Анализ графовых данных и глубокое обучение

Азимов Рустам

План

Познакомиться с методами машинного и глубокого обучения для работы с графовыми данными:

- **Построение эмбеддингов для вершин**: DeepWalk, Node2Vec
- **Графовые нейронные сети (GNN)**: GCN, GraphSAGE, GAT…

Инструменты

- PyG (PyTorch Geometric)
  - Самая популярная библиотека для GNN
- GraphGym - платформа для проектирования GNN
  - Реализовано множество модулей GNN
  - Упрощенный подбор гиперпараметров
  - Гибкая кастомизация
- NetworkX - полезная библиотека для различных манипуляция над графами и сетевого анализа

Полезные ссылки

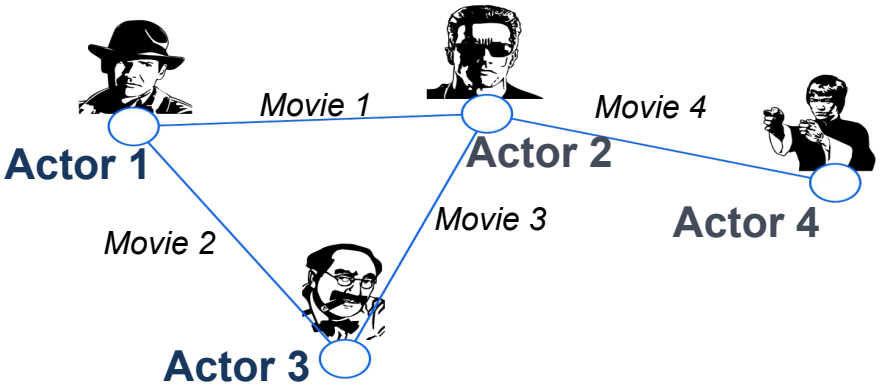
- Graph Representation Learning Book - Will Hamilton
- CS224W: Machine Learning with Graphs

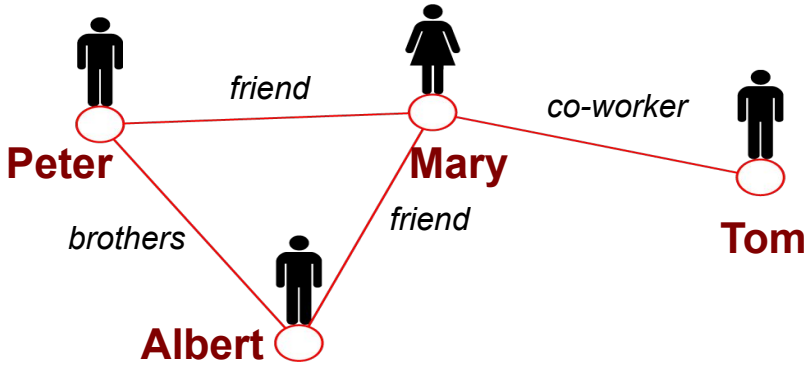
Машинное обучение для анализа графов

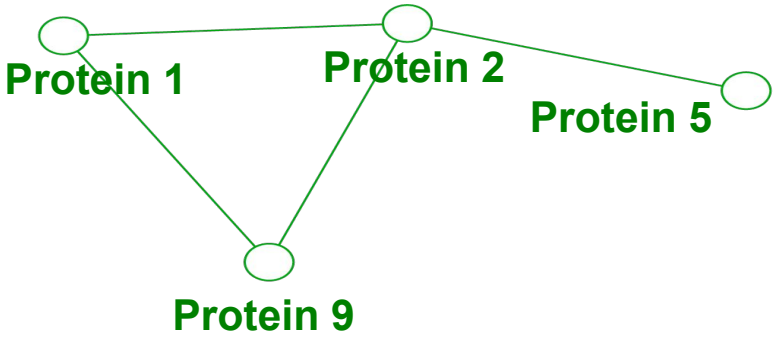
Почему графы?

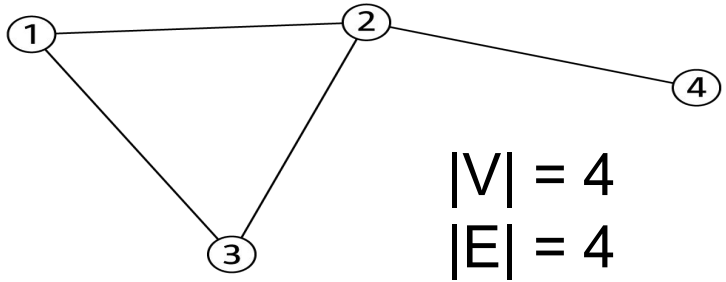
- Графы - это универсальный язык для описания и анализа сущностей с отношениями/взаимодействиями
- Данные во многих областях естественным образом представляются в виде графов

Графы - универсальный язык







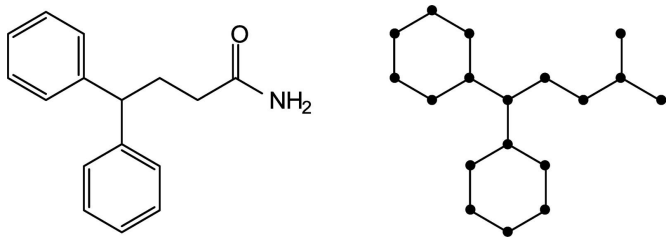


$|V| = 4$

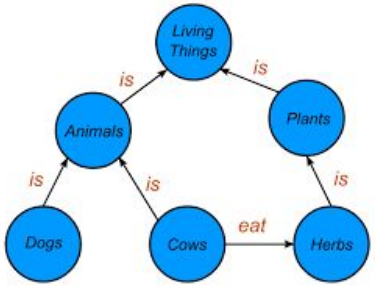
$|E| = 4$

Графы в реальной жизни

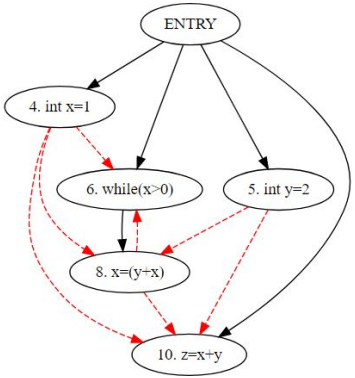
Социальные сети



Молекулы



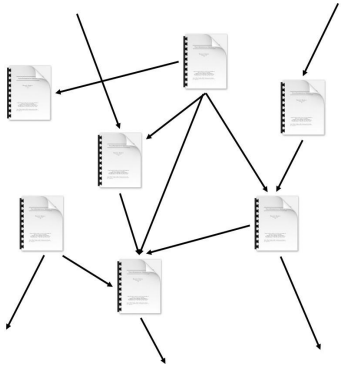
Графы знаний



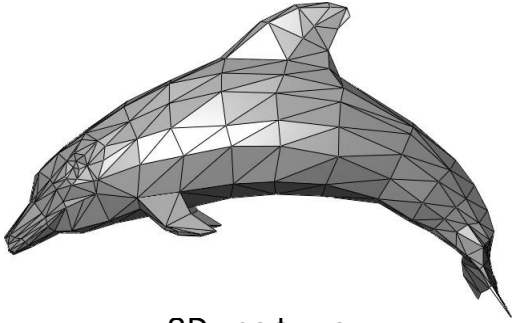
Программы

Графы в реальной жизни

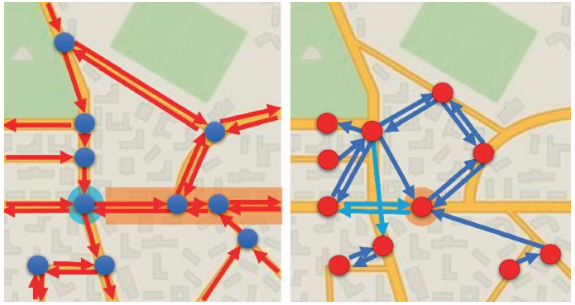
Компьютерные сети



Граф цитирований

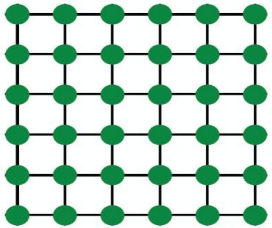


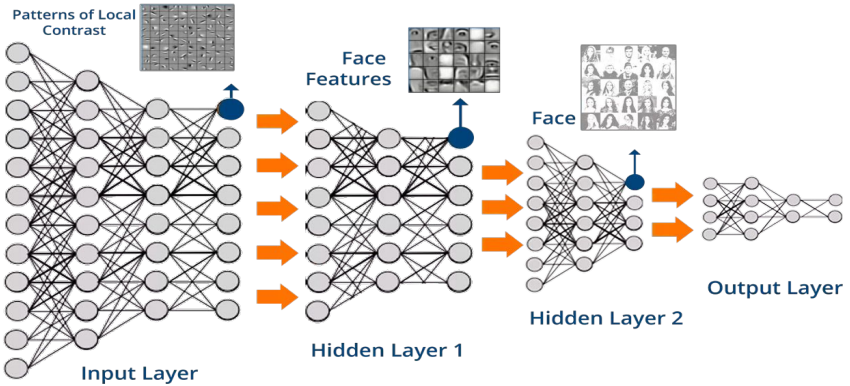
3D графика



Навигация

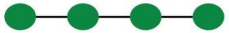
Современные ML инструменты

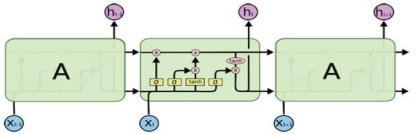




Изображения

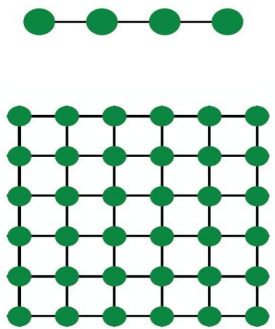
Текст/Аудио



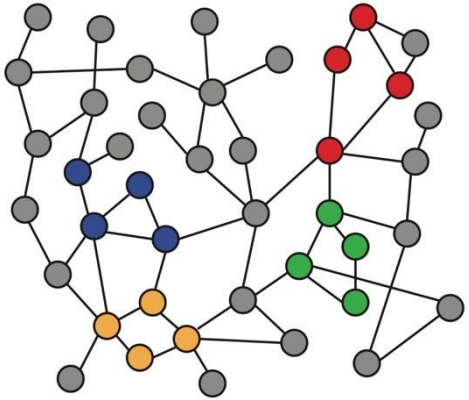


Современные ML инструменты

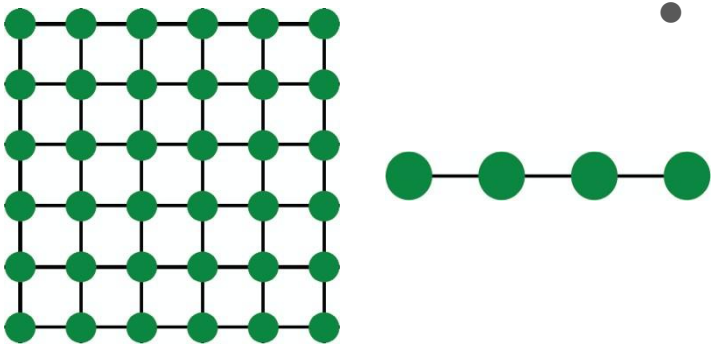
- Инструменты для глубокого обучения изначально спроектированы для анализа лишь подмножества графов
    - Последовательности
    - Сетки



Сложность анализа графов



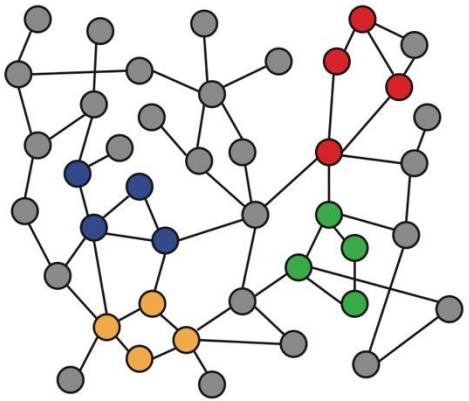
vs



- Сложная топологическая структура (нет пространственной локальности как в сетках)
- Нету начальной вершины или порядка обхода графа
- Граф часто динамичный
- Вершины могут иметь мультимодальные признаки

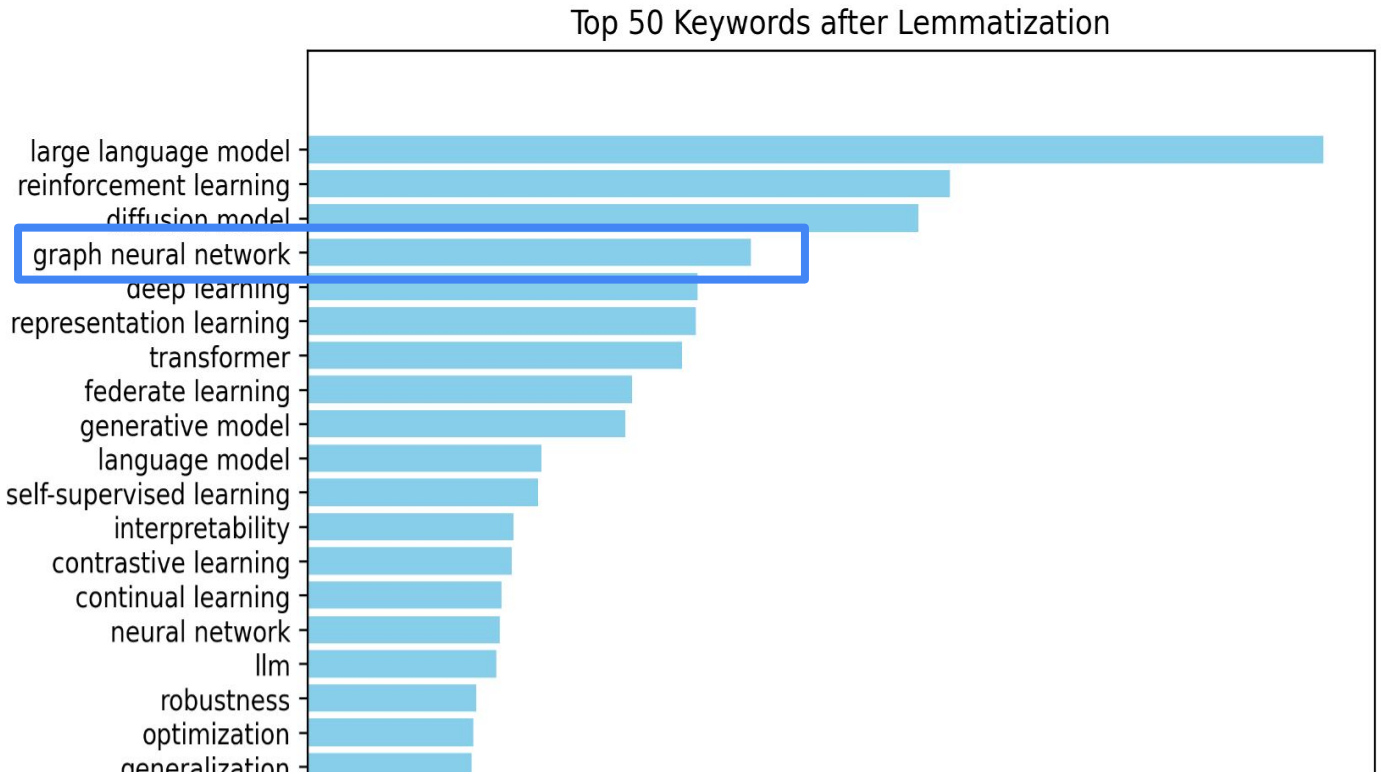
В следующих лекциях

- Как мы можем спроектировать нейронные сети намного более общего применения?
    - Произвольные графы



Популярность GNN

ICLR 2024

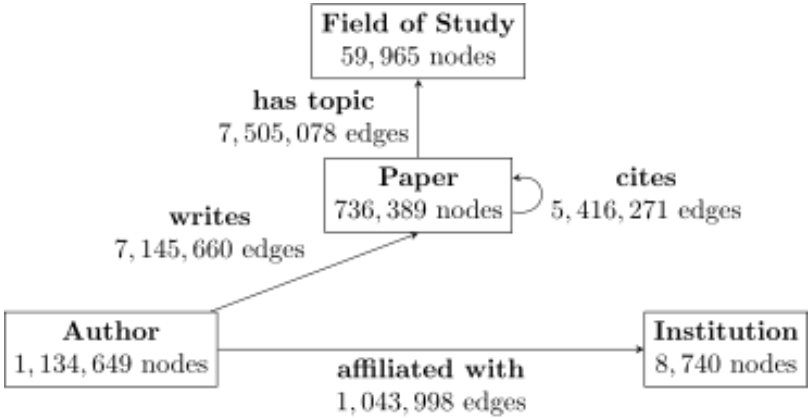


Выбор представления графовых данных

Геторогенные графы

- Во многих областях данные могут быть представлены в виде геторогенного графа $G = (V, E, R, T)$
    - Вершины $v_i \in V$
    - Рёбра $(v_i, r, v_j) \in E$
    - Типы вершин $T(v_i)$
    - Типы отношений $r \in R$
    - У вершин и рёбер могут быть атрибуты/признаки

Пример гетерогенного графа



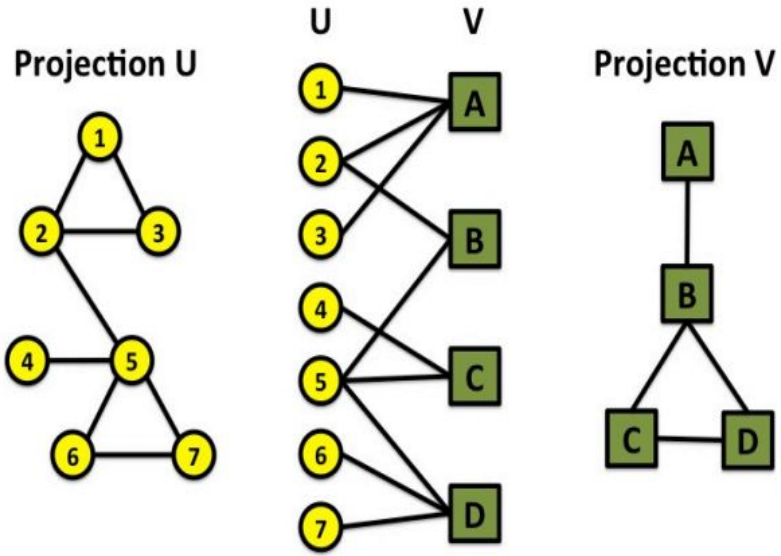
OPEN GRAPH BENCHMARK

- ogbn-mag (Microsoft Academic Graph)
- Типы вершин: ***author***, ***paper***, ***institution*** и ***field of study***
- Типы рёбер: ***writes***, ***affiliated with***, ***cites*** и ***has topic***

Выбор подходящего представления

- Как построить граф?
  - Что сделать вершинами?
  - Что сделать рёбрами?
  - Направленный vs ненаправленный
  - Нужны ли веса на рёбрах?
  - Какие типы вершин/рёбер?
  - Какие признаки хранятся в вершинах/рёбрах?
  - Нужен ли особый вид графа?
- От сделанного выбора зависит природа вопросов, на которые можно будет ответить в результате анализа графа

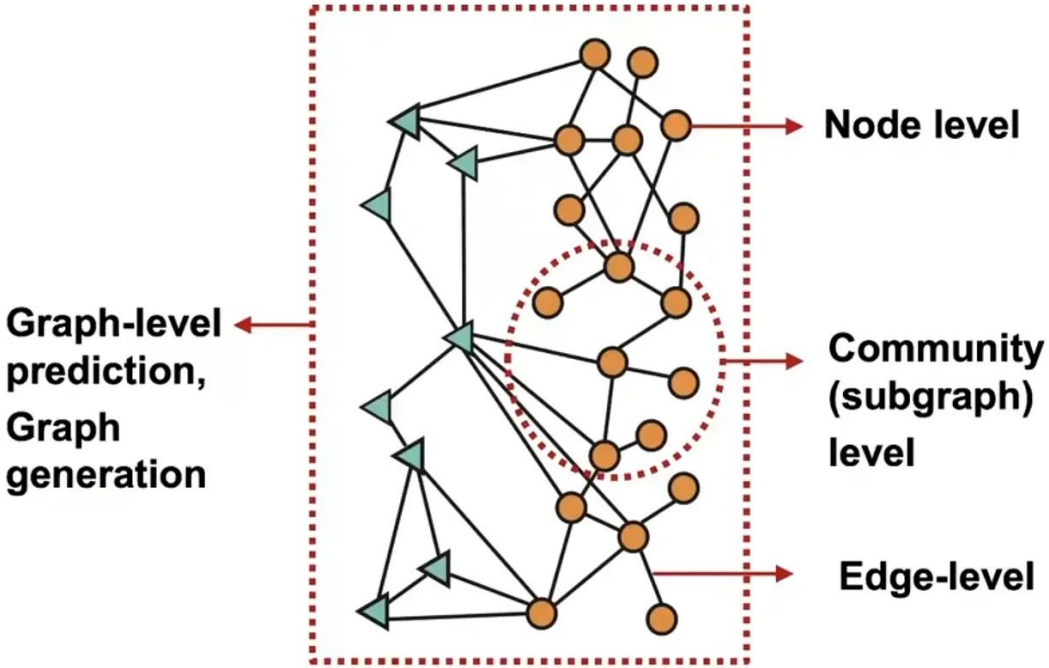
Двудольные графы



- Примеры двудольных графов
  - Авторы-Статьи
  - Пользователи-Фильмы
  - Покупатели-Товары
- Можно провести дополнительные рёбра и получить новые графы
  - Соавторы
  - Пользователи/покупатели со схожими вкусами

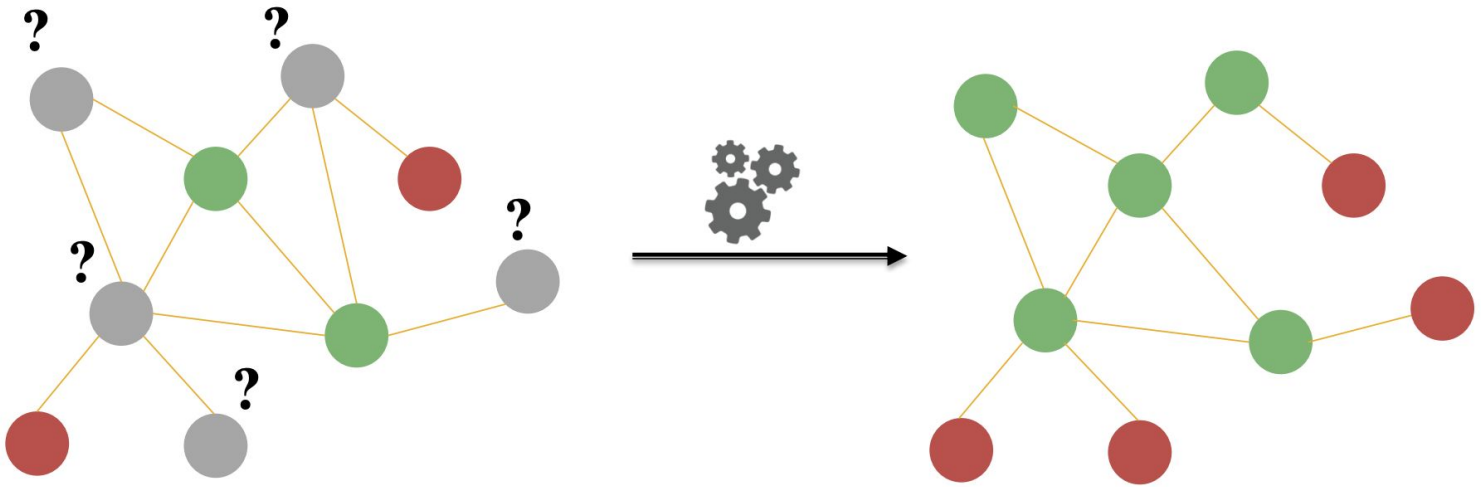
Виды задач машинного обучения на графах

Виды graph ml задач



Node-level tasks

Node classification

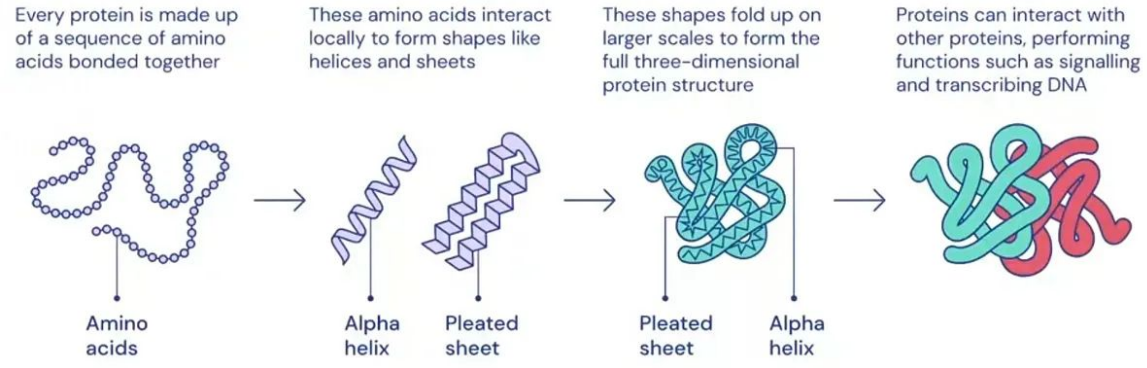


Node classification

- Предсказываем признаки отдельных вершин
- Например, категоризация
  - Покупателей
  - Товаров
  - Транзакций
  - Лекарств

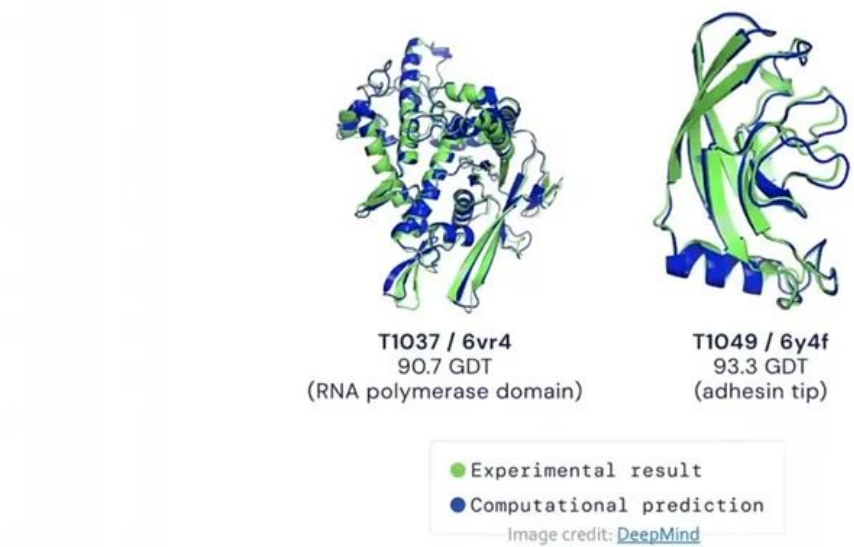
Protein folding

- Белки, составленные из аминокислот, под действием магнитных и прочих воздействий сворачиваются в сложные 3D фигуры
- От этого зависят многие важные биологические функции
  - Взаимодействие лекарств с белками и изменение процессов в организме для выздоровления



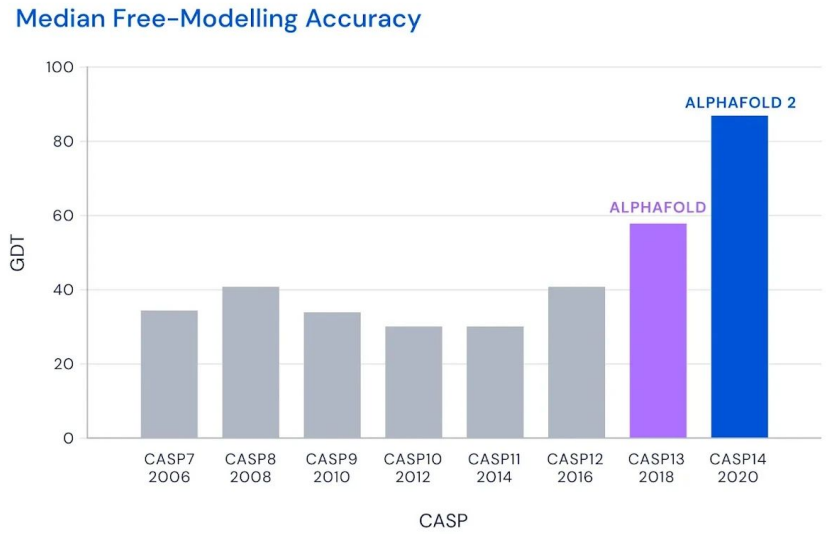
Protein folding

- Задача - предсказать 3D структуру белка, основываясь только на последовательности аминокислот



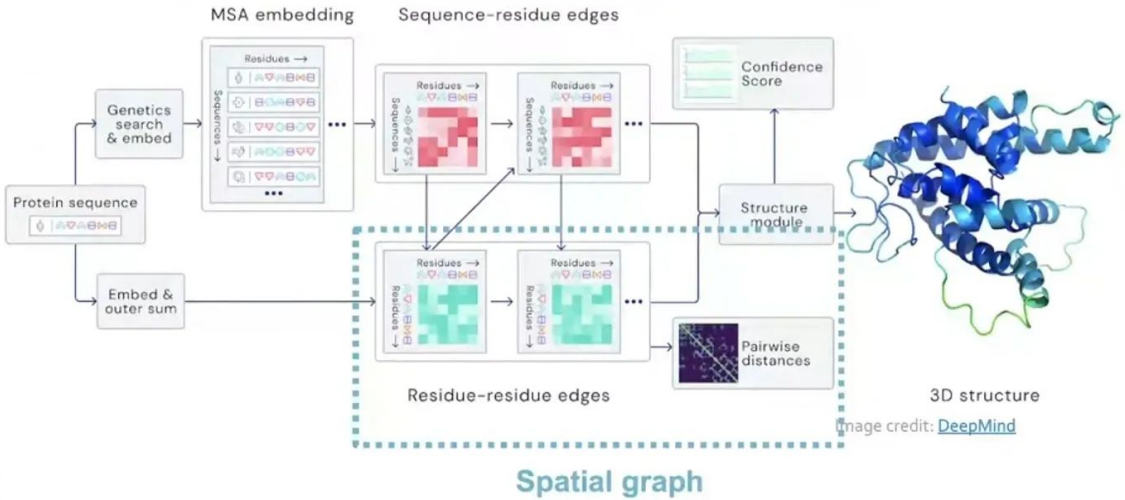
AlphaFold

- Начиная с 1970-ых годов пытаются решить данную задачу
- Использование GNN позволило сделать прорыв и решить задачу с 90% точностью



AlphaFold

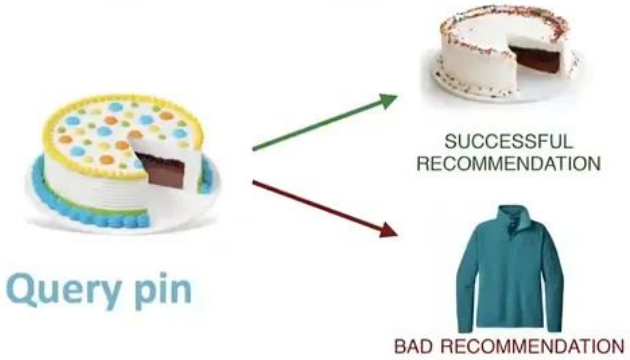
- Идея - представить белок в виде графа (пространственного графа)
- Вершины - аминокислоты
- Рёбра - пространственная близость аминокислот



Edge-level tasks

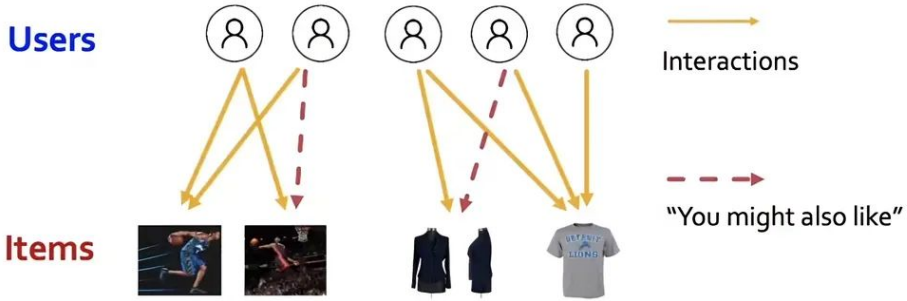
Link prediction

Task: Recommend related pins to users



Task: Learn node embeddings z_i such that

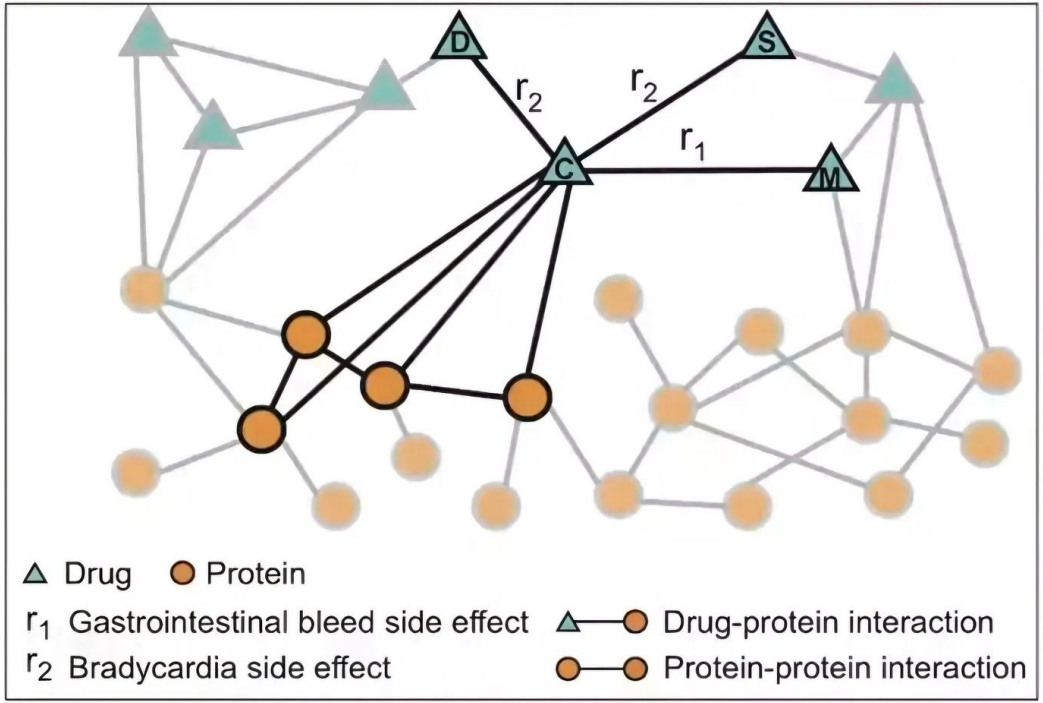
$$d(z_{cake1}, z_{cake2}) < d(z_{cake1}, z_{sweater})$$



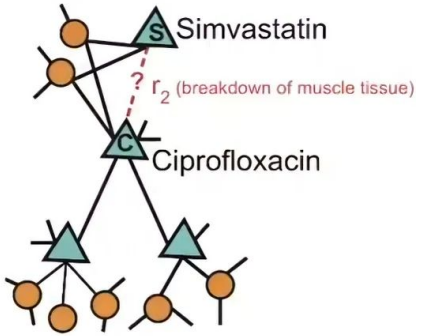
Recommendation system

- Многие компании используют GNNs для более точных рекомендаций
  - Pinterest
  - LinkedIn
  - Instagram
- Например, изображения (вершины) в Pinterest кодируются на основе взаимодействия пользователей и визуального контента
- Похожие вершины получают близкие представления (эмбеддинги)
- В итоге качество рекомендаций становится выше, чем после анализа только изображений

Взаимодействия лекарств



Query: How likely will Simvastatin and Ciprofloxacin, when taken together, break down muscle tissue?



Предсказание побочных эффектов

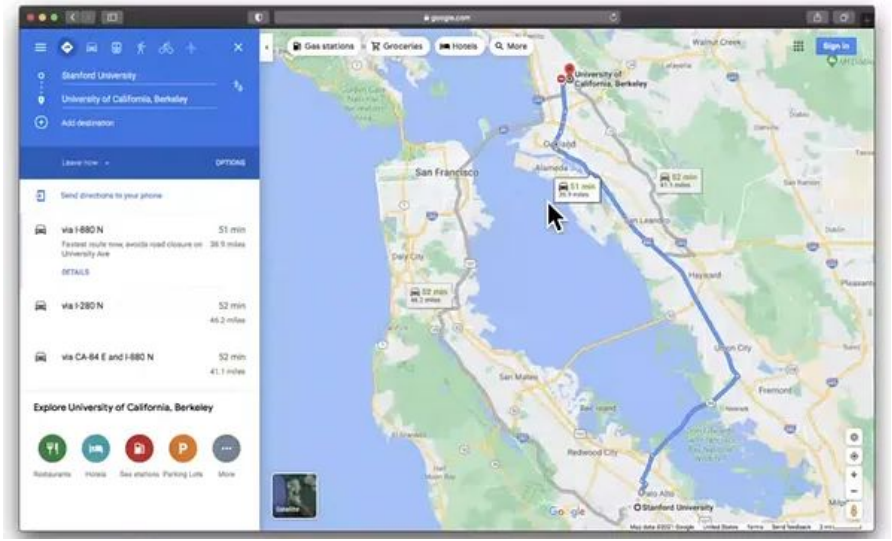
Rank	Drug i	Drug j	Side effect r	Evidence
1	Pyrimethamine	Aliskiren	Sarcoma	Stage *et al.* 2015
2	Tigecycline	Bimatoprost	Autonomic neuropathy	
3	Omeprazole	Dacarbazine	Telangiectases	
4	Tolcapone	Pyrimethamine	Breast disorder	Bicker *et al.* 2017
5	Minoxidil	Paricalcitol	Cluster headache	
6	Omeprazole	Amoxicillin	Renal tubular acidosis	Russo *et al.* 2016
7	Anagrelide	Azelaic acid	Cerebral thrombosis	
8	Atorvastatin	Amlodipine	Muscle inflammation	Banakh *et al.* 2017
9	Aliskiren	Tioconazole	Breast inflammation	Parving *et al.* 2012
10	Estradiol	Nadolol	Endometriosis	

Zitnik et al., Modeling Polypharmacy Side Effects with Graph Convolutional Networks, Bioinformatics 2018

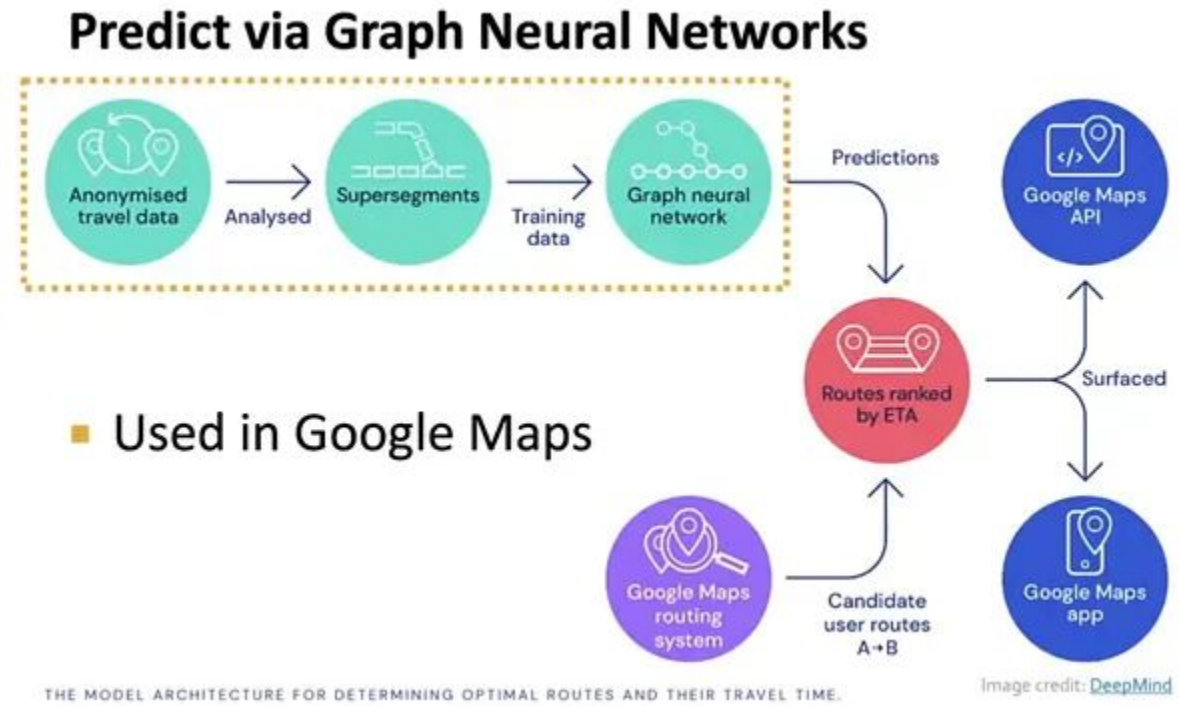
Graph-level tasks

Sub-graph level

- Вершины - сегменты дорог
- Время поездки предсказывается с помощью GNN

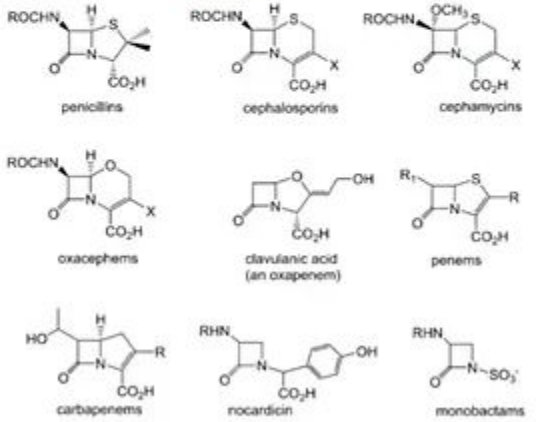


Traffic prediction



Graph-level task

- **Antibiotics are small molecular graphs**
  - **Nodes:** Atoms
  - **Edges:** Chemical bonds



Konaklieva, Monika I. "Molecular targets of β-lactam-based antimicrobials: beyond the usual suspects." Antibiotics 3.2 (2014): 128-142.

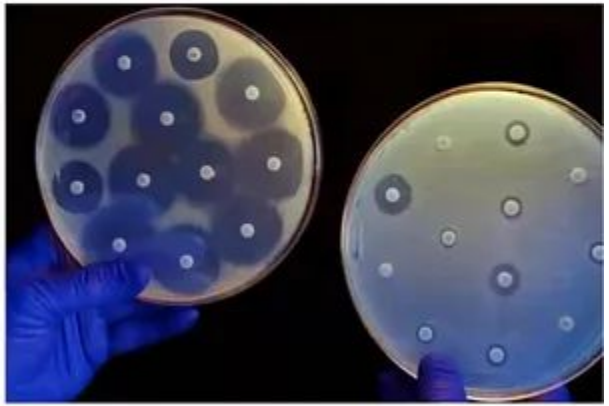
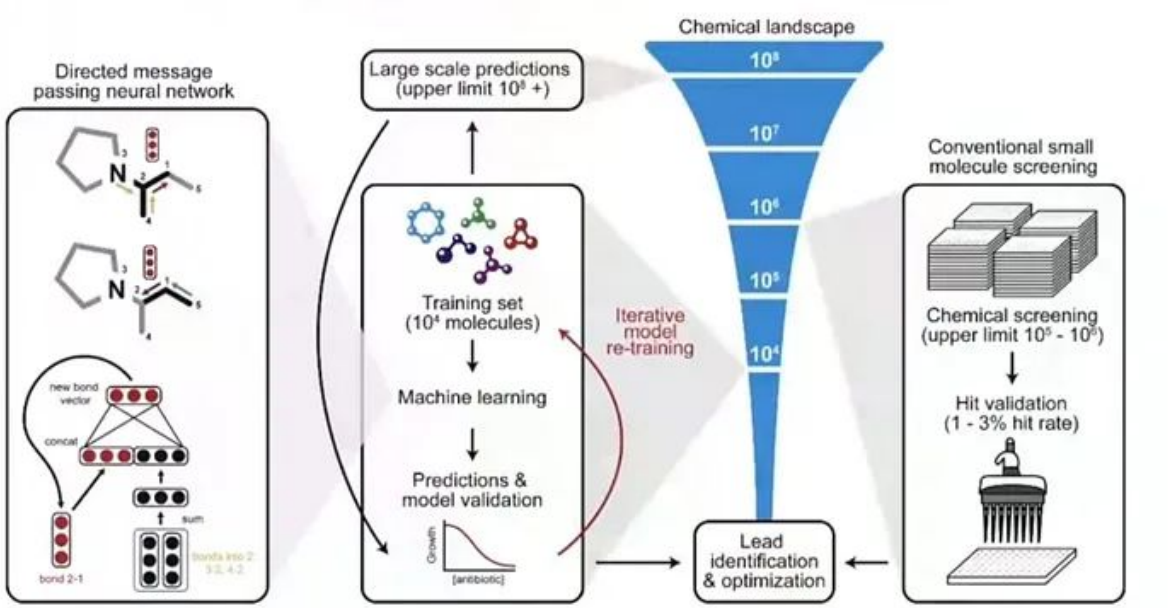


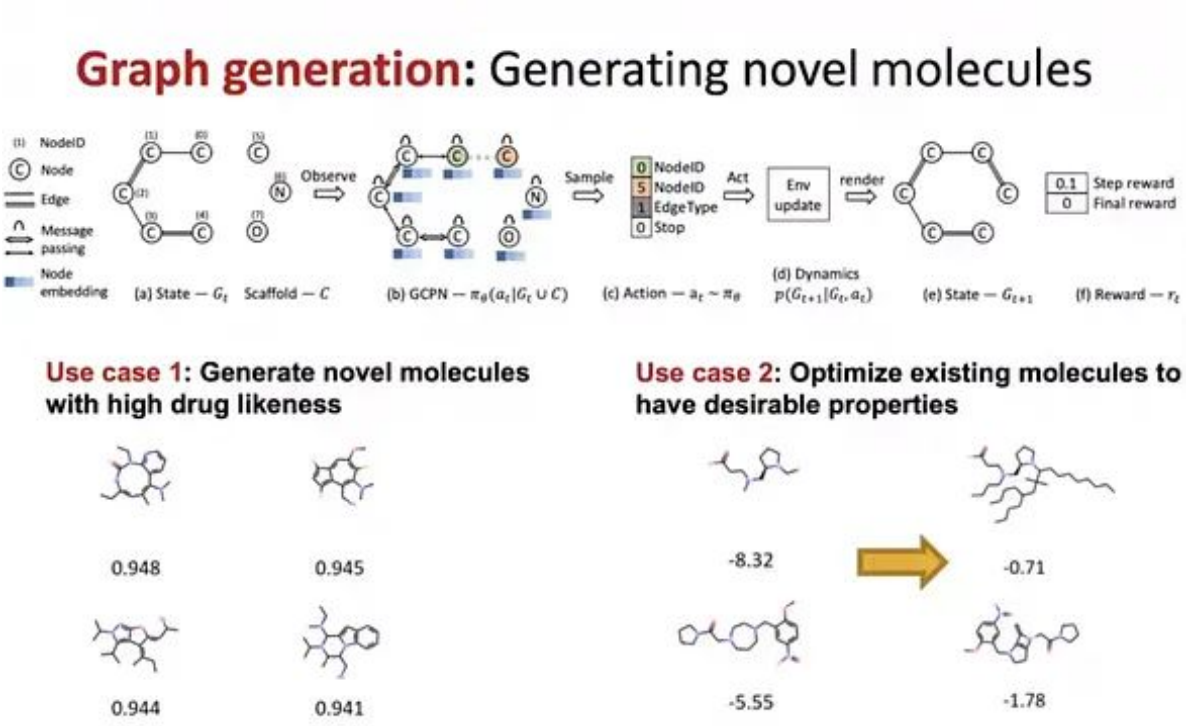
Image credit: CNN

Antibiotic discovery

- Задача - предсказать нужные свойства графов (молекул)



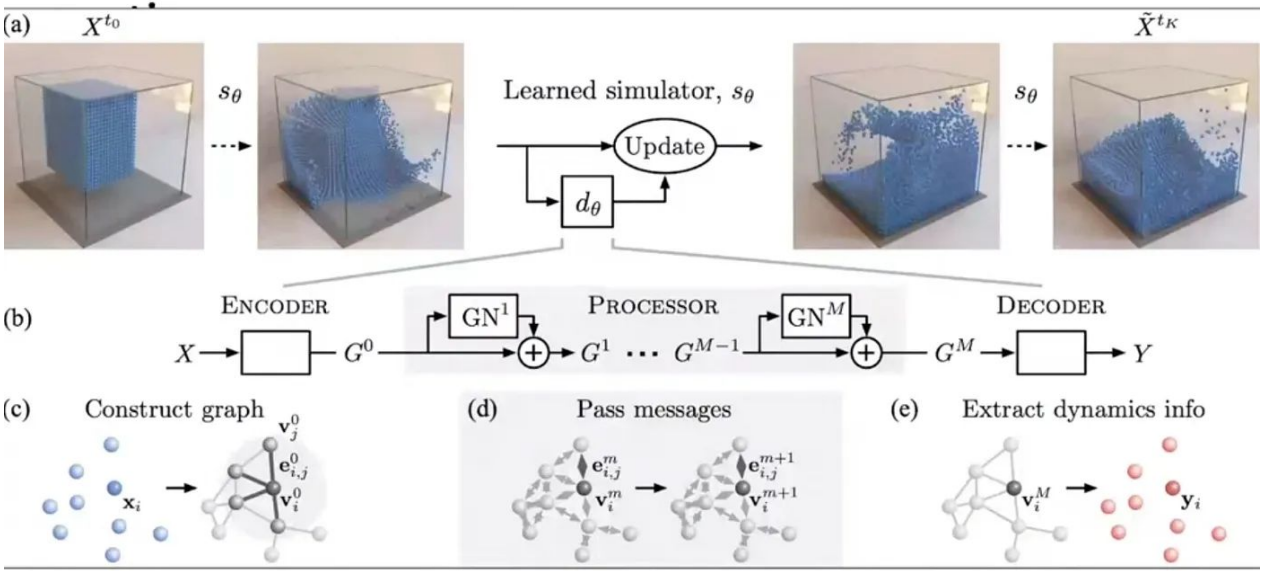
Генерация новых молекул



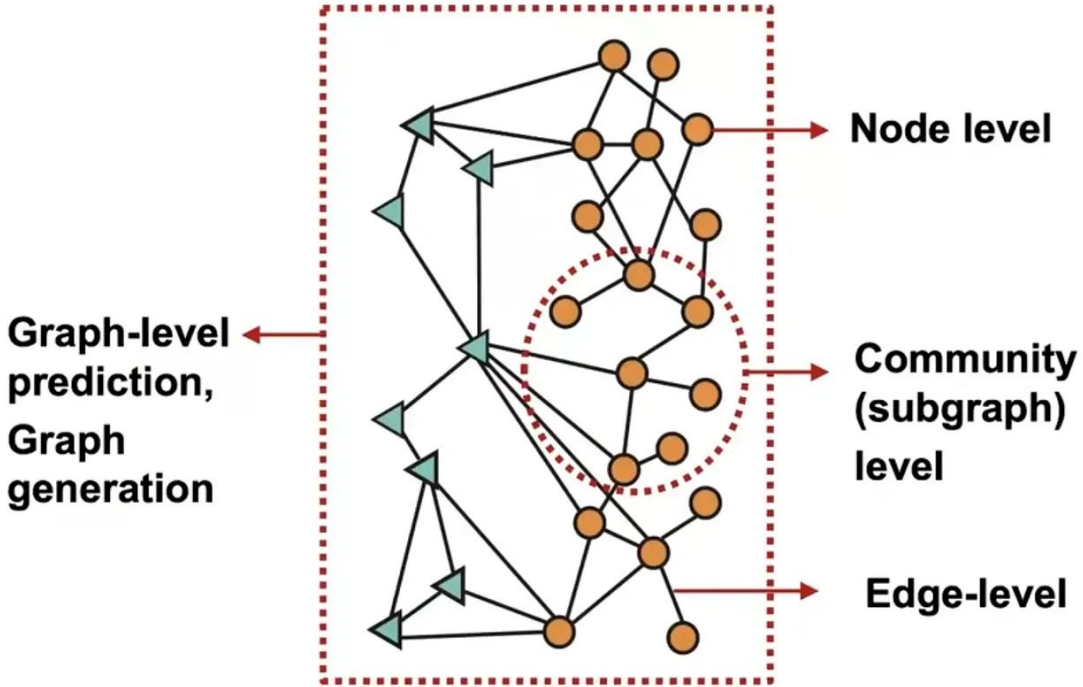
Симуляция изменений графа

A graph evolution task:

- **Goal**: Predict how a graph will evolve over



Заключение


Node level
Graph-level prediction, Graph generation
Community (subgraph) level
Edge-level