ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ
УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
ВЫСШАЯ ШКОЛА ЭКОНОМИКИ

ОТЧЕТ ПО НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ

Решение проблемы фаз в кристаллографии методами машинного обучения

Студент [redacted]

[redacted]

Группа [redacted]

Научный руководитель [redacted]

Москва 2020

Содержание

1 ВВЕДЕНИЕ

Проблема восстановления фаз из интенсивностей преобразования Фурье, нерешаемая в общем слуае, является одной из центральных проблем рентгеноструктурного анализа. Существуют как экспериментальные, так и расчетные методы определения фаз рентгеновских отражений, хорошо работающие для малых молекул. Однако для крупных молекул такие методы работают плохо или не работают вообще[3]. Также все существующие методы имеют невысокую точность и требуют участия человека в решении кристаллов, что затрудняет автоматизацию. Таким образом, задача разработки метода, способного без участия человека решать кристаллы с приемлемой точностью все еще актуальна.

2 ОБЗОР ЛИТЕРАТУРЫ

Машинное обучение и нейронные сети в частности помогли решить множество проблем, ранее считавшиеся крайне сложными. В частности, с помошью методов машинного обучения была решена [4] задача классификации картин порошковой рентгеновской дифракции, определения параметров кристаллической решетки[1]. Также машинное обучение широко применяется в науке о материалах[6].

Актуальные методы решения проблемы фаз в рентгенструктурном анализе можно разделить на три типа: методы в прямом пространстве, методы в обратном пространстве, методы в ???(как по русски dual-space?)[2]. Рассмотрим основные методы решения проблемы фаз.

2.1 Метод Паттерсона

Структура малых молекул может быть решена даже при отсутствии информации о фазах. В то время как фазы определяют положения пиков электронной плотности поперек элементарной ячейки и, следовательно, положения атомов, одно только сильное дифракционное пятно дает четкое указание на то, что элементы должны присутствовать с соответствующим интервалом. Таким образом, только величины структурного фактора содержат информацию о расположении атомов в структуре.

Доступ к этой информации можно получить, вычислив функцию Паттерсона[5]. Функция Паттерсона получается путем расчета карты с использованием квадратов величин структурных факторов и всех фаз, установленных на ноль. Вместо пиков в положениях атомов карта Паттерсона показывает пики в каждой позиции, которая соответствует межатомному вектору в структуре. Функция Паттерсона была эффективным инструментом для решения проблем малых молекул; однако его полезность быстро па-

дает с увеличением числа атомов. Для структуры из N атомов функция Паттерсона будет содержать $N(N-1)$ межатомных векторов, многие из которых перекрываются. Этот подход становится непригодным для структур из более чем 20–50 атомов, если не существует подмножества атомов с высоким атомный номер.

3 ОСНОВНАЯ ЧАСТЬ

Все модели обучались на суперкомпьютере НИУ ВШЭ «Харизма». Для создания моделей использовался язык python и билиотеки машинного обучения Tensorflow и Keras. Все модели использовали бинарную кроссэнтропию, взвешенную по интенсивности отражения, в качестве функции ошибки и точность предсказаний в качестве метрики, если не указано иного.

3.1 Получение данных. Датасеты

Экспериментальные кристаллографические данные были получены из Acta Cristallographica Section E в виде cif-файлов. Данные были обработаны с помощью python и библиотек pandas и CCTBX для получения пригодного для обучения датасета: были отобраны отражения с индексами $h,k,l \in [-10,10]$, удалены эквивалентные отражения(для группы $P-1$ эквивалентными являются отражения удовлетворяющие условию $h_1 = -h_2, k_1 = -k_2, l_1 = -l_2$). Размер полученного датасета составил 7000 файлов - 6000 в обучающем датасете, 1000 в тестовом, по 4630 отражений в каждом. Проверка корректности расчета фаз отражений библиотекой cctbx была показана проверкой корреляции рассчетных интенсивностей с экспериментальными(рис. 1).

По итогам этой проверки стало ясно что интенсивности в файлах очень зависят от экспериментальных условий, а не только от структурных факторов - у каждого из файлов распределение $I_{exp}(I_{calc})$хорошо приближалось линейной регрессией, но коэффициенты $\frac{I_{exp}}{I_{calc}}$ сильно отличались для каждого из файлов. и чтобы исключить это влияние была необходима нормализация. Нормализация по максимуму (рис. 2), среднему (рис. 3 и медиане (рис. 4 не дала результатов, и коэффициенты $\frac{I_{exp}}{I_{calc}}$ все так же сильно отличаются. Это может быть связано с далеким от нормального распределением значений структурных факторов.

Лучший результат был получен при использовании линейной регрессии - все данные уложились близко к одной линии, что положительно влияет на качество работы НС.

Рассчетные данные были получены из базы данных CCDC также в виде cif-файлов, обработаны идентично экспериментальным и отправлены на суперкомпьютер. Размер датасета составил 150 000 фалов, по 4630 отражений в каждом. Нормализация данных проводилась на максимальное значение интенсивности из присутствовавших в фай-

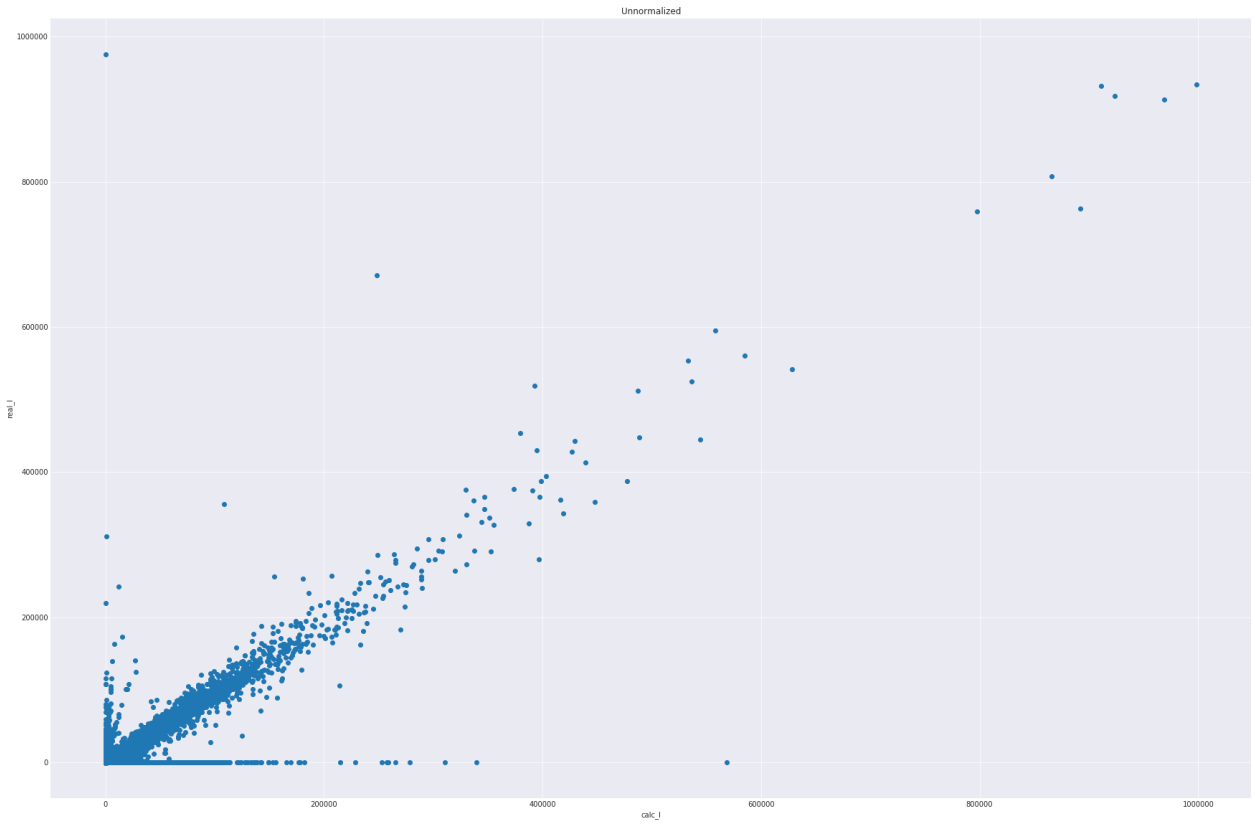

Рис. 1: Распределение реальных интенсивностей от рассчетных

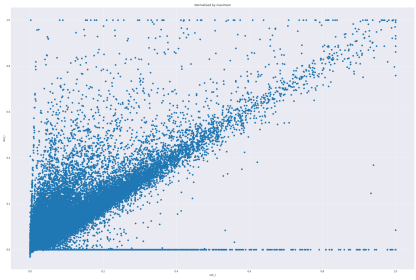

Рис. 2: нормализация по максимуму

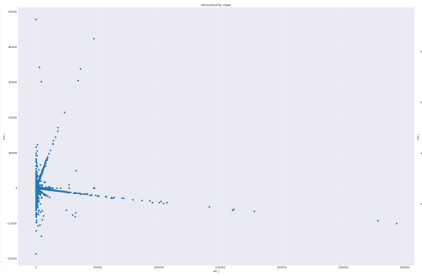

Рис. 3: нормализация по среднему значению

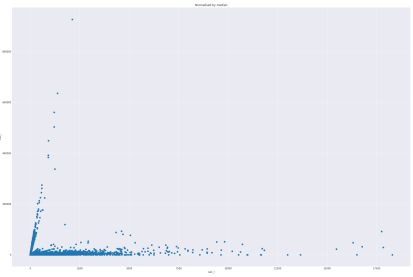

Рис. 4: Нормализация по медиане

ле отражений. Отрицательной стороной использования данных из CCDC вместо Acta Crystallographica E стало отсутствие в этих файлах тепловых факторов - параметров тепловх колебаний молекул, что снижает качество расчета структурных факторов, а значит и параметров отражений. Другая проблема с обучением на расчетных данных - отсутствие

3.2 Обучение на экспериментальных данных

Полученные из Acta E и обработанные данные были отправлены на С/К «Харизма», было запущено обучение простейшей модели(рис. 6) - полносвязного перцетпрона из

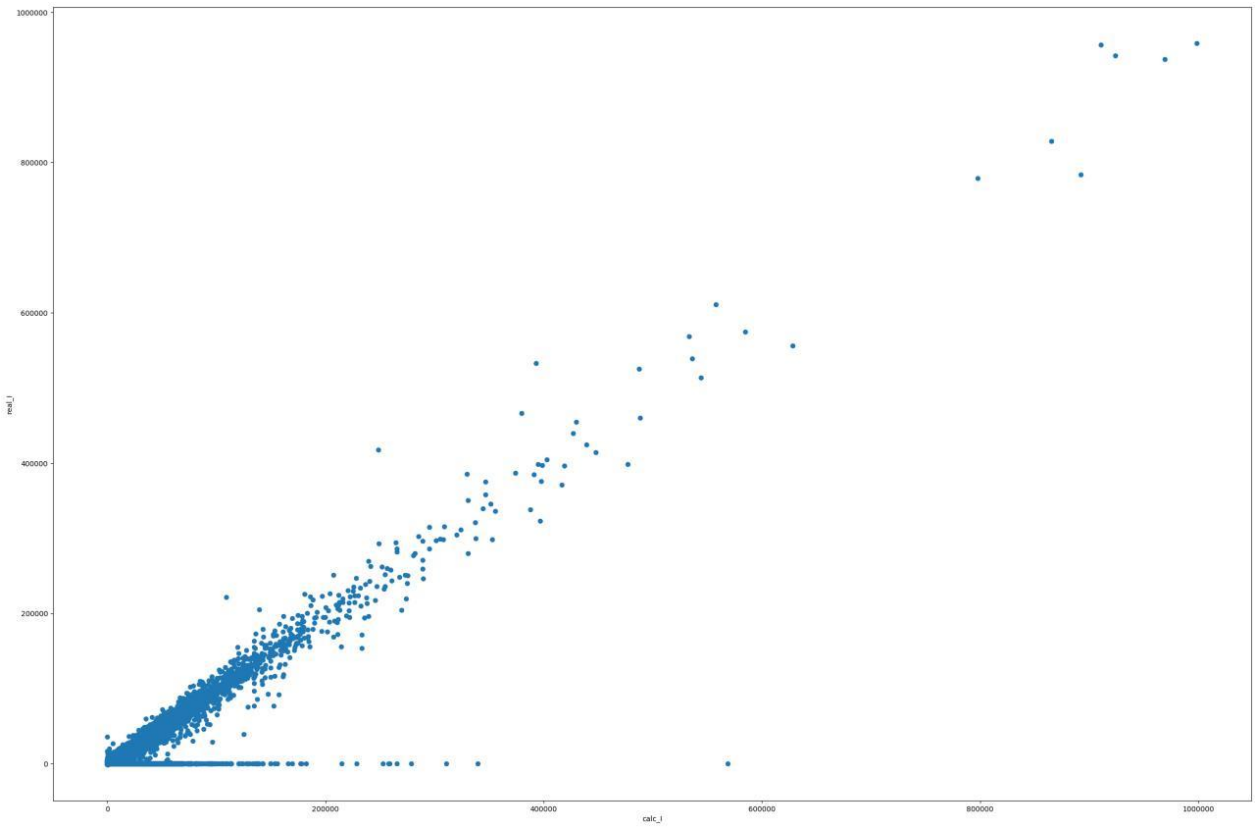

Рис. 5: Нормализация с помощью линейной регрессии

трех слоев, с размером слоя равным размеру входных и выходных данных.

Результат обучения отрицательный - модель очень сильно переобучается, демонстрируя хорошие показатели на обучающих данных и очень плохие - на тестовом.Этот эффект наблюдается как на графике(рис. 7) точности от эпохи обучения, так и на графике (рис. 8) функции потерь от эпохи обучения.

Из этого был сделан вывод о том, что данных для обучения модели недостаточно, и необходимо увеличить датасет. Другим возможным путем решения проблемы переобучения является снижение запоминающей способности сети путем добавления дропаутов, регуляризации, снижения количества нейронов в скрытых слоях сети и числа скрытых слоев.

3.3 Обучение нейронной сети на данных CCDC

Было проведено обучение нейросети на данных CCDC с разными гиперпараметрами и размером датасета. Для ускорения расчетов в части экспериментов были отброшены отражения с наибольшими индексами h,k,l, оставлены лишь удовлетворяющие условию $h^2+k^2+l^2<50$. После отбрасывания в каждом файле осталось по 709 отражений. Получены следующие результаты:

Рис. 6: Архитектура простой полносвязной нейросети

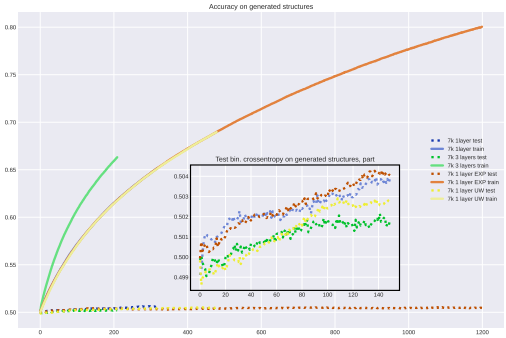

Рис. 7: Точность простой полносвязной нейросети в процессе обучения

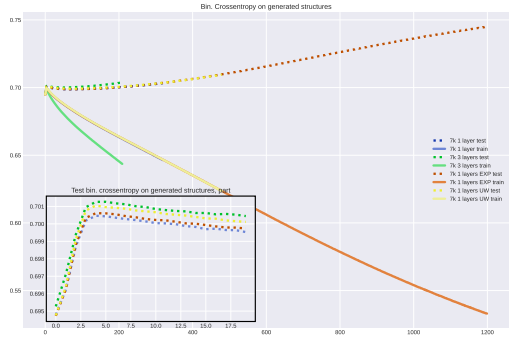

Рис. 8: Функция потерь простой полносвязной нейросети в процессе обучения

Для части датасета в 50 тысяч файлов, с отброшенными дальними отражениями была проведена проверка относительного качества обучения двух моделей: с одним внутренним слоем из 709 нейронов и с тремя внутренними слоями из 709 нейронов. Графики обучения (рис. 9, 10) демонстрируют сильное переобучение в обоих случаях, однако в случае с одним слоем оно менее выражено. Это говорит о том, что обе модели имеют избыточную запоминающую способность.

Был проверена зависимость качества обучения от размера датасета. Было проведено

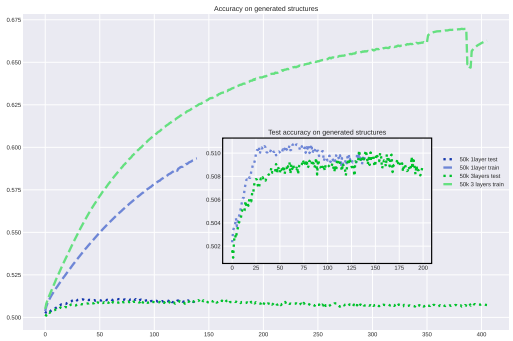

Рис. 9: Точность нейросети с одним и тремя слоями

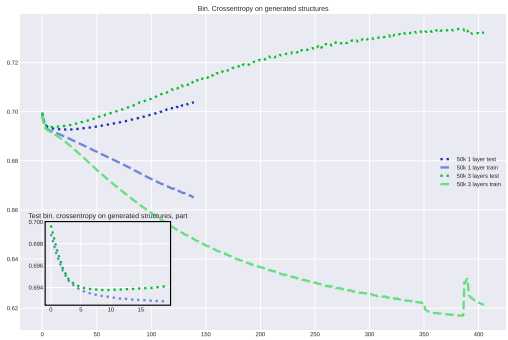

Рис. 10: Функция ошибки нейросети с одним и тремя слоями

обучение на разном количестве данных - 7000 файлов, 50 тысяч файлов, 150 тысяч файлов, с отбросом дальних отражений. На графиках обучения(рис. 11, 12) видно что больший размер датасета приводит к меньшему переобучению,

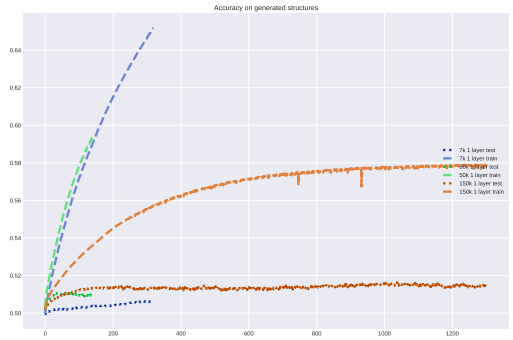

Рис. 11: Точность нейросети с одним слоем на разном количестве данных

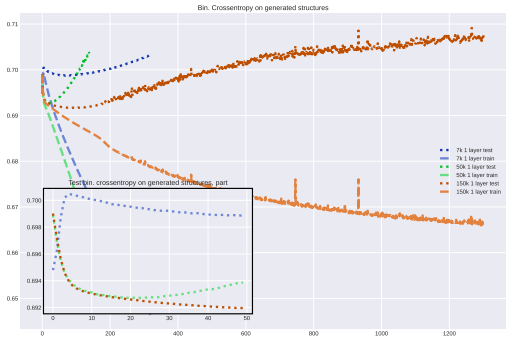

Рис. 12: Функция ошибки нейросети с одним слоем на разном количестве данных

Тот же результат был получен и при обучении нейросети из 3 слоев. Также было проведено обучение на 7000, 50 тысячах и 150 тысячах фалов. Заметно переобучение, оно уменьшается при увеличении размеров датасета, что можно видеть по рисункам 13, 14.

3.4 Зависимость качества обучения от типа весов

Функцю ошибки, на основе которой проводится обучение модели, можно взвесить по какой-либо характеристике, как и метрику. Была исследована зависимость качества обучения от весов применемой в модели метрики. Результаты представлены на рис. 15, 16

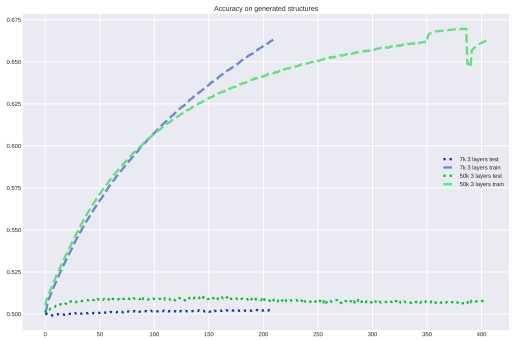

Рис. 13: Точность нейросети с одним слоем на разном количестве данных

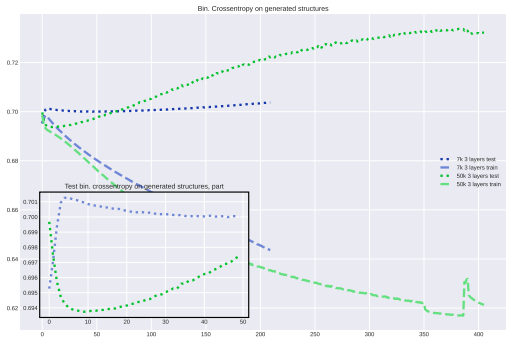

Рис. 14: Функция ошибки нейросети с одним слоем на разном количестве данных

Лучший результат дает обучнение с функцией ошибки взвешенной по интенсивности отражений. Взвешивание по экспоненте от интенсивности отражений, а также отсутствие взвешивания(равный вес всех отражений) дали незначительно худшие реультаты.

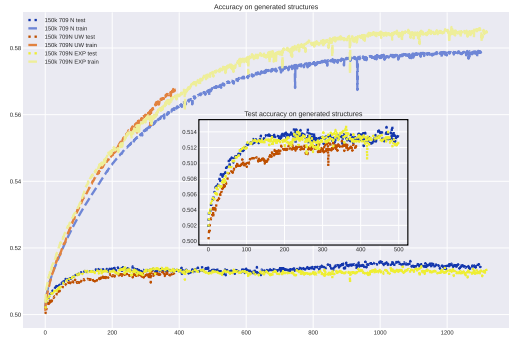

Рис. 15: Точность нейросети в зависимости от весов функции ошибки

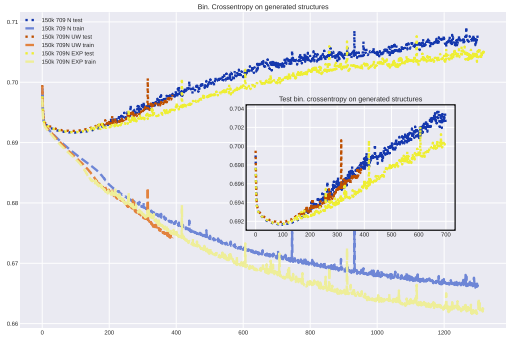

Рис. 16: Функция ошибки нейросети в зависимости от весов функции ошибки

Была исследована зависимость качества обучения от размера слоя нейросети для нейросети с 1 слоем. Полученные графики обучения (рис. 17, 18) говорят о том, что при уменьшении размера слоя с одной стороны снижается степень переобучения, с другой при уменьшении размера слоя до менее чем 100 нейронов происходит заметное снижение качества обучения. Таким образом, оптимальный размер скрытого слоя нейросети вероятно составляет 100 нейронов. Этот результат заставил нас провести исследование степени сжимаемости данных с помощью автокодировщика.

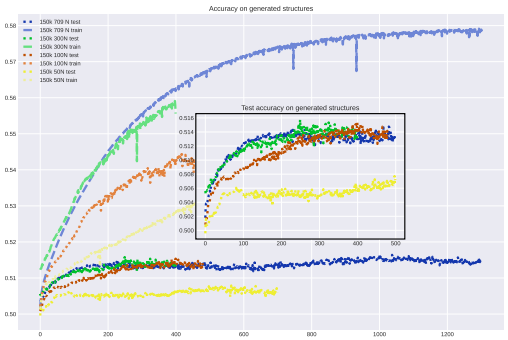

Рис. 17: Точность нейросети в зависимости от размеров скрытого слоя

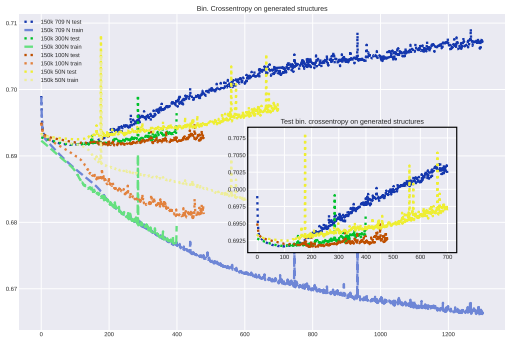

Рис. 18: Функция ошибки нейросети в зависимости от размеров скрытого слоя

3.5 Применение автокодировщиков для определения сжимаемости данных

Автокодировщик - нейронная сеть, задачей которой является получить на выходном слое нейрнонов значения максимально близкие к соответствующим значениям входного слоя. При этом промежуточный слой такой нейронной сети меньше, чем входной и выходной слои. Это заставляет нейросеть анализировать закономерности данных, выделять в них паттерны, т.е. сжимать данные. Таким образом по минимальному числу нейронов в скрытом слое с которым автокодировщик будет достаточно точно восстанавливать данные можно оценить количество информации на самом деле содержащееся в данных, которое в свою очередь определяет степень возможного сжатия данных и достаточный размер скрытых слоев нейронной сети.

Были обучены автокодировщики с размером скрытого слоя 50, 100, 250, 500, 1000 нейронов. Обучение проводилось без отброса дальних отражений, т.о. размер входа нейросети составил 4630 отражений на файл. Поскольку значения выходного слоя больше не являются бинарными, вместо бинарной кросссэентропии в качестве функции ошибки была применена среднеквадратичная ошибка(MSE, mean squared error). Полученные результаты представлены на рис. 19. Видно что большее количество нейронов в скрытом слое дает большую точность, но при этом если точность при 50 и 100 нейронах в скрытом слое заметно различается, то различия точности между 500 и 1000 нейронами почти нулевое.

Было проверено распределение ошибки по отражениям, результат представлен на рис. 20, 21. Отражения на графиках отсортированы по $R = h^2 + k^2 + l^2, h, k, l$. На рис. 21 применено сглаживание - показаны средние данные для окна в 50 точек. Видно что абсолютное значение среднеквадратичной ошибки выше для ближних, более интенсивных, отражений. Относительная ошибка для ближних отражений также чуть выше для малых h, k, l.

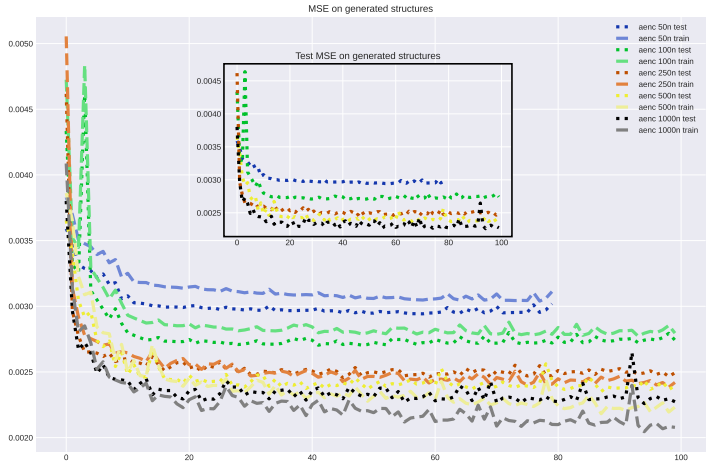

Рис. 19: График обучения автоэнкодеров с разным размером скрытого слоя

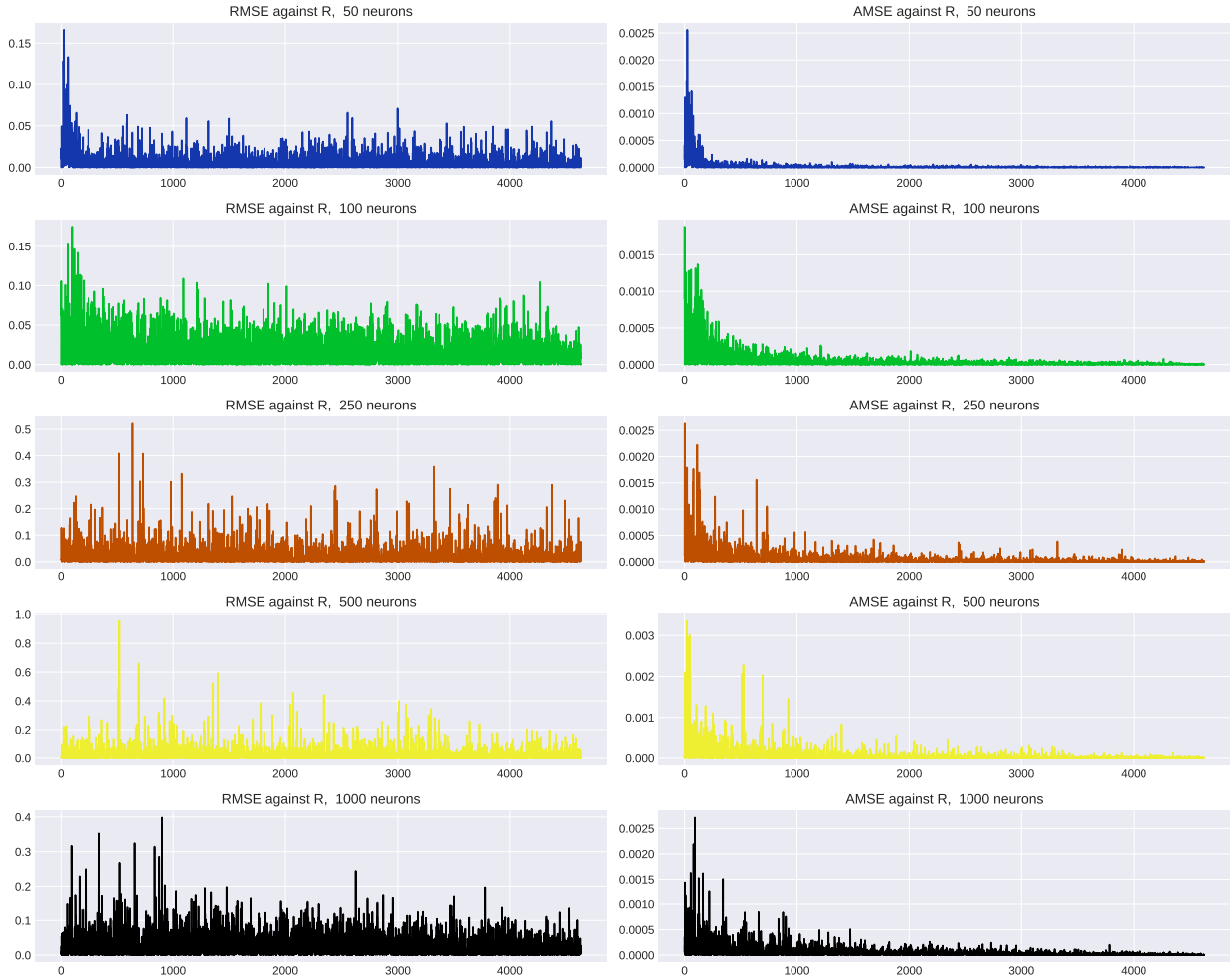

Рис. 20: Распределение ошибки по отражениям

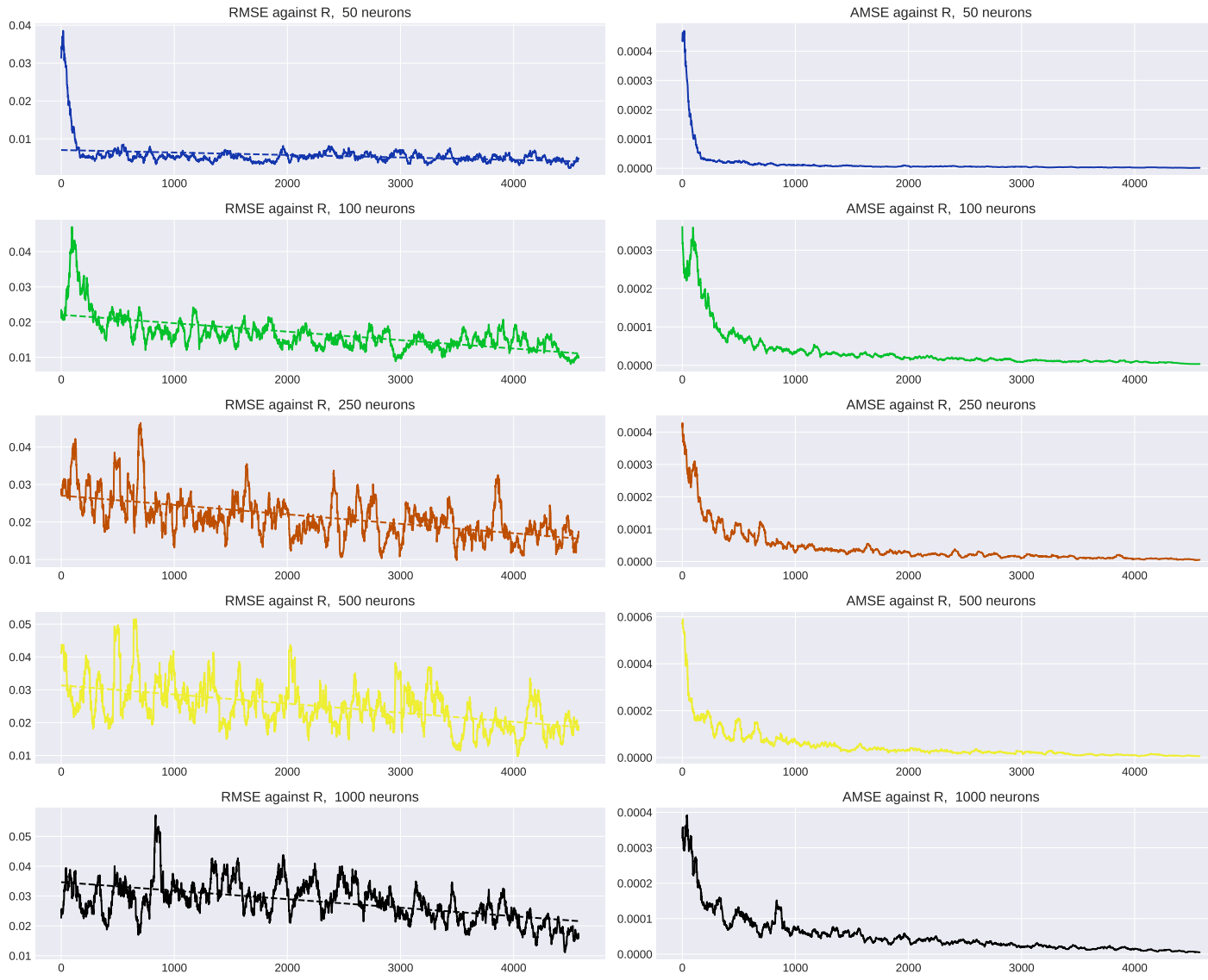

Рис. 21: распределение ошибки по отражениям

4 РЕЗУЛЬТАТЫ И ВЫВОДЫ

Были получены и обработаны кристаллографические данные. Построена модель машинного обучения для решения проблемы фаз для рентгеновских отражений, продемонстрировано серьезное переобучение. Это может быть объяснено тремя причинами:

- В интенсивностях отражений не содержится достаточно данных для определения фаз. В этом случае помочь может добавление параметров кристаллической решетки и других данных во входные данные нейронной сети
- Недостаточно данных для обучения. В этом случае решить пробему поможет увеличение размеров датасета, например с помощью OQMD или путем аугментации датасета.
- Нейронная сеть имеет слишком высокую запоминающую способность. В этом случае проблема может быть решена подбором оптимальной архитектуры: добавлением дропаутов, пакетных нормализаций, уменьшением размера слоев нейронной сети.

Еще одним вариантом улучшения результатов может быть использование вместо интенсивностей функции паттерсона[5] - картины электронной плотности полученной исходя из предположения о том что фазы всех отражений идентичны и равны нулю.

Список литературы

[1] Levin I. et al Fancher C. Han Z. «Use of Bayesian Inference in Crystallographic Structure Refinement via Full Diffraction Profile Analysis». В: Scientific Reports (2016), с. 2045—2322. DOI: https://doi.org/10.1038/srep31625.

[2] A. El Haouzi и др. «The Phase Problem in the Analysis of X-ray Diffraction Data in Terms of Electron-Density Distributions». В: Acta Crystallographica Section A Foundations of Crystallography 52.2 (март 1996), с. 291—301. DOI: 10.1107/s0108767395014942. URL: https://doi.org/10.1107/s0108767395014942.

[3] Robert W. Harrison. «Phase problem in crystallography». В: J. Opt. Soc. Am. A 10.5 (май 1993), с. 1046—1055. DOI: 10.1364/JOSAA.10.001046. URL: http://josaa.osa.org/abstract.cfm?URI=josaa-10-5-1046.

[4] Sun S. et al. Oviedo F. Ren Z. «Fast and interpretable classification of small X-ray diffraction datasets using data augmentation and deep neural networks.» В: npj Comput Mater 5 60 (2019).

[5] A. L. Patterson. «A Fourier Series Method for the Determination of the Components of Interatomic Distances in Crystals». В: Physical Review 46.5 (сент. 1934), с. 372—376. DOI: 10.1103/physrev.46.372. URL: https://doi.org/10.1103/physrev.46.372.

[6] Rampi Ramprasad и др. «Machine learning in materials informatics: recent applications and prospects». В: npj Computational Materials 3.1 (дек. 2017). DOI: 10.1038/s41524-017-0056-5. URL: https://doi.org/10.1038/s41524-017-0056-5.