«Московский физико-технический институт (национальный исследовательский университет)»
Физтех-школа прикладной математики и информатики
Кафедра «Интеллектуальные системы»

Северилов Павел Андреевич

Оценка качества прогнозирования структуры белка с использованием графовых свёрточных нейронных сетей

03.03.01 – Прикладные математика и физика

Выпускная квалификационная работа бакалавра

Научный руководитель:
д.ф.-м.н. Стрижов Вадим Викторович

Москва
2020

Содержание

Аннотация

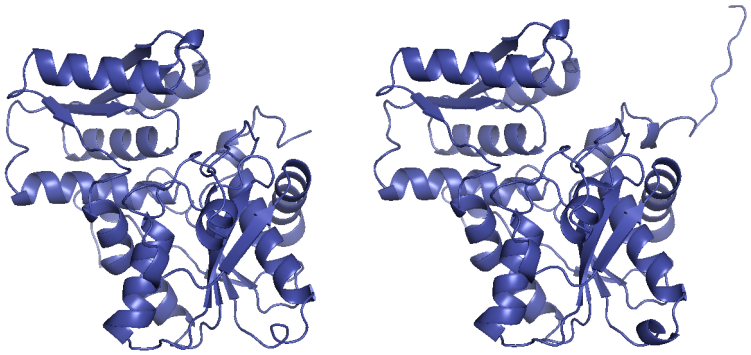
Последовательность аминокислот сворачивается в нативную структуру белка. Моделируется структура, в которую произойдет сворачивание. Определить качество смоделированной структуры по отношению к нативной вычислительно дорого. В работе решается задача оценки качества структуры смоделированного белка (Quality Assessment), т.е. строится регрессия смоделированных белковых структур на значение метрики схожести её и нативной структуры белка $\mathrm{CAD}_{\mathrm{score}}$. В данной работе впервые исследуется графовый подход к задаче вместе с использованием преобразования свёртки. Преимущество предложенного подхода по сравнению с ранее представленными заключается в том, что графовое представление позволяет одновременно учитывать и первичную, и третичную структуры белка. В работе методами спектральной теории графов проанализирован спектр графовой свёртки и применены графовые свёрточные нейронные сети к задаче Quality Assessment. Эксперименты проводятся на данных с соревнований CASP по решению данной задачи, которые представляют собой трёхмерные координаты и химические свойства атомов белка. Построены графовые представления для смоделированных структур в виде матриц смежности и матриц координат атомов белков. Параметры нейросети оптимизируются на наборах данных CASP9-CASP11. Проведен анализ корреляций Пирсона и Спирмена предсказаний модели и истинных значений качества структуры на данных CASP12. Качество, достигаемое моделью, сравнимо с качеством альтернативных моделей, дающих наилучшее качество в задаче.

Ключевые слова: *белковые структуры, графы, графовые свёртки, графовые нейронные сети, свёрточные нейронные сети, спектральный анализ.*

Введение

Белки являются наиболее универсальными макромолекулами в живых системах и выполняют важнейшие функции практически во всех биологических процессах [1]. Форма белковой структуры определяет выполняемые ей функции [1]. Понимание белковых структур и выполняемых ими задач имеют важное значение для медицинских, фармацевтических и генетических исследований [2]. Решение задачи определения, в какую *нативную структуру* свернётся последовательность аминокислот в белке, занимает большое количество времени и ресурсов.

Каждые два года проводятся соревнования Critical Assessment of protein Structure Prediction (CASP [3]) по решению задачи прогнозирования структуры. Вычислительные методы, которые её решают состоят из двух этапов: моделирование структуры белка из их аминокислотных последовательностей и оценивание качества прогнозирования. В данной работе рассматривается только второй этап. Под качеством прогнозирования понимается численное значение близости *смоделированной* и нативной структур (например, метрики $\text{CAD}_{\text{score}}$ [4], LDDT [5], GDT [6]). Вычислять напрямую данные метрики вычислительно дорого, поэтому данная проблема рассматривается как отдельная задача.



(a) True T0861 (b) Predicted Atome2_CBS_TS4

Рис. 1: Пример нативной и смоделированной структуры белка

Белковая структура состоит из одной или нескольких цепочек более мелких молекул – аминокислотных остатков. Последовательность

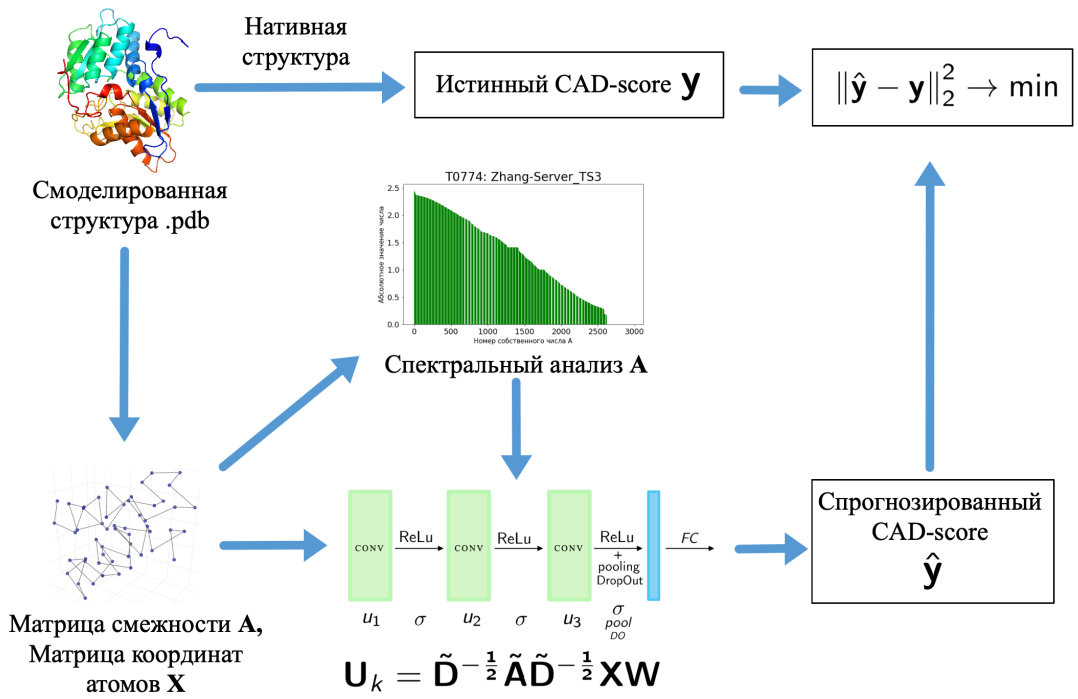


Рис. 2: Диаграмма оценки качества структуры белка

остатков $S = \{a_i\}_{i=1}^{N}$ представляет его первичную структуру, где a_i является одним из 22 типов аминокислот. Взаимодействия между соседними остатками и окружающей средой определяют, как цепочка будет сворачиваться в сложные структуры, которые представляют вторичную структуру и третичную структуру белка [2].

Поэтому для задач прогнозирования и оценки качества белковых структур требуется учитывать как пространственную информацию об атомах, третичную структуру, так и признаки в виде последовательностей аминокислот, первичную структуру белка. В работах [7,8] для оценки качества прогнозирования белковых структур используются LSTM или 1D-CNN, которые представляют белки в виде последовательности аминокислот с пространственными признаками. В работах [9,10] прогнозируется качество структуры белков с использованием 3D-CNN, но не учитывается первичная структура белка. В данных работах не учитываются одновременно первичная и третичная структуры белка. На основе графового представления учитываются как последовательности аминокислот, так и пространственные, геометрические структуры белков.

Работа [2] – единственная, в которой используется графовое представление структуры белка для решения задачи оценки качества прогнозирования структуры. В ней графовые нейронные сети на основе алгоритма, описанного в [11], показывают результаты, превосходящие остальные современные методы, дающие наилучшее качество в задаче. В модели из [2] не используются свёртки. Основные результаты в задаче оценивания качества структуры белка полагаются на свёрточные нейронные сети [10].

В данной работе впервые исследуются графовые свёртки применительно к задаче Quality Assessment. Методами спектральной теории графов определена свёртка на графах и проанализирован её спектр. На основе полученного преобразования графовой свёртки построена модель графовой свёрточной нейронной сети для задачи оценки качества прогнозирования структуры белка и протестирована на данных CASP9-12. Данные для экспериментов брались с соревнований CASP прошлых лет, которые представлены в виде информации об атомах и их пространственном расположении в виде координат для нативных и смоделированных структур белков. По этим данным построено графовое представление каждой смоделированной структуры – матрица координат $\mathbf{X}$ и матрица смежности $\mathbf{A}$. На рисунке 2 представлена диаграмма решения задачи оценки качества структуры белка.

1 Постановка задачи оценки качества структуры белка

Дана выборка

$$\mathfrak{D} = \{\mathbf{x}_i, y_i\}_{i=1}^{m},$$

где $\mathbf{x}_i \in \mathbb{R}^{n_i \times 3}$ – молекулы, каждая из которых описана множеством трёхмерных координат всех ее n_i атомов, $y_i \in \mathbb{R}$ – оценка близости смоделированной и нативной структуры белка. Оценка близости измеряется различными метриками: $\text{CAD}_{\text{score}}$ [4], LDDT [5], GDT [6]. В данной работе выбран $\text{CAD}_{\text{score}}$.

1.1 Вычисление CAD score

Обозначим через P множество всех пар элементов последовательности аминокислот (остатков) (i, j), имеющих ненулевую площадь контакта $N_{(i,j)}$ в нативной структуре. Затем для каждой пары остатков $(i, j) \in P$ вычисляется площадь контакта $M_{(i,j)}$ смоделированной структуры.

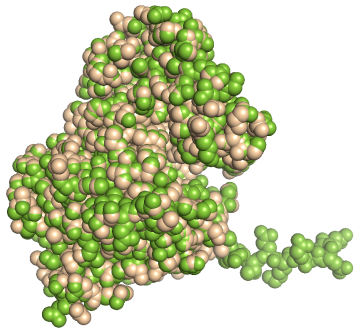


Рис. 3: Пересечение нативной и смоделированной структур

Для каждой пары остатков $(i, j) \in P$ определяется разность площадей контакта $\mathrm{CAD}_{(i,j)}$ как абсолютная разница площадей контакта между остатками i и j в нативной N и смоделированной структуре M:

$$\mathrm{CAD}_{(i,j)} = \left| N_{(i,j)} - M_{(i,j)} \right|.$$

Для вычислительной стабильности берется ограниченный CAD: $\mathrm{CAD}^{\mathrm{bounded}}_{(i,j)} = \min\left(\mathrm{CAD}_{(i,j)}, N_{(i,j)}\right)$. Таким образом, $\mathrm{CAD}_{\mathrm{score}}$ для всей структуры определяется как

$$\mathrm{CAD}_{\mathrm{score}} = 1 - \frac{\sum_{(i,j)\in P} \mathrm{CAD}^{\mathrm{bounded}}_{(i,j)}}{\sum_{(i,j)\in P} N_{(i,j)}}. \tag{1}$$

На рисунке 3 представлен пример пересечения нативной структуры T0861 (жёлтый) и её модели Atome2_CBS_TS4 (зелёный) при $\mathrm{CAD}_{\mathrm{score}} = 0.829$.

1.2 Задача регрессии белковых структур на $\mathrm{CAD}_{\mathrm{score}}$

Пусть $\mathbf{X} = \bigcup_{i=1}^{m} \mathbf{x}_i$. Рассматривается множество параметрических моделей $\mathfrak{F}$, взятых из класса графовых свёрточных нейронных сетей:

$$\mathfrak{F} = \{\mathbf{f}_k \colon (\mathbf{w}, \mathbf{X}) \to \hat{\mathbf{y}} \mid k \in \mathfrak{K}\},$$

где $\mathbf{w} \in \mathbb{W}$ – параметры модели, $\hat{\mathbf{y}} = \mathbf{f}(\mathbf{X}, \mathbf{w}) \in \mathbb{R}^m$ – вектор оценок предсказаний CAD-scores.

Решается задача регрессии для предсказания численного значения y_i $\mathrm{CAD}_{\mathrm{score}}$ белка на основе его смоделированной пространственной структуры $\mathbf{x}_i$.

Параметры модели $\mathbf{w} \in \mathbb{W}$ минимизируют функцию ошибки на обучении. Определим функцию ошибки:

$$\mathfrak{L}(\mathbf{y}, \mathbf{X}, \mathbf{w}) = \|\hat{\mathbf{y}} - \mathbf{y}\|_2^2,$$

где $\hat{\mathbf{y}} = \mathbf{f}(\mathbf{X}, \mathbf{w})$ – $\mathrm{CAD}_{\mathrm{score}}$ предсказанный моделью $\mathbf{f}$, $\mathbf{y}$ – данный в выборке $\mathrm{CAD}_{\mathrm{score}}$. Таким образом, решается данная задача оптимизации:

$$\mathbf{w}^* = \underset{\mathbf{w} \in \mathbb{W}}{\operatorname{argmin}}(\mathfrak{L}(\mathbf{w}))$$

Для оценивания качества модели анализируются коэффициенты корреляции Пирсона (R), Спирмена (ρ) [2, 9, 10]. Для каждой нативной структуры белка вычисляются коэффициенты корреляции Пирсона (R^{target}), Спирмена (ρ^{target}) между истинными и прогнозируемыми $\mathrm{CAD}_{\mathrm{score}}$ для смоделированных структур, соответствующих данной нативной структуре белка. Затем коэффициенты корреляции усредняются по всем T нативным структурам. Обозначим $\mathbf{y}_i \in \mathbb{R}^{m_i}$ и $\hat{\mathbf{y}}_i \in \mathbb{R}^{m_i}$ соответственно вектор истинных значений и вектор предсказаний $\mathrm{CAD}_{\mathrm{score}}$ для смоделированных структур белка, соответствующих нативной структуре i. Здесь m_i – количество смоделированных структур для i-ой нативной структуры. Тогда коэффициенты корреляции записываются:

$$R = R\,(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{T} \sum_{i=1}^{T} R_i^{\mathrm{target}} = \frac{1}{T} \sum_{i=1}^{T} \mathrm{PEARSON}\,(\mathbf{y}_i, \hat{\mathbf{y}}_i)$$

$$\rho = \rho\,(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{T} \sum_{i=1}^{T} \rho_i^{\mathrm{target}} = \frac{1}{T} \sum_{i=1}^{T} \mathrm{SPEARMAN}\,(\mathbf{y}_i, \hat{\mathbf{y}}_i)$$

Здесь PEARSON$(\cdot,\cdot)$ и SPEARMAN$(\cdot,\cdot)$ – корреляции Пирсона и Спирмена соответственно:

$$\text{PEARSON}\left(\mathbf{y}_i, \hat{\mathbf{y}}_i\right) = \frac{\sum_{l=1}^{m_i}\left(\mathbf{y}_{il} - \bar{\mathbf{y}}_i\right)\left(\hat{\mathbf{y}}_{il} - \bar{\hat{\mathbf{y}}}_i\right)}{\sqrt{\sum_{l=1}^{m_i}\left(\mathbf{y}_{il} - \bar{\mathbf{y}}_i\right)^2 \sum_{l=1}^{m_i}\left(\hat{\mathbf{y}}_{il} - \bar{\hat{\mathbf{y}}}_i\right)^2}}$$

$$\text{SPEARMAN}\left(\mathbf{y}_i, \hat{\mathbf{y}}_i\right) = \frac{\sum_{l=1}^{m_i}\left(\text{rank}\left(\mathbf{y}_{il}\right) - \frac{m_i+1}{2}\right)\left(\text{rank}\left(\hat{\mathbf{y}}_{il}\right) - \frac{m_i+1}{2}\right)}{\frac{1}{12}\left(m_i^3 - m_i\right)}$$

1.3 Построение матриц смежности

Т.к. данные о белках не содержат информации о соединениях между атомами, т.е. нет матрицы смежности, для всех взятых смоделированных структур белков вычисляются матрицы смежности A по следующим правилам:

- не соединяются водород с водородом,
- атом не соединяется с водородом, если расстояние между ними не менее 1.21Å,
- не соединяются атомы, которые находятся далеко в последовательности (номера остатков отличаются больше, чем на 1),
- не соединяются атомы, создающие дисульфидные связи,
- соединяются атомы, расстояние между которыми $r \in (r_{\min}, r_{\max}]$, где $r_{\min} = 0.01$Å, $r_{\max} = \left(0.6 \cdot \left(\rho_{\text{atom1}} + \rho_{\text{atom2}}\right)\right)^2$, ρ_{atom} – радиус атома (максимально возможное $r_{\max} = 5.76$ – при $\rho_{\text{atom1}} = \rho_{\text{atom2}} = 2.0$).

По попарным расстояниям между атомами на Рис. 4 видно, что соединения могут иметь атомы, обозначенные самым светлым желтым, т.к. максимально возможное расстояние между атомами, при котором они могут иметь соединение по представленным правилам составления

матрицы смежности равно 5.76. Т.е. матрица смежности будет сильно разреженной.

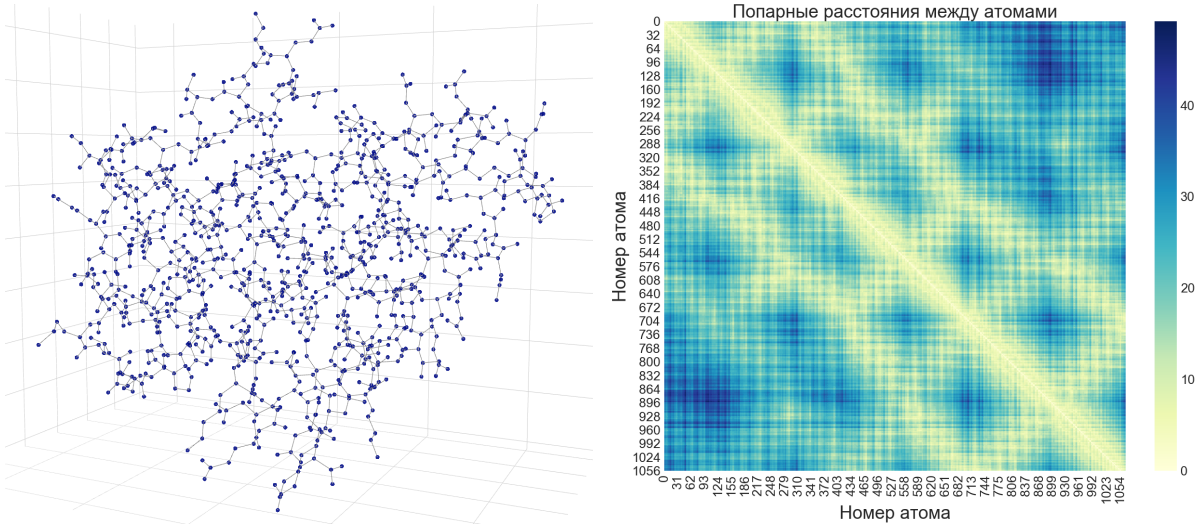


Рис. 4: Трехмерное представление с помощью координат $\mathbf{X}$ и полученной матрицы смежности $\mathbf{A}$ и попарные расстояния между атомами смоделированной структуры BAKER-ROSETTASERVER_TS3 для нативной структуры T0870 из набора данных CASP12

2 Спектральный анализ

Для обобщения свёрточных нейронных сетей на графы требуется определить свёрточные фильтры на графах. Существует два подхода: пространственный и спектральный [12, 13]. Как показано в [14], пространственный подход не имеет общего математического определения трансляции на графах, в то время как спектральный метод имеет хорошее математическое обоснование. Поэтому рассматривается спектральная теория графов.

Элементы аминокислотной последовательности рассматриваются как отдельные узлы, чьи связи (ребра) описывают пространственные отношения между ними.

В общем случае граф $\mathbf{G}$ определяется набором $(\mathbf{V}, \mathbf{A})$, где $\mathbf{V} \in \mathbb{R}^{n \times c}$ определяет вершины или узлы графа. Матрица смежности $\mathbf{A} \in \mathbb{R}^{n \times n}$ определяет соединения между n узлами (ребра), где $\mathbf{A}_{ij}$ – наличие связи

между узлами i и j. Используя это определение графа, белковые структуры можно определить как графы, признаки элементов аминокислотной последовательности которых закодированы в элементах $\mathbf{V}$ узлов, а пространственная близость между элементами закодирована в матрице смежности $\mathbf{A}$.

2.1 Преобразование графовой свёртки

Определение 1 *Графовый Лапласиан [15] – матрица* $\mathbf{L} = \mathbf{I}_n - \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}$*, где* $\mathbf{A}$ *– матрица смежности графа* $\mathbf{G}$*,* $\mathbf{D}$ *– диагональная матрица степеней вершин,* $\mathbf{D}_{ii} = \sum_j (\mathbf{A}_{ij})$*,* $\mathbf{I}_n$*– единичная матрица.*

Матрица $\mathbf{L}$ является вещественной симметричной положительной полуопределенной, поэтому может быть представлена в виде $\mathbf{L} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\mathsf{T}$, где $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n] \in \mathbb{R}^{n\times n}$ – это матрица собственных векторов, упорядоченных по собственным значениям, $\mathbf{\Lambda} \in \mathbb{R}^{n\times n}$ – диагональная матрица собственных значений (спектр), $\mathbf{\Lambda}_{ii} = \lambda_i$. Спектральное разложение Лапласиана позволяет определить преобразование Фурье для графов: собственные векторы соответствуют модам Фурье, а собственные значения – частотам.

Определение 2 *Графовое преобразование Фурье [16] для сигнала* $\mathbf{x} \in \mathbb{R}^n$ *задается* $\mathscr{F}(\mathbf{x}) = \mathbf{U}^\mathsf{T}\mathbf{x} \equiv \hat{\mathbf{x}} \in \mathbb{R}^n$*, а обратное графовое пребразование Фурье:* $\mathscr{F}^{-1}(\hat{\mathbf{x}}) = \mathbf{U}\hat{\mathbf{x}}$*, где* $\mathbf{x}$ *– вектор признаков всех вершин.*

Данное преобразование является ключевым в определении графовой свёртки. Оно проецирует входной графовый сигнал на ортонормированное пространство, где базис формируется собственными векторами графового Лапласиана. Элементы преобразованного сигнала $\hat{\mathbf{x}}$ являются координатами сигнала в новом пространстве, так что входной сигнал может быть представлен как $\mathbf{x} = \sum_i \hat{x}_i \mathbf{u}_i$, что является обратным графовым преобразованием Фурье.

Теорема 1 ***(Теорема о свёртках)*** *[17] Преобразование Фурье свёртки двух сигналов является покомпонентным произведением их преобразований Фурье, т.е.*

$$\mathscr{F}(\mathbf{f} * \mathbf{g}) = \mathscr{F}(\mathbf{f}) \odot \mathscr{F}(\mathbf{g}).$$

Следуя из теоремы 1, спектральная свёртка на графах определяется для сигнала $\mathbf{x}$ и фильтра $\mathbf{g} \in \mathbb{R}^n$ как

$$\mathbf{x} * \mathbf{g} = \mathscr{F}^{-1}(\mathscr{F}(\mathbf{x}) \odot \mathscr{F}(\mathbf{g})) = \mathbf{U}\left(\mathbf{U}^\mathsf{T}\mathbf{x} \odot \mathbf{U}^\mathsf{T}\mathbf{g}\right) = \mathbf{U}\mathbf{g}_\theta\mathbf{U}^\mathsf{T}\mathbf{x}, \qquad (2)$$

где $\mathbf{g}_\theta = diag\left(\mathbf{U}^\mathsf{T}\mathbf{g}\right)$ – спектральные коэффициенты фильтра.

Спектральные методы отличаются выбором фильтра $\mathbf{g}_\theta$. Соотношение (2) вычислительно дорогое, т.к. спектральное разложение требует $O\left(n^3\right)$ операций, а перемножение с матрицей собственных векторов $\mathbf{U}$ требует $O\left(n^2\right)$ операций. Chebyshev Spectral CNN (ChebNet) [18] обходит эти проблемы аппроксимацией $\mathbf{g}_\theta$ с помощью полиномов Чебышева $\mathbf{T}_k(\mathbf{x})$, убирая необходимость считать собственные векторы Лапласиана $\mathbf{L}$.

Определение 3 *Полиномы Чебышева* $\mathbf{T}_k(\mathbf{x})$ *k-ого порядка задаются рекуррентным соотношением* $\mathbf{T}_k(\mathbf{x}) = 2\mathbf{x} \cdot \mathbf{T}_{k-1}(\mathbf{x}) - \mathbf{T}_{k-2}(\mathbf{x}), \mathbf{T}_0(\mathbf{x}) = 1, \mathbf{T}_1(\mathbf{x}) = \mathbf{x}$. *Образуют ортогональный базис в* $L^2\left([-1,1], \dfrac{dx}{\sqrt{1-x^2}}\right)$.

Представляя $\mathbf{g}_\theta$ в виде

$$\mathbf{g}_\theta = \sum_{k=0}^{K} \theta_k \mathbf{T}_k(\tilde{\mathbf{\Lambda}}),$$

где $\tilde{\mathbf{\Lambda}} = 2\mathbf{\Lambda}/\lambda_{\max} - \mathbf{I}_n \in [-1,1]$, $\lambda_{\max}$ – максимальное собственное число $\mathbf{L}$, а также замечая, что

$$\left(\mathbf{U}\mathbf{\Lambda}\mathbf{U}^\mathsf{T}\right)^k = \mathbf{U}\mathbf{\Lambda}^k\mathbf{U}^\mathsf{T}$$

(собственные векторы образуют ортонормированный базис $\mathbf{U}^\mathsf{T}\mathbf{U} = \mathbf{I}$), получаем:

$$\mathbf{U}\mathbf{g}_\theta\mathbf{U}^\mathsf{T}\mathbf{x} = \mathbf{U}\left(\sum_{i=0}^{K} \theta_k \mathbf{T}_k(\tilde{\mathbf{\Lambda}})\right)\mathbf{U}^\mathsf{T}\mathbf{x} = \sum_{k=0}^{K} \theta_k \mathbf{T}_k(\tilde{\mathbf{L}})\mathbf{x}, \qquad (3)$$

где $\tilde{\mathbf{L}} = 2\mathbf{L}/\lambda_{\max} - \mathbf{I}_n$.

Graph Convolutional Network (GCN) [19] используют первое приближение ChebNet. Предполагая $\lambda_{\max} \approx 2$ и беря первые 2 слагаемых в сумме ($K = 1$), соотношение (3) упрощается до

$$\mathbf{x} * \mathbf{g} \approx \tilde{\theta}_0 \mathbf{x} + \tilde{\theta}_1 \left(\mathbf{L} - \mathbf{I}_n\right) \mathbf{x} = \tilde{\theta}_0 \mathbf{x} - \tilde{\theta}_1 \mathbf{D}^{-\frac{1}{2}} \mathbf{A} \mathbf{D}^{-\frac{1}{2}} \mathbf{x}. \tag{4}$$

Приняв $\theta = \tilde{\theta}_0 = -\tilde{\theta}_1$, получаем:

$$\mathbf{x} * \mathbf{g} \approx \theta \left(\mathbf{I}_n + \mathbf{D}^{-\frac{1}{2}} \mathbf{A} \mathbf{D}^{-\frac{1}{2}}\right) \mathbf{x}. \tag{5}$$

Оператор в скобках может может привести к вычислительной нестабильности и взрыву или затуханию градиентов, т.к. собственные значения данного оператора $\in [0, 2]$. Для решения проблемы в [19] предлагается *трюк перенормировки*:

$$\mathbf{I}_n + \mathbf{D}^{-\frac{1}{2}} \mathbf{A} \mathbf{D}^{-\frac{1}{2}} \to \tilde{\mathbf{D}}^{-\frac{1}{2}} \tilde{\mathbf{A}} \tilde{\mathbf{D}}^{-\frac{1}{2}}, \text{ где } \tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}_n, \ \tilde{\mathbf{D}}_{ii} = \sum_j \tilde{\mathbf{A}}_{ij}.$$

Дан граф $\mathbf{G}$ и матрица с информацией об узлах $\mathbf{X} \in \mathbb{R}^{n \times c}$ (n – число узлов и c – число признаков в каждом узле). Исходя из (5) и применяя трюк перенормировки, определяется слой свёртки графа:

$$\mathbf{U} = \tilde{\mathbf{D}}^{-\frac{1}{2}} \tilde{\mathbf{A}} \tilde{\mathbf{D}}^{-\frac{1}{2}} \mathbf{X} \mathbf{W}, \tag{6}$$

где $\mathbf{W} \in \mathbb{R}^{c \times t}$ – матрица параметров свёртки с t фильтрами, а $\mathbf{U} \in \mathbb{R}^{n \times t}$ – выходная матрица. На рисунке 5 изображена схема свёрточного слоя на основе полученного преобразования.

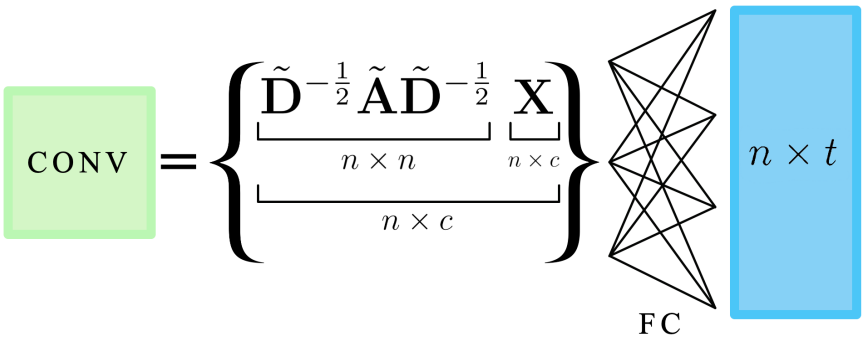


Рис. 5: Схема свёртки графа с матрицей $\mathbf{X}$ размера $n \times c$, t – число фильтров в свёртке, FC – полносвязный слой. Синий прямоугольник – выходная матрица размером $n \times t$

2.2 Функция преобразования регрессионной модели

Архитектура сети составляется по аналогии с моделью GCN [19]. На основе выражения (6) определяются свёрточные слои (рисунок 5). Нелинейная функция σ выбрана ReLu.

Сеть состоит из 3 свёрточных слоёв, макспулинга pool по вершинам графа и полносвязного слоя FC – скалярного умножения с вектором $\mathbf{w}_4$. Параметры свёрток t взяты равными 64, 64, 64 соответственно для первого, второго, третьего свёрточных слоёв. На рисунке 6 представлена схема тестируемой в работе нейронной сети.

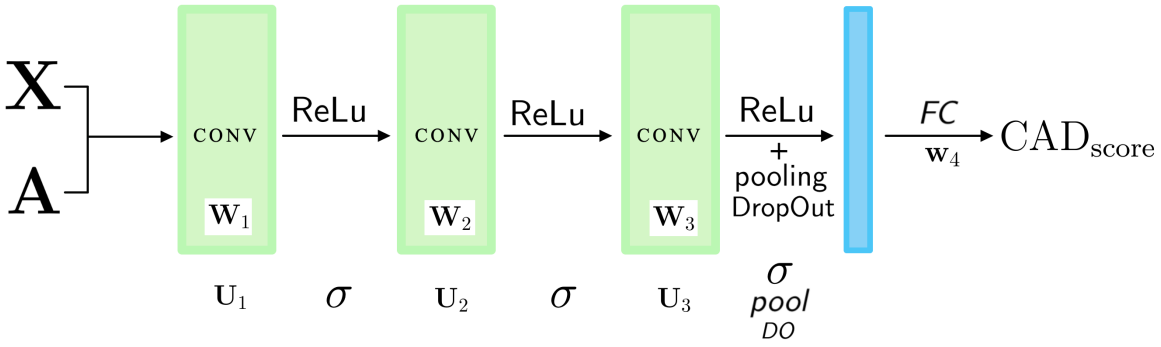


Рис. 6: Схематическое представление архитектуры свёрточной нейронной сети SpectralQA, использованной в данной работе

Таким образом, преобразование $f : \mathbf{X} \to \mathrm{CAD}_{\mathrm{score}}$ полученной нейросети записывается в виде

$$f = \langle \mathbf{w}_4, \mathrm{DO} \circ \mathrm{pool} \circ \sigma\left(\mathbf{U}_3\right) \circ \sigma\left(\mathbf{U}_2\right) \circ \sigma\left(\mathbf{U}_1\right) \rangle,$$

где $\mathbf{U}_k = \tilde{\mathbf{D}}^{-\frac{1}{2}} \tilde{\mathbf{A}} \tilde{\mathbf{D}}^{-\frac{1}{2}} \mathbf{X} \mathbf{W}_k$, DO – дропаут, pool – максимум по всем узлам графа.

3 Вычислительный эксперимент

3.1 Описание данных

Данные для эксперимента берутся с соревнований CASP разных лет. Используются наборы данных CASP9–CASP12 (таблица 1). Данные пред-

ставляют собой пары нативная-смоделированная структуры, каждая из которых описана координатами и химическими свойствами атомов структуры. Обучение модели происходит на данных CASP9–CASP11, тестирование – на CASP12. Для процессов обучения и тестирования по формуле (1) вычисляются $\mathrm{CAD}_{\mathrm{score}}$ для всех смоелированных структур на основе нативных структур.

Таблица 1: Наборы данных белковых структур

Набор	Нативные структуры	Модели структур	Разбиение
CASP 9	117	35963	Train, Validation
CASP 10	103	15450	
CASP 11	84	12291	
CASP 12	37	5501	Test
Суммарно	341	69205	

3.2 Собственное пространство матриц смежности

Для каждой полученной матрицы смежности $\mathbf{A}$ и матрицы после прохождения свёртки $\mathbf{U}_k$ производится сингулярное разложения для получения собственных чисел матрицы. На Рис. 7 и 8 представлены собственные числа для смоделированной структуры STRINGS_TS3, соответствующей нативной T0759.

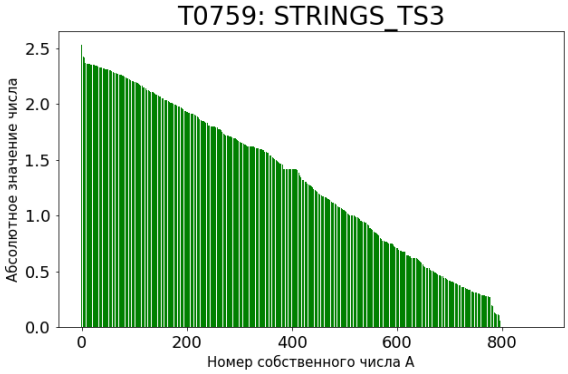


Рис. 7: Собственные числа $\mathbf{A}$

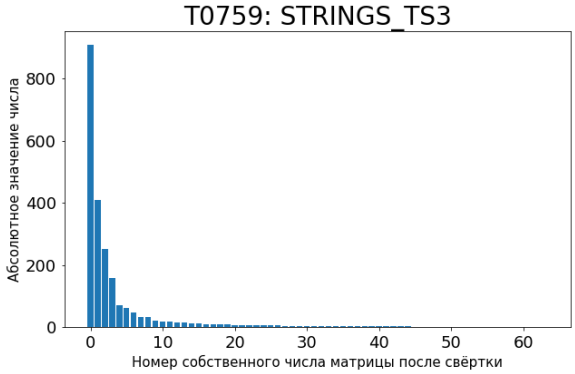


Рис. 8: Собственные числа $\mathbf{U}_k$

Для оценки размерности собственного пространства матриц используется правило сломанной трости [20]. Набор собственных чисел сравнивается с порогомами: для матрицы $\mathbf{A}$ с порогом A, для $\mathbf{U}_k$ – с порогом U. По правилу сломанной трости j-ый собственный вектор $\mathbf{A}$ сохраняется в списке главных компонент, если $\lambda_j > A$. Аналогично для $\mathbf{U}_k$.

Для каждой нативной структуры из данных CASP11 и CASP12 было выбрано случайным образом по одной смоделированной структуре. Для каждой из выбранных смоделированных структур посчитаны собственные числа для матриц $\mathbf{A}$ и $\mathbf{U}_k$. За размерность собственных пространств матриц взято количество собственных чисел, больших порога. Были рассмотрены пороги $U = 10$ и $A \in \{0.5, 1.0, 2.0\}$.

Результаты представлены на Рис. 9, на котором каждая точка соответствует одной смоделированной структуре. Размерность собственного пространства матрицы после прохождения через свёртку сжимается в 50-100 раз. Это может быть объяснено сильной разреженностью матриц смежности белковых структур.

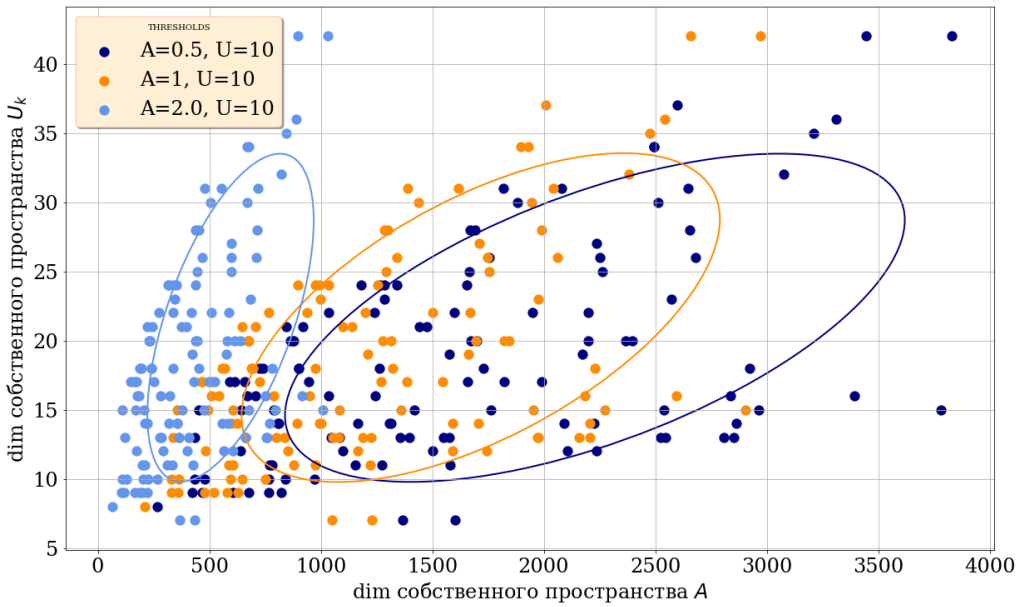


Рис. 9: Собственные пространства для порогов $U = 10$ и $A \in \{0.5, 1.0, 2.0\}$.

3.3 Анализ корреляций Пирсона и Спирмена

При обучении нейросети анализируются усредненные по T нативным структурам коэффициенты корреляции Пирсона и Спирмена. Процесс обучения представлен на рисунках 10 и 11

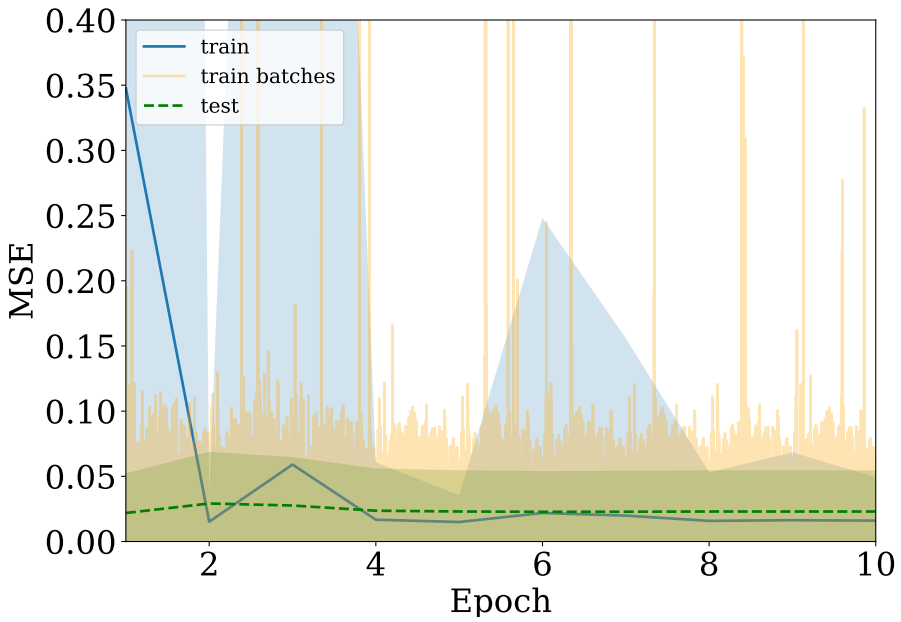


Рис. 10: График MSE ошибки SpectralQA на обучающей и тестовой выборке

Графики корреляций Пирсона и Спирмена стабилизируются возле одного значения (рисунок 11). Большая дисперсия объясняется тем, что для некоторых нативных структур сложно смоделировать структуру, из-за чего $\mathrm{CAD}_{\mathrm{score}}$ будет равным 0 для плохих смоделированных структур в силу выражения (1), а не близким к нулю значением. Этим же и объясняется невысокое значение корреляции.

В таблице 2 представлены результаты тестирования модели на данных соревнования CASP12. Корреляция здесь берется между всеми предсказаниями и истинными значениями, а не как при обучении усредненная по нативным структурам. Из сравнения данных в таблице видно, что модель из данной работы дает качество, сравнимое с качеством альтернативных моделей, дающих наилучшее качество в задаче.

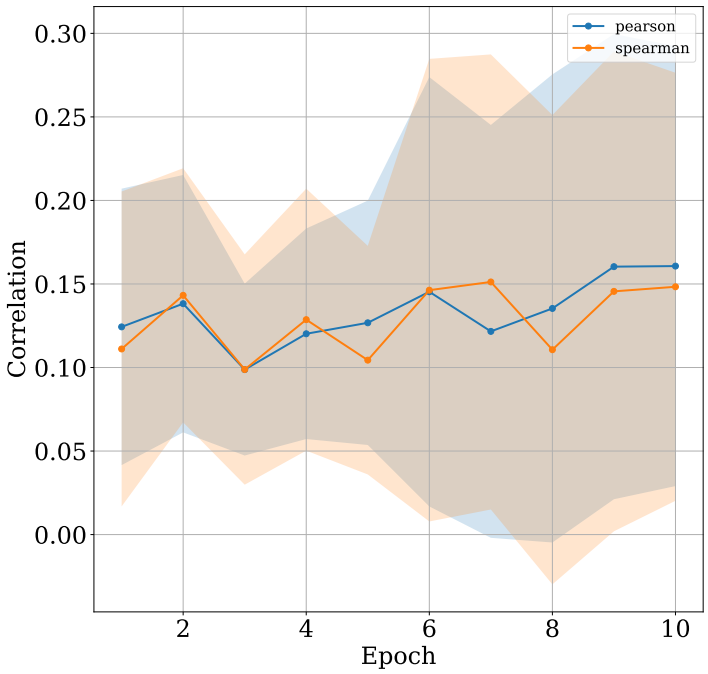


Рис. 11: Корреляция Пирсона и Спирмена при обучении

Таблица 2: Сравнение корреляции Пирсона и Спирмена существующих современных алгоритмов с моделью SpectralQA на данных CASP12

Метод	Spearmann ρ	Pearson R
ProQ3D	0.801	0.750
VoroMQA	0.803	0.766
SBROD	0.685	0.762
Ornate	**0.828**	**0.781**
SpectralQA (данная работа)	0.746	0.647

Заключение

Предложено решение задачи оценки качества прогнозирования структуры белка с использованием графовых сверток. Проведен анализ графовых свёрток на данной задаче. Проведен анализ корреляций Пирсона и Спирмена предсказаний полученной модели и истинных значений качества структуры на данных CASP12. Качество, достигаемое моделью, сравнимо с качеством альтернативных моделей, дающих наилучшее качество в задаче. В дальнейших исследованиях предлагается в основе архитектуры сети использовать другие существующие улучшения спектральных свёрток (CayleyNet, Adaptive Graph Convolution Network). Также предлагается учитывать в данных дополнительные химические свойства атомов и в матрице смежности учитывать не только наличие связи, но и расстояния между атомами при наличии связи.

Список литературы

[1] Berg J.M., Tymoczko J.L., Stryer L. Biochemistry, Fifth Edition. — W.H. Freeman, 2002. — ISBN: 9780716730514. — URL: https://books.google.ru/books?id=uDFqAAAAMAAJ.

[2] GraphQA: Protein Model Quality Assessment using Graph Convolutional Network / Federico Baldassarre, David Menéndez Hurtado, Arne Elofsson, Hossein Azizpour. — 2019.

[3] Protein Structure Prediction Center. — http://predictioncenter.org/.

[4] Olechnovic Kliment, Kulberkytė Eleonora, Venclovas Ceslovas. CAD-score: a new contact area difference-based function for evaluation of protein structural models. // Proteins. — 2013. — Vol. 81 1. — P. 149–62.

[5] lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests / Valerio Mariani, Marco Biasini, Alessandro Barbato, Torsten Schwede // Bioinformatics. — 2013. — Vol. 29. — P. 2722 – 2728.

[6] LGA: A method for finding 3D similarities in protein structures.

[7] Hurtado David, Uziela Karolis, Elofsson Arne. Deep transfer learning in the assessment of the quality of protein models. — 2018. — 04.

[8] AngularQA: Protein Model Quality Assessment with LSTM Networks / Matthew Conover, Max Staples, Dong Si et al. // Computational and Mathematical Biophysics. — 2019. — 01. — Vol. 7. — P. 1–9.

[9] Deep convolutional networks for quality assessment of protein folds / Georgy Derevyanko, Sergei Grudinin, Y. Bengio, Guillaume Lamoureux // Bioinformatics (Oxford, England). — 2018. — 01. — Vol. 34.

[10] Pagès Guillaume, Charmettant Benoit, Grudinin Sergei. Protein model quality assessment using 3D oriented convolutional neural networks // Bioinformatics. — 2019. — 02. — Vol. 35, no. 18. — P. 3313–3319. — http://oup.prod.sis.lan/bioinformatics/article-pdf/35/18/3313/30024731/btz122.pdf.

[11] Relational inductive biases, deep learning, and graph networks / Peter W. Battaglia, Jessica B. Hamrick, Victor Bapst et al. // ArXiv. — 2018. — Vol. abs/1806.01261.

[12] A Comprehensive Survey on Graph Neural Networks / Zonghan Wu, Shirui Pan, Fengwen Chen et al. // CoRR. — 2019. — Vol. abs/1901.00596. — 1901.00596.

[13] Graph Neural Networks: A Review of Methods and Applications / Jie Zhou, Ganqu Cui, Zhengyan Zhang et al. // CoRR. — 2018. — Vol. abs/1812.08434. — 1812.08434.

[14] Spectral networks and locally connected networks on graphs / Joan Bruna, Wojciech Zaremba, Arthur Szlam, Yann Lecun // International Conference on Learning Representations (ICLR2014), CBLS, April 2014. — 2014.

[15] Chung F. R. K. Spectral Graph Theory. — American Mathematical Society, 1997.

[16] The Emerging Field of Signal Processing on Graphs: Extending High-Dimensional Data Analysis to Networks and Other Irregular Domains. / David I. Shuman, Sunil K. Narang, Pascal Frossard et al. // IEEE Signal Process. Mag. — 2013. — Vol. 30, no. 3. — P. 83–98.

[17] Mallat Stphane. A Wavelet Tour of Signal Processing, Third Edition: The Sparse Way. — 3rd edition. — USA : Academic Press, Inc., 2008. — ISBN: 0123743702.

[18] Defferrard Michaël, Bresson Xavier, Van gheynst Pierre. Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering // Advances in Neural Information Processing Systems 29 / Ed. by D. D. Lee, M. Sugiyama, U. V. Luxburg et al. — Curran Associates, Inc., 2016. — P. 3844–3852. — URL: http://papers.nips.cc/paper/6081-convolutional-neural-networks-on-graphs-with-fast-localized-spectral-filtering.pdf.

[19] Kipf Thomas N., Welling Max. Semi-Supervised Classification with Graph Convolutional Networks // arXiv:1609.02907 [cs, stat]. — 2017. — Feb. — arXiv: 1609.02907. URL: http://arxiv.org/abs/1609.02907 (online; accessed: 2019-12-10).

[20] Cangelosi Richard, Goriely Alain. Component retention in principal component analysis with application to cDNA microarray data // Biology direct. — 2007. — 02. — Vol. 2. — P. 2.

[21] An End-to-End Deep Learning Architecture for Graph Classification / Muhan Zhang, Zhicheng Cui, Marion Neumann, Yixin Chen. — 2018.

[22] R.Evans J.Jumper J.Kirkpatrick L.Sifre T.F.G.Green C.Qin A.Zidek A.Nelson A.Bridgland H.Penedones S.Petersen K.Simonyan S.Crossan D.T.Jones D.Silver K.Kavukcuoglu D.Hassabis A.W.Senior. De novo structure prediction with deep-learning based scoring // Thirteenth Critical Assessment of Techniques for Protein Structure Prediction (Abstracts) 1-4. — 2018. — Dec. — URL: https://deepmind.com/blog/article/alphafold.