ФГАОУВО «Московский физико-технический институт
(национальный исследовательский университет)»
Физтех-школа прикладной математики и информатики
Кафедра «Интеллектуальные системы»
при Вычислительном центре им. А. А. Дородницына РАН

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

1 Введение **4**

2 Постановка задачи **6**

2.1 CAD score 6

2.2 Задача регрессии CAD-score 7

2.3 Матрица смежности 8

3 Спектральный анализ **10**

3.1 Архитектура сети 13

4 Вычислительный эксперимент **14**

4.1 Собственное пространство матриц смежности 15

5 Результаты **17**

6 Заключение **19**

Список литературы **20**

Аннотация

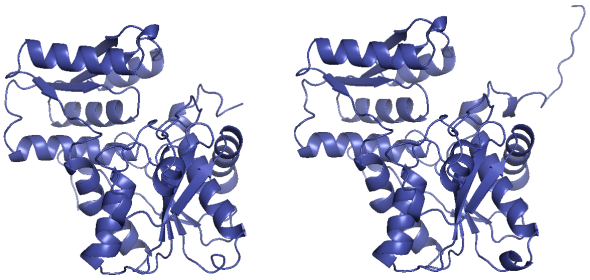
Решается задача оценки качества (QA – Quality Assessment) прогнозирования белковых структур. В работе показывается применимость к рассматриваемой задаче графовых свёрточных нейронных сетей, основанных на спектральной теории. Описание белковых структур представляется в виде графов. Спектральная теория графов определяет свёртку в нейронных сетях. Нейросеть в работе получает на вход матрицы координат атомов и матрицы смежности смоделированных белковых структур. Она предсказывает близость смоделированной и реальной структуры белка в виде $\mathrm{CAD}_{\mathrm{score}}$. Нейросеть обучается на наборах данных CASP7-CASP11 и тестируется на данных CASP12. На CASP12 достигается уровень ошибки MSE равный 0.051. Дополнительный анализ корреляционных коэффициентов Пирсона и Спирмена подтверждает применимость метода для различных белковых структур. Эксперименты в данной работе показывают новые направления в задаче QA.

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

1 Введение

Белки являются наиболее универсальными макромолекулами в живых системах и выполняют важнейшие функции практически во всех биологических процессах [1]. (?Понимание белковых структур и выполняемых задач помогают контролировать биологические процессы.?)Форма белковой структуры определяет выполняемые ей функции (?её функционал?) [1]. Но из имеющихся последовательностей аминокислот в белке трудно определить, в какую форму сворачивается структура. Идентификация структуры занимает большое количество времени и ресурсов, к тому же, не всегда возможна.

Каждые два года проводятся соревнования Critical Assessment of protein Structure Prediction (CASP [2]) по решению задачи предсказания структуры. Вычислительные методы, которые её решают состоят из двух этапов: генерация конформаций белка из их аминокислотных последовательностей и оценивание качества предсказания. В данной работе рассматривается только второй этап.



(a) True T0861 (b) Predicted Atome2_CBS_TS4

Рис. 1: Пример реальной и смоделированной структуры белка

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

Поэтому для задач с участием белковых структур модель должна учитывать как пространственную информацию об атомах, третичную структуру, так и признаки в виде последовательностей аминокислот, первичную структуру белка. В работах [3, 4] для моделирования белков используются LSTM или 1D-CNN, которые представляют белки в виде последовательности с пространственными признаками. В работах [5, 6] моделируется пространственная структура белков с использованием 3D-CNN, но не учитывается структура последовательностей. На основе графов моделируются как последовательности, так и геометрические структуры белков.

В работе [7] графовые нейронные сети на основе алгоритма, описанного в [8], показывают результаты, превосходящие остальные современные методы. Основные результаты в этой области полагаются на рточные нейронные сети (CNN) [6].

Поэтому предлагается использование графовых свёрточных нейронных сетей.

На рисунке 2 представлен общий ход работы.

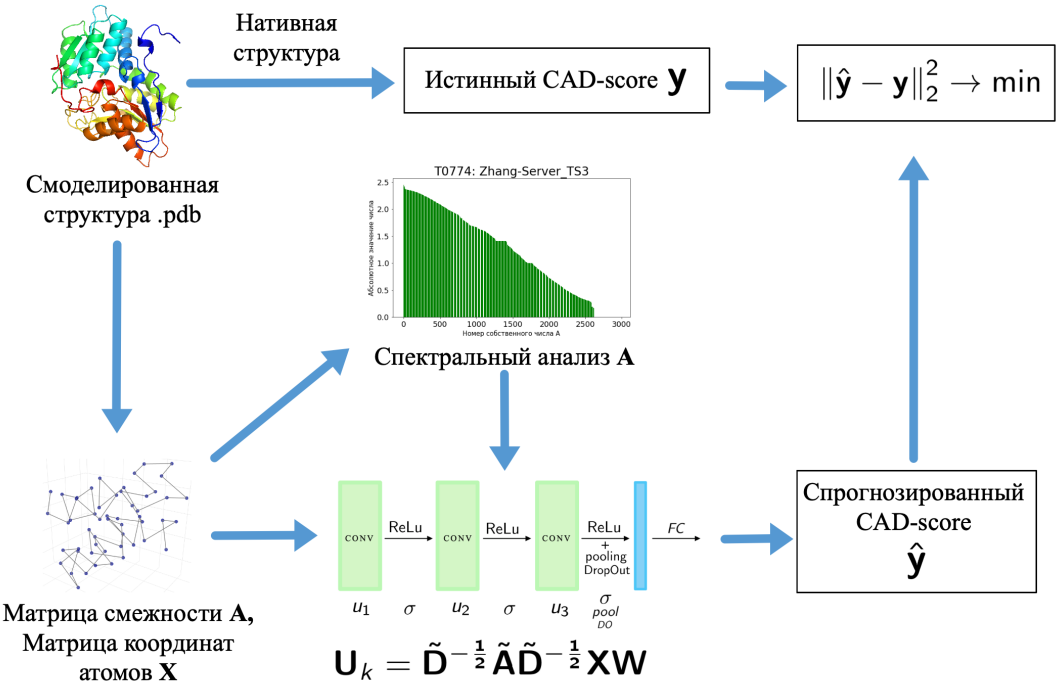


Рис. 2: Общая схема эксперимента

2 Постановка задачи

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 найтивной структуры белка. Оценка близости измеряется различными метриками: $\text{CAD}_{\text{score}}$ [9], LDDT [10], GDT [11]. В данной работе выбран $\text{CAD}_{\text{score}}$.

2.1 CAD score

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

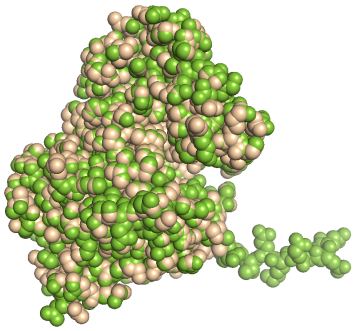


Рис. 3: Пересечение реальной и смоделированной структур

Для каждой пары остатков $(i, j) \in G$ определяется разность площадей контакта $\text{CAD}_{(i,j)}$ как абсолютная разница площадей контакта между остатками i и j в реальной T и смоделированной структуре M:

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

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

структуры определяется как

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

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

2.2 Задача регрессии CAD-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 $\mathrm{CAD}_{\mathrm{score}}$ y_i белка на основе его смоделированной пространственной структуры $\mathbf{x}_i$.

Параметры модели $\mathbf{w} \in \mathbb{W}$ подбираются в соответствии с минимизацией функции 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) [5–7]. 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 $\mathbf{y}_i$ и $\hat{\mathbf{y}}_i$ соответственно вектор истинных значений и вектор предсказаний $\mathrm{CAD}_{\mathrm{score}}$ для смоделированных структур белка, соответствующих нативной структуре i. Тогда коэффициенты корреляции записываются:

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

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

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

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

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

2.3 Матрица смежности

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

- не соединяются водород с водородом,
- атом не соединяется с водородом, если расстояние между ними $\geq$ 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 $r_{\min} = 0.01\text{Å}$, $r_{\max} = (0.6 \cdot (\rho_{\text{atom1}} + \rho_{\text{atom2}}))^2$, ρ_{atom} – радиус атома (максимально возможное $r_{\max} = 5.76$ – при $\rho_{\text{atom1}} = \rho_{\text{atom2}} = 2.0$).

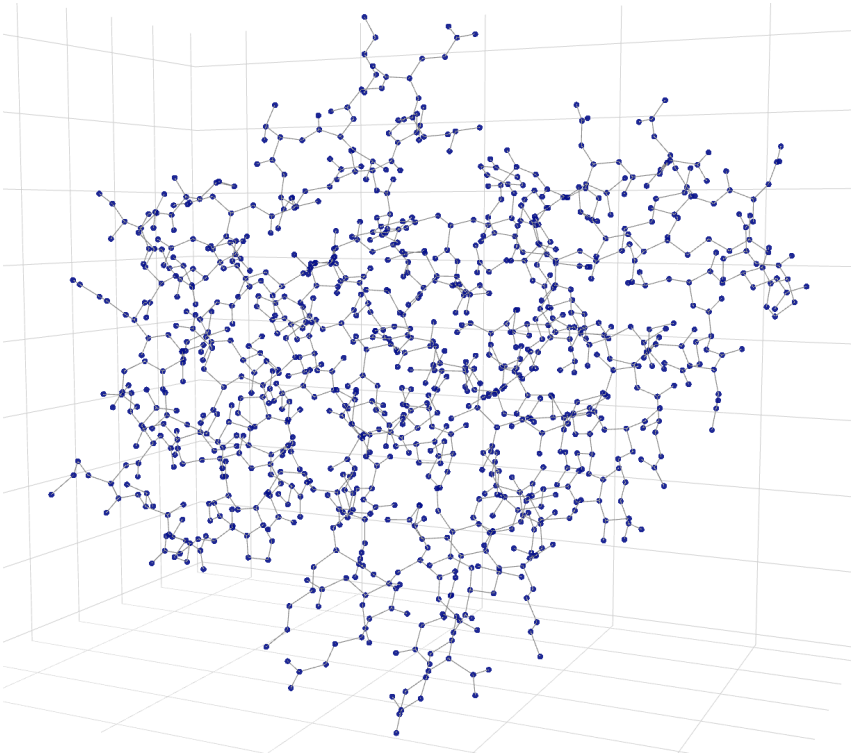


Рис. 4: Трехмерное представление с помощью координат $\mathbf{X}$ и полученной матрицы смежности $\mathbf{A}$ смоделированной структуры BAKER-ROSETTASERVER_TS3 для нативной структуры T0870 из набора данных CASP12

По попарным расстояниям между атомами на Рис. 5 видно, что могут иметь соединения 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.

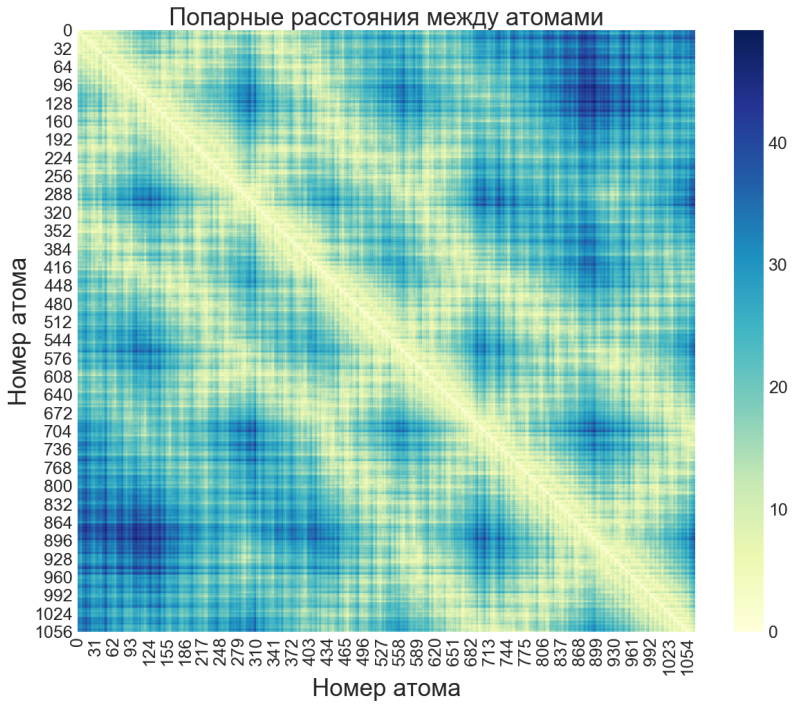


Рис. 5: Попарные расстояния между атомами смоделированной структуры BAKER-ROSETTASERVER_TS3 для нативной структуры T0870 из набора данных CASP12

3 Спектральный анализ

Для обобщения свёрточных нейронных сетей на графы необходимо определить свёрточные фильтры на графах. Существует два известных 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 обоснавание. 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 $\mathbf{A}_{ij}$ – сила 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 $\mathbf{A}$.

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 $\mathbf{U} = [\mathbf{u_0}, \mathbf{u_1}, \ldots, \mathbf{u_{n-1}}] \in \mathbb{R}^{n\times n}$ – 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 прebразование 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.*

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

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* $L^2\left([-1,1], \frac{dx}{\sqrt{1-x^2}}\right)$

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

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

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.

3.1 Архитектура сети

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

Сеть состоит из 3 свёрточных слоёв, макспуллинга по вершинам графа и нескольких полносвязных слоёв. Параметры свёрток t взяты равными 64, 64, 64 соответственно для первого, второго, третьего свёрточных

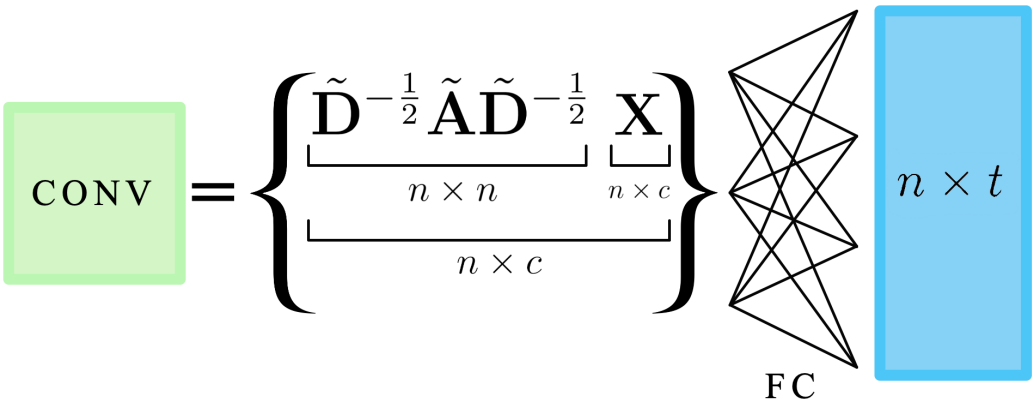


Рис. 6: Схема свёртки графа с матрицей $\mathbf{X}$ размера $n \times c$, t – число фильтров в свёртке, FC – полносвязный слой. Синий прямоугольник – выходная матрица размером $n \times t$

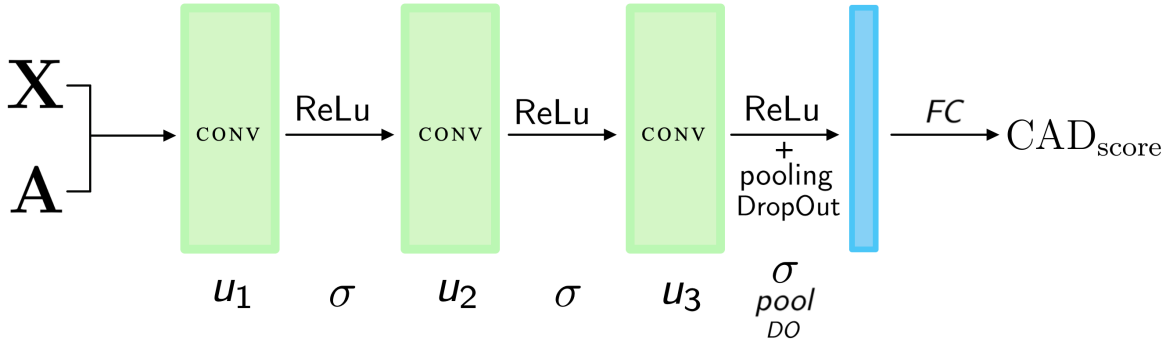


Рис. 7: Схематическое представление архитектуры свёрточной нейронной сети, использованной в данной работе

слоёв. На рисунке 7 представлена схема тестируемой в работе нейронной сети. σ – нелинейная функция активации

$$f = FC \circ DO \circ pool \circ \sigma \circ u_3 \circ \sigma \circ u_2 \circ \sigma \circ u_1$$

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

4 Вычислительный эксперимент

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

дели происходит на данных CASP9–CASP11 (xxx таргетов, $xxxxx$ моделей), тестирование – на CASP12(xxx таргетов, $xxxxx$ моделей). Для процессов обучения и тестирования по формуле (1) вычисляются $\mathrm{CAD}_{\mathrm{score}}$ для всех смоелированных структур на основе нативных структур.

Набор	Нативные структуры	Модели структур	Разбиение
CASP 9	117	35963	Train, Validation
CASP 10	103	15450	
CASP 11	84	12291	
CASP 12	37	5538	Test
Суммарно	559	129601	

Таблица 1: Наборы данных

4.1 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. 8 и 9 представлены собственные числа для смоделированной структуры STRINGS_TS3, соответствующей нативной T0759.

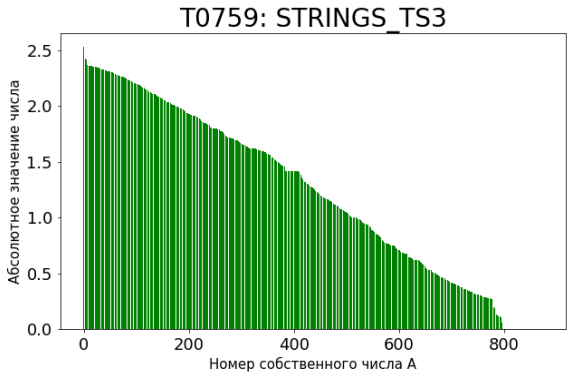


Рис. 8: Собственные числа $\mathbf{A}$

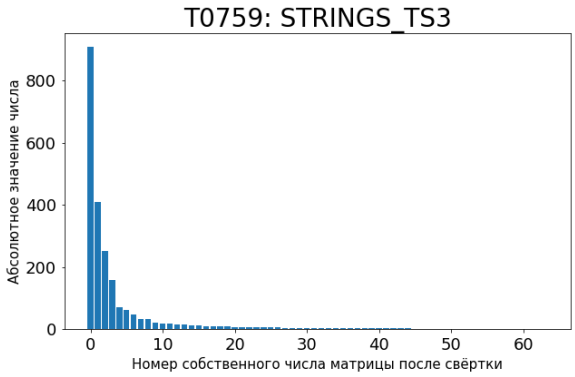


Рис. 9: Собственные числа $\mathbf{U}_k$

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

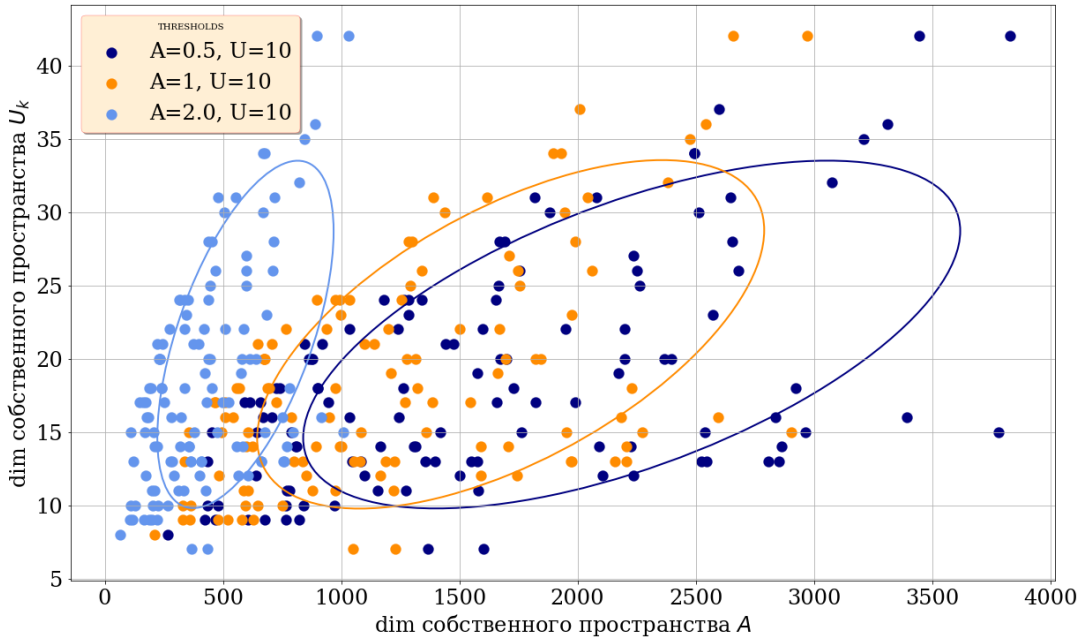


Рис. 10: Собственные пространства для порогов $U = 10$ и $A \in \{0.5, 1.0, 2.0\}$.

5 Результаты

При обучении нейросети анализируются усредненные по T нативным структурам коэффициенты корреляции Пирсона и Спирмена

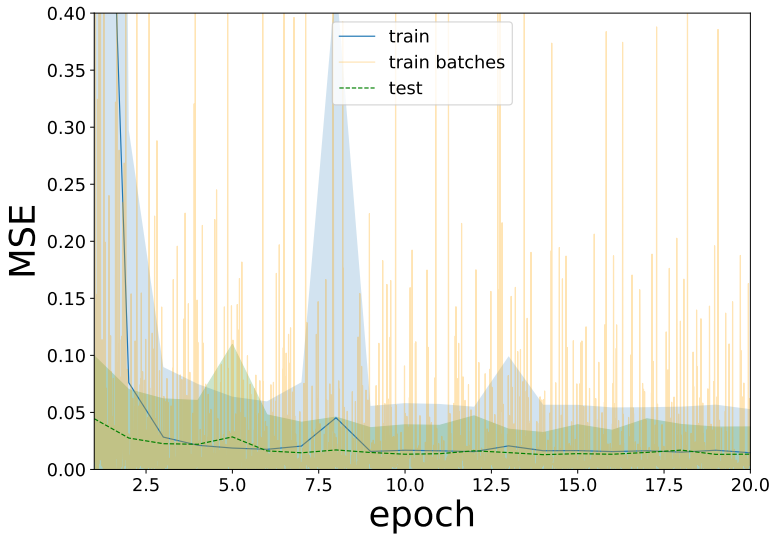


Рис. 11: График MSE ошибки GCN на обучающей и тестовой выборке

Графики корреляций Пирсона и Спирмена стабилизируются

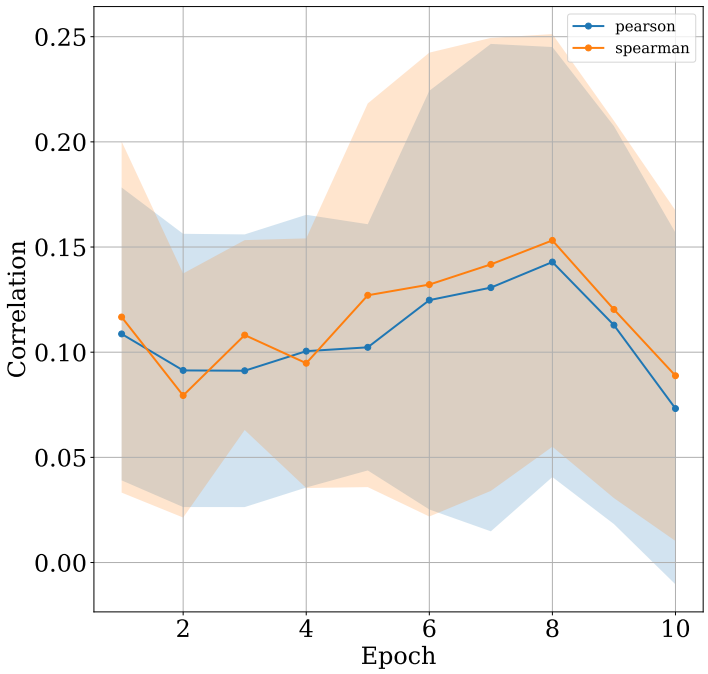


Рис. 12: Корреляция Пирсона, Кендалла, Спирмена

В таблице 2 представлены результаты тестирования модели на данных соревнования CASP12. Из сравнения данных в таблице видно, что модель из данной работы дает качество, сравнимое с качеством альтернативных моделей, дающих наилучшее качество в задаче.

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)	0.79	**0.83**

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

6 Заключение

Предложено решение задачи оценки качества прогнозирования структуры белка с использованием графовых сверток. Проведен анализ графовых свёрток на данной задаче. Полученная модель дает качество, сравнимое 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] Protein Structure Prediction Center. — http://predictioncenter.org/.

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

[4] 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.

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

[6] 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.

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

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

[9] 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.

[10] 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.

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

[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.