

RAG-GNN: графовые нейронные сети с дополненным поиском для эмбедингов сетей белковых взаимодействий

Источник: Frontiers in AI — Medicine

Оригинал: <https://www.frontiersin.org/articles/10.3389/frai.2026.1851917>

RAG

белковые взаимодействия

биоинформатика

графовые нейронные сети

машинное обучение

Мы представляем **RAG-GNN** — сквозную (end-to-end) обучаемую архитектуру, которая дополняет энкодер на основе графовых нейронных сетей (**GNN**, Graph Neural Network) для сетей белковых взаимодействий совместно оптимизированным модулем плотного поиска по корпусу документов, индексированных с помощью **TF-IDF** (term frequency-inverse document frequency), механизмом управляемого слияния (gated fusion mechanism) и контрастивным выравниванием представлений узлов и документов.

Данное исследование позиционируется как контролируемое методологическое изучение того, обеспечивает ли дополнение поиском (retrieval augmentation) измеримую выгоду по сравнению с сопоставимым вариантом, использующим только GNN (ablation study), а не как инструмент прецизионной медицины или поиска терапевтических мишеней.

В рамках тематического исследования сигнальных путей рака (379 белков, 3 498 взаимодействий, 14 категорий путей) **RAG-GNN** улучшает силуэтный коэффициент функциональной кластеризации с $-0,237 \pm 0,065$ (только GNN) до $-0,144 \pm 0,066$ ($+0,093 \pm 0,022$ по 10 итерациям; **ARI**, Adjusted Rand Index

+0,021 ± 0,015), в то время как выученная проекция поиска достигает среднего значения точности при 10 (**precision at 10**) = 0,242, что на 152% выше относительно случайного базового уровня (0,096).

Контрфактические эксперименты подтверждают, что случайные и отсутствующие контексты поиска снижают производительность, однако контрольная группа с перемешанными документами (реальные документы, переназначенные неверным белкам) показывает результаты, сопоставимые с правильным поиском; это указывает на то, что прирост отражает общий биологический сигнал в совокупном корпусе **TF-IDF**, а не специфическое семантическое соответствие узлам.

Эвристическая декомпозиция информации с доверительными интервалами методом бутстрепа показывает, что топология и поиск кодируют подавляющее большинство общей информации (95,6%), при этом поиск вносит вклад преимущественно за счет реорганизации этого общего сигнала.

Бенчмаркинг по сравнению с восемью методами эмбединга выявляет специфическую для задач взаимодополняемость: методы, ориентированные на топологию, более эффективны для предсказания связей (link prediction), тогда как дополнение поиском улучшает функциональную кластеризацию в рамках контролируемого исследования абляции.

Анализ подсети **DDR1** представлен как подтверждающее восстановление биологических данных, уже присутствующих в исходном корпусе, а не как проспективное открытие.