УДК: 004.032.26

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

N.A. Rekut[1] A. N. Beznosikov[1]

[1]MIPT

1 Introduction

Molecular representation remains a fundamental challenge in cheminformatics. Traditional approaches like SMILES strings [1] or graph-based models [2] fail to capture both structural and physicochemical properties effectively. Our work bridges this gap by introducing a bimodal architecture that combines substructure-aware language models with graph networks, achieving state-of-the-art performance while maintaining chemical interpretability.

2 Methodology

2.1 Molecular Representation

The core innovation of our approach lies in the chemically-grounded preprocessing pipeline. Unlike traditional methods that process molecules as either strings or graphs, we first decompose them into synthetically meaningful substructures using BRICS fragmentation [3]. This mirrors how chemists mentally break down complex molecules into functional groups and scaffolds.

Each substructure undergoes descriptor computation across three categories: topological, e.g. Wiener indices [4], physicochemical, e.g. uff energy, and hybrid, such as ECFP [5] connection points encode fragment linkage information.

2.2 Neural Architecture

Our model processes molecules through parallel language and graph pathways:

2.2.1 Language Pathway

The RoBERTa-based language model [6] treats each molecule as a document where substructures' descriptors form sentences and then the [CLS] token embedding represents the whole molecule. We implement position-aware tokenization where descriptor values are adjusted based on their ordinal position in the sequence, allowing the model to distinguish between identical descriptors appearing in different substructures.

2.2.2 Graph Pathway

For structural representation, we compare two approaches:

- **GIN Network**: Implements message passing with learnable neighborhood aggregation weights ($\epsilon = 0.01$) [7]
- **Graphormer**: Incorporates spatial encoding through shortest-path distances and edge gating mechanisms [8]

Both variants use contrastive learning with 20% feature masking to improve robustness.

2.3 Multimodal Fusion

The alignment of language and graph representations occurs through a shared projection space. The projection blocks consist of two linear layers with batch normalization and ReLU activation. The complete loss function combines:

$$L = \underbrace{0.4L_{\text{lm}}}_{\text{language}} + \underbrace{0.3L_{\text{graph}}}_{\text{structure}} + \underbrace{0.3L_{\text{align}}}_{\text{cross-modal}} \tag{1}$$

where L_{align} minimizes the cosine distance between matched molecule pairs.

3 Results and Discussion

Table 1: Classification Performance Comparison (ROC-AUC)

Dataset	Our Model	Best Baseline
BBBP	0.88	0.74 [9]
Tox21	0.79	0.74 [10]
HIV	0.81	0.62 [11]

The results demonstrate three key advantages:

- **Chemical Accuracy**: 14.2% improvement on BBBP shows better capture of blood-brain barrier penetration patterns
- **Efficiency**: Matches Uni-Mol [12] performance with significantly fewer parameters
- **Robustness**: Consistent gains across diverse tasks from toxicity (Tox21) to antiviral activity (HIV)

4 Conclusion

We present a novel molecular representation framework that combines the interpretability of descriptor-based approaches with the expressive power of graph neural networks. Key innovations include: chemically meaningful substructure decomposition, position-aware descriptor tokenization, earnable projection space for multimodal alignment.

Future work will explore applications in generative chemistry and reaction prediction.

References

[1] David Weininger. "SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules". In: *Journal of chemical information and computer sciences* 28.1 (1988), pp. 31–36.

[2] Thomas N Kipf and Max Welling. "Semi-supervised classification with graph convolutional networks". In: *arXiv preprint arXiv:1609.02907* (2016).

[3] Jorg Degen et al. "On the art of compiling and using 'drug-like' chemical fragment spaces". In: *ChemMedChem* 3.10 (2008), pp. 1503–1507.

[4] Harry Wiener. "Structural determination of paraffin boiling points". In: *Journal of the American Chemical Society* 69.1 (1947), pp. 17–20.

[5] David Rogers and Mathew Hahn. "Extended-connectivity fingerprints". In: *Journal of chemical information and modeling* 50.5 (2010), pp. 742–754.

[6] Yinhan Liu et al. "Roberta: A robustly optimized bert pretraining approach". In: *arXiv preprint arXiv:1907.11692* (2019).

[7] Keyulu Xu et al. "How powerful are graph neural networks?" In: *arXiv preprint arXiv:1810.00826* (2018).

[8] Chengxuan Ying et al. "Do transformers really perform badly for graph representation?" In: *Advances in neural information processing systems* 34 (2021), pp. 28877–28888.

[9] Yuyang Wang et al. "Molecular contrastive learning of representations via graph neural networks". In: *Nature Machine Intelligence* 4.3 (2022), pp. 279–287.

[10] Yu Rong et al. "Self-supervised graph transformer on large-scale molecular data". In: *Advances in neural information processing systems* 33 (2020), pp. 12559–12571.

[11] Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. "ChemBERTa: large-scale self-supervised pretraining for molecular property prediction". In: *arXiv preprint arXiv:2010.09885* (2020).

[12] Gengmo Zhou et al. "Uni-mol: A universal 3d molecular representation learning framework". In: *arXiv preprint arXiv:2303.16982* (2023).