Thinking like a CHEMIST: Combined Heterogeneous Embedding Model Integrating Structure and Tokens

Автор: *Рекут Н.А.*

Научный руководитель: *к.ф.м. Безносиков А.Н.*

Введение и мотивация

- **Проблематика:** традиционные методы представления молекул (SMILES) имеют ряд ограничений.
- **Актуальность:** современные химические задачи требуют более точных и интерпретируемых представлений молекул
- **Текущее состояние сферы:** языковые модели (ChemBERTa, SmilesBERT, MolFormer-XL), графовые модели (GEM, MolCLR), бимодальные архитектуры, основанные на SMILES и графовых нейросетях.

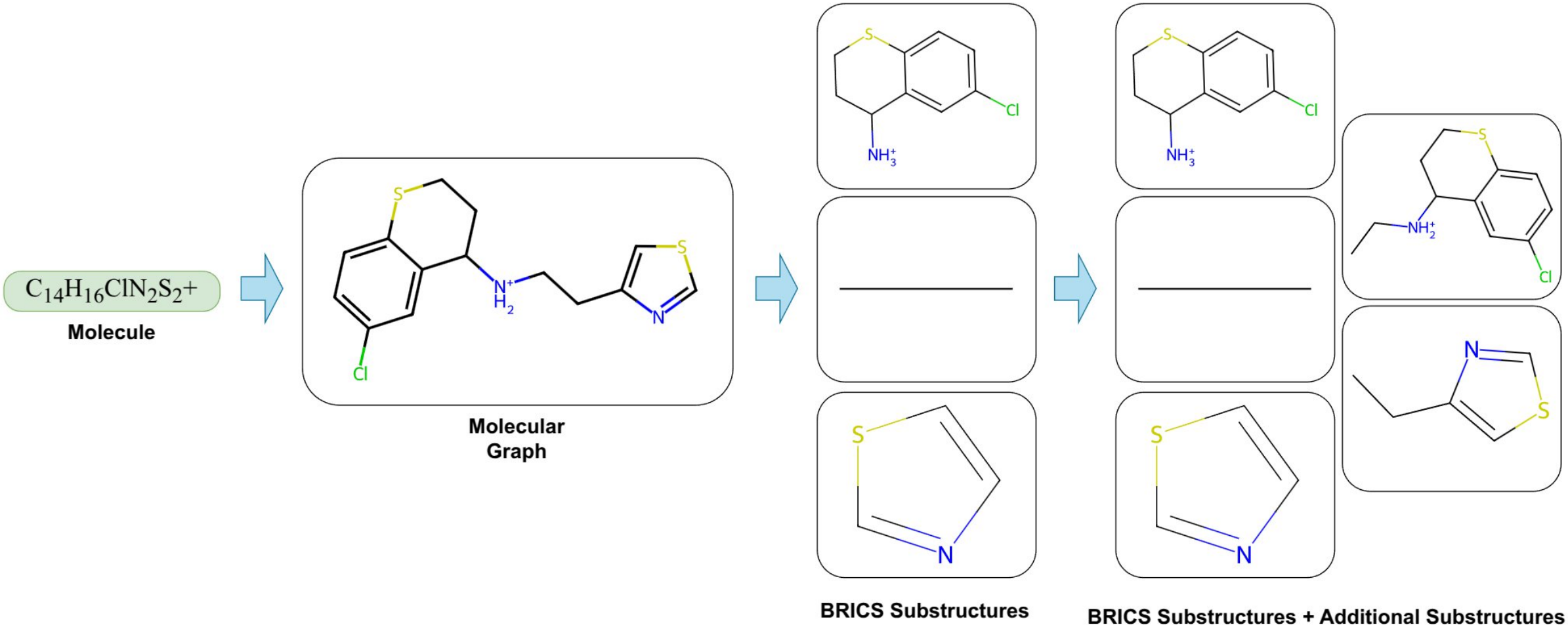
Постановка задачи и методы

- **Задача:** разработать бимодальную архитектуру, объединяющую языковую и графовую модели для получения представлений молекул
- **Методы:** принимающая на вход последовательность дескрипторов языковая модель (RoBERTa), графовая модель (GCN/GIN/Graphormer)

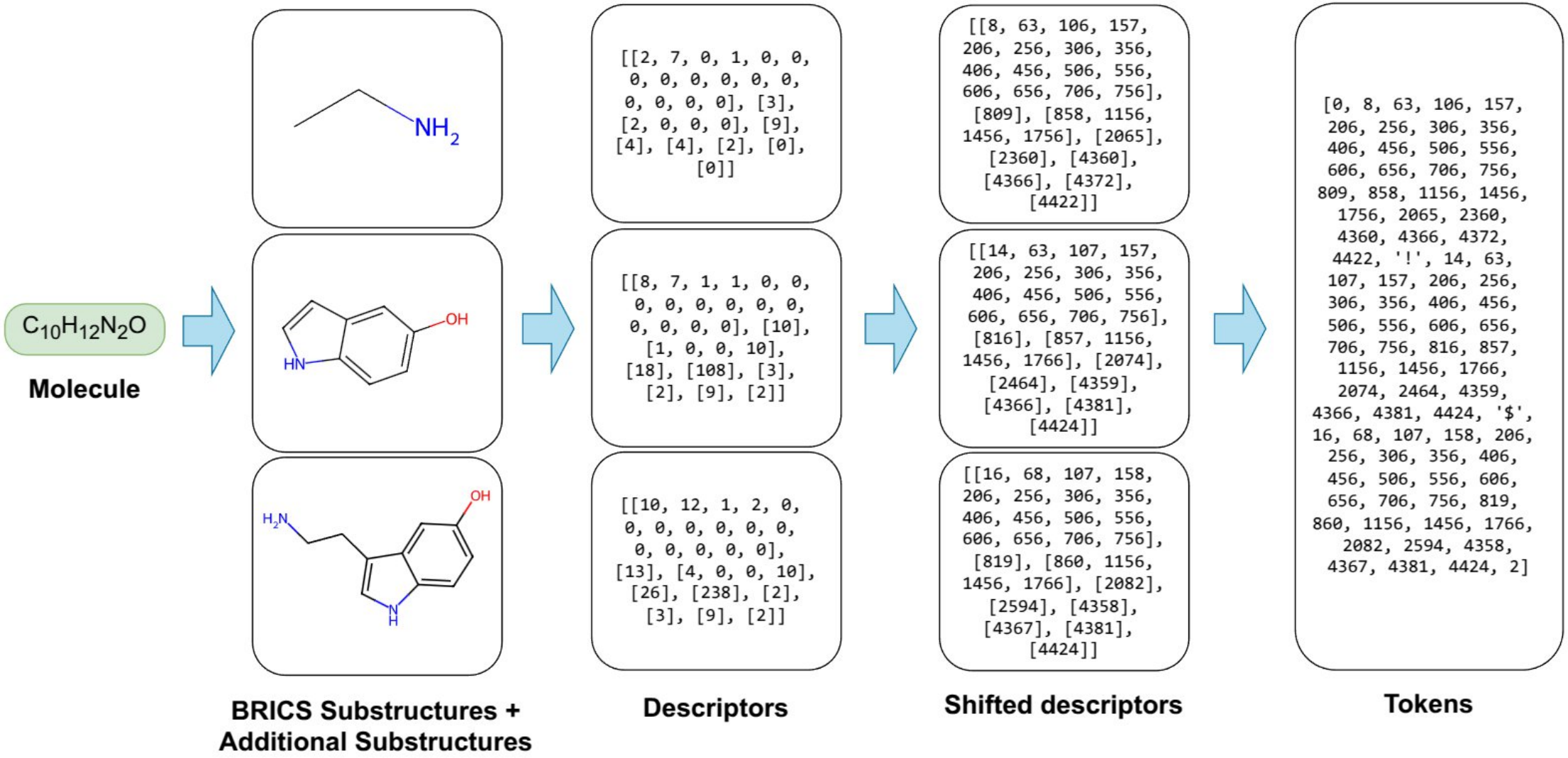
Химическая часть проекта: SMILES и ECFP

- Пример SMILES-кодировки: COc(c1)cccc1C#N (молекулярная формула C8H7NO)
- Дескрипторы — некоторые численные детерменированные представления молекул. Бывают топологическими, физико-химическими, квантово-химическими и др.

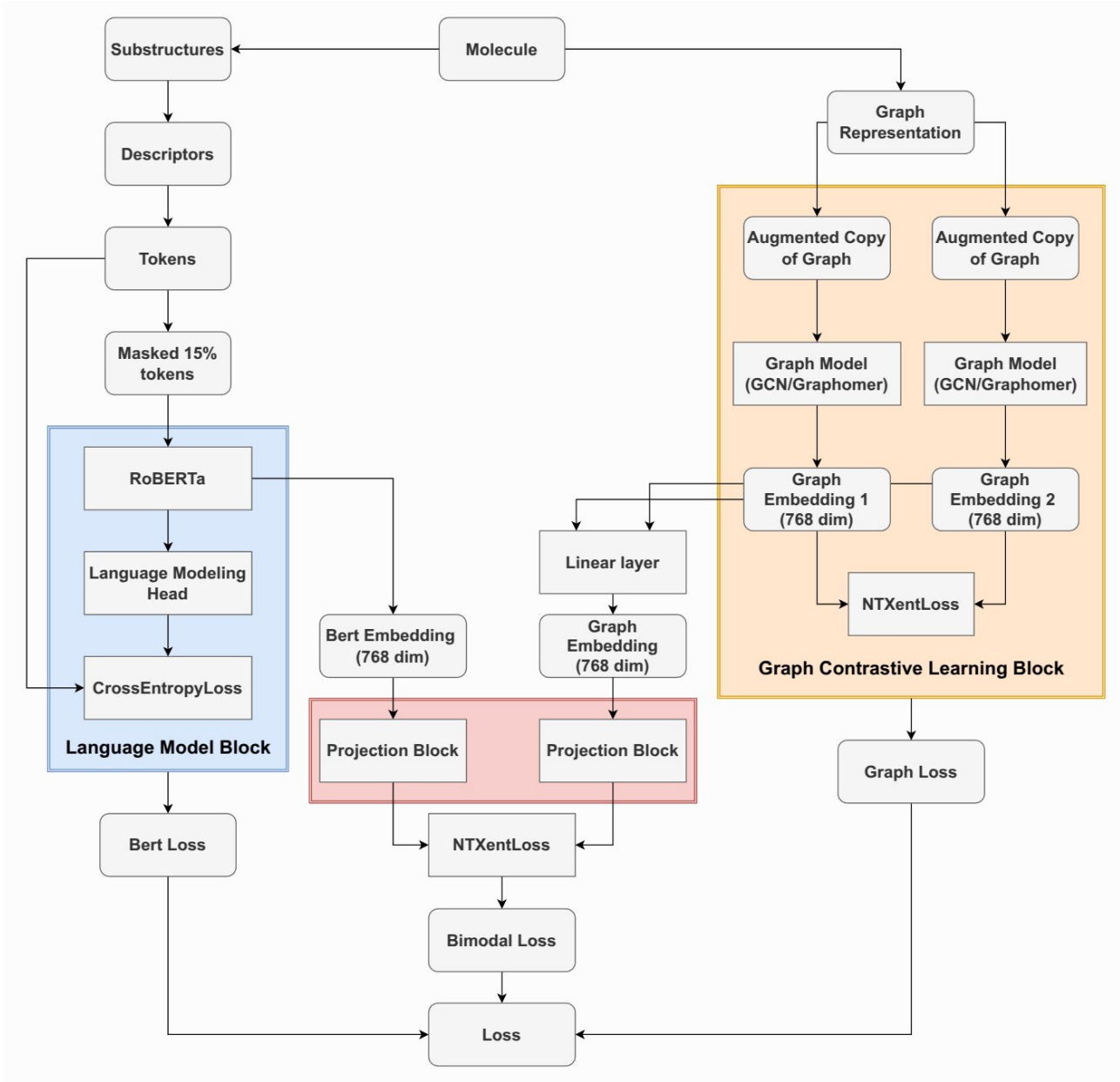
Представление данных для LM: BRICS-разбиение



Представление данных для LM: дескрипторы и токенизация



Архитектура

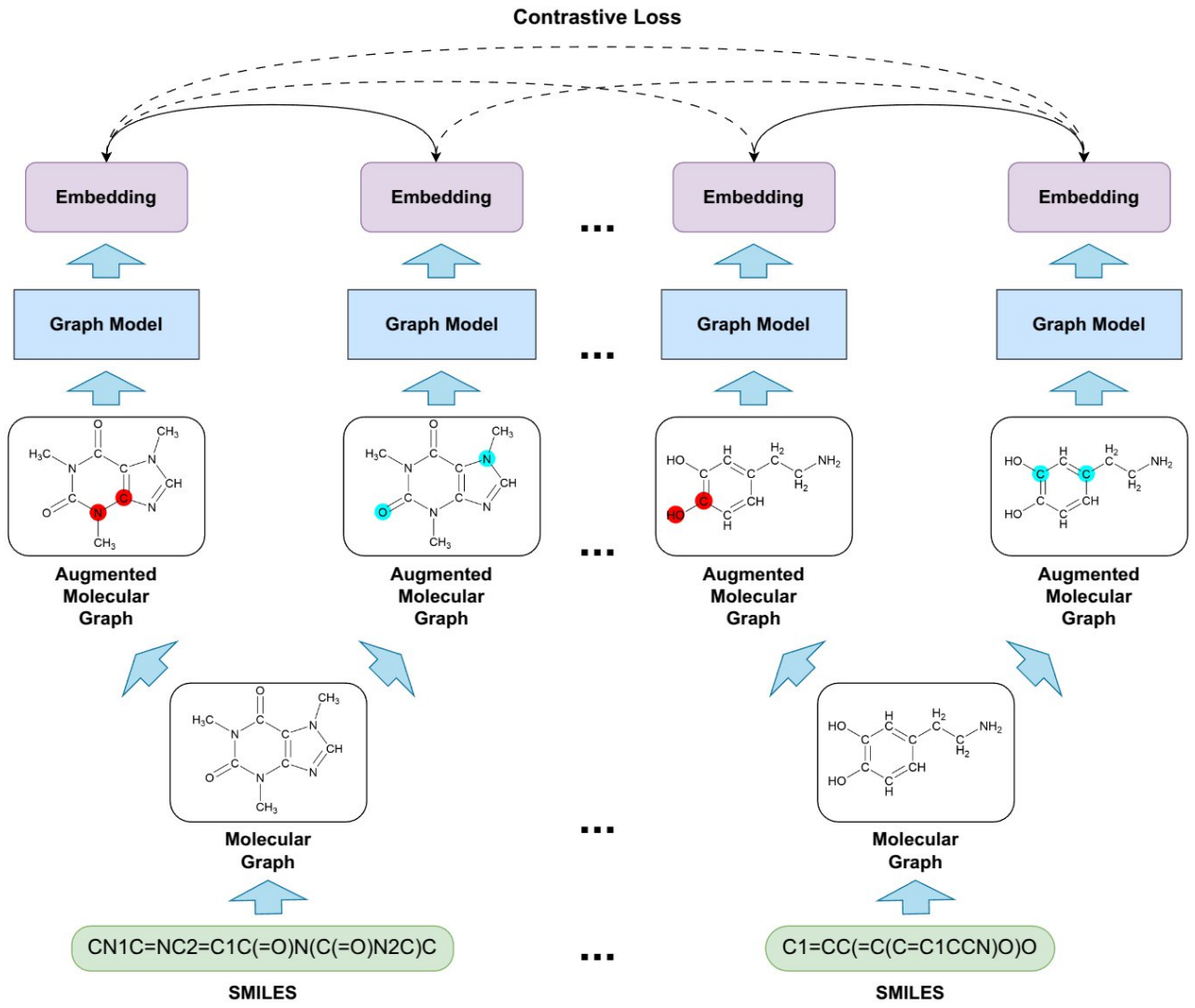


Обучение BERT

- Считаем дескрипторы токенами, массивы подструктур предложениями, а молекулы — текстом.
- Маскируем 15% токенов, предсказываем вероятности маскированных токенов
- Выходной эмбеддинг является эмбеддингом CLS-токена

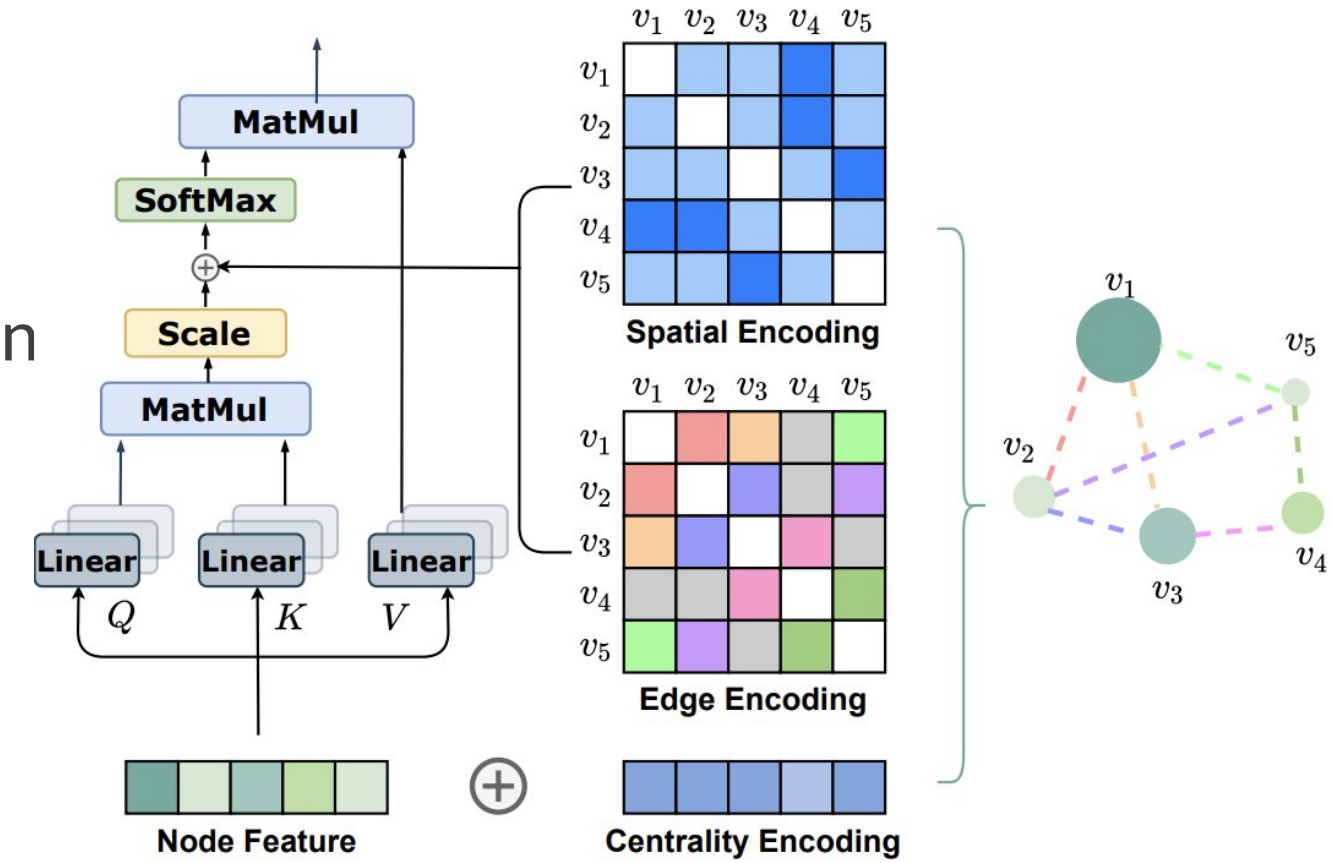
Обучение GCN/GIN

- Маскируем несколько атомов (10%-15%)
- Учим модель сближать эмбеддинги аугментаций одной и той же молекулы и отдалять аугментации различныхх



Обучение Graphormer

- Centrality encoding
- Spatial encoding
- Edge encoding in the attention



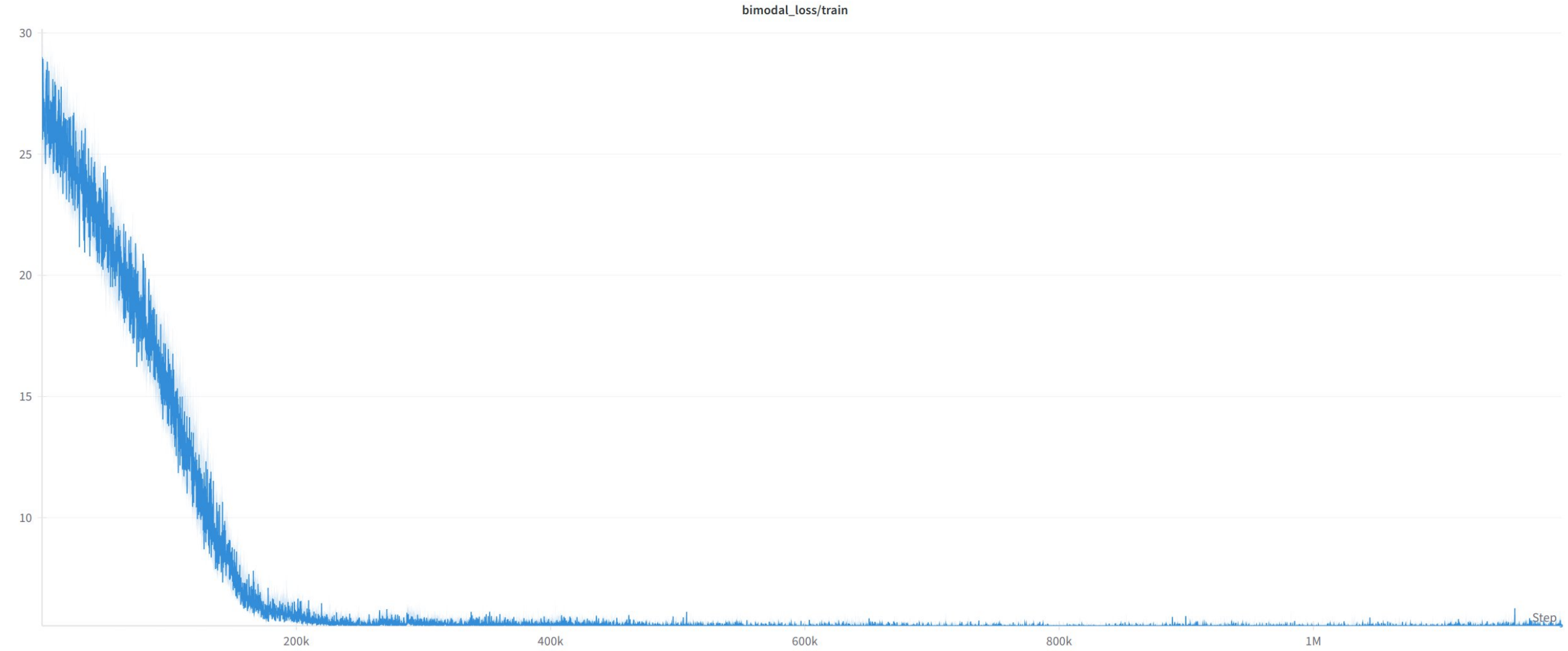
Получение результирующего эмбеддинга

- Косинусоидальная близость

$$CosineSimilarityLoss = \sum_i (input\_label_i - \cos\_sim(\mathrm{emb}_{i1}, \mathrm{emb}_{i2}))^2$$

- Линейная модель, предсказывающая маскированные токены
- Линейная модель, предсказывающая физические свойства

Немного графиков



Метрики обучения

Models	Datasets						
	BBBP	Tox21 (NR-AR)	ClinTox (FDA APPROVED)	ClinTox (CT TOX)	BACE	MUV	HIV
MolCLR(GCN) [53]	0.723±0.025	0.704±0.002	0.668±0.035	0.694±0.032	0.711±0.090	0.676±0.019	0.787±0.005
MolCLR(GIN) [53]	0.742±0.020	0.740±0.003	0.872±0.031	0.775±0.037	0.814±0.07	0.796±0.017	0.761±0.006
ChemBERTa [6]	0.647±0.053	0.753±0.009	-	0.736±0.015	0.721±0.022	0.667±0.015	0.625±0.012
Uni-Mol [62]	0.729±0.006	0.796±0.005	0.895±0.018	0.711±0.023	0.857±0.002	0.821±0.013	0.808±0.003
GEM [15]	0.724±0.004	0.781±0.001	0.875±0.013	0.692±0.019	0.856±0.011	0.817±0.005	0.806±0.009
GROVER (base) [40]	0.700±0.001	0.743±0.001	0.812±0.030	0.664±0.032	0.826±0.007	0.673±0.018	0.625±0.009
GROVER (large) [40]	0.695±0.001	0.735±0.001	0.75±0.037	0.683±0.041	0.810±0.014	0.673±0.018	0.682±0.011
Molformer [56]	0.916±0.002	-	0.907±0.006	0.812±0.031	0.844±0.017	-	-
MolFormer-XL [41]	0.917±0.001	0.847±0.001	0.933±0.004	0.901±0.012	0.862±0.009	-	0.812±0.003
SubD-BERT (ours)	0.893±0.018	0.829±0.007	**0.947**±0.013	**0.926**±0.017	0.811±0.022	0.753±0.015	0.692±0.011
BERT+GIN (ours)	**0.937**±0.002	**0.852**±0.003	0.912±0.009	**0.924**±0.014	0.855±0.015	**0.832**±0.011	0.786±0.007
BERT+GCN (ours)	0.891±0.005	0.830±0.002	0.903±0.016	0.793±0.031	0.738±0.012	0.794±0.017	0.736±0.010
BERT+Graphormer (ours)	0.862±0.009	0.815±0.003	0.878±0.019	0.837±0.021	**0.892**±0.015	0.819±0.013	**0.851**±0.060
XGBoost (descriptors, ours)	0.821	0.663	0.856	0.871	0.695	0.650	0.562
LightGBM (descriptors, ours)	0.832	0.653	0.886	0.853	0.682	0.581	0.546
SVM (descriptors, ours)	0.612	0.617	0.525	0.679	0.547	0.559	0.534

Ограничения и перспективы развития

Текущие недостатки:

- Проблемы с датасетами малых размеров
- Неэффективность для неорганических соединений и полимеров

Перспективы развития:

- Обучение на большей выборке (100 млн. молекул)
- Возможное изменение BRICS-rules