Оценка качества прогнозирования структуры белка с использованием графовых сверточных нейронных сетей.*

Северилов П.А.[1]
severilov.pa@phystech.edu
[1]Московский физико-технический институт (МФТИ)

Решается задача оценки качества прогнозирования белковых структур. Спектральная теория графов определяет свертку в нейронных сетях при работе с данными в виде графов. Описание белковых структур представлено в виде графов. В работе рассматриваются результаты применения графовых сверточных нейронных сетей к задаче оценивания предсказания белковой структуры.

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

1 Введение

Белки являются наиболее универсальными макромолекулами в живых системах и выполняют важнейшие функции практически во всех биологических процессах [3]. (?Понимание белковых структур и выполняемых задач помогают контроллировать биологические процессы.?) Форма белковой структуры определяет (дикутет) её функционал [3]. 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) по решению задачи предсказания структуры. Вычислительные методы, которые решают её состоят из двух этапов: генерация конформаций белка из их аминокислотных последовательностей и оценивание качества предсказания. В данной работе рассматривается только второй этап.

Белковая структура состоит из одной или нескольких цепочек более мелких молекул – аминокислотных остатков. Последовательность остатков $\mathrm{S} = \{a_i\}_{i=1}^{N}$ 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, т.е. третичную структуру, так и признаки в виде последовательностей, т.е. первичную структуру белка. В работах [6, 9] для моделирования белков используются LSTM или 1D-CNN, которые представляют белки в виде последовательности с пространственными признаками. В работах [8, 14] 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 [1] графовые нейронные сети на основе алгоритма, описанного в [2], показывают результаты, превосходящие остальные современные методы.

*Научный руководитель: В.В. Стрижов

Основные результаты в этой области полагаются на сверточные нейронные сети (CNN) [14]. Поэтому предлагается использование графовых сверточных нейронных сетей.

2 Постановка задачи

2.1 Задача регрессии

Пусть $\mathfrak{D} = (\mathbf{X}, \mathbf{y})$ — заданная выборка, где $\mathbf{X} \in \mathbb{R}^{m \times n \times 3}$ — тензор объект-признак, объекты $\mathbf{x}_i \in \mathbb{R}^{1 \times n_i \times 3}, i = \overline{1, m}$ — это молекулы, каждая из которых описана множеством 3-мерных координат всех ее атомов, а $\mathbf{y} = [y_1, \ldots, y_m]^\mathsf{T} \in \mathbb{R}^{m \times 1}$ — оценка близости предсказанной и реальной структуры белка. Оценка близости измеряется различными метриками: CAD-score [12], LDDT, GDT. В данной работе выбран CAD-score.

Рассмотривается множество параметрических моделей $\mathfrak{F}$, взятых из класса графовых сверточных 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, а $\hat{\mathbf{y}} \in \mathbb{R}^{m \times 1}$ - вектор оценок предсказаний (CAD-scores).

Рассматривается задача регрессии для предсказания численного значения CAD-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 $\mathfrak{L}(\mathbf{y}, \mathbf{X}, \mathbf{w}) = (\hat{\mathbf{y}} - \mathbf{y})^2$, где $\hat{\mathbf{y}} = \mathbf{f}(\mathbf{X}, \mathbf{w})$ – CAD-score, предсказанный моделью $\mathbf{f}$, $\mathbf{y}$ – данный в выборке CAD-score.

2.2 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.

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

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

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

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

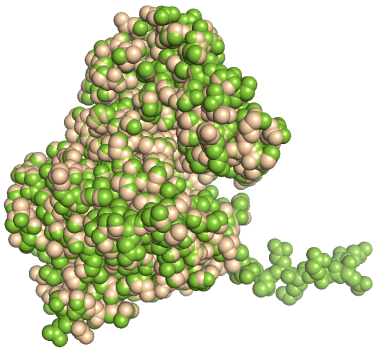


Рис. 1 Пересечение реальной структуры T0861 (жёлтый) и её модели Atome2_CBS_TS4 (зелёный) при $\mathrm{CAD}_{\mathrm{score}} = 0.829$

3 Теоретическая часть (?Спектральный анализ?)

Для обобщения сверточных нейронных сетей на графы необходимо определить сверточные фильтры на графах. Существует два известных подхода: пространственный и спектральный [17, 19]. Как показано в [4] 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.

3.1 Спектральный анализ

Определение 1. *Графовый Лапласиан [5] – 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 $\mathbf{L} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$, 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* $\mathscr{F}(\mathbf{x}) = \mathbf{U}^T\mathbf{x} \equiv \hat{\mathbf{x}} \in \mathbb{R}^n$, *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 свертки. 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 свертках)*** *[11] Преобразование Фурье свертки 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, спектральная свертка на графах определяется для сигнала $\mathbf{x}$ и фильтра $\mathbf{g} \in \mathbb{R}^n$ как

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

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

Спектральные методы отличаются выбором фильтра $\mathbf{g}_\theta$. Соотношение 1 вычислительно дорогое, т.к. спектральное разложение требует $O(n^3)$ операций, а перемножение с матрицей собственных векторов $\mathbf{U}$ требует $O(n^2)$ операций. Chebyshev Spectral CNN (ChebNet) [7] 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 порядкка 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 $\mathbf{g}_\theta = \sum_{k=0}^{K-1}\theta_k\mathbf{T}_k(\tilde{\mathbf{\Lambda}})$, 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 $\left(\mathbf{U}\mathbf{\Lambda}\mathbf{U}^T\right)^k = \mathbf{U}\mathbf{\Lambda}^k\mathbf{U}^\top$ (собственные векторы образуют ортонормированный базис $\mathbf{U}^T\mathbf{U} = \mathbf{I}$), получаем:

$$\mathbf{U}\mathbf{g}_\theta\mathbf{U}^T\mathbf{x} = \mathbf{U}\left(\sum_{i=0}^{K-1}\theta_k\mathbf{T}_k(\tilde{\mathbf{\Lambda}})\right)\mathbf{U}^T\mathbf{x} = \sum_{k=0}^{K-1}\theta_k\mathbf{T}_k(\tilde{\mathbf{L}})\mathbf{x}, \tag{2}$$

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

Graph Convolutional Network (GCN) [10] используют первое приближение ChebNet. Предполагая $\lambda_{\max} \approx 2$ и $K = 1$, соотношение 2 упрощается до

$$\mathbf{x} * \mathbf{g} \approx \tilde{\theta}_0 \mathbf{x} + \tilde{\theta}_1 (\mathbf{L} - \mathbf{I}_N) \mathbf{x} = \tilde{\theta}_0 \mathbf{x} - \tilde{\theta}_1 \mathbf{D}^{-\frac{1}{2}} \mathbf{A} \mathbf{D}^{-\frac{1}{2}} \mathbf{x} \tag{3}$$

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

$$\mathbf{x} * \mathbf{g} \approx \theta \left(\mathbf{I}_N + \mathbf{D}^{-\frac{1}{2}} \mathbf{A} \mathbf{D}^{-\frac{1}{2}} \right) \mathbf{x} \tag{4}$$

Оператор в скобках может может привести к вычислительной нестабильности и взрыву/затуханию градиентов, т.к. собственные значения данного оператора $\in [0, 2]$. Для решения проблемы в [10] предлагается *трюк перенормировки*:

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

3.2 Сверточный слой

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 4 и применяя трюк перенормировки, можно определить слой свертки графа таким образом:

$$\mathbf{Z} = \sigma \left(\tilde{\mathbf{D}}^{-1} \tilde{\mathbf{A}} \tilde{\mathbf{D}} \mathbf{X} \mathbf{W} \right),$$

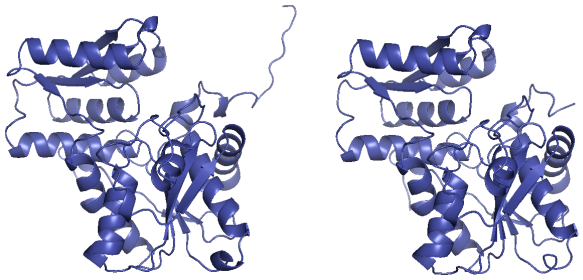
где $\mathbf{W} \in \mathbb{R}^{c \times t}$ – матрица параметров свертки с t фильтрами, σ – нелинейная функция активации, а $\mathbf{Z} \in \mathbb{R}^{n \times t}$ – выходная матрица.

4 Данные

Берутся с соревнований CASP. Для реальной структуры белка берется еще смоделированная структура. Для них вычисляется CAD-score. Модель на тесте предсказывает CAD-score для смоделированной структуры, не имея возможности напрямую вычислить CAD-score по реальной структуре.

Датасет	Таргеты	Модели	
CASP 7	95	19591	
CASP 8	122	34789	Train/Val
CASP 9	117	34946	
CASP 10	103	26254	
CASP 11	83	16094	
CASP 12	40	6924	Test

Таблица 1 Датасеты



(a) Predicted Atome2_-CBS_TS4 (b) True

Рис. 2 T0861

4.1 Представление белков в виде 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 $\mathbf{A}$.

4.2 Матрица смежности

Т.к. данные о белках не содержат информации о соединениях между атомами, т.е. нет матрицы смежности, построены соединения по следующим правилам:

- не соединяются водород с водородом
- атом не соединяется с водородом, если расстояние между ними $\geqslant 1.21$ Å
- 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, $r_{max} = (0.6 \cdot (\rho_{atom1} + \rho_{atom2}))^2$, ρ_{atom} – радиус атома (максимально возможное $r_{max} = = 5.76$ – при $\rho_{atom1} = \rho_{atom2} = 2.0$)

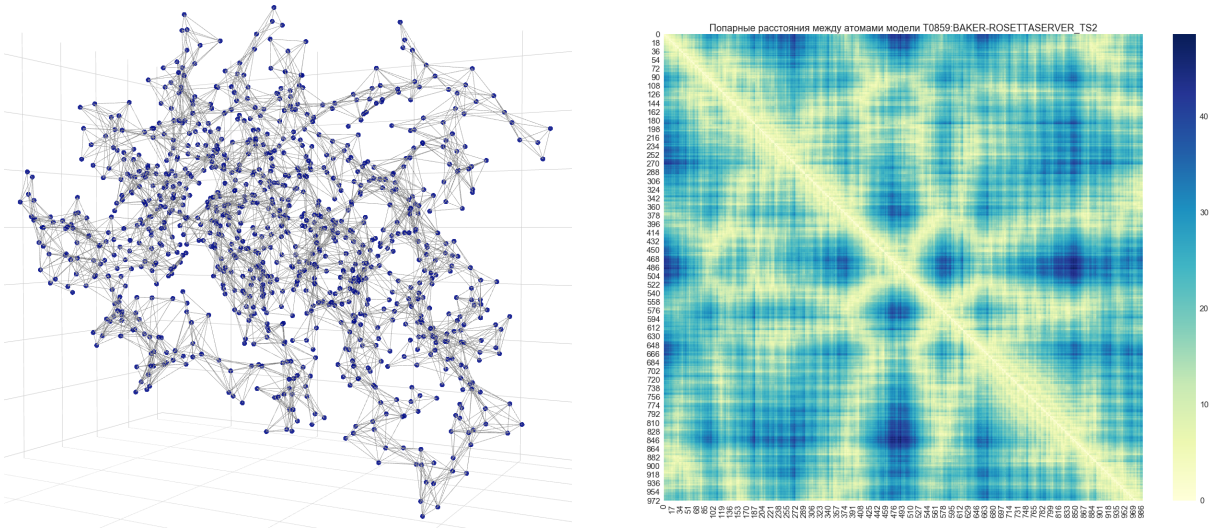


Рис. 3 3D представление с помощью координат и полученной матрицы смежности и попарные расстояния между атомами модели модели T0859 BAKER-ROSETTASERVER_TS2 (CASP12)

По попарным расстояниям между атомами на Рис. 3 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 – 5.76 . Т.е. матрица смежности будет сильно разреженной

5 Вычислительный эксперимент

Варианты архитектур для классификации, основанных на спектральной теории.

1. ChebNet [7]
2. GCN [10]

3. CayleyNet (???)
4. Adaptive Graph Convolution Network (AGCN) (???)
5. GRAPH WAVELET NEURAL NETWORK (???)

Основа архитектуры берется из существующей модели, где изменяется выходной слой для работы с задачей регрессии.

5.1 SpectralQA

На основе модели GCN

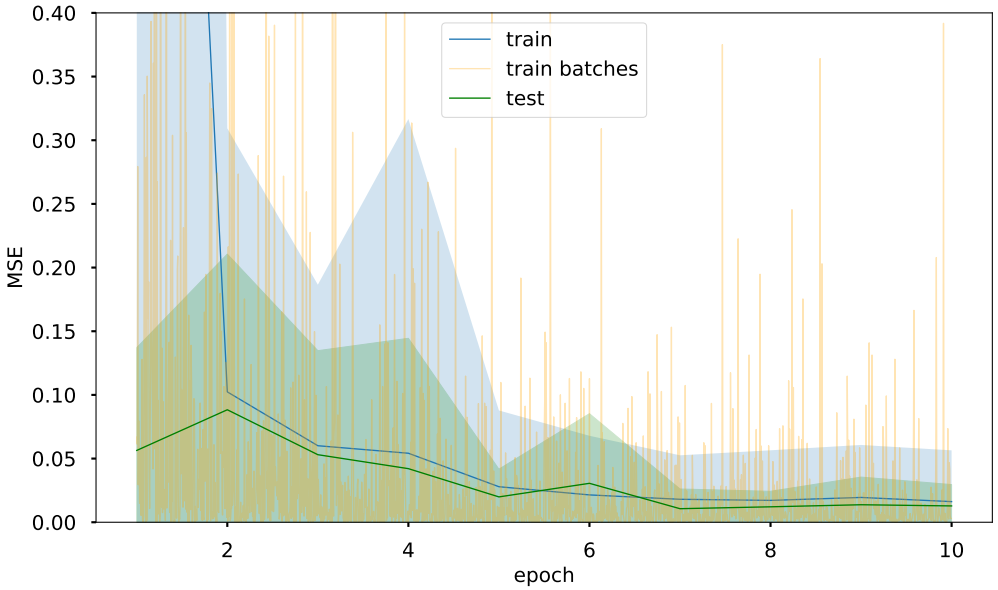


Рис. 4 График MSE ошибки GCN на обучающей и тестовой выборке

Часть данных	Количество молекул	MSE
Train	190	
Validation	31	
Test		

6 Результаты и выводы

Сравнение с существующими методами QA:

	ρ	r	z-score
ProQ3D	0.801	11.961	1.670
VoroMQA	0.803	17.171	1.410
SBROD	0.685	23.579	1.282
Ornate	0.828	0.781	1.780
SpectralQA (МОЯ)			

Таблица 2 Сравнение корреляции Пирсона, Спирмена и z-score существующих современных алгоритмов с моделью SpectralQA на данных CASP12

7 Заключение

Впервые для задачи оценки качества прогнозирования структуры белка применены графовые сверточные нейронные сети, в которых свертки определены на основе спектральной теории. В качестве улучшения, можно в основе архитектуры сети пробовать другие существующие улучшения спектральных сверток: CayleyNet, Adaptive Graph Convolution Network (AGCN), GRAPH WAVELET NEURAL NETWORK. Также в будущей работе предлагается учитывать в данных дополнительные свойства атомов и в матрице смежности учитывать не только наличие связи, но и расстояния между атомами при наличии связи.

Литература

[1] Federico Baldassarre, David Menéndez Hurtado, Arne Elofsson, and Hossein Azizpour. Graphqa: Protein model quality assessment using graph convolutional network. 2019.

[2] Peter W. Battaglia, Jessica B. Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinícius Flores Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, Çaglar Gülçehre, H. Francis Song, Andrew J. Ballard, Justin Gilmer, George E. Dahl, Ashish Vaswani, Kelsey R. Allen, Charles Nash, Victoria Langston, Chris Dyer, Nicolas Manfred Otto Heess, Daan Wierstra, Pushmeet Kohli, Matthew M Botvinick, Oriol Vinyals, Yujia Li, and Razvan Pascanu. Relational inductive biases, deep learning, and graph networks. *ArXiv*, abs/1806.01261, 2018.

[3] J.M. Berg, J.L. Tymoczko, and L. Stryer. *Biochemistry, Fifth Edition.* W.H. Freeman, 2002.

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

[5] F. R. K. Chung. *Spectral Graph Theory.* American Mathematical Society, 1997.

[6] Matthew Conover, Max Staples, Dong Si, Miao Sun, and Renzhi Cao. Angularqa: Protein model quality assessment with lstm networks. *Computational and Mathematical Biophysics*, 7:1–9, 01 2019.

[7] Michaël Defferrard, Xavier Bresson, and Pierre Van gheynst. Convolutional neural networks on graphs with fast localized spectral filtering. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 3844–3852. Curran Associates, Inc., 2016.

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

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

[10] Thomas N. Kipf and Max Welling. Semi-Supervised Classification with Graph Convolutional Networks. *arXiv:1609.02907 [cs, stat]*, February 2017. arXiv: 1609.02907.

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

[12] Kliment Olechnovic, Eleonora Kulberkytė, and Ceslovas Venclovas. Cad-score: a new contact area difference-based function for evaluation of protein structural models. *Proteins*, 81 1:149–62, 2013.

[13] Angelo Oliveira and Renato José Sassi. Behavioral malware detection using deep graph convolutional neural networks, Nov 2019.

[14] Guillaume Pagès, Benoit Charmettant, and Sergei Grudinin. Protein model quality assessment using 3D oriented convolutional neural networks. *Bioinformatics*, 35(18):3313–3319, 02 2019.

[15] 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 R.Evans, J.Jumper. De novo structure prediction with deep-learning based scoring. *Thirteenth Critical Assessment of Techniques for Protein Structure Prediction (Abstracts) 1-4*, Dec 2018.

[16] David I. Shuman, Sunil K. Narang, Pascal Frossard, Antonio Ortega, and Pierre Vandergheynst. The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. *IEEE Signal Process. Mag.*, 30(3):83–98, 2013.

[17] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S. Yu. A comprehensive survey on graph neural networks. *CoRR*, abs/1901.00596, 2019.

[18] Muhan Zhang, Zhicheng Cui, Marion Neumann, and Yixin Chen. An end-to-end deep learning architecture for graph classification. 2018.

[19] Jie Zhou, Ganqu Cui, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, and Maosong Sun. Graph neural networks: A review of methods and applications. *CoRR*, abs/1812.08434, 2018.