 923 | < 400 | 95103 | <0.01 | 1293 |
| 1900-1969 | 23048 | 400-800 | 44485 | 0.01-0.02 | 11540 |
| 1970-1979 | 20486 | 800-1200 | 21949 | 0.02-0.03 | 23984 |
| 1980-1983 | 10234 | 1200-1600 | 10050 | 0.03-0.04 | 21260 |
| 1984-1987 | 11851 | 1600-2000 | 5815 | 0.04-0.05 | 17923 |
| 1988-1991 | 12315 | 2000-2400 | 3445 | 0.05-0.06 | 14205 |
| 1992-1995 | 14183 | 2400-2800 | 2196 | 0.06-0.07 | 8787 |
| 1996-1999 | 12977 | 2800-3200 | 1696 | 0.07-0.08 | 6771 |
| 2000-2003 | 16053 | 3200-3600 | 1119 | 0.08-0.09 | 5492 |
| 2004-2007 | 19631 | 3600-4000 | 825 | 0.09-0.10 | 3182 |
| 2008-2011 | 21714 | > 4000 | 5321 | 0.10-0.15 | 7831 |
| 2012-2017 | 28589 | | | >0.15 | 2715 |
| | | | | Нет данных | 67021 |

ющей 192004 структуры. Дифрактограммы были вычислены для излучения CuKα1, содержали 10000 точек с шагом 0,01 по углу дифракции 2Тета и были нормированы по интенсивности к 1000. Профиль линий моделировался функцией псевдо-Войгта со средней полушириной линий 0,1 град. по 2Тета, фон – линейной функцией. В табл. 1 приведены некоторые данные о структурах базы ICSD, характеризующие качество их определения.

Методом исследования стала нейросетевая классификация дифрактограмм по кристаллическим системам и пространственным группам симметрии с помощью сверточной ИНС. Для глубокого обучения нейронной сети применены 80 % дифрактограмм, а остальные 20 % использовались для независимого тестирования обученной ИНС. Распределение структур в базе ISCD по кристаллическим системам и особенно по пространственным группам очень неравномерное, например, в 4-х группах находится более 10000 структур в каждой, а в 12-ти – менее, чем по 10. Это было учтено при создании тестовой выборки, в которую из каждой системы и группы случайным образом отбирали по 20 % структур плюс 1 (чтобы представители нескольких групп с числом структур менее 5 не были пустыми).

Архитектура экспериментальной нейронной сети основана на структуре сверточных сетей, подобных [14] и используемых для классификации изображений, и близка к описанной в [13]. В частности, сеть состоит из трех слоев свертки с макс-пулингом и трех полносвязных слоев. При этом для каждой из классификаций (по 7 кристаллическим системам и по 230 пространственным группам) после слоев свертки использовались три своих полносвязных слоя. Функция потерь рассчитывалась как перекрестная энтропия между номерами классов, представленных в виде унитарного кода и выходами софтмакс-слоя. При этом перекрестные энтропии, рассчитанные по первой и второй классификации, суммировались. Веса на всех слоях

инициализировались по нормальному закону с нулевым математическим ожиданием и среднеквадратичным отклонением 0.1. Оптимизация весов производилась алгоритмом адаптивного момента (ADAM), представленным в [15] с начальным коэффициентом обучения 0.0005. Продолжительность обучения составляла 50 эпох, размер подвыборки (batch size) устанавливался случайным в диапазоне от 100 до 500. Программный код был написан на языке Python 3.5.2 [16] с использованием библиотеки машинного обучения Tensorflow 1.8 [17]. Для обучения применяли графический ускоритель NVidia GeForce GTX 1070, основная система базировалась на процессоре Intel Core i7 2600K.

В ходе тестирования проверялись различные параметры сети, в том числе варьировались типы активационных функций, добавлялись слои с выпадением, менялось количество фильтров в сверточной части и количество нейронов в полносвязных слоях. Кроме того, испытывались различные типы оптимизаторов и их начальные коэффициенты обучения. Архитектура сети позволяет достичь достаточной точности при сравнительно небольшом количестве связей. Точность классификации рассчитывала как частота распознавания, т.е. отношение количества верно классифицированных измерений к их общему числу:

$$f = \frac{\sum_{i=1}^{N} F(\hat{y}_i, y_i)}{N}, \quad (1)$$

где $N$ – число измерений в выборке; $y_i$ – истинный номер класса; $\hat{y}_i$ – предсказанный номер класса; $F(a, b) = 1$, если $a = b$, иначе 0.

В результате экспериментального выбора подходящих параметров сети точность классификации на 20%-й случайной тестовой выборке из ICSD составила 87,2 % для кристаллических систем и 75,6 % для 230 пространственных групп. При заметном увеличении количества фильтров (от 120 до 150) или их уменьшении (до 80) точность классификации снижается (до 87,1 и 75,8 %, а также до 85,8 и 73,4 %, соответственно). Увеличение количества слоев свертки с трех до четырех не дает существенных преимуществ, однако увеличивает время счета. При использовании больших коэффициентов обучения на первых этапах сходимость значительно быстрее; например, после первой эпохи точность может быть порядка 75 и 65 %, однако окончательный результат, как правило, оказывается хуже. Дальнейшее увеличение размерностей полносвязных слоев также не повышает точности классификации, но значительно увеличивает количество параметров и сложность модели. Что касается активационных функций, то при использовании функции ReLU вместо сигмоиды в полносвязных слоях наблюдалась нестабильная работа, т.е. несколько первых эпох происходило быстрое увеличение точности, однако после этого сеть начинала относить все измерения к одному из классов, очевидно, вследствие слишком большого момента, накопленного на начальных эпохах, веса становились слишком большими, что приводило к некорректной классификации.

*Обсуждение результатов*

Для анализа результатов работы сети были построены матрицы ошибок классификации по тестовой выборке и гистограммы распределения неверно классифицированных измерений. Таблица 2 содержит матрицу результатов и ошибок сети для 7 кристаллических систем, расположенных по возрастанию симметрии элементарных ячеек. Жирным выделено количество

Таблица 2. Матрица результатов и ошибок классификации для 7 кристаллических систем

Table 2. Matrix of results and classification errors for 7 crystalline systems

| № системы | Результаты нейросетевой классификации | | | | | | | Всего | Точность % |
|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | | |
| **1** | **869** | 481 | 71 | 16 | 7 | 5 | 12 | 1461 | 59,5 |
| **2** | 356 | **4914** | 718 | 47 | 55 | 16 | 8 | 6114 | 80,4 |
| **3** | 99 | 735 | **6653** | 255 | 118 | 77 | 42 | 7979 | 83,4 |
| **4** | 16 | 65 | 252 | **5327** | 51 | 78 | 161 | 5950 | 89,5 |
| **5** | 19 | 79 | 141 | 89 | **3266** | 149 | 123 | 3866 | 84,5 |
| **6** | 10 | 30 | 128 | 69 | 160 | **3914** | 110 | 4421 | 88,5 |
| **7** | 1 | 1 | 13 | 13 | 18 | 7 | **8548** | 8601 | 99,4 |

Таблица 3. Нумерация и параметры ячейки кристаллических систем

Table 3. Numbering and cell parameters of crystalline systems

| № системы | Кристаллографическая система | Соотношения осей и углов элементарной ячейки |
|---|---|---|
| 1 | Триклинная | $a \neq b \neq c$, $\alpha \neq \beta \neq \gamma \neq 90^o$ |
| 2 | Моноклинная | $a \neq b \neq c$, $\alpha = \beta = 90^o \neq \gamma$ |
| 3 | Орторомбическая | $a \neq b \neq c$, $\alpha = \beta = \gamma = 90^o$ |
| 4 | Ромбоэдрическая | $a = b = c$, $\alpha = \beta = \gamma \neq 90^o$ |
| 5 | Гексагональная | $a = b \neq c$, $\alpha = \beta = 90^o$, $\gamma = 120^o$ |
| 6 | Тетрагональная | $a = b \neq c$, $\alpha = \beta = \gamma = 90^o$ |
| 7 | Кубическая | $a = b = c$, $\alpha = \beta = \gamma = 90^o$ |

правильно классифицированных систем, в предпоследней графе указано общее количество членов каждой системы в тестовой выборке, а в последней – точность классификации, рассчитанная согласно (1). Нумерация семи кристаллических систем и соответствующие им соотношения параметров элементарных ячеек приведены в табл. 3.

В результате оптимизации выбранных параметров сети общая точность классификации на 20%-й случайной тестовой выборке из ICSD несколько повысилась и составила 87,9 % для кристаллических систем и 77,2 % для 230 пространственных групп. На рис. 1 представлена гистограмма распределения полного состава членов в системах и ошибок их классификации, соответствующая табл. 2 и 3. На рис. 1 видно, что для наиболее симметричной кубической системы ошибки классификации практически отсутствуют, а большая часть ошибок приходится на первые три низкосимметричные системы.

Из данных табл. 2 видно, что подавляющее число ошибок классификации этих систем приходится на соседние системы: у триклинной на моноклинную (481), у моноклинной на орторомбическую (718), а у орторомбической на моноклинную (735). Это объясняется тем, что у многих веществ этих систем параметры решетки близки к соответствующим параметрам одной из соседних систем, например, часть углов ячейки близка к 90 градусам. При этом со-

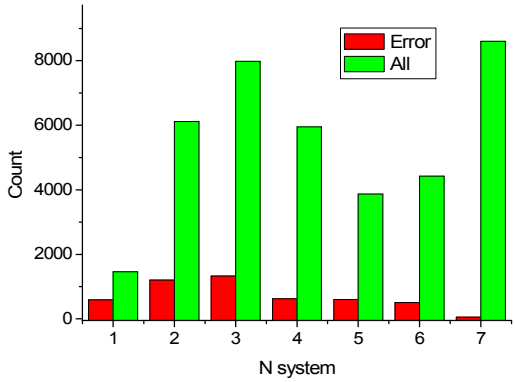


Рис. 1. Гистограмма распределения ошибок классификации (error) в тестовой выборке для 7 кристаллических систем

Fig. 1. Histogram of the classification error distribution (error) in the test sample for 7 crystal systems

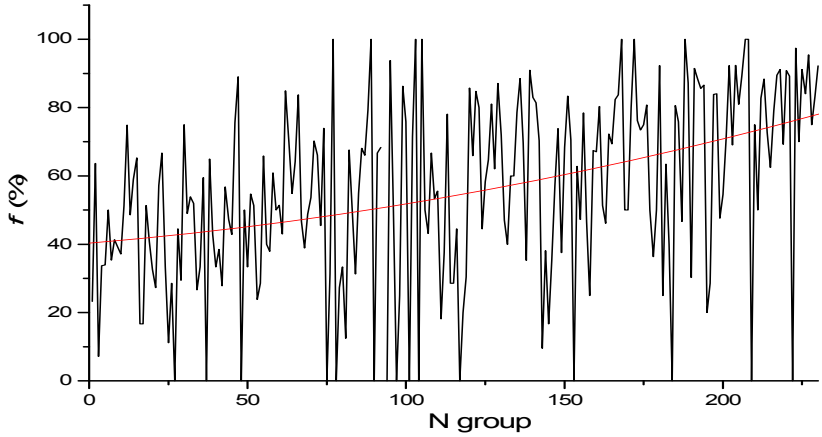


Рис. 2. График распределения частоты распознавания пространственных групп (в %) в тестовой выборке; красная линия – квадратичная регрессия

Fig. 2. Graph of the distribution of the recognition frequency of space groups (%) in the test set; red line – square regression

ответствующие дифракционные линии расщепляются (или сливаются), а сеть оказалась недостаточно чувствительной для учета этих небольших различий при классификации. Для систем средней симметрии данное явление тоже имеет место, хотя и в меньшей степени. Из этого можно сделать вывод о том, что для повышения точности классификации следует улучшать способы сетевой фильтрации, а при анализе реальных дифрактограмм с помощью ИНС выполнять их съемку с узкими гониометрическими щелями для уменьшения инструментальной ширины линий.

На рис. 2 представлен график точности классификации 230 пространственных групп в тестовой выборке, вычисленной индивидуально для каждой группы по (1) и выраженной в %. В целом, вероятность распознавания пространственных групп возрастает с ростом симметрии, в среднем от 40 до почти 80 % (красная линия на рис. 2). Однако распределение этой вероятности

по конкретным группам крайне неравномерная и заметно зависит от числа членов в них, хотя четкой корреляции нет.

Можно отметить, что свыше 10 групп классифицировано в тестовой выборке с вероятностью, близкой к 100 %, в т.ч. для вышеупомянутых 4-х наибольших групп. В то же время сеть не смогла классифицировать ни одной дифрактограммы только в 15 пространственных группах ($f$=0), причем только таких, в которых количество структур в базе ICSD менее 20-ти (всего таких групп 30). Очевидно, этого количества оказалось недостаточно для обучения сети, но, по-видимому, его можно улучшить, например, путем эмуляции достаточного количества структур для таких групп.

Следует отметить, что практическое определение пространственной группы выполняется путем анализа погасаний части рентгеновских рефлексов, обусловленных симметрией кристаллической структуры. Сложность задачи в том, что для определения конкретной группы симметрии из 230 возможных имеется всего 122 различных набора правил погасания (дифракционных групп). По ним однозначно определяется только 61 группа, а остальным наборам погасаний отвечают по две, три или четыре пространственные группы [18]. Если пространственная группа определяется неоднозначно, дальнейшее структурное исследование приходится проводить, учитывая варианты, основанные на каждой из возможных групп симметрии. Однозначное же определение сетью большинства представителей из 215 пространственных групп симметрии (для которых $f$>0 на рис. 3) показывает, что сеть обладает уникальной классификационной способностью. Очевидно, это связано с тем, что сеть использует при обучении и классификации не только позиции рефлексов, но и их интенсивности, т.е. неявным образом учитывает связь симметрии с кристаллической структурой вещества.

Отметим также, что распознавание обучающей выборки уже обученной сетью имеет частоту правильной классификации 96,4 % для 230 пространственных групп и 98,3 % для 7 кристаллических систем. Очевидно, столь высокая вероятность вызвана эффектом некоторого «переобучения» сети, когда большое количество весовых коэффициентов нейронов оказались «подогнанными» к индивидуальным дифрактограммам. По-видимому, это должно было снизить точность классификации независимой тестовой выборки. Таким образом, определено еще одно направление дальнейшего повышения точности классификации, основанное на устранении эффекта переобучения сети.

В качестве конечного результата этих исследований ожидается разработка способа высокоавтоматизированного определения с помощью обученной сети кристаллических систем и пространственных групп симметрии новых веществ, необходимых для дальнейшего определения их кристаллической структуры методами глобальной оптимизации в прямом пространстве, в частности, с помощью разработанного нами мультипопуляционного генетического алгоритма [12].

## 2. Нейросетевая классификация тестовых структурных моделей

*Объекты и методы исследования*

Методом исследования являлась нейросетевая классификация тестовых структурных моделей, генерируемых генетическим алгоритмом [19, 20] в процессах поиска атомной кри-

сталлической структуры веществ по их порошковым дифрактограммам. Эволюционные генетические алгоритмы (ГА) являются одним из средств ab initio поиска кристаллической структуры химических соединений в прямом пространстве [21]. Суть ГА заключается в моделировании операций естественного биологического отбора: парного скрещивания, мутации и селекции множества (популяции) тестовых структурных моделей для передачи лучших моделей – потомков на новые поколения эволюции с целью поэтапного формирования структуры. Проблема ГА – ухудшение сходимости эволюционного процесса из-за недостаточно эффективной селекции структурных моделей с уже сформированными адекватными атомными фрагментами при повышении сложности определяемых структур. Для улучшения селекции нами предлагается нейросетевая классификация структурных моделей предварительно обученной сверточной ИНС. Глубокое обучение ИНС предлагается проводить на множествах структурных моделей, создаваемых в процессах поиска по ГА известных структур химических соединений, близких к определяемой структуре по составу и симметрии. Критерием обучения и классификации является попадание одного или нескольких атомов структурной модели в ближайшую окрестность их позиций в структуре вещества (~0,1 ангстрем). В отличие от рассмотренной выше задачи нейросетевой классификации симметрии обучение и классификацию моделей предложено проводить не на самих расчетных полнопрофильных дифрактограммах, а на их разностных профилях с экспериментальной дифрактограммой вещества – таких же, как в методе Ритвельда [22].

Объектами исследования стали модели атомных кристаллических структур тестовых триклинных соединений $K_4PbO_4$ и $K_4SnO_4$ с 27 степенями свободы атомных координат, генерируемые в процессе их поиска по методу ГА [20]. Дифрактограммы структурных моделей были вычислены так же, как описано выше, но число точек уменьшено до 5000 (с шагом 0,015 по углу дифракции 2Тета). Для глубокого обучения сверточной ИНС использовалось около 120 тысяч разностных профилей дифрактограмм структурных моделей $K_4PbO_4$. Обучение выполнялось в течение 100 эпох, по 1,2 тысячи дифрактограмм на эпохе. Обученная в последовательности этих эпох сеть применялась для выявления атомов в структурных моделях другого соединения ($K_4SnO_4$), которые расположены в указанной окрестности их кристаллографических позиций в структуре данного соединения.

Архитектура экспериментальной сверточной нейронной сети близка к вышеописанной для задачи нейросетевой классификации симметрии кристаллических веществ. Однако, с учетом специфики данной задачи выход сети изменен, в частности, вместо использования SoftMax – слоя в выходном слое – используются нейроны с сигмоидальной функцией. Дело в том, что для задач двоичной классификации на выходе достаточно только одного бита «да/нет». Для регистрации же попадания атомов в окрестность их истинных позиций в структуре необходимо задавать число бит, соответствующее числу атомов структуры (в данном случае N = 9), каждый из которых может быть либо 0 (не попал), либо 1 (попал). При этом возможно попадание одновременно нескольких атомов в истинные позиции. Поэтому в данном случае использование перекрестной энтропии в качестве функции потерь невозможно и на выходе установлена сигмоида, которая возвращает число между 0 и 1. Это число можно рассматривать как «уверенность» сети в том, что данный атом попал в нужную точку. В процессе обучения сети на структурных моделях обучающей выборки по

методике обратного распространения ошибки рассчитывается и минимизируется среднеквадратичная ошибка, т. е.:

$$E = \frac{\sum_{i=1}^{N}(Y_i - U_i)^2}{N},$$

где $U_i$ – метки попадания атомов из выборки, т. е. строка из 0 и 1; $Y_i$ – выходы сети, вещественные числа в интервале [0, 1]; N – число атомов.

Таким образом, чем ближе Y к U для конкретных экземпляров моделей, тем выше качество обучения. Однако, для итоговой оценки качества распознавания необходима другая мера, например точность распознавания попаданий каждого из атомов в отдельности:

$$f_i = \frac{\sum_{j}^{M} F(round(Y_{ij}), U_{ij})}{M},$$

где i – номер атома, i=1…N; M – число измерений в выборке; j – номер измерения; F возвращает единицу, если округленное значение совпадает со значением из выборки (округление до 0 или 1 происходит в сторону ближайшего целого).

*Обсуждение результатов*

Общая выборка из 150 тысяч разностных дифрактограмм структурных моделей соединения $K_4PbO_4$ была разбита случайным образом на две части, обучающую и валидационную, в соотношении 80:20. Обучающая выборка использовалась для обучения ИНС, а валидационная – для первичного тестирования и более точной настройки сети. Обученная в нарастающей последовательности эпох сеть применялась для тестирования качества распознавания адекватных позиций атомов структурных моделей соединения $K_4SnO_4$. При этом, множество тестируемых структурных моделей было разбито на 100 одинаковых порций, в соответствии с количеством эпох обучения. Результаты тестирования для атомов, расположенных в 9 конкретных кристаллографических позициях этой структуры, представлены на рис. 3. По оси абсцисс указаны номера эпох, по оси ординат – доли выявленных сетью позиций атомов относительно существующих в данной порции структурных моделей в данной позиции. Горизонтальная пунктирная линия указывает общую долю атомов в данной позиции во всем множестве структурных моделей $K_4SnO_4$ и позволяет визуально оценивать качество их распознавания сетью на эпохах обучения.

Из графиков можно заключить, какое количество эпох обучения является оптимальным для данной задачи. После первых эпох обучения сети распознавание еще практически не работает, далее доля распознанных атомов во всех позициях постепенно нарастает и достигает максимума после 50-70 эпох обучения. Здесь процент распознавания для всех атомов выше 50 %. После этого он несколько понижается, что, по-видимому, обусловлено эффектом переобучения сети. Следует отметить, что даже точность распознавания порядка 20 % может существенно повысить эффективность структурного анализа по ГА, т.к. большинство ошибочно отобранных сетью моделей будут отсеиваться по R-фактору, а адекватные модели, имея статистически меньший R-фактор, наоборот, будут преимущественно накапливаться в популяции при селекции.

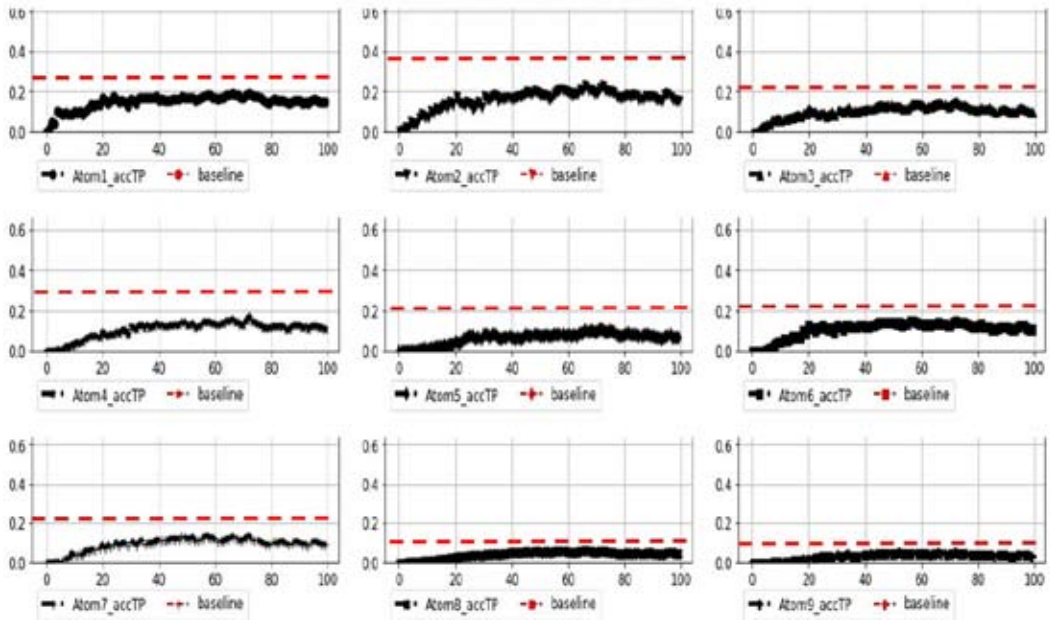


Рис. 3. Графики распознавания ИНС атомов в структурных моделях ГА, расположенных в своих кристаллографических позициях в структуре $K_4SnO_4$; порядок графиков (слева направо): 1 – для атома Sn, 2-5 – для 4-х атомов K, 6-9 – для 4-х атомов O

Fig. 3. Recognition of atoms, located in true crystallographic positions in structural models GA for $K_4SnO_4$, by neural network; order of graphs (from left to right): 1 – for Sn atom, 2-5 – for 4 K atoms, 6-9 – for 4 O atoms

## Заключение

Точность классификации сетью симметрии независимого множества тестовых структур ICSD по их расчетным порошковым дифрактограммам составила для кристаллических систем 87,9 %, а для пространственных групп 77,2 %. Аналогично точность классификации множества структурных моделей ГА независимого тестового соединения $K_4SnO_4$ по критерию расположения атомов в структуре заметно выше 50 %. Полученные результаты показывают, что обученные глубокие сверточные ИНС могут быть весьма эффективными для классификации кристаллических структур химических соединений по структурно обусловленным особенностям их полнопрофильных порошковых дифрактограмм.

## Список литературы

1. Inorganic Crystal Structure Database. FIZ Karlsruhe. http://www.fiz-karlsruhe.de/icsd.html.

2. Cambridge Structural Database. Cambridge Crystallographic Data Centre. http://www.ccdc.cam.ac.uk/products/csd/

3. Powder Diffraction Theory and Practice, ed. R.E. Dinnebier and S.J.L. Billinge. Royal Society of Chemistry, 2008. 507P.

4. Young R.A. The Rietveld Method. Oxford University Press. 1995. 298P.

5. Le Cun Yann, Bengio Y., Hinton G. Deep learning. *Nature 2015.* Vol. 521. P. 436–444.

6. Tatlier M. Artificial neural network methods for the prediction of framework crystal structures of zeolites from XRD data. *Neural Computing and Applications 2011.* Vol. 20(3), P. 365-371.

7. Griffen D.T. Quantitative phase analysis of clay minerals by X-ray powder diffraction using artificial neural networks. I. Feasibility study with calculated powder patterns. *Clay Minerals 1999*. Vol. 34, P. 117-126.

8. Habershon S., Cheung E.Y., Harris K.D.M., Johnston R.L. Powder Diffraction Indexing as a Pattern Recognition Problem: A New Approach for Unit Cell Determination Based on an Artificial Neural Network. *J. Phys. Chem. A 2004.* Vol. 108(5), P. 711–716.

9. Wang B., Yager K., Yu D., Hoai M. X-ray Scattering Image Classification Using Deep Learning. *Winter Conference on Applications of Computer Vision 2017.* United States, P. 697-704.

10. Ke T.W., Brewster A.S., Yu S.X., Ushizima D., Yang Ch., Sauter N.K. A convolutional neural network-based screening tool for X-ray serial crystallography. *Journal of Synchrotron Radiation 2018*. Vol. 25(3), P. 655-670.

11. Ziletti A., Kumar D., Scheffler M., The face of crystals: insightful classification using deep learning. 2017. arXiv preprint arXiv:1709.02298.

12. Ziletti A., Kumar D., Scheffler M., Ghiringhelli L.M. Insightful classification of crystal structures using deep learning. *Nature Communications 2018*. Vol. 9, P. 1-10.

13. Park W.B., Chung J. et al. Classification of crystal structure using a convolutional neural network. *IUCrJ 2017.* Vol. 4, 486–494.

14. Krizhevsky A., Sutskever I., Hinton G.E. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems 2012*. Vol. 10. P. 1097-1105.

15. Kingma D.P., Ba J.L., ADAM: a method for stochastic optimization. *Published as a conference paper at ICLR 2015*. Version, v9.

16. https://www.python.org/downloads/release/python-352/

17. https://www.tensorflow.org/versions/r1.8/

18. Порай-Кошиц М.А. Основы структурного анализа химических соединений. М.: Высшая школа, 1989. 192с. [Poray-Koshic M.A. Basics of structural analysis of chemical compounds. Moscow: High School, 1989. 192 p. (In Russ.)]

19. Залога А.Н., Бураков С.В., Семенкин Е.С., Якимов И.С. Мультипопуляционный генетический алгоритм моделирования кристаллической структуры вещества из рентгенодифракционных данных. *Журнал Сибирского федерального университета. Серия Химия. 2014*. Т. 7(4), С. 572-580. [Zaloga A.N., Burakov S.V., Semenkin E.S., Yakimov I.S. Multipopulation genetic algorithm for modeling the crystal structure of a substance from X-ray diffraction data. *Journal of the Siberian Federal University. Chemistry series 2014*. Vol. 7 (4), P. 572-580. (In Russ. )]

20. Zaloga A.N., Burakov S.V., Semenkin E.S., Yakimov I.S. Research on convergence of multipopulation binary- and real-coded genetic algorithms for solution of crystal structure from X-Ray powder diffraction data. *Crystal Research and Technology 2015*. Vol. 50 (9-10), P. 724–728.

21. David W.I.F., Shankland K. Structure determination from powder diffraction data. *Acta Cryst. 2008*. Vol. 64, P. 52–64.

22. Young R.A. The Rietveld Method. *Oxford University Press. 1995*. P. 298.