На правах рукописи 

Бобков Сергей Алексеевич

КЛАССИФИКАЦИЯ ДИФРАКЦИОННЫХ ИЗОБРАЖЕНИЙ БИОМОЛЕКУЛ ПО ТИПАМ 3D СТРУКТУРЫ С ПОМОЩЬЮ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

Специальность 05.13.18 —
«Математическое моделирование, численные методы и комплексы программ»

Автореферат

диссертации на соискание учёной степени

кандидата физико-математических наук

Москва — 2018

Работа выполнена в Национальном исследовательском центре «Курчатовский институт»

Научный руководитель: доктор физико-математических наук
Ильин Вячеслав Анатольевич

Научный консультант: кандидат физико-математических наук
Вартаньянц Иван Анатольевич

Официальные оппоненты: **Ососков Геннадий Алексеевич,**
доктор физико-математических наук, профессор,
Объединенный Институт Ядерных Исследований,
главный научный сотрудник

Доленко Сергей Анатольевич,
кандидат физико-математических наук,
МГУ имени М.В. Ломоносова, Научно-исследовательский институт ядерной физики имени Д.В. Скобельцына,
заведующий лабораторией

Ведущая организация: Федеральное государственное учреждение «Федеральный исследовательский центр Институт прикладной математики им. М.В. Келдыша Российской академии наук»

Защита состоится 10 октября 2018 г. в 15.00 на заседании диссертационного совета Д 212.130.09 на базе Федерального государственного автономного образовательного учреждения высшего профессионального образования «Национальный исследовательский ядерный университет «МИФИ» по адресу: 115409, г. Москва, Каширское шоссе д.31.

С диссертацией можно ознакомиться в библиотеке НИЯУ МИФИ и на сайте http://ods.mephi.ru.

Автореферат разослан _____ 2018 года.

Ученый секретарь
диссертационного совета
Д 212.130.09, д-р физ.-мат. наук Леонов А. С.

Общая характеристика работы

Диссертационная работа направлена на решение проблемы классификации дифракционных изображений, получаемых в экспериментах по изучению структуры биомолекул методом когерентной рентгеновской дифракционной микроскопии, при обработке экспериментальных данных в режиме квазионлайн.

Актуальность темы.

Появление техники рентгеновской микроскопии открыло возможность исследований структуры вещества с высоким разрешением. Эта техника заняла свое место среди самых мощных инструментов изучения структуры и привела к фундаментальным открытиям во многих научных дисциплинах, от биологии до физики твердого тела.

При рассеянии рентгеновского излучения на кристаллах, регулярное расположение атомов приводит к когерентному вкладу в брэгговские пики, что многократно усиливает сигнал на детекторе [1]. К сожалению, большинство белков и вирусов не кристаллизуются, поэтому были разработаны подходы к определению структуры таких объектов без использования брэгговского рассеяния. Для повышения сигнала на детекторе приходится повышать интенсивность падающего излучения, что приводит к повреждению образцов, и, в результате, к снижению разрешения при определении трехмерной структуры.

Эти проблемы успешно решаются в новом методе — микроскопии на отдельных объектах (Single Particle Imaging, SPI) [2; 3], которая развивается в последние 15 лет. Этот метод позволяет определять трехмерную структуру по дифракционным картинам от отдельных экземпляров исследуемого объекта в случайных ориентациях. Метод SPI открывает возможности для изучения структуры биомолекул с субнанометровым разрешением [4–7].

Для SPI требуется использование рентгеновского излучения с когерентной фазой по всему объему изучаемого объекта, а также требуется высокая интенсивность. Требуемые характеристики обеспечиваются в экспериментах на рентгеновских лазерах на свободных электронах (ЛСЭ) [3; 7].

В SPI экспериментах идентичные экземпляры исследуемого объекта впрыскиваются в луч рентгеновского лазера в случайных ориентациях. Образцы разрушаются в результате воздействия мощного рентгеновского излучения. Но благодаря сверхкороткой длительности импульсов ($\sim$ 10 фемтосекунд), ди-

фракционная картина измеряется до момента, когда изменение взаимного положения атомов в объекте станет значимым [3; 4; 8]. На основе статистически значимого набора двумерных дифракционных изображений в разных ориентациях можно восстановить структуру исследуемого объекта [3; 7; 9]. Для этого необходимо определить относительную ориентацию образцов на изображениях и объединить дифракционные изображения в трехмерную дифракционную картину [10], и далее восстановить фазы рассеянного излучения [11–13], которые не фиксируются детектором.

Но не все дифракционные изображения, измеряемые детектором, подходят для восстановления структуры. В SPI экспериментах, которые проводятся на лазере на свободных электронах LCLS (Стэнфорд) с 2010 года, примерно 98% получаемых изображений пустые, т.е. ни один объект не попадает в импульс лазера [14]. Дифракционные изображения одиночных экземпляров исследуемого объекта, которые подходят для восстановления структуры, составляют всего лишь 1% от общего количества детектируемых изображений. Еще 1% оставшихся изображений получаются от капель воды, от рассеяния на нескольких образцах исследуемых объектов или от примесных объектов. Существуют методы фильтрации, которые позволяют быстро и надежно исключить пустые изображения из анализа [15]. Выделение же изображений одиночных экземпляров исследуемого объекта от дифракционных изображений других объектов является более сложной задачей, которая в экспериментах на лазерах на свободных электронах, в настоящее время, выполняется вручную.

Далее мы будем использовать термин «классификация по типам структуры» для задачи разделения дифракционных изображений на несколько групп: одиночные экземпляры исследуемого объекта, несколько экземпляров исследуемого объекта, примесные и другие объекты.

В 2017 году в Гамбурге был запущен новый европейский рентгеновский лазер на свободных электронах (European XFEL, далее мы будем использовать аббревиатуру EuXFEL) [16], который позволит регистрировать до 27000 дифракционных изображений в секунду. Если предположить, что соотношение типов изображений будет соответствовать экспериментам на LCLS, то за 12 часов эксперимента будет собираться 23 Тбайт данных после фильтрации пустых изображений. Ручная классификация таких объемов потребует огромных трудозатрат. Отсюда возникает необходимость создания метода классификации,

который позволит выделять изображения одиночных образцов исследуемого типа в режиме квази-онлайн.

Ключевым аспектом диссертационной работы является моделирование свойств дифракционных изображений релевантных для классификации по типам структуры за счет использования метода сжатия. Разработанный метод использует угловые корреляционные функции для учёта структурных особенностей дифракционных изображений.

Целью данной работы является разработка метода классификации по типам 3D структуры, который обеспечит обработку в режиме квази-онлайн для дифракционных изображений, получаемых от биологических объектов в экспериментах на лазерах на свободных электронах, включая EuXFEL.

Для достижения поставленной цели были решены следующие **задачи**:

1) исследование существующих подходов к классификации данных SPI экспериментов на основе методов машинного обучения, в том числе нейронных сетей;
2) разработка метода сжатия дифракционных изображений в характеристический вектор на основе моделирования данных SPI экспериментов в части ключевых структурных особенностей исследуемых объектов. В результате сжатия размерность изображений должна уменьшаться на несколько порядков. Также в методе сжатия должны учитываться технические параметры детектора и экспериментальной установки, влияющие на точность классификации;
3) разработка методов классификации дифракционных изображений по типам структуры исследуемых объектов с применением методов машинного обучения и разработанного метода сжатия в характеристический вектор;
4) верификация разработанных методов классификации на наборах дифракционных изображений, полученных в экспериментах по изучению структуры биологических объектов на LCLS (Стэнфорд) в 2011–2016 годах, и сравнение результатов классификации, выполненных в различных подходах машинного обучения;
5) разработка сценария классификации дифракционных изображений, в котором будет обеспечена классификация в режиме квази-онлайн при обработке данных экспериментов на EuXFEL.

Основные положения, выносимые на защиту:

1. Разработанный метод сжатия дифракционных изображений в характеристический вектор существенно повышает точность классификации по типам структуры, а уменьшение размерности данных на 4 порядка значительно повышает скорость обучения и классификации.
2. Разработанный метод классификации дифракционных изображений по типам структуры исследуемых объектов на основе математического метода опорных векторов с использованием разработанного метода сжатия в характеристический вектор, обеспечивает обработку данных SPI экспериментов в режиме квази-онлайн.
3. Разработанный метод классификации по типам структуры позволяет сформировать сценарий программно-аппаратной реализации для классификации дифракционных изображений по типу структуры исследуемых объектов в режиме квази-онлайн в экспериментах на EuXFEL.

Научная новизна:

1. Разработан оригинальный метод сжатия дифракционных изображений в характеристический вектор с использованием корреляционных функций с учетом конструктивных особенностей детектора, который сокращает размерность данных на несколько порядков без потери информации о ключевых особенностях пространственной структуры исследуемых объектов.
2. Впервые разработан метод классификации дифракционных изображений по типам 3D структуры исследуемых объектов, который обеспечивает классификацию в режиме квази-онлайн в SPI экспериментах. На его основе впервые сформирован сценарий классификации дифракционных изображений в режиме квази-онлайн для экспериментов на EuXFEL.
3. Впервые получены оценки эффективности классификации по типам структуры с использованием различных методов машинного обучения и нейронных сетей.

Практическая значимость данной работы заключается в том, что разработанный метод классификации дифракционных изображений по типам структуры позволяет отбирать содержательные дифракционные картины в экс-

периментах на EuXFEL в режиме квази-онлайн, что является важным шагом к получению результатов восстановления трехмерной структуры исследуемых объектов практически сразу после окончания эксперимента.

Достоверность полученных результатов обеспечивается применением научно обоснованных подходов к построению и сравнению методов классификации дифракционных изображений, в том числе на основе математических методов машинного обучения и нейронных сетей, а также верификацией разработанных методов на данных SPI экспериментов на лазере на свободных электронах LCLS (Стэнфорд), полученных в 2011-2016 годах.

Апробация работы. Основные результаты работы докладывались на семинарах в ведущих университетах и научных центрах:

- ЛИТ ОИЯИ (28 февраля, 2018, г. Дубна, Россия, URL: http://lit.jinr.ru/Bobkov_rus.pdf);
- DESY (15 февраля, 2018, г. Гамбург, Германия);
- «Технологии машинного обучения для слабоструктурированных данных большого объема» Университет ИТМО (26 октября, 2017, г. Санкт-Петербург, Россия);
- «Методы суперкомпьютерного моделирования» ИКИ РАН (1-3 октября, 2014, г. Таруса, Россия, URL: http://www.iki.rssi.ru/seminar/2014100103/);

а также на конференциях:

- International Conference «Supercomputer Simulations in Science and Engineering» (6-10 сентября, 2016, г. Москва, Россия, URL: http://http://ssse2016.ac.ru);
- The 6th International Conference «Distributed Computing and Grid-technologies in Science and Education» (30 июня - 5 июля, 2014, г. Дубна, Россия, URL: http://grid2014.jinr.ru);
- V Международная конференция «Математическая биология и биоинформатика» (19-24 октября, 2014, Институт математических проблем биологии РАН, г. Пущино, Россия, URL: http://icmbb.impb.ru/);
- International scientific conference «Science of the future» (17-20 сентября, 2014, г. Санкт-Петербург, Россия, URL: http://www.p220conf.ru);
- 11-я Курчатовская молодежная научная школа (12-15 ноября, 2013, г. Москва, Россия);

Личный вклад. Автору принадлежит идея и программная реализация метода сжатия дифракционных изображений в характеристический вектор с использованием корреляционных функций с учетом конструктивных особенностей детектора, а также идея и программные реализации методов классификации по типам структуры исследуемых объектов на основе математических методов машинного обучения и нейронных сетей в режиме квази-онлайн. Верификация этих методов на наборах данных SPI экспериментов, а также сравнение результатов с различными подходами машинного обучения и нейронных сетей были проведены автором лично. Автор предложил методику формирования сценариев классификации данных в режиме квази-онлайн в экспериментах EuXFEL.

Публикации. Основные результаты по теме диссертации опубликованы в 6 печатных изданиях. Четыре из них опубликованы в научных журналах, рекомендованных ВАК: одна статья в журнале, индексируемом WoS и Scopus; две в журналах, индексируемых Scopus; и одна в журнале, индексируемом РИНЦ. Еще опубликованы 2 тезиса докладов на научных конференциях. По результатам диссертационной работы получено 3 свидетельства о Государственной регистрации программ для ЭВМ.

Диссертация состоит из введения, пяти глав, заключения и двух приложений. Полный объём диссертации составляет 135 страниц с 37 рисунками и 17 таблицами. Список литературы содержит 170 наименований.

Содержание работы

Во **введении** обоснована актуальность исследований, проводимых в рамках диссертационной работы, определены цели и задачи работы, сформулирована научная новизна и практическая значимость полученных результатов.

В **первой главе** указана роль этапа классификации в полном цикле восстановлении структуры биомолекул на основе дифракционных изображений, получаемых в SPI экспериментах.

В **§ 1.1** описаны современные методы исследования структуры наноразмерных объектов. Рассмотрена связь этих методов с развитием оптических инструментов и источников излучения [17].

В **§ 1.2** описан метод исследования структуры с разрешением в субнанометровом диапазоне — когерентная рентгеновская дифракционная микроскопия (Coherent X-ray Diffractive Imaging, CXDI) [5].

В **§ 1.3** описан метод восстановления трёхмерной структуры одиночных биомолекул — микроскопия на отдельных объектах (Single Particle Imaging, SPI), которая основана на методе CXDI. В SPI эксперименте (Рис. 1) экземпляры исследуемого объекта вводятся в луч лазера в случайной ориентации, и собирается дифракционная картина на детекторе. Общая схема восстановления структуры в SPI экспериментах представлена на рисунке 2.

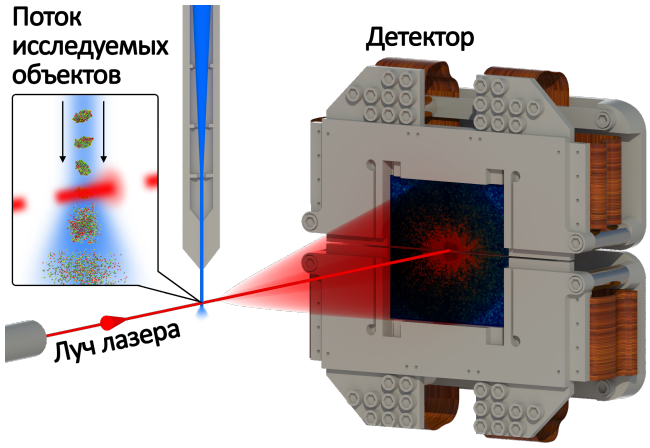


Рис. 1 — Схема проведения SPI эксперимента для определения трёхмерной структуры одиночных биомолекул.

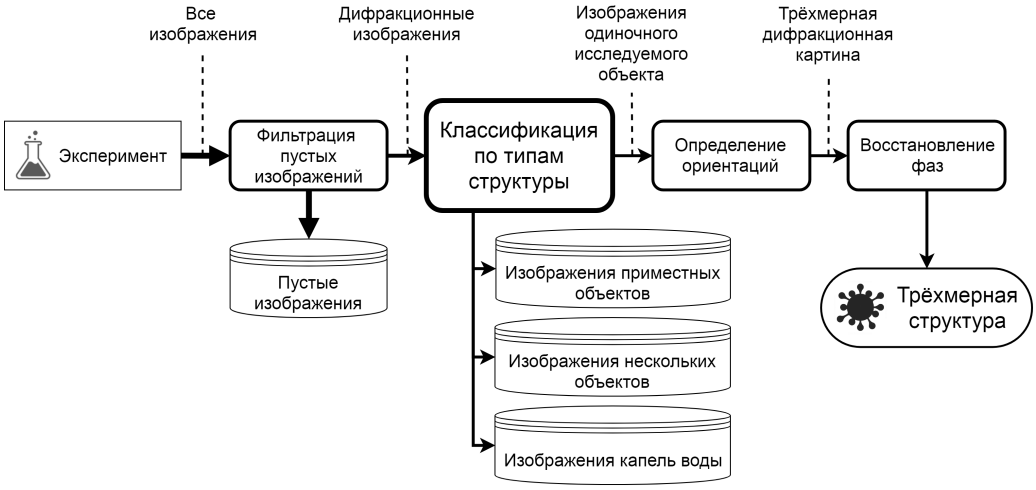


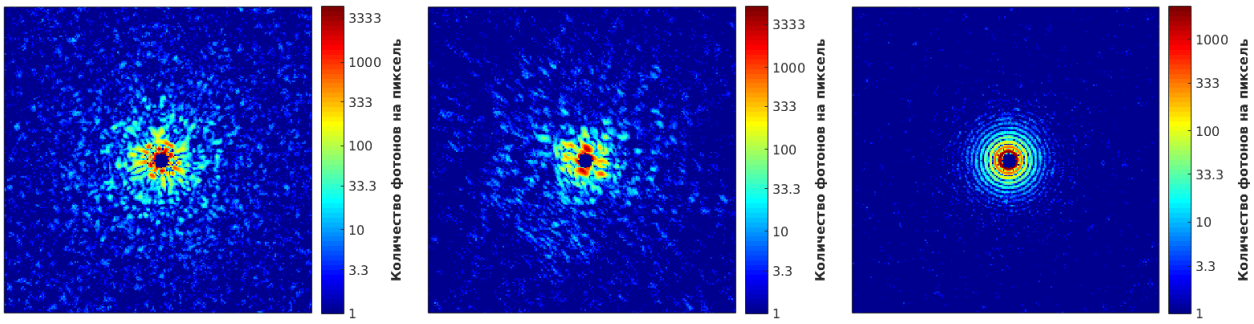
Рис. 2 — Общая схема восстановления структуры исследуемого объекта на основе дифракционных изображений в SPI экспериментах.

В **§ 1.4** указано значение этапа классификации при восстановлении структуры — диссертационные исследования относятся к этому этапу обработки данных SPI экспериментов.

Точность классификации по типам структуры напрямую влияет на предел разрешения, с которым можно восстановить 3D структуру исследуемого объекта.

Во **второй главе** представлены методы машинного обучения перспективные для классификации дифракционных изображений по типам структуры.

В **§ 2.1** дан анализ особенностей дифракционных изображений, получаемых в SPI экспериментах, которые необходимо учитывать при классификации по типам структуры. На рисунке 3 представлены примеры дифракционных изображений от разных типов объектов, демонстрирующие характерный вид дифракционных изображений от биологических объектов и капель воды.



а) Аденовирус б) BTV в) Капля воды

Рис. 3 — Пример дифракционных изображений от трех типов объектов из набора данных моделирования: пентоновый белок аденовируса человека (аденовирус), комбинация белков VP3 и VP7 из ядра вируса катаральной лихорадки (BTV) и капля воды. Цветовая шкала показывает количество детектируемых фотонов на 1 пиксель детектора.

В **§ 2.2** представлено описание подходов к классификации дифракционных изображений в SPI экспериментах, используемых в настоящее время.

В **§ 2.3** приведены методы машинного обучения, которые перспективны для классификации дифракционных изображений по типам структуры.

В пункте 2.3.4 описаны методы классификации на основе искусственных нейронных сетей: перцептрон с тремя скрытыми слоями и свёрточная нейронная сеть.

В **§ 2.4** описан подход к классификации данных с использованием сжатия в характеристический вектор. В SPI экспериментах измеряются дифракционные изображения, которые состоят из миллиона пикселей. В результате, в задаче классификации эти изображения образуют параметрическое простран-

ство размерности порядка 10^6. Как показано в диссертации, классификации по типам структуры достаточно отобразить это параметрическое пространство в пространство меньшей размерности, сохранив только ключевую информацию о пространственных особенностях структуры исследуемых объектов.

В **§ 2.5** на основе моделирования структуры данных SPI (дифракционных изображений) показано, что угловые корреляционные функции учитывают информацию о пространственных особенностях структуры исследуемых объектов независимо от ориентации.

Третья глава посвящена разработке метода сжатия в характеристический вектор для дифракционных изображений в SPI экспериментах, а также дается описание разработанных методов классификации дифракционных изображений по типам структуры. Под классификацией по типам структуры понимается задача разделения дифракционных изображений на несколько групп: одиночные экземпляры исследуемого объекта, несколько экземпляров объекта, примесные и другие объекты.

В **§ 3.1** представлен метод сжатия дифракционных изображений в характеристический вектор. Он основан на анализе корреляционных функций, которые определяются по формуле:

$$C(q_1, q_2, \Delta) = \langle I(q_1, \varphi) \cdot I(q_2, \varphi + \Delta) \rangle_\varphi \,, \tag{1}$$

где $I(q, \varphi)$ - интенсивность дифракционной картины в полярной системе координат с радиусом q и углом φ. Символ $\langle \ldots \rangle_\varphi$ обозначает усреднение по углу φ.

На дифракционных изображениях вводится полярная система координат с началом в центре симметрии дифракционной картины. Шаг радиальной координаты q равен размеру пикселя детектора.

Для всего набора изображений определяется диапазон $q_{\min} \leq q \leq q_{\max}$, где дифракционная картина наиболее информативна. Для дифракционных изображений, полученных в экспериментах на LCLS, был выделен диапазон значений от $q_{\min} = 51$ до $q_{\max} = 255$ пикселей, $q_{\max}$ соответствует радиусу окружности, которая касается ближайшей границы детектора.

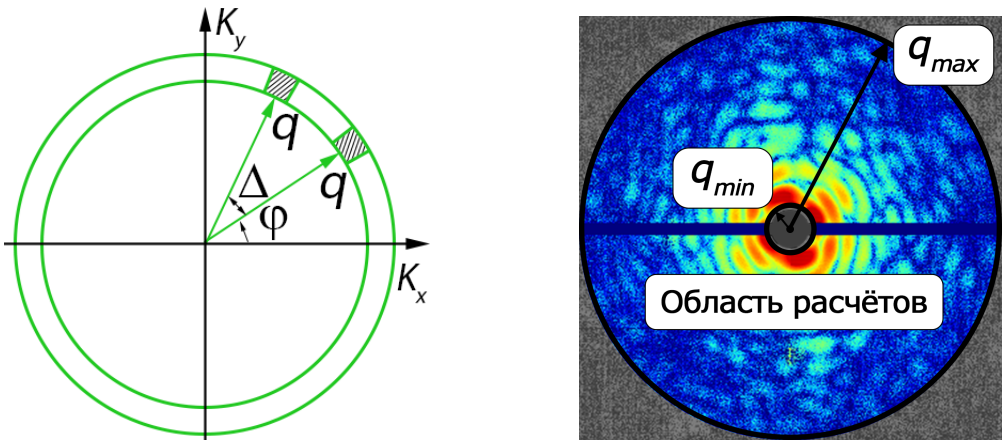
Далее определяются коэффициенты $C(q, \Delta)$ в полярных координатах q и φ для диапазона $q_{\min} \leq q \leq q_{\max}$ (Рис. 4), полученные из (1) с $q = q_1 = q_2$:

$$C(q, \Delta) = \langle I(q, \varphi) I(q, \varphi + \Delta) \rangle_\varphi \,. \tag{2}$$

Функция $C(q,\Delta)$ усредняется по q и нормируется:

$$\bar{C}(\Delta) = \frac{\langle C(q,\Delta)\rangle_q}{\langle I(q,\varphi)^2\rangle_{\varphi,q}}. \tag{3}$$

Функция $\bar{C}(\Delta)$ периодическая и четная, поэтому при разложении в ряд Фурье



а) Схема расчета $C(q,\Delta)$ б) Диапазон значений q

Рис. 4 — Расчет коэффициентов $C(q,\Delta)$ в полярных координатах q и φ для диапазона $q_{\text{min}} \leq q \leq q_{\text{max}}$.

только компоненты при косинусах отличны от нуля:

$$\bar{C}(\Delta) = 2\sum_{n=1}^{\infty} \bar{C}^n \cos(n\Delta), \tag{4a}$$

$$\bar{C}^n = \frac{1}{\pi}\int_0^{\pi} \bar{C}(\Delta)\cos(n\Delta)d\Delta. \tag{4b}$$

Фурье компоненты $\bar{C}^n$ для $n = 1,\ldots,N$ используются как первая часть характеристического вектора. При анализе данных экспериментов на LCLS, мы используем $N = 50$.

Компоненты $\bar{C}^n$ не содержат информацию о зависимости распределения интенсивности дифракционной картины от координаты q. Такая информация может быть получена из функции $C(q,0) = \langle I(q,\varphi)^2\rangle_\varphi$. Она также нормируется:

$$\bar{C}_q = \frac{C(q,0)}{\langle I(q,\varphi)\rangle_\varphi^2} = \frac{\langle I(q,\varphi)^2\rangle_\varphi}{\langle I(q,\varphi)\rangle_\varphi^2}. \tag{5}$$

Компоненты $\bar{C}_q$ для $q = q_{\text{min}},\ldots,q_{\text{max}}$ используются как вторая часть характеристического вектора.

Два набора компонент $\bar{C}^n$ и $\bar{C}_q$ могут значительно отличаться по своим значениям. Это может привести к тому, что влияние одного набора может быть пренебрежительно мало по сравнению со другим набором. Чтобы компенсировать эту возможную разбалансированность, оба набора умножаются на веса a и b, которые определяются по стандартным отклонениям компонент $\bar{C}^n$ и $\bar{C}_q$:

$$a = \frac{1}{\sigma_\alpha}, b = \frac{1}{\sigma_\beta}. \tag{6}$$

$$\sigma_\alpha = \left[\frac{1}{N}\sum_{n=1}^{N}(\sigma_\alpha^n)^2\right]^{1/2}, \sigma_\beta = \left[\frac{1}{Q}\sum_{q=q_{min}}^{q_{max}}\left(\sigma_\beta^q\right)^2\right]^{1/2}, \tag{7}$$

$$\sigma_\alpha^n = \frac{1}{M}\sum_{i=1}^{M}\left(\bar{C}_{\psi_i}^n - \left\langle\bar{C}_{\psi_i}^n\right\rangle_{\psi_i}\right)^2, \sigma_\beta^q = \frac{1}{M}\sum_{i=1}^{M}\left(\bar{C}_{q,\psi_i} - \left\langle\bar{C}_{q,\psi_i}\right\rangle_{\psi_i}\right)^2. \tag{8}$$

В (8) суммирование проводится по набору из M дифракционных изображений. В (7) суммирование проводится по N компонентам вектора $\bar{C}^n$ и Q компонентам вектора $\bar{C}_q$.

В результате, характеристический вектор для каждого изображения определяется по формуле:

$$F = (a\bar{C}^1, \ldots, a\bar{C}^N, b\bar{C}_{q_{\text{min}}}, \ldots, b\bar{C}_{q_{\text{max}}}), \tag{9}$$

длина которого равна $N + Q$.

Таким образом, для дифракционных изображений, получаемых в экспериментах на LCLS, длина характеристических векторов составляла всего лишь 254 компоненты, что почти на 4 порядка меньше, чем размерность пространства параметров изображений, состоящего из полного количества пикселей на детекторе.

В **§ 3.2** представлены алгоритмы аппроксимации дифракционной картины в области зазоров детектора и подходы к определению сдвига центра дифракционной картины относительно центра детектора при сжатии изображений в характеристические векторы (Рис. 5).

В **§ 3.3** описано применение метода главных компонент (PCA) для кластеризации дифракционных изображений.

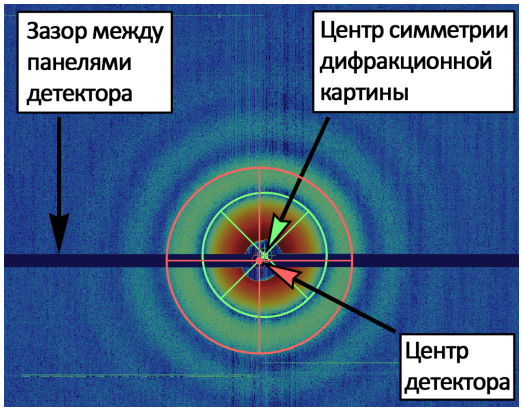


Рис. 5 — Пример зазора между панелями детектора PnCCD (используемого в экспериментах на LCLS) и пример сдвига центра симметрии дифракционной картины относительно центра детектора.

В **§ 3.4** сформулированы методы классификации по типам структуры с использованием сжатия в характеристический вектор и математических методов кластеризации: метода к-средних и метода спектральной кластеризации.

В **§ 3.5** представлен метод классификации по типам структуры на основе математического метода опорных векторов (SVM) с использованием сжатия дифракционных изображений в характеристический вектор. SVM позволяет определить вероятность корректной классификации для каждого изображения в зависимости от расстояния для разделяющей поверхности (Рис. 6). За счет отсеивания изображений с низкой вероятностью корректной классификации можно повысить точность классификации.

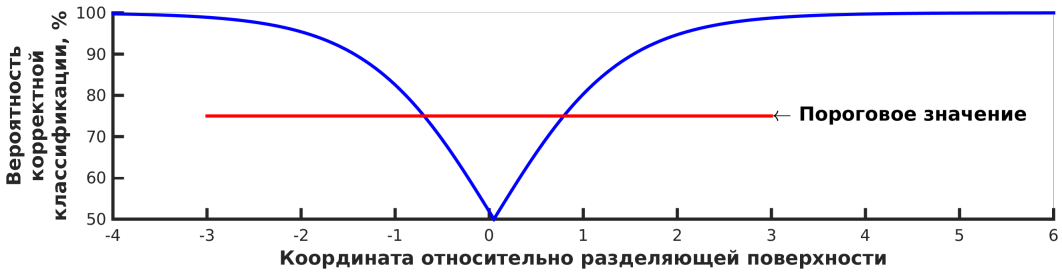


Рис. 6 — Пример зависимости вероятности корректной классификации от координаты относительно разделяющей поверхности. Показано пороговое значение в 75%.

В **§ 3.6** описаны методы классификации по типам структуры на основе математического метода искусственных нейронных сетей: перцептрона с тремя скрытыми слоями [18] и свёрточной нейронной сети [19].

В **§ 3.7** приведено описание стандартных критериев точности и полноты классификации, которые применяются для количественной оценки результатов. Описана стандартная схема перекрестной проверки результатов [20], которая позволяет сравнивать результаты разных методов классификации.

В **четвертой главе** представлено сравнение результатов применения разработанных методов классификации на различных наборах модельных и экспериментальных данных.

В **§ 4.1** приведены результаты классификации по типам структуры для двух наборов модельных данных.

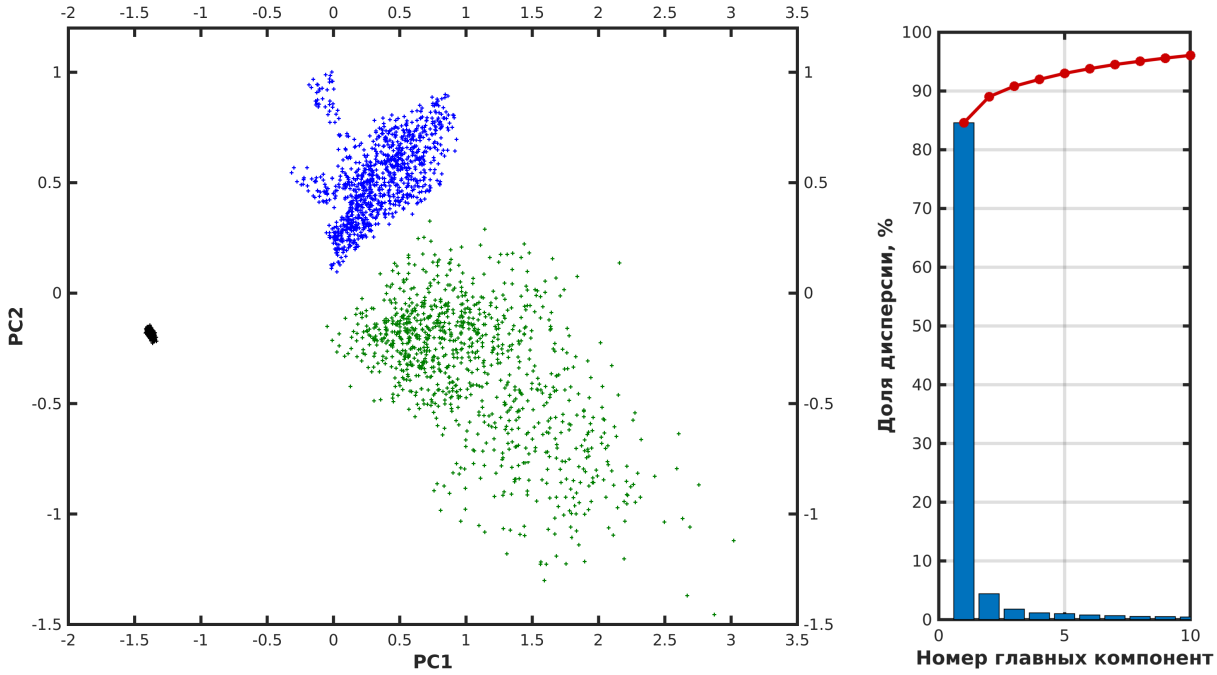
В пункте 4.1.1 описан первый набор модельных данных, который содержит 3000 изображений от трех типов объектов: пентоновый белок аденовируса человека (аденовирус), комбинация белков VP3 и VP7 из ядра вируса катаральной лихорадки (BTV) и капля воды (Рис. 3).

В пункте 4.1.2 представлены результаты классификации по типам структуры для первого набора модельных данных. Сначала все изображения были сжаты в характеристические векторы. На первом этапе векторы были спроецированы на двумерную плоскость методом главных компонент (Рис. 7). Видно, что векторы разделяются на три отдельные группы. Данные группы соответствуют типам структуры исследуемых объектов. Важно отметить, что применение метода главных компонент к изображениям без сжатия не приводит к явному разделению изображений на группы.

Затем, первый набор модельных данных разделялся на кластеры с использованием метода к-средних и метода спектральной кластеризации. При использовании сжатия, полученные кластеры соответствуют разбиению по типам структуры, но при кластеризации изображений без сжатия, полученные кластеры включают изображения нескольких типов структуры.

Далее, изображения первого набора классифицировались на основе метода опорных векторов с использованием сжатия в характеристический вектор. Все изображения были правильно классифицированы в соответствии с типом структуры.

В пунктах 4.1.3 – 4.1.6 представлены результаты исследования точности классификации второго набора модельных данных, который включает в себя 7000 изображений для семи белков с разной симметрией: циклическая, двугранная, четырехгранная, восьмигранная, икосаэдральная, спиральная и отсутствие



а) Каждому изображению соответствует точка на плоскости главных компонент, черные точки соответствует изображениям капли воды, синие - аденовирусу, зеленые - BTV

б) Распределение дисперсии данных по направлениям главных компонент

Рис. 7 — Проекция характеристических векторов изображений из набора модельных данных на плоскость из первых двух главных компонент (PC1 и PC2) методом главных компонент

симметрии. Результаты показывают, что метод классификации на основе математического метода опорных векторов с использованием сжатия изображений в характеристический вектор позволяет классифицировать с высокой точностью наборы изображений белков с разной симметрией.

В **§ 4.2** представлены результаты классификации по типам структуры на экспериментальных данных.

В пункте 4.2.1 приведено описание используемых наборов данных из открытой базы данных Coherent X-ray Imaging Data Bank (CXIDB) [21]. Они были получены в экспериментах на LCLS в 2011–2016 годах.

Всего использовалось четыре блока дифракционных изображений: CXIDB 13-14, CXIDB 10-11, CXIDB 20-25-27 и CXIDB 25. В названиях блоков перечислены идентификаторы наборов в базе данных CXIDB, на основе которых составлены блоки. Изображения перемешаны, однако сохранена информация о типах объектов на изображениях, которая была определена экспертами

CXIDB. Значения точности и полноты классификации определялись согласно схеме перекрестной проверки.

В пункте 4.2.3 приведено сравнение результатов классификации блока CXIDB 13-14 при использовании метода опорных векторов со сжатием в характеристический вектор в двух вариантах: с учетом зазоров детектора и смещения центра симметрии дифракционной картины, и без учета этих эффектов. Учет этих особенностей позволил повысить точность и полноту классификации в среднем на 2.5%.

В пункте 4.2.4 сравниваются результаты классификации наборов экспериментальных данных, полученные с использованием следующих методов:

- классификация на основе метода опорных векторов с порогом 75%,
- классификация на основе метода опорных векторов без порога,
- классификация на основе метода к-средних,
- классификация на основе метода спектральной кластеризации,
- классификация на основе перцептрона с тремя скрытыми слоями,
- классификация на основе свёрточной нейронной сети.

Первые четыре метода использовали сжатие изображений в характеристический вектор. Свёрточные нейронные сети и перцептрон с тремя скрытыми слоями применялись непосредственно к дифракционным изображениям без использования сжатия.

Результаты классификации представлены в таблице 1. Наилучшие значения точности и полноты достигаются при классификации на основе метода опорных векторов с использованием сжатия в характеристический вектор.

В пункте 4.2.5 подведен итог сравнения разработанных методов классификации и сделан вывод, что метод классификации на основе математического метода опорных векторов с использованием сжатия в характеристический вектор наиболее эффективен, так как:

1) он обеспечивает самую высокую точность классификации (более 90%).
2) введение порога вероятности корректной классификации на уровне 75% позволяет повысить точность на 3-5%, при допустимом уменьшении полноты;

Таблица 1 — Сравнение точности и полноты классификации на экспериментальных данных

Методы классификации		Точность (Т) и полнота (П) классификации, %			
		CXIDB 13-14	CXIDB 10-11	CXIDB 20-25-37	CXIDB 25
Классификация на основе метода опорных векторов с порогом 75%	Т	94.1%	99.8%	99.9%	95.0%
	П	80.8%	99.8%	98.8%	79.6%
Классификация на основе метода опорных векторов без порога	Т	90.8%	99.8%	99.6%	90.7%
	П	90.8%	99.8%	99.6%	91.5%
Классификация на основе метода к-средних	Т	82.9%	71.2%	86.1%	64.4%
	П	78.2%	53.6%	80.1%	99.8%
Классификация на основе спектральной кластеризации	Т	85.8%	79.0%	84.1%	68.9%
	П	83.4%	53.4%	73.7%	99.3%
Классификация на основе перцептрона с тремя скрытыми слоями	Т	87.1%	97.7%	99.6%	86.6%
	П	86.9%	97.6%	99.6%	91.2%
Классификация на основе свёрточной нейронной сети	Т	88.5%	99.4%	99.8%	85.8%
	П	88.3%	99.3%	99.8%	95.5%

3) сжатие в характеристические векторы позволяет сократить размерность пространства данных на 4 порядка при сохранении ключевых особенностей пространственной структуры исследуемых объектов.

В **пятой главе** представлено сравнение разработанных методов классификации по типам структуры с точки зрения обработки данных SPI экспериментов в режиме квази-онлайн.

В **§ 5.1** приведены результаты исследования зависимости точности классификации от размера обучающей выборки (таблица 2). Здесь мы вводим специальную терминологию: *максимальная точность* — это точность при обучении на выборке в 90% от полного набора, которая определена на основе перекрестной проверки (таблица 1). Мы определяем оптимальные размеры обучающей выборки, при которых точность составляет 99% от *максимальной точности*.

Таблица 2 — Оптимальный размер обучающей выборки при классификации дифракционных изображений, полученных в экспериментах на LCLS

Метод классификации	CXIDB 13-14	CXIDB 10-11	CXIDB 20-25-37	CXIDB 25
Классификация на основе метода опорных векторов с порогом 75%	40	30	160	80
Классификация на основе метода опорных векторов без порога	300	40	600	300
Классификация на основе перцептрона с тремя скрытыми слоями	640	920	520	2720
Классификация на основе свёрточной нейронной сети	580	420	520	2680
Всего изображений	958	2149	2665	4506

Метод классификации на основе метода опорных векторов со сжатием в характеристический вектор позволяет достичь 99% *максимальной точности* при меньших размерах обучающей выборки, чем другие методы классификации по типам структуры, а использование порога вероятности корректной клас-

сификации дополнительно уменьшает оптимальный размер обучающей выборки до 7 раз.

В **§ 5.2** приведено сравнение временных затрат на ручную разметку обучающей выборки, обучение и классификацию. Сравнение проведено на примере аппаратной конфигурации: центрального процессора Intel Xeon E5-2680v3 (далее обозначен как CPU), графического процессора NVIDIA Tesla K80 (далее GPU), а оптимальные размеры обучающей выборки берутся из таблицы 2.

Перед началом обучения требуется вручную разметить изображения обучающей выборки по типам структуры. Мы повторили работу эксперта и оцениваем время разметки на уровне 5 секунд на одно изображение. В этом случае, для метода классификации на основе метода опорных векторов с использованием сжатия в характеристический вектор время разметки составляет от 3 до 12 минут при условии использования порога вероятности корректной классификации в 75%. В то же время, в методах классификации на основе нейронных сетей, ручная разметка требует до нескольких часов, что обусловлено большим оптимальным размером обучающей выборки.

Сравнение временных затрат на обучение и классификацию приведено в таблице 3. При использовании сжатия в характеристический вектор, на обучение требуется не более 16 секунд. Для обучения нейронных сетей без сжатия, требуется более 5 минут для перцептрона с тремя скрытыми слоями и более двух часов для свёрточной нейронной сети. Время классификации 1000 изображений составляет менее 52 секунд в однопоточном режиме на CPU для всех методов, а при переходе на GPU время классификации уменьшается в несколько раз.

В **§ 5.3** представлены результаты исследования возможности классификации по типам структуры для потока дифракционных изображений в SPI экспериментах на EuXFEL в режиме квази-онлайн. Так как EuXFEL позволит фиксировать до 27000 изображений в секунду, а в сегодняшних экспериментах на LCLS остается около 2% изображений после фильтрации пустых, необходимо обеспечить производительность на уровне 540 изображений в секунду.

Требуемая производительность может быть обеспечена при распределении изображений по параллельным потокам вычислений на основе аппаратных ресурсов, приведенных в таблице 4. Требования к ресурсам были рассчитаны

Таблица 3 — Сравнение временных затрат при обучении и классификации

Методы классификации	Время обучения, сек.		Время классификации 1000 изображений, сек.	
	CPU	GPU	CPU	GPU
Классификация на основе метода опорных векторов с порогом 75%	4.5	1.0	51.3	11.3
Классификация на основе метода опорных векторов без порога	15.9	3.5	51.3	11.3
Классификация на основе метода к-средних	2.3	0.51	51.2	11.2
Классификация на основе метода спектральной кластеризации	2.3	0.50	51.9	11.7
Классификация на основе перцептрона с тремя скрытыми слоями	19.4	9.4	51.2	11.3
Классификация на основе свёрточной нейронной сети	8021	1542	23.9	19.7

на основе анализа временных затрат (таблица 3), а затем были проверены на практике.

В результате, классификация в режиме квази-онлайн может быть реализована по следующему сценарию:

1) характеристические вектора изображений рассчитываются со скоростью их получения в потоке;
2) после старта рабочего сеанса, накапливается обучающая выборка оптимального размера (таблица 2);
3) обучающая выборка размечается экспертом;
4) проводится обучение, временные затраты приведены в таблице 3;
5) после завершения обучения, вновь фиксируемые дифракционные изображения классифицируются со скоростью их поступления в потоке экспериментальных данных после фильтрации. Дифракционные изображения, сохраненные до завершения обучения, будут классифицированы за счет некоторой избыточной производительности ис-

Таблица 4 — Пример аппаратных ресурсов, которые обеспечивают классификацию дифракционных изображений со скоростью их поступления в экспериментах на EuXFEL

Методы классификации	Классификация на CPU	Классификация на GPU
Классификация на основе метода опорных векторов, метода к-средних, метода спектральной кластеризации с использованием сжатия	Сервер (Intel Xeon) с 20 ядрами на частоте 2.5 ГГц	Сервер с четырьмя GPU Nvidia K80
Классификация изображений без сжатия на основе перцептрона с тремя скрытыми слоями и сверточной нейронной сети	Сервер (Intel Xeon) с 12 ядрами на частоте 2.5 ГГц	Сервер с шестью GPU Nvidia K80

пользуемых аппаратных ресурсов, или после завершения эксперимента.

Таким образом, термин квази-онлайн означает наличие задержки на разметку и обучение, после чего данные обрабатываются в режиме онлайн. К концу эксперимента все полученные изображения будут классифицированы.

Результаты показывают, что для классификации дифракционных изображений в режиме квази-онлайн в SPI экспериментах на EuXFEL оптимально использовать метод классификации на основе математического метода опорных векторов с использованием сжатия в характеристические векторы, так как он показывает наилучшую скорость и точность классификации, а оптимальный размер обучающей выборки меньше, чем в других методах.

В **приложении 1** описаны методы уменьшения шума и устранения дефектов детектора, которые были применены к изображениям блока CXIDB 20-25-37. Входящие в этот блок изображения были получены разными группами ученых в разное время. На изображениях присутствуют особенности, обусловленные разными параметрами экспериментов. Все изображения блока CXIDB 20-25-37 были обработаны для устранения таких особенностей.

В **приложении 2** приведено описание комплекса программ, который был разработан в ходе диссертационных исследований. Он включает реализацию описанных в диссертации методов классификации по типам структуры и метода сжатия дифракционных изображений в характеристические векторы.

В **заключении** приведены основные результаты работы:

1. Разработан метод сжатия дифракционных изображений в характеристический вектор, который уменьшает размерность пространства параметров изображений на 4 порядка, сохраняя информацию о ключевых особенностях пространственной структуры исследуемых объектов.
2. Разработан метод классификации дифракционных изображений по типам структуры исследуемых объектов на основе математического метода опорных векторов с использованием сжатия в характеристический вектор. При классификации изображений, полученных в SPI экспериментах на LCLS, точность классификации по типам структуры составила более 90%.
3. Разработанные методы классификации по типам структуры верифицированы на наборах дифракционных изображений, полученных в экспериментах на LCLS в 2011–2016 годах. При сравнении этих методов определен наиболее эффективный — метод классификации на основе математического метода опорных векторов с использованием сжатия в характеристический вектор и порога вероятности корректной классификации порядка 75%.
4. Разработана методика формирования сценариев классификации по типам структуры с учетом характеристик аппаратных ресурсов, позволившая получить рекомендации по выполнению классификации по типам структуры в режиме квази-онлайн в экспериментах на EuXFEL. Классификация изображений по типам структуры идет со скоростью их поступления после этапа разметки обучающей выборки и этапа обучения.

Основные результаты диссертации опубликованы в работах:

Публикации в журналах Scopus, WoS, RSCI WoS, РИНЦ

1. Sorting algorithms for single-particle imaging experiments at X-ray free-electron lasers / S. A. Bobkov [и др.] // Journal of Synchrotron Radiation. — 2015. — Т. 22. — С. 1345–1352. — DOI: 10.1107/S1600577515017348. — Индексируется WoS и Scopus.
2. Классификация дифракционных изображений биологических макромолекул с разными типами симметрии в экспериментах по когерентной рентгеновской дифракционной микроскопии / С. А. Бобков [и др.] // Математическая биология и биоинформатика. — 2016. — Т. 11, № 2. — С. 299–310. — DOI: 10.17537/2016.11.299. — Индексируется Scopus.
3. *Бобков С. А.* Сравнительный анализ подходов к классификации дифракционных изображений биологических частиц, получаемых в экспериментах по когерентной рентгеновской дифракционной микроскопии // Математическая биология и биоинформатика. — 2017. — Нояб. — Т. 12, № 2. — С. 411–434. — DOI: 10.17537/2017.12.411. — Индексируется Scopus.
4. Метод представления дифракционных изображений XFEL для классификации, индексации и поиска / С. А. Бобков [и др.] // Компьютерные исследования и моделирование. — 2015. — Т. 7. — С. 631–639. — Индексируется RSCI WoS и РИНЦ.

Иные

1. *Бобков С. А.*, *Теслюк А. Б.* Метод кластеризации и классификации дифракционных изображений // 11-я Курчатовская молодежная научная школа, Сборник Аннотаций. — 2013. — С. 136.
2. Сравнительный анализ методов классификации многомерных данных для дифракционных изображений / С. А. Бобков [и др.] // Сборник “Proceedings of the 5th International Conference on Mathematical Biology and Bioinformatics”. — 2014. — С. 93.

Свидетельства о государственной регистрации программ для ЭВМ

1. *Бобков С. А.* Свидетельство о государственной регистрации программы для ЭВМ ’CCF Diag’ №2014611286. — 2013.
2. *Бобков С. А.* Свидетельство о государственной регистрации программы для ЭВМ ’CDI Center Detection Lib’ №2015662872. — 2013.
3. *Бобков С. А.* Свидетельство о государственной регистрации программы для ЭВМ ’Cuda CDI Correlation Lib for Python’ №2015663289. — 2015.

Список литературы

1. Fundamentals of Crystallography. 2002 / C. Giacovazzo [и др.].
2. High resolution 3D x-ray diffraction microscopy / J. Miao [и др.] // Physical Review Letters. — 2002. — Т. 89, № 8. — С. 088303.
3. *Gaffney K. J.*, *Chapman H. N.* Imaging Atomic Structure and Dynamics with Ultrafast X-ray Scattering // Science. — 2007. — Т. 316. — С. 1444.
4. Femtosecond diffractive imaging with a soft-X-ray free-electron laser / H. N. Chapman [и др.] // Nat Phys. — 2006. — Нояб. — Т. 2, № 12. — С. 839–843.
5. *Chapman H. N.*, *Nugent K. A.* Coherent lensless X-ray imaging // Nature Photonics. — 2010. — Дек. — Т. 4, № 12. — С. 833–839.
6. Single mimivirus particles intercepted and imaged with an X-ray laser / M. M. Seibert [и др.] // Nature. — 2011. — Т. 470, № 7332. — С. 78–81.
7. *Mancuso A. P.*, *Yefanov O. M.*, *Vartanyants I. A.* Coherent diffractive imaging of biological samples at synchrotron and free electron laser facilities // J. Biotechnology. — 2010. — Т. 149. — С. 229.
8. Potential for biomolecular imaging with femtosecond X-ray pulses / R. Neutze [и др.] // Nature. — 2000. — Авг. — Т. 406, № 6797. — С. 752–757.
9. Femtosecond X-ray protein nanocrystallography / H. N. Chapman [и др.] // Nature. — 2011. — Т. 470. — С. 73.

10. *Loh N.-T. D.*, *Elser V.* Reconstruction algorithm for single-particle diffraction imaging experiments // Physical Review E. — 2009. — Авг. — Т. 80, № 2.

11. *Fienup J. R.* Phase retrieval algorithms: a comparison // Applied Optics. — 1982. — Авг. — Т. 21, № 15. — С. 2758.

12. *Bauschke H. H.*, *Combettes P. L.*, *Luke D. R.* Phase retrieval, error reduction algorithm, and Fienup variants: a view from convex optimization // Journal of the Optical Society of America A. — 2002. — Июль. — Т. 19, № 7. — С. 1334.

13. *Elser V.* Phase retrieval by iterated projections // Journal of the Optical Society of America A. — 2003. — Янв. — Т. 20, № 1. — С. 40–55.

14. Coherent soft X-ray diffraction imaging of coliphage PR772 at the Linac coherent light source / H. K. Reddy [и др.] // Scientific data. — 2017. — Т. 4. — С. 170079.

15. Automated identification and classification of single particle serial femtosecond X-ray diffraction data / J. Andreasson [и др.] // Optics Express. — 2014. — Февр. — Т. 22, № 3. — С. 2497–2510.

16. The European x-ray free-electron laser / M. Altarelli [и др.] // Technical Design Report, DESY. — 2006. — Т. 97. — С. 1–26.

17. *Sakdinawat A.*, *Attwood D.* Nanoscale X-ray imaging // Nature Photonics. — 2010. — Т. 4, № 12. — С. 840–848.

18. *Rosenblatt F.* Principles of neurodynamics. perceptrons and the theory of brain mechanisms : тех. отч. / Cornell Aeronautical Lab Inc Buffalo NY. — 1961.

19. *LeCun Y.*, *Bengio Y.*, *Hinton G.* Deep learning // Nature. — 2015. — Т. 521, № 7553. — С. 436–444.

20. *Stone M.* Cross-validatory choice and assessment of statistical predictions // Journal of the royal statistical society. Series B (Methodological). — 1974. — С. 111–147.

21. *Maia F. R. N. C.* The Coherent X-ray Imaging Data Bank // Nature methods. — 2012. — Т. 9, № 9. — С. 854–855.