# Нейронные сети графов для прогнозирования свойств молекул: пример разрывов HOMO-LUMO Gaps с помощью PygPCQM4Mv2

Фернадо Леон, студент кафедры физико-математических
и естественных наук, Российский университет Дружбы Народов

В этом отчете представлена разработка и оценка Graph Attention Network (GAT) для прогнозирования разрыва HOMO-LUMO в молекулах из набора данных PCQM4Mv2. Исследование включает предварительную обработку данных молекулярного графа, реализацию модели GAT, обучение модели на наборе данных и анализ ее производительности на основе ключевых метрик оценки.

## 1 Введение

Разрыв HOMO-LUMO является важнейшим электронным свойством в молекулярной химии, влияющим на оптическое и электронное поведение. Точное предсказание этого свойства имеет значительные последствия для материаловедения и открытия лекарств. Графические модели глубокого обучения, такие как GAT, показали себя многообещающими в эффективной обработке молекулярных структур.

В хемоинформатике традиционные методы машинного обучения часто полагаются на созданные вручную молекулярные дескрипторы и признаки, которые могут ограничивать способность модели обобщать различные молекулярные структуры. С другой стороны, GAT обеспечивают более гибкий и выразительный подход, используя графические представления молекул. Они используют механизм внимания для назначения различной важности различным атомным и связевым взаимодействиям, что позволяет более тонко понимать молекулярные свойства. Это делает их особенно полезными в таких приложениях, как открытие лекарств, материаловедение и прогнозирование свойств, где захват сложных молекулярных взаимосвязей имеет решающее значение. По сравнению с традиционными графовыми сверточными сетями (GCN), GAT улучшают извлечение признаков за счет динамического взвешивания вкладов соседей, что приводит к повышению производительности в прогнозировании молекулярных свойств.

В этом отчете подробно описывается методология, используемая для обучения модели GAT для прогнозирования разрыва HOMO-LUMO.

## 2 PCQM4Mv2: Прогнозирование на уровне графа

Мы описываем наш набор данных OGB-LSC, который охватывает категорию задач прогнозирования на уровне графа ML в графах. Мы подчеркиваем практическую

значимость и разделение данных для набора данных. Набор данных OGB-LSC доступен на его официальном сайте.

## 2.1 Набор данных PCQM4Mv2

**Практическая значимость**. Точное прогнозирование зазоров HOMO-LUMO имеет решающее значение для продвижения исследований в области материаловедения и химии. Это свойство играет важную роль в определении электронного, оптического и проводящего поведения молекулы. Приложения охватывают различные области, такие как проектирование органических полупроводников, фотоэлектрических материалов и катализаторов. Традиционные вычислительные методы, такие как теория функционала плотности (DFT), хотя и точны, требуют больших вычислительных затрат. Подходы машинного обучения, особенно те, которые используют GNN, предлагают многообещающую альтернативу, предоставляя быстрые и точные прогнозы [1].

**Обзор датасета**. Набор данных, используемый в этом исследовании, представляет собой набор данных PCQM4Mv2, полученный из Open Graph Benchmark – Large Scale Challenge (OGB-LSC). Этот набор данных специально подобран для задачи прогнозирования на уровне графа, где цель состоит в том, чтобы предсказать молекулярные свойства на основе их графических представлений.

- **Число молекул**: Набор данных содержит более 3,8 миллионов молекулярных графов, что делает его одним из крупнейших общедоступных наборов данных по квантовой химии.
- **Узловые и граничные объекты**: Узлы соответствуют атомам молекулы с признаками, кодирующими атомные свойства (9 признаков). Края представляют собой химические связи с признаками, отражающими типы связей, порядок связей и стереохимию (3 признака).
- **Целевое свойство**: Целевым свойством служит щель HOMO-LUMO каждой молекулы, которая измеряется в электронвольтах (eV).

| **Набор данных** | **Тип задачи** | **Статистика** |
|---|---|---|
| PCQM4M | Графический уровень | #graphs: 3,746,619<br>#nodes (total): 52,970,652<br>#edges (total): 54,546,813 |

Table 1: Базовая статистика набора данных OGB-LSC

## 2.2 Методология

**Модель Архитектуры**. Предлагаемая модель представляет собой графовую нейронную сеть (GNN), разработанную для прогнозирования разрыва HOMO-LUMO gap молекул. Архитектура адаптирована для захвата графически структурированной природы молекулярных данных [2].

**GNN Layers**. Модель использует слои GATv2 для агрегации и распространения информации через молекулярный граф. Четыре слоя GATv2 накладываются друг

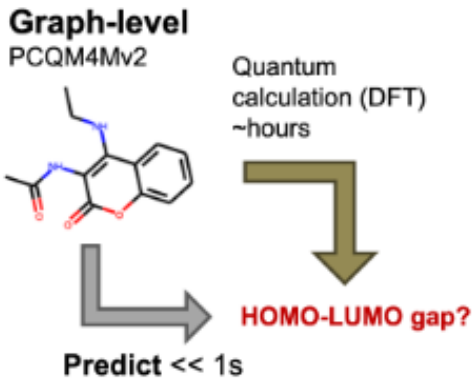


Figure 1: Обзор набора данных OGB-LSC, охватывающий задачу прогнозирования на уровне графа

на друга, за каждым из которых следует функция активации ELU и механизм нормализации BatchNorm и LayerNorm.

**Механизм объединения**. Операция «global mean pooling» и «global sum pooling» объединяют вложения на уровне узлов в единое представление на уровне графа, что позволяет учитывать разные аспекты молекулярной структуры.

**Окончательная регрессия головы**. Представление на уровне графа передается через глубокую нейросеть, состоящую из нескольких полностью связанных слоев с активацией ELU и dropout для регуляризации. Итоговый выходной слой формирует один скалярный прогноз, представляющий зазор HOMO-LUMO gap.

**Гиперпараметры**. Ключевые гиперпараметры модели GNN:

- Hidden Dimension: 256
- Number of Layers: 4
- Heads: 8
- Learning Rate: 0.001
- Batch Size: 128
- Dropout Rate: 0.4
- Number of Passes: 3

**Разделение набора данных**. Мы разделяем молекулы по их PubChem ID (CID) с соотношением 90/2/4/4 (train/validation/test-dev/test-challenge). Однако мы не используем молекулы подмножеств test-dev и test-challenge, поскольку test-dev — это квалифицировать нашу модель и сравнить ее с другими моделями, ранее представленными на сайте OGB-LSC, а с другой стороны, подмножество test-challenge предназначено для использования в конкурсах моделирования.

**Data Loaders**. Настроено для перемешивания набора данных, что улучшает обобщение модели. Где каждая партия содержит 128 молекулярных графа, параллельная загрузка данных с использованием 4 рабочих потоков и оптимизирует передачу данных в память GPU.

## 2.3 Процесс обучения

**Loss Function**: Функция потерь, используемая во время обучения, была среднеквадратичной ошибкой (MSE), которая подходит для задач регрессии. Эта функция измеряет среднеквадратичное отклонение между предсказанными и фактическими значениями, сильнее штрафуя большие ошибки. Выбор MSE гарантирует, что модель отдает приоритет минимизации значительных отклонений в прогнозах.

**Optimizer/Scheduler**: Модель была обучена с использованием оптимизатора Adam, популярного выбора для задач глубокого обучения из-за его механизма адаптивной скорости обучения. Начальная скорость обучения была установлена на уровне 0,001. Это постепенное снижение помогает модели более эффективно сходиться по мере обучения.

Процесс обучения состоял из 30 эпох, со следующими стратегиями для повышения стабильности и производительности. Параметры модели сохранялись всякий раз, когда улучшалась средняя абсолютная ошибка валидации (MAE), гарантируя, что модель с наилучшими показателями сохранялась для дальнейшей оценки.

# 3 Результаты

Предложенная модель GATv2 демонстрирует превосходную производительность в прогнозировании значений зазора HOMO-LUMO среднего диапазона. Модель достигает средней относительной процентной ошибки 5,19%, что указывает на сильную точность прогнозирования в целом.

Однако, как показано в результатах, модель испытывает трудности с экстремальными значениями зазора HOMO-LUMO, особенно для меньших значений. Относительная процентная ошибка значительно выше для этих случаев, что говорит о том, что модель испытывает трудности с захватом распределения этих менее распространенных молекулярных структур. Такое поведение может быть связано с дисбалансом в наборе данных или проблемами с представлением молекулярных графов с экстремальными электронными свойствами.

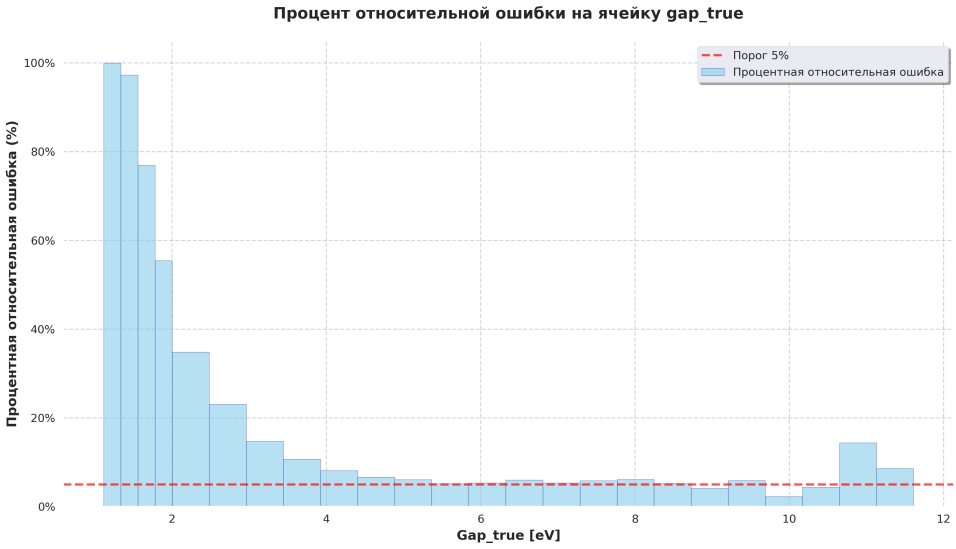


Figure 2: Диаграмма рассеяния: прогнозируемые значения против реальные ценности.

# 4 Вывод

В целом, разработанная модель GATv2 хорошо предсказывает значения разрыва HOMO-LUMO в среднем диапазоне, но испытывает трудности с прогнозированием экстремальных значений, особенно меньших значений разрыва. Это указывает на ограниченную обобщающую способность модели в отношении молекул с редкими электронными характеристиками.

Для дальнейшего улучшения точности модели предлагаются несколько стратегий:

Нормализация целевой переменной (разрыва HOMO-LUMO) – это может улучшить сходимость модели и помочь ей более точно предсказывать как низкие, так и высокие значения. Нормализация позволит модели работать с более равномерно распределенными данными, что снизит влияние дисбаланса между средними и экстремальными значениями.

Использование взвешенной функции потерь – для того чтобы повысить важность молекул с редкими значениями разрыва HOMO-LUMO. Это позволит модели уделять больше внимания графам с экстремальными значениями и уменьшить их процентную ошибку.

Изменение архитектуры – можно экспериментировать с дополнительными слоями или механизмами регуляризации, которые помогут модели лучше улавливать сложные закономерности в данных и учитывать влияние молекулярных графов с крайними значениями.

Дальнейшие исследования могут быть сосредоточены на тестировании этих улучшений и их влиянии на качество предсказаний. Важно стремиться к тому, чтобы модель работала стабильно во всем диапазоне значений HOMO-LUMO и могла надежно прогнозировать свойства как типичных, так и аномальных молекул.

# Список литературы

[1] Stanford Open Graph Benchmark. Pcqm4mv2: A benchmark for learning from molecular graphs, 2021. Accessed: 2025-01-16.

[2] Weihua Hu, Matthias Fey, Hongyu Ren, Maho Nakata, Yuxiao Dong, and Jure Leskovec. Ogb-lsc: A large-scale challenge for machine learning on graphs. *KDD Cup 2021. NeurIPS Datasets and Benchmarks Track*, 2021. Subjects: Machine Learning (cs.LG).

# 5 Приложение