Мультимасштабные графовые нейронные сети на SES-поверхности для предсказания сайтов связывания белков

В. С. Половников*

Представлен SES-GNN — мультимасштабная графовая нейронная сеть на поверхности, исключающей растворитель (SES), для предсказания сайтов связывания белков. Вместо однородного графа с фиксированным радиусом рёбер строится трёхмасштабное представление (0–2, 2–4, 4–8 Å) с отбором центроидов через Farthest Point Sampling. Архитектура — 12 слоёв TransformerConv. Мультимасштабный граф даёт прирост +0.015–0.017 ROC-AUC относительно однородного. Метод достигает ROC-AUC = 0.838 на собственном PPI-датасете. Покомпонентный анализ показывает, что признаки вершины несут основную информацию, но реализуются только через графовую структуру: без графа поточечный классификатор на тех же признаках существенно уступает, а топология сама по себе бесполезна.

Ключевые слова: предсказание сайтов связывания белков, SES-поверхность, графовые нейронные сети.

1. Введение

Белок-белковые взаимодействия (PPI) лежат в основе клеточной сигнализации, регуляции и сборки макромолекулярных комплексов, поэтому предсказание их сайтов — одна из центральных задач структурной биоинформатики. Задача: для каждой точки i поверхности белка предсказать бинарную метку $y_i \in \{0, 1\}$ — принадлежит ли точка сайту связывания. Комплементарность впадин и выступов позволяет заряженным и гидрофобным группам сблизиться, поэтому для предсказания важна *геометрия* поверхности, а не только её химия. Геометрическое глубокое обучение на молекулярных поверхностях показало эффективность [1].

* *Половников Владимир Сергеевич* — кандидат физико-математических наук, старший научный сотрудник кафедры математической теории интеллектуальных систем мех.-мат. факультета МГУ имени М.В. Ломоносова, e-mail: polovnikov@intsys.msu.ru, ORCID: 0000-0001-9784-9067.

Polovnikov Vladimir Sergeevich — PhD, Senior Researcher, Department of Mathematical Theory of Intelligent Systems, Faculty of Mechanics and Mathematics, Moscow State University.

Опубликовано на условиях лицензии CC-BY 4.0
Published under CC-BY 4.0

Сайт связывания имеет ядро и периферию и требует учёта мелкой геометрии (размер остатка, ∼5 Å) и широкого контекста (форма впадины, ∼20 Å). Используются два представления поверхности. *SAS* — поверхность, которую описывает центр пробной сферы радиуса 1.4 Å при качении по ван-дер-ваальсовой поверхности белка; $\Delta SASA$ — падение её площади при связывании (основа ASA-определений интерфейса). *SES* — поверхность, образуемая обращённой к белку частью пробной сферы при её качении; состоит из контактных, тороидальных и возвратных участков и точнее описывает геометрию впадин.

2. Подход SES-GNN

SES сэмплируется точками с плотностью ~ 1 точка/Å^2 (**10 000–20 000 точек** на цепь, шаг ~ 1 Å — меньше размера остатка); точки выбираются равномерно по площади из аналитически построенной SES. Однородный граф с фиксированным радиусом не может одновременно захватить локальную геометрию и широкий контекст. Мы строим рёбра на трёх масштабах: **локальный** (0–2 Å, все рёбра), **средний** (2–4 Å, до 10 центроидов), **глобальный** (4–8 Å, до 10 центроидов).

Выборка по наиболее удалённым точкам (FPS). Для вершины с соседями в диапазоне $[r_1, r_2]$ выбирается k центроидов: первый — случайно, каждый следующий — максимально удалён от выбранных, что даёт равномерное покрытие окрестности по направлениям. Кандидатов в среднем ∼77 на среднем масштабе и ∼374 на глобальном; FPS ограничен 10 центроидами на масштаб, но не более 32 рёбер на вершину: локальные рёбра (0–2 Å) занимают большую часть бюджета, поэтому выбирается ∼5 центроидов на среднем масштабе и ∼2 на глобальном. Итоговый граф содержит в среднем **27 рёбер на вершину** (∼20 локальных, ∼7 FPS).

Признаки. Вектор вершины $\mathbf{x}_i \in \mathbb{R}^{14}$: $\mathbf{x}_i = [\mathbf{n}, \mathbf{t}, \mathbf{g}, \mathbf{c}]$, где $\mathbf{n} = (n_x, n_y, n_z)$ — нормаль; $\mathbf{t} \in \{0,1\}^3$ — one-hot типа SES-участка; $\mathbf{g} \in \{0,1\}^6$ — multihot элементов поддерживающих атомов; $\mathbf{c} \in \mathbb{R}^2$ — полярность и гидрофобность по Kyte–Doolittle, агрегированные по поддерживающим атомам. Вектор ребра $\mathbf{r}_{ij} \in \mathbb{R}^3$: расстояние $\|p_i - p_j\|$, косинус угла между нормалями $\langle n_i, n_j \rangle$ и штрафующий вес $e^{-\|p_i - p_j\|^2/(2\sigma^2)}(1 + \langle n_i, n_j \rangle)/2$, где $\sigma = 3.0$ Å.

Архитектура. 12 слоёв TransformerConv, hidden = 32 суммарно (по 8 на голову, конкатенация), heads = 4, edge_dim = 3, residual connections, LayerNorm, GELU, dropout 0.1. Глобальные рёбра разрежены (∼2 на вершину), поэтому охват поверхности требует глубины: 12 слоёв оптимальны, 16 не улучшают (0.834 против 0.838). Обучение: AdamW (lr =

10^{-3}, weight decay 10^{-4}), cosine annealing, 150 эпох. Функция потерь — BCEWithLogitsLoss с hard negative mining: в каждом графе выбираются до двух негативов на позитив — с наибольшим текущим логитом; потеря считается только на этой подвыборке. Гиперпараметры одинаковы для всех конфигураций; основные результаты получены без поворотной аугментации, а отдельный эксперимент с поворотами даёт +0.009 val-ROC.

3. Результаты

Собственный PPI-набор: подмножество dMaSIF [1] (training_ppi.txt), сплит 5000/50/412 (train/val/test) комплексов. Определение интерфейса: для пары цепей (A,B) точка $p \in \mathrm{SES}(A)$ интерфейсна относительно цепи B, если $\min_{q \in \mathrm{SES}(B)} \|p - q\| \leq 4$ Å. Пары (A,B) и (B,A) — разные примеры.

Таблица 1: Зависимость ROC-AUC и F1 на тестовом множестве от глубины сети и конфигурации графа.

	Однородный, $R = 3$ Å	Мультимасштабный		
Слоёв	ROC-AUC	ROC-AUC	Δ ROC-AUC	Test-F1
6	0.789	0.806	+0.017	0.531
8	0.806	0.823	+0.017	0.547
10	0.819	0.834	+0.015	0.530
12	—	**0.838**	—	**0.612**
16	—	0.834	—	0.603

F1 — среднее по белкам при пороге, подобранном на валидации.

Мультимасштабный граф даёт **+0.015–0.017 ROC-AUC** над однородным. Покомпонентный анализ (500 train / 50 val, 4 слоя) показывает: основную информацию несут признаки вершин, но реализуются они только в совокупности с графовой структурой. Без графа поточечная логистическая регрессия на тех же 14D-признаках даёт ROC-AUC 0.556 против 0.711 у полной модели; чистая топология без признаков — 0.500, уровень случайного угадывания; вес рёбер добавляет +0.017. Иными словами, топология сама по себе бесполезна, но служит механизмом распространения признаков по поверхности.

4. Заключение

SES-GNN — мультимасштабная графовая нейронная сеть для предсказания сайтов связывания на SES-поверхности. При одинаковой глубине

10 слоёв мультимасштабный граф даёт +0.015 ROC-AUC относительно однородного (0.834 vs 0.819); 12-слойная конфигурация достигает **0.838** и **F1 = 0.612**. Покомпонентный анализ показывает, что SES-GNN использует признаки вершин и графовую структуру совместно: топология без признаков бесполезна, а признаки без графа уступают полной модели.

Список литературы

[1] F. Sverrisson, J. Feydy, B. E. Correia, M. M. Bronstein, "Fast end-to-end learning on protein surfaces", *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, 15267–15276. DOI: 10.1109/CVPR46437.2021.01502 .

Статья поступила 10 октября 2026 г.

Multiscale Graph Neural Networks on the SES Surface for Protein Binding Site Prediction

V. S. Polovnikov

We present SES-GNN, a multiscale graph neural network on the solvent-excluded surface (SES) for protein binding site prediction. Instead of a uniform graph with fixed edge radius, a three-scale representation (0–2, 2–4, 4–8 Å) with Farthest Point Sampling centroid selection is built. The architecture uses 12 TransformerConv layers. The multiscale graph gives a +0.015–0.017 ROC-AUC gain over the uniform one. The method achieves ROC-AUC = 0.838 on our own PPI dataset. Ablation shows that vertex features carry the main information but realize it only through the graph structure: a pointwise classifier without message passing falls substantially below the full model, while topology alone is useless.

Keywords: protein binding site prediction, SES surface, graph neural networks.

References

[1] F. Sverrisson, J. Feydy, B. E. Correia, M. M. Bronstein, "Fast end-to-end learning on protein surfaces", *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, 15267–15276. DOI: 10.1109/CVPR46437.2021.01502.

Received on October 10, 2026