OXFORD

Briefings in Bioinformatics, 2025, **26(6)**, bbaf629
https://doi.org/10.1093/bib/bbaf629
Problem Solving Protocol

ProtoMol: enhancing molecular property prediction via prototype-guided multimodal learning

Yingxu Wang[1,‡], Kunyu Zhang[2,‡], Jiaxin Huang[1], Nan Yin[3], Siwei Liu[4,*], Eran Segal[1,5,*]

[1]Department of Machine Learning, Mohamed bin Zayed University of Artificial Intelligence, AI Diyafah Street, 7909 Abu Dhabi, United Arab Emirates
[2]International College, Zhengzhou University, Daxue North Road, 450000 Henan, China
[3]Department of Computer Science and Engineering, Hong Kong University of Science and Technology, 999077, Hong Kong, China
[4]School of Natural and Computing Science, University of Aberdeen, 32 Elphinstone Road, AB24 3EU Scotland, United Kingdom
[5]Department of Molecular Cell Biology, Weizmann Institute of Science, 7610001, Rehovot, Israel

*Corresponding authors. Siwei Liu, School of Natural and Computing Science, University of Aberdeen, 32 Elphinstone Road, AB24 3EU Scotland, United Kingdom. E-mail: siwei.liu@abdn.ac.uk; Eran Segal, Department of Molecular Cell Biology, Weizmann Institute of Science, 7610001, Rehovot, Israel. E-mail: eran.segal@weizmann.ac.il

‡Yingxu Wang and Kunyu Zhang are equal contributions to this paper.

Abstract

Multimodal molecular representation learning, which jointly models molecular graphs and their textual descriptions, enhances predictive accuracy and interpretability by enabling more robust and reliable predictions of drug toxicity, bioactivity, and physicochemical properties through the integration of structural and semantic information. However, existing multimodal methods suffer from two key limitations: (i) they typically perform cross-modal interaction only at the final encoder layer, thus overlooking hierarchical semantic dependencies; (ii) they lack a unified prototype space for robust alignment between modalities. To address these limitations, we propose ProtoMol, a prototype-guided multimodal framework that enables fine-grained integration and consistent semantic alignment between molecular graphs and textual descriptions. ProtoMol incorporates dual-branch hierarchical encoders, utilizing Graph Neural Networks to process structured molecular graphs and Transformers to encode unstructured texts, resulting in comprehensive layer-wise representations. Then, ProtoMol introduces a layer-wise bidirectional cross-modal attention mechanism that progressively aligns semantic features across layers. Furthermore, a shared prototype space with learnable, class-specific anchors is constructed to guide both modalities toward coherent and discriminative representations. Extensive experiments on multiple benchmark datasets demonstrate that ProtoMol consistently outperforms state-of-the-art baselines across a variety of molecular property prediction tasks. Our source code is available at: https://github.com/zky04/Protomol.

Keywords: molecular property prediction; multi-modal learning; molecular graph

Introduction

Accurate molecular representation is fundamental to a broad spectrum of applications in bioinformatics, such as drug discovery, toxicity assessment, and functional genomics [1–3]. In recent years, multimodal learning has become an effective approach for integrating diverse molecular information, especially for predicting biologically relevant properties of small molecules and biomacromolecules [4, 5]. In typical bioinformatics pipelines, molecular data are represented in two complementary forms: structured graphs [6, 7], which depict atomic connectivity and chemical bonding, and descriptive textual annotations [8, 9], such as SMILES strings and expert summaries, that capture chemical, functional, or pharmacological characteristics. Graph-based models provide a faithful depiction of molecular structure, effectively capturing spatial interactions and stereochemical configurations essential for understanding biological function [10–12]. In contrast, textual descriptors provide a compact, information-rich sequence that can highlight key functional groups, bioactive motifs, and provide higher-level semantic abstraction within biological or pharmacological contexts [13–15]. By combining these modalities, it becomes possible to develop more comprehensive and interpretable predictive models, thereby advancing molecular property prediction and supporting hypothesis-driven biomedical research [16–18].

Early approaches in molecular representation learning primarily relied on textual descriptors to capture chemical information. Among these, the Simplified Molecular Input Line Entry System (SMILES) [8] emerged as a widely adopted format for encoding molecular structures as linear sequences of characters. This sequential representation enables the use of established sequence modeling architectures, including recurrent neural networks (RNNs) [19], convolutional neural networks (CNNs) [20], and Transformers [21–23], which leverage the syntactic regularity and semantic compactness of SMILES to extract features such as atom types, bond orders, and functional groups [24–26]. While SMILES provides a concise and expressive encoding, its inherent linear structure limits the preservation of molecular topology, making it difficult to accurately represent complex structural features, including stereochemistry, ring closures, and branching patterns that are essential for understanding biological function and pharmacological properties [24, 27, 28]. To overcome this limitation, graph-based models have been developed, where molecules are represented as graphs with atoms as nodes and chemical bonds as edges [7, 29, 30]. Graph neural networks (GNNs) [6, 11, 31] enable the propagation of

Received: September 3, 2025. **Revised:** October 16, 2025. **Accepted:** October 28, 2025
© The Author(s) 2025. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

information through message-passing mechanisms, allowing for the aggregation of both local atomic environments and long-range structural dependencies [11, 12]. This approach preserves spatial connectivity and topological relationships, providing a more biologically relevant representation compared with sequence-based methods [12, 32]. Recent research has increasingly focused on integrating textual and graph-based information to achieve more comprehensive and informative molecular representations [16, 33]. These multimodal approaches align the structural fidelity of molecular graphs with the semantic richness of textual descriptions through cross-modal interactions, joint embedding spaces, or aligned training objectives, thereby enabling the capture of a broader spectrum of molecular characteristics [16, 34]. To further enhance semantic consistency and model interpretability, prototype-based learning strategies have been introduced, where learnable, class-specific prototypes serve as semantic anchors to align heterogeneous representations from graphs and text [35–37]. Prototype-guided frameworks have demonstrated significant potential for improving the discriminative power and generalizability of molecular embeddings, especially in tasks such as few-shot property prediction, identification of rare or complex substructures, and discovery of chemically or biologically relevant functional groups [38–40]. As a result, these advances in multimodal and prototype-guided learning are driving progress in computational drug discovery, functional genomics, and other key areas of biomedical research.

However, current multimodal molecular representation learning methods still face two fundamental limitations that restrict their applicability in real-world bioinformatics and biomedical research. (i) Most existing approaches perform cross-modal interaction only at the final layer of graph and textual encoders, overlooking the hierarchical alignment across intermediate semantic levels [41, 42]. Molecular graphs naturally encode atomic connectivity and spatial topology, while textual descriptions provide functional and pharmacological annotations in sequential form [43, 44]. These modalities not only differ in representational structure and semantic density, but also encapsulate distinct aspects of biological relevance. Restricting their interaction to a single, final layer of encoders limits the model's ability to capture fine-grained, context-dependent correspondences that are crucial for tasks such as bioactivity prediction [45]. Without intermediate fusion, biologically relationships between modalities remain uncaptured, ultimately diminishing the interpretability, and predictive power of the model in real-world biomedical scenarios. (ii) Although prototype learning has been incorporated into multimodal molecular representation learning, most existing approaches still construct prototypes independently for each modality [46, 47]. Rather than establishing a unified prototype space that integrates both structural and semantic information following cross-modal fusion, these methods rely on separate, modality-specific prototype spaces. This decoupled design prevents the formation of consistent and biologically meaningful semantic anchors across modalities, which often results in distributional shifts and semantic misalignment even for the same molecular entity [48, 49]. As a result, such models may struggle to reliably identify or annotate functionally important substructures, rare chemotypes, or disease-associated motifs [50, 51]. These elements are crucial for a variety of bioinformatics applications, including drug discovery and biomarker identification. Without a unified semantic space to reconcile complementary molecular information, the resulting representations tend to be fragmented and poorly generalizable. This ultimately limits the effectiveness of these models in supporting translational bioinformatics and precision medicine.

To address these limitations, we propose ProtoMol, a prototype-guided multimodal framework specifically designed to enable fine-grained integration and unified semantic alignment between molecular graphs and their associated textual descriptions. Our framework is built on a dual-branch hierarchical encoder architecture, with a multilayer GNNs capturing the structural and topological characteristics of molecules, and a Transformer-based encoder processing rich functional and pharmacological annotations from molecular texts. The resulting layer-wise semantic representations effectively reflect both the chemical connectivity and the biological context of each molecule. To facilitate cross-modal understanding, we introduce a layer-wise bidirectional cross-modal attention mechanism, which progressively aligns features from both modalities at multiple semantic levels. Moreover, we establish a unified semantic prototype space composed of learnable, class-specific, and modality-invariant prototypes, serving as consistent semantic anchors across modalities. After cross-modal fusion, both graph and text representations are projected into this shared prototype space, where their alignment is supervised using a dual-objective training approach. This approach combines a prototype alignment loss, which enforces consistency between modalities based on Kullback–Leibler divergence [52–54], with a prototype contrastive loss that enhances intra-class compactness and inter-class separability. To verify the effectiveness of ProtoMol, we conduct extensive experiments on a range of molecular property prediction benchmarks. The results show that ProtoMol consistently outperforms current state-of-the-art approaches in most cases.

Our main contributions are as follows:

- We propose ProtoMol, a prototype-guided multimodal framework for multimodal molecular representation learning that jointly captures both structural and semantic information from molecular graphs and their corresponding textual descriptions, leading to improved predictive performance and enhanced interpretability.
- ProtoMol integrates dual-branch hierarchical encoders with a novel layer-wise bidirectional cross-modal attention mechanism and a unified semantic prototype space, enabling fine-grained semantic alignment and consistent, modality-invariant representation learning through prototype-level supervision.
- We demonstrate the practical effectiveness of ProtoMol through extensive experiments on diverse molecular property prediction benchmarks, showing that it consistently outperforms state-of-the-art baselines in most cases.

Materials and Methods

Preliminary

Given a molecular graph $G = (V, E, \mathbf{X})$, where V denotes the set of atoms, E represents the set of chemical bonds, and $\mathbf{X}$ is the atom feature matrix, together with a textual description $\mathbf{t}$ derived from the molecule's SMILES representation, we address the task of molecular property prediction. This task includes both classification and regression settings. For classification, the objective is to predict binary or multiclass molecular attributes, such as the toxicity or biological activity of molecules. For regression, the goal is to estimate continuous-valued properties, including the solubility and lipophilicity of molecules. The proposed ProtoMol method processes each molecular instance as a multimodal input pair $(G, \mathbf{t})$ and outputs a prediction $\hat{y} \in \mathcal{Y}$, where $\mathcal{Y}$ corresponds to either a discrete set of class labels or a continuous property value.

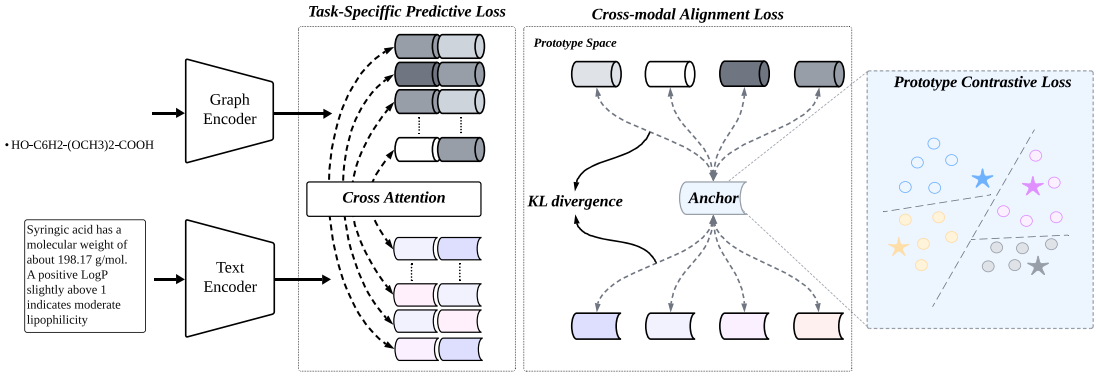


Figure 1. Overall architecture of the proposed ProtoMol.

Overview

In this section, we present ProtoMol, a prototype-guided multimodal framework designed to integrate molecular graphs and their corresponding textual descriptions for accurate molecular property prediction, as shown in Fig. 1. The overall architecture comprises three key components: (i) a dual-branch encoder, which captures hierarchical representations from both graph and text modalities using a multilayer GNN and a Transformer, respectively; (ii) a structured cross-modal interaction module, which employs layer-wise bidirectional attention to model semantic dependencies between modalities at multiple hierarchical levels; and (iii) a unified semantic prototype space that acts as modality-invariant anchors to facilitate robust cross-modal alignment. By jointly modeling hierarchical interactions and prototype-level alignment, ProtoMol enables fine-grained semantic fusion and promotes the learning of consistent and discriminative representations across modalities.

Dual-branch hierarchical representation learning

To effectively capture both structural and semantic information from molecular graphs and their associated textual descriptions, the proposed ProtoMol framework employs a dual-branch encoder that independently processes each modality within a hierarchical architecture.

Graph branch for structural representation learning

Given a molecular graph G, we employ a multilayer GNNs to encode both local structural features and global topological dependencies. Let $\mathbf{h}_v^{(l)} \in \mathbb{R}^{d_g}$ denote the representation of atom $v \in V$ at the lth GNN layer, d_g denotes the dimensionality of the node representation. The initial representation of atom v is obtained via a learnable multilayer perceptron (MLP):

$$\mathbf{h}_v^{(0)} = \mathrm{MLP}(\mathbf{x}_v), \tag{1}$$

where $\mathbf{x}_v \in \mathbb{R}^d$ denotes the raw input feature of atom v, d is the dimensionality of raw input feature. The GNN then iteratively updates each atom representation through a message-passing mechanism aggregating information from its local neighborhood [55]. At the lth layer, the representation $\mathbf{h}_v^{(l)}$ of atom v is computed as:

$$\mathbf{h}_v^{(l)} = \mathcal{C}^{(l-1)}\left(\mathbf{h}_v^{(l-1)},\ \mathcal{A}^{(l-1)}\left(\left\{\mathbf{h}_u^{(l-1)} \mid u \in \mathcal{N}(v)\right\}\right)\right),$$
$$l = 1, \ldots, L, \tag{2}$$

where $\mathcal{A}^{(l-1)}(\cdot)$ is the aggregation function (e.g. summation, mean, or attention-based pooling) over the neighbors $\mathcal{N}(v)$ of atom v, and $\mathcal{C}^{(l-1)}(\cdot)$ is the combination function that integrates the aggregated neighborhood information with the current atom representation [56].

To capture hierarchical structural information from molecular graphs, we apply a readout operation at each layer of the GNN to obtain a series of graph-level representations [55]. Specifically, at the lth layer, the graph-level representation is defined as

$$\mathbf{z}_g^{(l)} = \mathrm{READOUT}\left(\left\{\mathbf{h}_v^{(l)} \mid v \in V\right\}\right), \tag{3}$$

where $\mathrm{READOUT}(\cdot)$ denotes a permutation-invariant function such as summation, mean, or max pooling [55]. This results in a set of hierarchical graph-level representations:

$$\mathbf{Z}_g = \left\{\mathbf{z}_g^{(1)}, \ldots, \mathbf{z}_g^{(L)}\right\}, \quad \mathbf{z}_g^{(l)} \in \mathbb{R}^{d_g}.$$

Text branch for semantic representation learning

Given a textual description $\mathbf{t} = [t_1, t_2, \ldots, t_n]$ consisting of n tokens corresponding to a molecular graph G (e.g. "a benzene ring connected to a hydroxyl group and a carboxylic acid"), we leverage a pretrained large language model, such as Qwen [57], to generate context-aware semantic embeddings for the textual input. Pretrained language models are capable of capturing nuanced semantic relationships, contextual dependencies, and domain-specific chemical terminology by drawing upon knowledge acquired from large-scale text corpora [57, 58]. Specifically, each token t_i is mapped to a dense, context-dependent embedding vector $\mathbf{e}_i \in \mathbb{R}^{d_t}$ as follows:

$$\mathbf{e}_i = f_{\text{text}}(t_i), \quad i = 1, \ldots, n, \tag{4}$$

where d_t denotes the dimensionality of the token embedding, and $f_{\text{text}}(\cdot)$ represents the pretrained language model. To obtain an initial global representation that summarizes the overall semantic content of the molecular description, we apply mean pooling across all token embeddings:

$$\mathbf{z}_t^{(0)} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{e}_i, \tag{5}$$

where $\mathbf{z}_t^{(0)} \in \mathbb{R}^{d_t}$ denotes the aggregated, context-aware representation of the textual input, effectively capturing both local token-level semantics and the broader molecular context.

To capture hierarchical semantic information, we further refine the initial text representation $\mathbf{z}_t^{(0)}$ using a Transformer composed of L layers. At each layer, the representation is updated according to

$$\mathbf{z}_t^{(l)} = \text{Transformer}^{(l)}(\mathbf{z}_t^{(l-1)}), \quad l = 1, \ldots, L. \tag{6}$$

This process yields a set of hierarchical text-level representations:

$$\mathbf{Z}_t = \left\{\mathbf{z}_t^{(1)}, \ldots, \mathbf{z}_t^{(L)}\right\}, \quad \mathbf{z}_t^{(l)} \in \mathbb{R}^{d_t}.$$

Layer-wise bidirectional attention for cross-modal interaction

To capture fine-grained semantic dependencies between molecular graphs and their textual descriptions, we design a hierarchical bidirectional cross-modal attention module that enables progressive structural–semantic alignment between modalities. Different from standard cross-attention mechanisms that perform a single-stage fusion at the final layer [41, 45], our design introduces layer-wise bidirectional interaction between the hierarchical representations $\mathbf{Z}_g = \{\mathbf{z}_g^{(1)}, \ldots, \mathbf{z}_g^{(L)}\}$ and $\mathbf{Z}_t = \{\mathbf{z}_t^{(1)}, \ldots, \mathbf{z}_t^{(L)}\}$. At each layer l, the graph representation $\mathbf{z}_g^{(l)} \in \mathbb{R}^{d_g}$ is projected into the text representation space using a learnable transformation matrix $\mathbf{W}_{g\to t}^{(l)} \in \mathbb{R}^{d_t \times d_g}$:

$$\mathbf{z}_{g\to t}^{(l)} = \mathbf{W}_{g\to t}^{(l)} \mathbf{z}_g^{(l)}. \tag{7}$$

Conversely, the text representation $\mathbf{z}_t^{(l)} \in \mathbb{R}^{d_t}$ is projected into the graph representation space via a learnable transformation matrix $\mathbf{W}_{t\to g}^{(l)} \in \mathbb{R}^{d_g \times d_t}$:

$$\mathbf{z}_{t\to g}^{(l)} = \mathbf{W}_{t\to g}^{(l)} \mathbf{z}_t^{(l)}. \tag{8}$$

We then compute the layer-wise bidirectional attention from graph to text by using $\mathbf{z}_{g\to t}^{(l)}$ as the key and value, and $\mathbf{z}_t^{(l)}$ as the query:

$$\tilde{\mathbf{z}}_t^{(l)} = \text{CrossAttn}_{g\to t}^{(l)}\left(\mathbf{z}_t^{(l)}, \mathbf{z}_{g\to t}^{(l)}, \mathbf{z}_{g\to t}^{(l)}\right), \tag{9}$$

where the cross-attention operation is defined as:

$$\text{CrossAttn}\left(\mathbf{z}_t^{(l)}, \mathbf{z}_{g\to t}^{(l)}, \mathbf{z}_{g\to t}^{(l)}\right) = \text{softmax}\left(\frac{\mathbf{z}_t^{(l)}\left(\mathbf{z}_{g\to t}^{(l)}\right)^\top}{\sqrt{d_g}}\right)\mathbf{z}_{g\to t}^{(l)}. \tag{10}$$

Analogously, we compute the cross-attention from text to graph as:

$$\tilde{\mathbf{z}}_g^{(l)} = \text{CrossAttn}_{t\to g}^{(l)}\left(\mathbf{z}_g^{(l)}, \mathbf{z}_{t\to g}^{(l)}, \mathbf{z}_{t\to g}^{(l)}\right). \tag{11}$$

The final cross-modally enhanced representations for the graph and text modalities are obtained by applying a residual connection with the original modality-specific representations:

$$\hat{\mathbf{z}}_t^{(l)} = \mathbf{z}_t^{(l)} + \tilde{\mathbf{z}}_t^{(l)}, \quad \hat{\mathbf{z}}_g^{(l)} = \mathbf{z}_g^{(l)} + \tilde{\mathbf{z}}_g^{(l)}. \tag{12}$$

This layer-wise bidirectional attention mechanism facilitates progressive and structured fusion between graph and text semantics. In this process, lower layers primarily capture structural dependency transfer, while higher layers progressively refine semantic alignment. Through this hierarchical interaction, the module effectively bridges the representational gap between structural and linguistic modalities, thereby establishing a solid foundation for unified prototype learning in the subsequent stage.

Modality-invariant prototype space for cross-modal alignment

To bridge the semantic gap between graph and textual modalities, we construct a unified semantic prototype space that serves as a set of modality-invariant anchors for cross-modal alignment. This space consists of a collection of learnable, class-specific prototypes defined as

$$\mathcal{P} = \{\mathbf{p}_{c,n}\}_{c=1,n=1}^{C,N} \in \mathbb{R}^{C\times N\times d_p},$$

where each prototype $\mathbf{p}_{c,n} \in \mathbb{R}^{d_p}$ represents the nth semantic anchor for class c in the shared latent space. Here, C denotes the number of classes in the dataset, N is the number of prototypes for per class, and d_p is the dimensionality of each prototype vector. For regression tasks, we set $C = 1$ to reflect the continuous output space; for classification tasks, C equals the number of discrete class labels.

Given the cross-modally enhanced representations $\hat{\mathbf{Z}}_g = \{\hat{\mathbf{z}}_g^{(1)}, \ldots, \hat{\mathbf{z}}_g^{(L)}\}$ and $\hat{\mathbf{Z}}_t = \{\hat{\mathbf{z}}_t^{(1)}, \ldots, \hat{\mathbf{z}}_t^{(L)}\}$, we first aggregate these representations using mean pooling:

$$\bar{\mathbf{z}}_g = \frac{1}{L}\sum_{l=1}^{L}\hat{\mathbf{z}}_g^{(l)}, \quad \bar{\mathbf{z}}_t = \frac{1}{L}\sum_{l=1}^{L}\hat{\mathbf{z}}_t^{(l)}, \tag{13}$$

where $\bar{\mathbf{z}}_g \in \mathbb{R}^{d_g}$ and $\bar{\mathbf{z}}_t \in \mathbb{R}^{d_t}$ denote the aggregated graph and text representations, respectively. These aggregated representations are then projected into the unified prototype space using two modality-specific linear transformations:

$$\bar{\mathbf{z}}_{g\to p} = \mathbf{W}_{g\to p}\bar{\mathbf{z}}_g + \mathbf{b}_g, \quad \bar{\mathbf{z}}_{t\to p} = \mathbf{W}_{t\to p}\bar{\mathbf{z}}_t + \mathbf{b}_t, \tag{14}$$

where $\bar{\mathbf{z}}_{g\to p}, \bar{\mathbf{z}}_{t\to p} \in \mathbb{R}^{d_p}$ are the graph and text representations in the unified prototype space, respectively, and $\mathbf{W}_{g\to p} \in \mathbb{R}^{d_g\times d_p}$ and $\mathbf{W}_{t\to p} \in \mathbb{R}^{d_t\times d_p}$ are learnable projection matrices with modality-specific bias terms $\mathbf{b}_g \in \mathbb{R}^{d_p}$ and $\mathbf{b}_t \in \mathbb{R}^{d_p}$, respectively.

To promote prototype selectivity and semantic alignment, we compute the squared Euclidean distance between each projected representation and all prototypes. By flattening the prototype set $\mathcal{P} \in \mathbb{R}^{C\times N\times d_p}$ into a matrix of size $\mathbb{R}^{(C\cdot N)\times d_p}$, we obtain the following distances:

$$\mathbf{D}_g = \left[\left\|\bar{\mathbf{z}}_{g\to p} - \mathbf{p}_{c,n}\right\|_2^2\right]_{(c,n)=1}^{C,N}, \tag{15}$$

$$\mathbf{D}_t = \left[\left\|\bar{\mathbf{z}}_{t\to p} - \mathbf{p}_{c,n}\right\|_2^2\right]_{(c,n)=1}^{C,N}, \tag{16}$$

where $\mathbf{D}_g, \mathbf{D}_t \in \mathbb{R}^{C\cdot N}$ represent the prototype-wise squared distances for the graph and text modalities, respectively. For differentiability and numerical stability, these distances are converted to similarity scores using a log-ratio transformation:

$$\mathbf{s}_g = \log\left(\frac{\mathbf{D}_g + 1}{\mathbf{D}_g + \epsilon}\right), \quad \mathbf{s}_t = \log\left(\frac{\mathbf{D}_t + 1}{\mathbf{D}_t + \epsilon}\right), \tag{17}$$

where ϵ is a small constant to prevent division by zero.

To enhance discriminability, we retain only the top-K most responsive prototype dimensions through a sparsification step:

$$\hat{\mathbf{s}}_g = \mathrm{TopK}(\mathbf{s}_g, K), \quad \hat{\mathbf{s}}_t = \mathrm{TopK}(\mathbf{s}_t, K), \tag{18}$$

where $\hat{\mathbf{s}}_g = \{\hat{s}_g^{(1)}, \ldots, \hat{s}_g^{(K)}\}$ and $\hat{\mathbf{s}}_t = \{\hat{s}_t^{(1)}, \ldots, \hat{s}_t^{(K)}\}$ are the K-dimensional sparse vectors for the graph and text modalities, respectively, with all nontop-K entries masked to zero. These sparse vectors are then normalized into distributions over the selected prototypes using the softmax function:

$$\alpha_g^{(k)} = \frac{\exp(\hat{s}_g^{(k)})}{\sum_{j=1}^{K} \exp(\hat{s}_g^{(j)})}, \quad \alpha_t^{(k)} = \frac{\exp(\hat{s}_t^{(k)})}{\sum_{j=1}^{K} \exp(\hat{s}_t^{(j)})}, \tag{19}$$

where $\boldsymbol{\alpha}_g = \{\alpha_g^{(1)}, \ldots, \alpha_g^{(K)}\} \in \mathbb{R}^K$ and $\boldsymbol{\alpha}_t = \{\alpha_t^{(1)}, \ldots, \alpha_t^{(K)}\} \in \mathbb{R}^K$ represent the prototype distribution for the graph and text modalities, respectively. To promote semantic alignment across modalities, we minimize the Kullback–Leibler divergence between their prototype distributions:

$$\mathcal{L}_{\text{align}} = \mathrm{KL}(\boldsymbol{\alpha}_g \parallel \boldsymbol{\alpha}_t). \tag{20}$$

This prototype-level alignment objective encourages both modalities to activate the same subset of semantic prototypes, thereby promoting consistent and discriminative representations within the unified prototype embedding space.

Learning objective and optimization

To optimize the proposed framework for accurate molecular property prediction, we formulate a comprehensive training objective. In addition to the prototype alignment loss $\mathcal{L}_{\text{align}}$, we introduce two complementary components to further enhance model performance: (i) a supervised loss $\mathcal{L}_{\text{pred}}$ for downstream classification or regression tasks, and (ii) a prototype contrastive loss $\mathcal{L}_{\text{proto}}$, which encourages semantic separation between classes and compactness within each class in the prototype space.

Task-specific predictive loss

To enable task-specific supervision, we employ a predictive loss tailored to the type of downstream task. At the final prediction stage, the model leverages the modality-specific graph representations $\hat{\mathbf{z}}_g^{(i)} \in \mathbb{R}^{d_g}$, which incorporates semantic information transferred from the textual modality, and maps it into the output space through a layer-specific linear transformation:

$$\mathbf{o}^{(i)} = \mathrm{Linear}(\hat{\mathbf{z}}_g^{(i)}). \tag{21}$$

The outputs from all hierarchical layers are then aggregated to produce the final prediction:

$$\hat{\mathbf{o}} = \frac{1}{L}\sum_{i=1}^{L} \mathbf{o}^{(i)}, \tag{22}$$

where $\mathbf{o}^{(i)} \in \mathbb{R}^C$ for classification tasks (with C denoting the number of classes), or $\mathbf{o}^{(i)} \in \mathbb{R}$ for regression tasks. This design ensures that the final prediction $\hat{\mathbf{o}}$ captures both the structural dependencies modeled by the GNN and the semantic context transferred from the textual modality through hierarchical bidirectional attention, enabling a unified multimodal representation for molecular property prediction.

(1) For classification tasks, the loss $\mathcal{L}_{\text{pred}}^{\text{cl}}$ is computed as the cross-entropy between the predicted distribution and the ground-truth label $y \in \{1, \ldots, C\}$:

$$\mathcal{L}_{\text{pred}}^{\text{cl}} = -\sum_{c=1}^{C} \mathbb{I}(y = c) \log\left(\frac{\exp(\hat{\mathbf{o}}_c)}{\sum_{j=1}^{C} \exp(\hat{\mathbf{o}}_j)}\right), \tag{23}$$

where $\hat{\mathbf{o}}_c$ is the aggregated logit for class c, and $\mathbb{I}(\cdot)$ is the indicator function.

(2) For regression tasks, the final prediction $\hat{y} = \hat{\mathbf{o}} \in \mathbb{R}$ is treated as a continuous scalar output, and we employ the mean squared error (MSE) loss to minimize the difference between the predicted and true molecular property values y:

$$\mathcal{L}_{\text{pred}}^{re} = \|\hat{y} - y\|_2^2, \tag{24}$$

which directly penalizes deviations from the ground-truth value.

Prototype contrastive loss

To further enhance the discriminative capacity of the prototype space, we introduce a prototype contrastive loss that encourages modality-specific representations to align closely with semantically relevant prototypes while remaining dissimilar to those associated with other classes.

(1) For classification tasks, given the unified prototype space $\mathcal{P} = \{\mathbf{p}_{c,n}\}_{c=1,n=1}^{C,N}$, where $\mathbf{p}_{c,n} \in \mathbb{R}^{d_p}$ denotes the nth prototype of class c, each prototype is treated in turn as an anchor. For an anchor prototype $\mathbf{p}_{c,n}$, the remaining prototypes of the same class $\mathcal{P}_c \setminus \mathbf{p}_{c,n}$ are considered positive samples, while all prototypes belonging to different classes serve as negatives. The prototype-level contrastive loss for $\mathbf{p}_{c,n}$ is formulated as

$$\mathcal{L}_{\text{proto}}^{c,n} = -\log \frac{\frac{1}{N-1}\sum_{n'=1,\, n'\neq n}^{N} \exp\left(\frac{\mathbf{sim}(\mathbf{p}_{c,n}, \mathbf{p}_{c,n'})}{\tau}\right)}{\sum_{j=1}^{C}\sum_{m=1}^{N} \exp\left(\frac{\mathbf{sim}(\mathbf{p}_{c,n}, \mathbf{p}_{j,m})}{\tau}\right)}, \tag{25}$$

where $\mathbf{sim}(\cdot,\cdot)$ denotes cosine similarity and τ is a temperature hyperparameter. The overall prototype contrastive loss is then defined as the mean over all prototypes:

$$\mathcal{L}_{\text{proto}}^{\text{cl}} = \frac{1}{CN}\sum_{c=1}^{C}\sum_{n=1}^{N} \mathcal{L}_{\text{proto}}^{(c,n)}. \tag{26}$$

(2) For regression tasks, the objective is to predict a continuous value whose magnitude reflects the degree or intensity of the molecular property under consideration. For example, in the ESOL dataset, the goal is to predict aqueous solubility, where higher values correspond to greater solubility. To introduce semantic structure into the prototype space and prevent prototype collapse, we employ a two-step approach. First, we apply the K-means algorithm [59] to partition the prototypes $\mathcal{P}$ into two clusters, denoted as $\mathcal{P}_1$ and $\mathcal{P}_2$, which serve as

pseudo-classes reflecting different ranges of the target property. Then, we perform prototype-level contrastive learning within and between these clusters: prototypes belonging to the same cluster are treated as positives, while those from the other cluster are treated as negatives. The prototype contrastive loss is defined as

$$\mathcal{L}_{\text{proto}}^{(m,n)} = -\log \frac{\frac{1}{|\mathcal{P}_m|-1}\sum_{n'=1,\, n'\neq n}^{|\mathcal{P}_m|} \exp\left(\frac{\text{sim}(\mathbf{p}_{m,n},\mathbf{p}_{m,n'})}{\tau}\right)}{\sum_{j=1}^{M}\sum_{l=1}^{|\mathcal{P}_j|} \exp\left(\frac{\text{sim}(\mathbf{p}_{m,n},\mathbf{p}_{j,l})}{\tau}\right)}, \tag{27}$$

where $\textbf{sim}(\cdot,\cdot)$ denotes cosine similarity, τ is a temperature hyperparameter, and $|\mathcal{P}_m|$ is the number of prototypes in cluster m.

The overall prototype contrastive loss is then defined as the mean over all prototypes:

$$\mathcal{L}_{\text{proto}}^{\text{re}} = \frac{1}{|\mathcal{P}|}\sum_{m=1}^{M}\sum_{n=1}^{|\mathcal{P}_m|}\mathcal{L}_{\text{proto}}^{(m,n)}, \tag{28}$$

where $|\mathcal{P}|$ denotes the total number of prototypes.

Unified training objective

The overall training objective is formulated as a weighted combination of all loss components:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{align}}\mathcal{L}_{\text{align}} + \lambda_{\text{pred}}\mathcal{L}_{\text{pred}} + \lambda_{\text{proto}}\mathcal{L}_{\text{proto}}, \tag{29}$$

where λ_{align}, λ_{pred}, and λ_{proto} are tunable hyperparameters that balance the contributions of semantic alignment, task-specific supervision, and prototype-level contrastive learning, respectively. This unified loss framework enables end-to-end optimization of cross-modal consistency, predictive accuracy, and representation discriminability in a cohesive manner. The overall pipeline of ProtoMol is illustrated in Appendix 5.2.

Complexity analysis

The overall time complexity of ProtoMol is determined by the dominant components within each module. The dual-branch encoders contribute a complexity of $\mathcal{O}(L \cdot md_g^2)$ for the L-layer GNNs and $\mathcal{O}(L \cdot n^2 d_t)$ for the L-layer Transformer, where m and n denote the number of atoms and tokens, respectively. The structured cross-modal interaction module introduces an additional complexity of $\mathcal{O}(L \cdot d^2)$ due to the layer-wise bidirectional attention mechanism. The prototype alignment module incurs a complexity of $\mathcal{O}(C \cdot d_p)$, where C is the number of prototypes. Therefore, the overall time complexity of the proposed ProtoMol framework is $\mathcal{O}(L(md_g^2 + n^2 d_t + d^2) + Cd_p)$.

Results

Experimental settings

Datasets

To evaluate the effectiveness of the proposed ProtoMol, we conduct experiments on extensive benchmarks [1]. We categorized them into two task types: (i) For classification, we utilize eight molecular property prediction datasets, including ACE, BBBP, HIV, ClinTox, Tox21, MUV, SIDER, and ToxCast. These datasets cover a diverse range of biochemical properties, such as blood–brain barrier permeability (BBBP), molecular toxicity (ClinTox and Tox21), and biological activity (HIV and MUV), spanning both binary and multilabel classification settings; (ii) For regression, we adopt three datasets: ESOL, FreeSolv, and Lipophilicity, which involve predicting continuous-valued molecular properties including ESOL, hydration free energy (FreeSolv), and lipophilicity (Lipo). These tasks present unique challenges due to the fine-grained structural and semantic variations across molecules. The statistics of the above datasets is provided in Table 1. Additionally, we adopt a standard random stratified split of the dataset into training (80%), testing (10%), and validation (10%) sets.

Baselines

We compare the proposed ProtoMol with various competitive baselines across the evaluated datasets, including two textual methods: SMILES2vec [60] and SMILES-BERT [61]; fourteen graph-based methods: GraphPT [62], GraphSAGE [63], DGI [64], JOAOv2 [65], GraphCL [66], GraphLoG [67], MICRO-Graph [68], MGSSL [69], GraphFP [70], GROVE [17], SimSGT [71], MoAMa [72], Uni-Mol [73] and S-CGIB [74]; and three multimodal methods that integrate textual and graph-based representations: Tri-SGD [75], MMSG [76], and MDFCL [77]. More details about baselines are introduced in Appendix 5.1.

Evaluation metrics

Following the prior work [74], we adopt two widely recognized metrics tailored to the nature of the prediction task: the Area Under the Receiver Operating Characteristic Curve (ROC-AUC) for classification, and the Root Mean Squared Error (RMSE) for regression.

- **ROC-AUC.** For classification tasks, we report the ROC-AUC. The ROC-AUC quantifies the model's ability to distinguish between classes by integrating the trade-off between the True Positive Rate (TPR) and the False Positive Rate (FPR) across all possible thresholds. The TPR and FPR are defined as:

$$\text{TPR} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}},$$

$$\text{FPR} = \frac{\text{False Positive}}{\text{False Positive} + \text{True Negative}}.$$

  A higher ROC-AUC value indicates superior classification performance, reflecting a stronger capacity to discriminate between positive and negative classes.
- **RMSE.** For regression tasks, we utilize the RMSE to measure the accuracy of predicted continuous-valued molecular properties. RMSE is computed as:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2},$$

  where y_i and $\hat{y}_i$ denote the ground-truth and predicted values for the ith molecule, respectively, and N is the total number of samples. Lower RMSE values indicate higher regression accuracy, signifying better agreement between predicted and actual property values.

Implementation details

We implement the proposed ProtoMol using the PyTorch framework (https://pytorch.org/), and re-run all baselines within the same framework to ensure a fair comparison. All experiments are conducted on the NVIDIA RTX 4090 GPUs. We employ the Adam optimizer with a learning rate of 8×10^{-5} and a weight decay

Table 1. Statistics of the molecular property prediction benchmarks

Tasks	Dataset	Classes	Graphs	Avg. nodes	Avg. edges
Classification	HIV	2	41 127	25.5	27.5
	BACE	2	1513	34.1	36.9
	BBBP	2	2039	23.9	25.2
	Tox21	12	7831	18.6	19.8
	ToxCast	617	8575	18.7	19.4
	SIDER	27	1427	33.6	35.1
	ClinTox	2	1478	26.1	27.7
	MUV	17	93 087	24.2	26.0
Regression	Lipo	–	4200	27.0	28.5
	ESOL	–	1128	13.3	13.9
	FreeSolv	–	642	8.7	9.1

of 1×10^{-4}. A cosine annealing learning rate schedule is used, with the maximum number of epochs set to 100. The proposed ProtoMol and all baselines are trained for a total of 100 epochs with a batch size of 128. Additionally, for ProtoMol, we select GIN [55] as the backbone for the graph encoder to effectively capture complex molecular structures, and Qwen-2.5 7B [57] as the backbone for the textual encoder to extract rich semantic information from textual representations. All reported results are averaged over five independent runs.

Performance comparison

We present the results of the proposed ProtoMol with all baseline models on two types of molecular property prediction tasks: classification and regression in Tables 2 and 3.

From these tables, we observe that: (i) graph-based methods, such as S-CGIB [74] and Uni-Mol [73], consistently outperform text-based approaches by more effectively capturing intrinsic molecular topology and relational structure. While text-based models like SMILES2vec and SMILES-BERT encode molecules as linear sequences, they are inherently limited in their ability to represent the complex connectivity patterns, stereochemistry, and spatial dependencies present in molecular graphs. In contrast, graph-based approaches operate directly on molecular graphs and leverage message passing mechanisms to aggregate both local and global structural information, which enables the construction of richer and more chemically meaningful representations for molecules. (ii) Multimodal methods, such as MDFCL [77] and MMSG [76], which integrate both graph and textual information, consistently surpass graph-based models by harnessing the complementary strengths of each modality. These approaches combine the structural inductive bias of GNNs with the semantic richness provided by textual encoders, allowing the models to jointly reason over atomic connectivity, subgraph motifs, and higher-level chemical semantics present in SMILES or other text representations. (iii) The proposed ProtoMol consistently achieves superior performance over all baselines across most tasks, attributable to its architecture that integrates hierarchical cross-modal interaction with prototype-guided semantic alignment. Different from conventional methods that typically perform fusion at the final encoder layer, ProtoMol introduces a structured, layer-wise bidirectional cross-modal attention mechanism. This design enables progressive and fine-grained semantic integration between graph and textual representations at multiple levels, thereby preserving the distinctive strengths of each modality throughout the representation learning process. Furthermore, ProtoMol employs a unified semantic prototype space comprising learnable, class-specific anchors that act as shared semantic groundings for both modalities. This space not only facilitates consistent cross-modal alignment but also underpins prototype-guided contrastive learning, which enhances intra-class compactness and inter-class separability in the joint embedding space. More analysis can be found in Appendix. 5.3.

Additionally, we present comparative visualization results for ProtoMol, SMILES-BERT [61], S-CGIB [74], and MDFCL [77] on the HIV and BBBP datasets, as shown in Figs 2 (a-d) and 3 (a-d). These results demonstrate that ProtoMol yields substantially clearer class boundaries and more accurate classification outcomes compared to the baseline methods. Moreover, a quantitative evaluation using the Davies–Bouldin (DB) Index [78] shows that ProtoMol attains lower DB values (1.14 on HIV and 0.87 on BBBP) than all baselines, indicating stronger inter-class separability and more discriminative feature representations.

Flexible architecture

To assess the impact of different backbone choices for the graph and textual encoders in ProtoMol, we systematically evaluate several state-of-the-art architectures. For the graph encoder, we consider a range of message passing methods, including Graph Convolutional Network (GCN) [79], Graph Attention Network (GAT) [80], Graph Isomorphism Network (GIN) [55], and Graph Multiset Transformer (GMT) [81]. For the textual encoder, we experiment with pretrained language models such as BERT [82], LLaMA3-8B [58], and Qwen2.5-7B [57].

As illustrated in Figs 4 and 5, we observe that both GIN and Qwen2.5-7B consistently outperform other backbone settings, underscoring their superior capacity to capture the complex structural patterns and semantic nuances critical for molecular property prediction. These findings further justify our design choice of adopting GIN and Qwen2.5-7B as the core components of ProtoMol, as they provide a strong foundation for learning robust and discriminative multimodal molecular representations.

Ablation study

We conduct ablation studies to examine the contributions of each component in the proposed ProtoMol. The ablations are categorized into two groups: (1) structural components, which govern cross-modal representation learning; and (2) objective components, which determine the optimization behavior of the model.

(1) For structural components, we conduct two variants of ProtoMol: ProtoMol w/o CA: it removes the hierarchical cross-modal

Table 2. Classification results (ROC-AUC %) for molecular property prediction on the BACE, BBBP, HIV, ClinTox, Tox21, MUV, SIDER, and ToxCast datasets

Methods	BACE	BBBP	HIV	ClinTox	Tox21	MUV	SIDER	ToxCast
SMILES2vec	81.4±3.6	78.0±2.4	75.8±3.2	74.1±2.2	73.4±2.0	72.3±3.4	56.7±1.9	62.2±1.9
SMILES-BERT	83.3±3.3	80.5±2.2	77.9±2.9	76.2±2.1	75.1±1.9	73.7±3.2	58.0±1.8	64.5±1.8
GraphPT	76.0±0.5	67.1±0.5	72.7±0.7	60.1±1.2	73.4±0.6	67.9±0.6	61.2±0.7	61.7±1.2
GraphSAGE	74.3±1.4	64.7±1.1	70.6±1.7	61.6±1.3	70.3±1.6	70.8±1.6	60.2±0.8	60.0±0.8
DGI	77.8±0.5	68.4±0.6	73.6±0.5	58.6±0.8	72.7±0.2	72.1±1.2	59.0±0.6	62.8±0.5
JOAOv2	74.4±1.7	72.0±0.2	77.1±1.5	65.2±0.8	74.0±1.9	68.5±1.6	59.9±1.7	63.1±1.9
GraphCL	77.8±0.5	68.4±0.6	73.6±0.5	61.6±1.3	73.3±0.6	72.1±1.2	61.8±0.6	62.8±0.5
GraphLoG	76.6±1.0	66.8±0.3	73.8±0.3	53.8±0.9	71.6±0.5	72.5±2.0	59.1±0.5	61.5±0.4
GraphFP	80.3±3.1	72.1±1.2	75.7±1.4	76.8±1.8	77.4±1.4	71.8±1.3	65.9±3.1	69.2±1.9
MICRO-Graph	63.6±1.6	67.2±1.9	76.7±1.1	77.6±1.6	71.8±1.7	70.5±1.6	60.3±1.0	60.8±1.2
MGSSL	82.0±3.8	79.5±2.0	77.5±2.9	75.8±1.8	74.8±1.6	73.4±3.5	57.5±1.5	63.9±1.6
GROVE	81.1±0.1	87.1±0.1	75.0±0.1	72.5±0.1	68.6±0.2	67.7±0.1	57.5±0.2	64.4±0.1
SimSGT	79.8±1.3	71.5±1.8	78.1±1.1	74.1±1.1	76.2±1.3	72.8±1.5	59.7±1.3	65.8±0.8
MoAMa	81.3±1.1	85.9±0.6	78.1±0.6	77.1±1.7	78.3±0.6	72.4±1.8	62.7±0.4	68.0±1.1
Uni-Mol	85.7±0.2	72.9±0.6	80.8±0.3	91.9±1.8	79.6±0.5	**82.1±1.3**	65.9±1.3	69.6±0.1
S-CGIB	86.5±0.8	88.8±0.5	78.3±1.3	78.6±2.0	80.9±0.2	77.7±1.2	64.0±1.0	71.0±0.3
Tri_SGD	90.3±1.0	89.4±0.5	80.9±0.9	83.0±0.8	80.7±0.6	79.4±1.0	67.8±0.8	70.9±0.9
MMSG	90.8±0.5	89.7±0.2	81.0±0.3	85.0±0.7	81.0±1.0	79.3±0.7	67.8±0.5	71.0±1.4
MDFCL	86.4±1.0	78.4±1.0	78.7±1.3	**93.7±2.0**	80.5±0.7	78.5±0.7	67.7±0.6	70.9±0.8
ProtoMol	**91.4±0.3**	**90.3±0.6**	**81.2±0.2**	84.3±0.3	**81.2±0.3**	80.7±0.2	**68.1±0.5**	**71.2±0.4**

Bold results indicate the best performance.

Table 3. Regression results (RMSE) for molecular property prediction on the ESOL, FreeSolv, and Lipophilicity datasets

Methods	ESOL	FreeSolv	Lipophilicity
SMILES2vec	2.802±0.108	2.978±0.372	1.068±0.086
SMILES-BERT	2.612±0.085	2.354±0.061	0.832±0.079
GraphPT	2.954±0.087	4.023±0.039	0.982±0.052
GraphSAGE	2.368±0.070	3.192±0.023	1.085±0.061
DGI	2.953±0.049	3.033±0.026	0.970±0.023
JOAOv2	2.144±0.009	3.842±0.012	1.116±0.024
GraphCL	1.390±0.363	3.166±0.027	1.014±0.018
GraphLoG	1.542±0.026	2.335±0.052	0.932±0.052
GraphFP	2.136±0.096	2.528±0.016	1.371±0.058
MICRO-Graph	0.842±0.055	1.865±0.061	0.851±0.073
MGSSL	2.936±0.071	2.940±0.051	1.106±0.077
GROVE	1.237±0.403	2.712±0.327	0.823±0.027
SimSGT	0.932±0.026	1.953±0.038	0.771±0.041
MoAMa	1.125±0.029	2.072±0.053	1.085±0.024
Uni-Mol	0.788±0.029	1.620±0.035	0.603±0.010
S-CGIB	0.816±0.019	1.648±0.074	0.762±0.042
Tri_SGD	0.654±0.106	1.607±0.114	0.606±0.023
MMSG	0.688±0.030	1.638±0.525	**0.577±0.014**
MDFCL	0.663±0.026	1.607±0.032	0.608±0.037
ProtoMol	**0.629±0.014**	**1.522±0.044**	0.583±0.039

Bold results indicate the best performance.

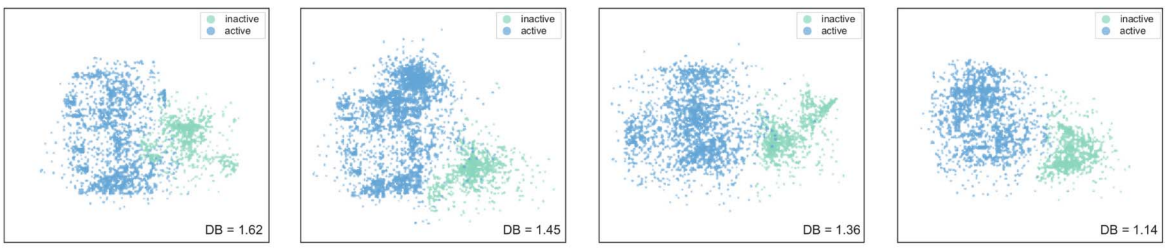


Figure 2. Visualization results of baseline methods and ProtoMol on the HIV dataset.

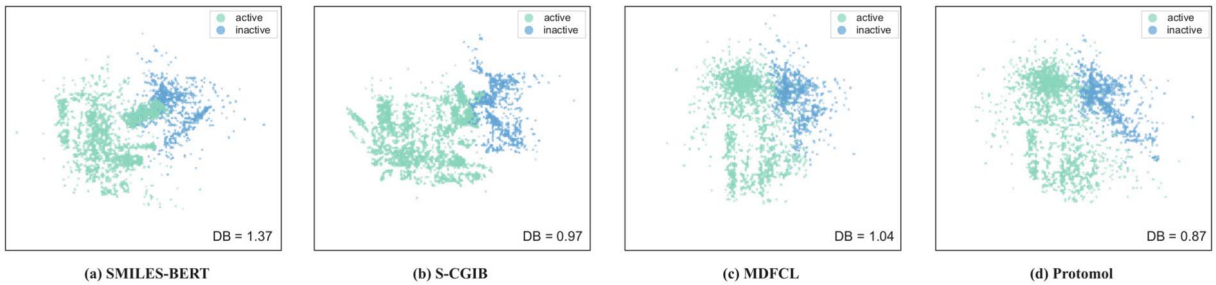


Figure 3. Visualization results of baseline methods and ProtoMol on the BBBP dataset.

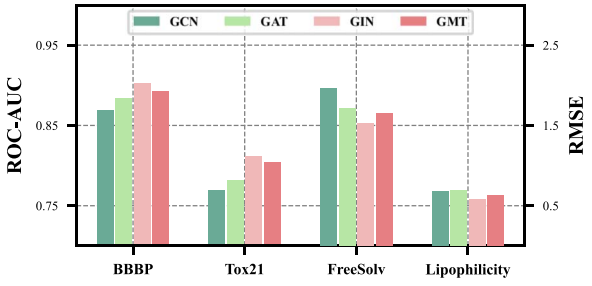


Figure 4. The performance of different backbones for the graph encoder on the BBBP and Tox21 datasets for classification tasks, and on the FreeSolv and Lipophilicity datasets for regression tasks.

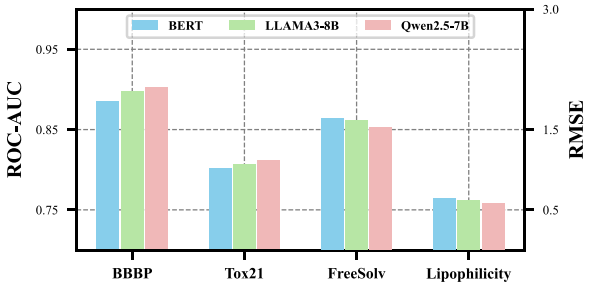


Figure 5. The performance with different backbones for textual encoder on the BBBP and Tox21 datasets for classification tasks, and on the FreeSolv and Lipophilicity datasets for regression tasks.

attention and performs cross-modal interaction only at the final layer, and ProtoMol w/o UP: it removes the unified prototype space and applies separate prototype learning for the graph and text modalities independently. Experimental results for the structural components are reported in Tables 4 and 5. From the results, we find that ProtoMol outperforms ProtoMol w/o CA, demonstrating the essential role of hierarchical cross-modal attention. This mechanism enables progressive, layer-wise semantic integration between graph and textual modalities, leading to unified and informative molecular representations that substantially improve predictive accuracy. ProtoMol also surpasses ProtoMol w/o UP, highlighting the significance of the unified prototype space. By introducing shared semantic anchors across modalities, this component enhances cross-modal alignment and enables the model to jointly capture complementary structural and semantic information. (2) For objective components, we evaluate three variants of ProtoMol: ProtoMol w/o AL, which removes the alignment loss $\mathcal{L}_{\text{align}}$ that enforces semantic coherence between graph and text representations; ProtoMol w/o CL, which eliminates the prediction loss $\mathcal{L}_{\text{pred}}$ providing task-specific supervision; and ProtoMol w/o PR, which excludes the prototype-guided contrastive loss $\mathcal{L}_{\text{proto}}$ designed to strengthen inter-class separation and enhance the discriminative structure of the embedding space. As shown in Tables 4 and 5, ProtoMol w/o AL, ProtoMol w/o CL, and ProtoMol w/o PR all exhibit noticeably degraded performance compared to the full ProtoMol, confirming the necessity of the alignment, prediction, and prototype-guided contrastive losses. These objectives collectively ensure semantic consistency, provide reliable supervised signals, and promote well-structured embedding spaces, thereby improving the discriminative power and generalization capability of ProtoMol across a wide range of molecular property prediction tasks.

Sensitivity analysis

We perform a sensitivity analysis to assess how the key hyperparameters of ProtoMol, including the alignment loss coefficient λ_{align}, the prediction loss coefficient λ_{pred}, and the prototype-guided contrastive loss coefficient λ_{proto}, as well as the number of prototypes per class N and the number of top-K activated prototype entries, collectively influence its overall performance. In particular, λ_{align} modulates the strength of semantic alignment between graph and textual representations, λ_{pred} controls the contribution of the task-specific prediction objective, and λ_{proto} determines the impact of prototype-guided contrastive learning on the structure of the joint embedding space. These parameters are critical in balancing supervision quality and model robustness. Meanwhile, the number of prototypes per class N controls the granularity of semantic abstraction within each category, and the number of top-K activated prototype entries governs how selectively the model attends to representative prototypes during alignment.

Figure 6a–e illustrates how λ_{align}, λ_{pred}, λ_{proto}, as well as the number of prototypes per class N and the number of top-K activated prototype entries, influence the performance of ProtoMol on the BACE, HIV, ESOL, and FreeSolv datasets. Each loss coefficient is varied within the range $\{0.1, 0.3, 0.5, 0.7, 0.9\}$, while N and K are varied within $\{3, 4, 5, 6, 7\}$. From the results, we observe that: (i) The performance of ProtoMol in Fig. 6a steadily improves as λ_{align} increases. Increasing the strength of semantic alignment between graph and textual modalities enables the model to learn more consistent and robust multimodal representations. This enhanced alignment bridges the modality gap and promotes more effective information integration, ultimately leading to superior predictive performance across a wide range of molecular property prediction tasks. Thus, we set λ_{align} to 0.9 as the default. (ii) As shown in Fig. 6b, increasing λ_{pred}, which governs the contribution of the supervised information, consistently improves performance across both classification and regression tasks, highlighting the importance of balancing supervision strength within the overall objective for effective model optimization. Thus, we set λ_{pred} to 0.9 as the default. (iii) Figure 6c demonstrates

Table 4. The results of ablation studies on the classification task across the BACE, BBBP, HIV, ClinTox, Tox21, MUV, SIDER, and ToxCast datasets

Methods	BACE	BBBP	HIV	ClinTox	Tox21	MUV	SIDER	ToxCast
ProtoMol w/o CA	90.1±0.4	88.7±0.6	80.7±0.6	79.6±0.4	79.5±0.6	78.1±0.7	67.2±0.7	70.4±0.8
ProtoMol w/o UP	89.4±0.3	85.6±0.7	80.1±0.6	81.4±0.6	79.3±0.4	75.2±0.9	64.7±0.8	70.2±0.7
ProtoMol w/o AL	88.5±0.5	87.9±0.6	78.4±0.8	77.9±0.8	78.1±0.7	77.0±0.8	61.0±0.7	68.7±0.7
ProtoMol w/o CL	84.3±0.7	85.3±0.7	78.9±0.4	78.2±0.5	77.3±0.5	70.3±1.2	62.1±0.8	66.0±0.7
ProtoMol w/o PR	87.3±0.5	82.9±0.7	78.1±0.6	76.9±0.5	75.7±0.4	73.1±1.1	60.2±0.9	67.1±0.5
ProtoMol	**91.4±0.3**	**90.3±0.6**	**81.2±0.2**	**84.3±0.3**	**81.2±0.3**	**80.7±0.2**	**68.1±0.5**	**71.2±0.4**

Bold results indicate the best performance.

Table 5. The results of ablation studies on the regression task across the ESOL, FreeSolv, and Lipophilicity datasets

Methods	ESOL	FreeSolv	Lipophilicity
ProtoMol w/o CA	0.712±0.097	1.657±0.082	0.638±0.084
ProtoMol w/o UP	0.691±0.072	1.657±0.088	0.627±0.069
ProtoMol w/o AL	0.705±0.083	1.865±0.074	0.695±0.082
ProtoMol w/o CL	0.810±0.107	1.698±0.094	0.666±0.072
ProtoMol w/o PR	0.735±0.079	1.739±0.082	0.817±0.061
ProtoMol	**0.629±0.014**	**1.522±0.044**	**0.583±0.039**

Bold results indicate the best performance.

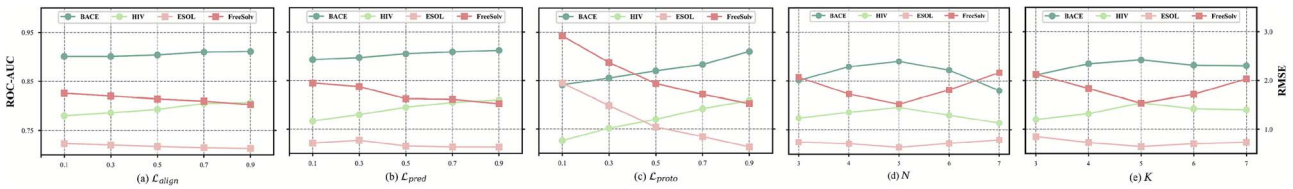


Figure 6. Hyperparameter sensitivity analysis of λ_{align}, λ_{pred}, and λ_{proto}, number of prototype N, and number of K entries on the BACE and HIV datasets for classification tasks, and on the ESOL and FreeSolv datasets for regression tasks.

that higher values of λ_{proto} further improve performance, as the prototype-guided contrastive loss encourages the formation of well-structured and discriminative embedding spaces. This improved embedding structure facilitates greater intra-class compactness and inter-class separability, enhancing the model's generalization ability across diverse molecular property prediction tasks. Thus, we set λ_{proto} to 0.9 as the default. (iv) Figure 6d and e illustrates that the performance of ProtoMol initially improves and then gradually declines as the number of prototypes per class N and the number of top-K activated prototype entries increase. Moderate settings ($N = 5$, $K = 5$) yield the best results, suggesting that an optimal trade-off between semantic diversity and representational sparsity is essential for effective prototype learning. Insufficient values of N or K restrict the expressiveness of the prototype space, whereas excessively large values introduce redundancy and noise, undermining discriminative capability. Therefore, we set $N = 5$ and $K = 5$ as the default configuration in all experiments.

Conclusion

In this work, we introduced ProtoMol, a prototype-guided cross-modal molecular representation learning framework that enables fine-grained semantic integration between molecular graphs and textual descriptions. By leveraging dual-branch hierarchical encoders and a structured, layer-wise bidirectional cross-modal attention mechanism, ProtoMol captures complex structural and semantic dependencies at multiple levels. The introduction of a unified semantic prototype space further ensures robust modality-invariant alignment and enhances the discriminative power of molecular embeddings. Extensive experiments across diverse molecular property prediction benchmarks demonstrate that ProtoMol consistently outperforms state-of-the-art methods in both classification and regression settings, advancing the field of multimodal molecular representation learning. In the future, we plan to extend ProtoMol to support additional modalities, such as protein structures and biological pathways, and to explore its application in large-scale, real-world bioinformatics scenarios including drug–target interaction prediction and functional annotation.

Key Points

- We introduce ProtoMol, a prototype-guided multimodal framework that jointly encodes molecular graphs and textual descriptions, achieving fine-grained semantic integration and unified alignment across structural and semantic modalities.
- ProtoMol features dual-branch hierarchical encoders with a layer-wise bidirectional cross-modal attention mechanism, alongside a unified, learnable prototype space that serves as modality-invariant anchors, facilitating consistent and discriminative molecular representations.
- Comprehensive experiments on diverse molecular property prediction benchmarks demonstrate that ProtoMol consistently surpasses state-of-the-art methods, underscoring the critical role of hierarchical cross-modal interaction and prototype-level alignment in enhancing predictive accuracy and model generalization.

Supplementary data

Supplementary data is available at *Briefings in Bioinformatics* online.

Conflict of interest: None declared.

Funding

None declared.

Data availability

The datasets used in this study are available at https://github.com/zky04/Protomol.

References

1. Zhenqin W, Ramsundar B, Feinberg EN. *et al.* MoleculeNet: a benchmark for molecular machine learning. *Chem Sci* 2017; **9**: 513–30.
2. McGibbon M, Shave S, Dong J. *et al.* From intuition to AI: evolution of small molecule representations in drug discovery. *Brief Bioinform* 2024; **25**:bbad422.
3. Lipinski CA, Lombardo F, Dominy BW. *et al.* Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv Drug Deliv Rev* 1997; **23**:3–25. https://doi.org/10.1016/S0169-409X(96)00423-1
4. Wang R, Yang M, Shen Y. Bridging molecular graphs and large language models. *Proceedings of the AAAI Conference on Artificial Intelligence* 2025; **39**:21234–42. https://doi.org/10.1609/aaai.v39i20.35422
5. Radford A, Kim J W, Hallacy C. *et al.* Learning Transferable Visual Models From Natural Language Supervision[J]. arXiv preprint arXiv:2103.00020, 2021.
6. Wang H, Zhang A, Zhong Y. *et al.* Chain-aware graph neural networks for molecular property prediction. *Bioinformatics* 2024; **40**:btae574. https://doi.org/10.1093/bioinformatics/btae574
7. Duvenaud D, Maclaurin D, Aguilera-Iparraguirre J. *et al.* Convolutional networks on graphs for learning molecular fingerprints[C]. *Proceedings of the 29th International Conference on Neural Information Processing Systems-Volume 2*. 2015: 2224–32.
8. Weininger D. Smiles, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J Chem Inf Comput Sci* 1988; **28**:31–6. https://doi.org/10.1021/ci00057a005
9. Flores-Hernandez H, Martinez-Ledesma E. A systematic review of deep learning chemical language models in recent era. *J Chem* 2024; **16**:129. https://doi.org/10.1186/s13321-024-00916-y
10. Schütt KT, Kindermans PJ, Sauceda HE. *et al.* SchNet: A continuous-filter convolutional neural network for modeling quantum interactions[J]. arXiv preprint arXiv:1706.08566, 2017.
11. Zhou G, Gao Z, Ding Q. *et al.* Uni-Mol: A Universal 3D Molecular Representation Learning Framework[C]. *The Eleventh International Conference on Learning Representations*.
12. Zheng S, He J, Chang Liu Y. *et al.* Predicting equilibrium distributions for molecular systems with deep learning. *Nat Mach Intell* 2024; **6**:558–67. https://doi.org/10.1038/s42256-024-00837-3
13. Gómez-Bombarelli R, Wei JN, Duvenaud D. *et al.* Automatic chemical design using a data-driven continuous representation of molecules. *ACS Central Sci* 2018; **4**:268–76. https://doi.org/10.1021/acscentsci.7b00572
14. Chithrananda S, Grand G, Ramsundar B. ChemBERTa: large-scale self-supervised pretraining for molecular property prediction[J]. arXiv preprint arXiv:2010.09885, 2020.
15. Sadeghi S, Bui A, Forooghi A. *et al.* Can large language models understand molecules? *BMC Bioinform* 2024; **25**:225. https://doi.org/10.1186/s12859-024-05847-x
16. Liu S, Nie W, Wang C. *et al.* Multi-modal molecule structure–text model for text-based retrieval and editing. *Nat Mach Intell* 2023; **5**:1447–57. https://doi.org/10.1038/s42256-023-00759-6
17. Rong Y, Bian Y, Tingyang X. *et al.* Self-supervised graph transformer on large-scale molecular data. *Adv Neural Inf Process Syst* 2020; **33**:12559–71.
18. Shao M, Wang Y, Shen X. *et al.* Enhanced Molecular Property Prediction with SMILES and Graph Aligned Contrastive Learning[C]. *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Singapore: Springer Nature Singapore, 2025, 117–29.
19. Elman JL. Finding structure in time. *Cognit Sci* 1990; **14**:179–211. https://doi.org/10.1207/s15516709cog1402_1
20. LeCun Y, Bottou L, Bengio Y. *et al.* Gradient-based learning applied to document recognition. *Proc IEEE* 2002; **86**:2278–324. https://doi.org/10.1109/5.726791
21. Vaswani A, Shazeer N, Parmar N. *et al.* Attention Is All You Need[J]. arXiv preprint arXiv:1706.03762, 2017.
22. Yoshikai Y, Mizuno T, Nemoto S. *et al.* Difficulty in chirality recognition for transformer architectures learning chemical structures from string representations. *Nat Commun* 2024; **15**:1197. https://doi.org/10.1038/s41467-024-45102-8
23. Guo T, Nan B, Liang Z. *et al.* What can large language models do in chemistry? A comprehensive benchmark on eight tasks. *Adv Neural Inf Process Syst* 2023; **36**:59662–88.
24. Ucak UV, Ashyrmamatov I, Lee J. Improving the quality of chemical language model outcomes with atom-in-SMILES tokenization[J]. *J Cheminform*, 2023;**15**:55. https://doi.org/10.1186/s13321-023-00725-9
25. Tang X, Zhao Q, Wang J. *et al.* MolFCL: predicting molecular properties through chemistry-guided contrastive and prompt learning. *Bioinformatics* 2025; **41**:btaf061.
26. Rogers D, Hahn M. Extended-connectivity fingerprints. *J Chem Inf Model* 2010; **50**:742–54. https://doi.org/10.1021/ci100050t
27. Ganeeva V, Khrabrov K, Kadurin A. *et al.* Chemical language models have problems with chemistry: a case study on molecule captioning task. In: *The Second Tiny Papers Track at ICLR* 2024; **2024**.
28. Krenn M, Häse F, Nigam AK. *et al.* Self-referencing embedded strings (selfies): a 100% robust molecular string representation. *Mach Learn Sci Technol* 2020; **1**:045024.
29. Zhao B, Xu W, Guan J. *et al.* Molecular Property Prediction Based on Graph Structure Learning[J]. arXiv preprint arXiv:2312.16855, 2023. https://doi.org/10.1093/bioinformatics/btae304
30. Boulougouri M, Vandergheynst P, Probst D. Molecular set representation learning. *Nat Mach Intell* 2024; **6**:754–63. https://doi.org/10.1038/s42256-024-00856-0
31. Gilmer J, Schoenholz SS, Riley PF. *et al.* Message passing neural networks[M]. *Machine learning meets quantum physics*. Cham: Springer International Publishing, 2020, 199–214.
32. Schütt KT, Sauceda HE, Kindermans P-J. *et al.* SchNet–a deep learning architecture for molecules and materials. *J Chem Phys* 2018; **148**:241722. https://doi.org/10.1063/1.5019779
33. Moayedpour S, Broadbent J, Riahi S. *et al.* Representations of lipid nanoparticles using large language models for transfection

efficiency prediction. *Bioinformatics* 2024; **40**:btae342. https://doi.org/10.1093/bioinformatics/btae342
34. Wang J, Guan J, Zhou S. Molecular property prediction by contrastive learning with attention-guided positive sample selection. *Bioinformatics* 2023; **39**:btad258. https://doi.org/10.1093/bioinformatics/btad258
35. He G, Liu S, Liu Z. *et al.* Prototype-based contrastive substructure identification for molecular property prediction. *Brief Bioinform* 2024; **25**:bbae565. https://doi.org/10.1093/bib/bbae565
36. Zhang R, Chao W, Yang Q. *et al.* MolFeSCue: enhancing molecular property prediction in data-limited and imbalanced contexts using few-shot and contrastive learning. *Bioinformatics* 2024; **40**:btae118.
37. Snell J, Swersky K, Zemel RS. Prototypical Networks for Few-shot Learning[J]. arXiv preprint arXiv:1703.05175, 2017.
38. Hou L, Xiang H, Zeng X. *et al.* Attribute-guided prototype network for few-shot molecular property prediction. *Brief Bioinform* 2024; **25**:bbae394. https://doi.org/10.1093/bib/bbae394
39. Seo S, Kim S, Park C. Interpretable prototype-based graph information bottleneck. *Adv Neural Inf Process Syst* 2023; **36**:76737–48.
40. Chen C, Li O, Tao C. *et al.* This Looks Like That: Deep Learning for Interpretable Image Recognition[J]. arXiv preprint arXiv:1806.10574, 2018.
41. Han X, Zhang Z, Bai C. *et al.* Integrated multimodal hierarchical fusion and meta-learning for enhanced molecular property prediction. *Brief Bioinform* 2025; **26**:bbaf251. https://doi.org/10.1093/bib/bbaf251
42. Lu J, Batra D, Parikh D. *et al.* ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks[J]. arXiv preprint arXiv:1908.02265, 2019.
43. Zhang R, Lin Y, Yijia W. *et al.* MvMRL: a multi-view molecular representation learning method for molecular property prediction. *Brief Bioinform* 2024; **25**:bbae298.
44. Morgan HL. The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. *J Chem Doc* 1965; **5**:107–13. https://doi.org/10.1021/c160017a018
45. Rollins ZA, Cheng AC, Metwally E. MolPROP: molecular property prediction with multimodal language and graph fusion. *J Cheminf* 2024; **16**:56.
46. Fan Y, Xu W, Wang H. *et al.* PMR: Prototypical Modal Rebalance for Multimodal Learning[J]. arXiv preprint arXiv:2211.07089, 2022.
47. Vinyals O, Blundell C, Lillicrap T. *et al.* Matching Networks for One Shot Learning[J]. arXiv preprint arXiv:1606.04080, 2016.
48. Wang Z, Zhao Y, Huang H. *et al.* Connecting multi-modal contrastive representations. *Adv Neural Inf Process Syst* 2023; **36**:22099–114.
49. Frome A, Corrado GS, Shlens J. *et al.* DeViSE: a deep visual-semantic embedding model[C]. *Proceedings of the 27th International Conference on Neural Information Processing Systems-Volume 2*. 2013, 2121–29.
50. Amara K, Rodríguez-Pérez R, Jiménez-Luna J. Explaining compound activity predictions with a substructure-aware loss for graph neural networks. *J Chem* 2023; **15**:67. https://doi.org/10.1186/s13321-023-00733-9
51. Bemis GW, Murcko MA. The properties of known drugs. 1. Molecular frameworks. *J Med Chem* 1996; **39**:2887–93. https://doi.org/10.1021/jm9602928
52. Zhuang W, Huang X, Zhang X. *et al.* Math-puma: progressive upward multimodal alignment to enhance mathematical reasoning. *Proceedings of the AAAI Conference on Artificial Intelligence* 2025; **39**:26183–91. https://doi.org/10.1609/aaai.v39i24.34815
53. Zhang C, Stepputtis S, Sycara K. *et al.* Dual prototype evolving for test-time generalization of vision-language models. *Adv Neural Inf Process Syst* 2024; **37**:32111–36.
54. Li J, Pan Z, Xiong C. *et al.* Prototypical contrastive learning of unsupervised representations. 2020.
55. Xu K, Hu W, Leskovec J. *et al.* How powerful are graph neural networks?[J]. arXiv preprint arXiv:1810.00826, 2018.
56. Wang Y, Liang V, Yin N. *et al.* SGAC: A Graph Neural Network Framework for Imbalanced and Structure-Aware AMP Classification[J]. arXiv preprint arXiv:2412.16276, 2024.
57. Yang A, Li A, Yang B. *et al.* Qwen3 technical report[J]. arXiv preprint arXiv:2505.09388, 2025.
58. Grattafiori A, Dubey A, Jauhri A. *et al.* The llama 3 herd of models[J]. arXiv preprint arXiv:2407.21783, 2024.
59. Hartigan JA, Wong MA. Algorithm as 136: a k-means clustering algorithm. *J R Stat Soc Ser C Appl Stat* 1979; **28**:100–8.
60. Goh GB, Hodas NO, Siegel C. *et al.* Smiles2vec: An interpretable general-purpose deep neural network for predicting chemical properties[J]. arXiv preprint arXiv:1712.02034, 2017.
61. Wang S, Guo Y, Wang Y. *et al.* Smiles-bert: large scale unsupervised pre-training for molecular property prediction[C]. *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics* 2019, 429–36.
62. Hu W, Liu B, Gomes J. *et al.* Strategies for pre-training graph neural networks[J]. arXiv preprint arXiv:1905.12265, 2019.
63. Hamilton WL, Ying R, Leskovec J. Inductive Representation Learning on Large Graphs[J]. arXiv preprint arXiv:1706.02216, 2017.
64. Veličković P, Fedus W, Hamilton WL. *et al.* Deep graph infomax[J]. arXiv preprint arXiv:1809.10341, 2018.
65. You Y, Chen T, Shen Y. *et al.* Graph Contrastive Learning Automated[J]. arXiv preprint arXiv:2106.07594, 2021.
66. You Y, Chen T, Sui Y. *et al.* Graph Contrastive Learning with Augmentations[J]. arXiv preprint arXiv:2010.13902, 2020.
67. Xu M, Wang H, Ni B. *et al.* Self-supervised Graph-level Representation Learning with Local and Global Structure[J]. arXiv preprint arXiv:2106.04113, 2021.
68. Zhang S, Hu Z, Subramonian A. *et al.* Motif-driven contrastive learning of graph representations[J]. arXiv preprint arXiv:2012.12533, 2020.
69. Zhang Z, Liu Q, Wang H. *et al.* Motif-based graph self-supervised learning for molecular property prediction. *Adv Neural Inf Process Syst* 2021; **34**:15870–82.
70. Luong K-D, Singh AK. Fragment-based pretraining and finetuning on molecular graphs. *Adv Neural Inf Process Syst* 2023; **36**:17584–601.
71. Liu Z, Shi Y, Zhang A. *et al.* Rethinking tokenizer and decoder in masked graph modeling for molecules. *Adv Neural Inf Process Syst* 2023; **36**:25854–75.
72. Inae E, Liu G, Jiang M. Motif-aware attribute masking for molecular graph pre-training[J]. arXiv preprint arXiv:2309.04589, 2023.
73. Zhou G, Gao Z, Ding Q. *et al.* Uni-Mol: A Universal 3D Molecular Representation Learning Framework[C]. *The Eleventh International Conference on Learning Representations*.
74. Hoang VT, Lee O. Pre-training Graph Neural Networks on Molecules by Using Subgraph-Conditioned Graph Information Bottleneck[J]. arXiv preprint arXiv:2412.15589, 2024.
75. Xiaohua L, Xie L, Lei X. *et al.* Multimodal fused deep learning for drug property prediction: integrating chemical language and molecular graph. *Comput Struct Biotechnol J* 2024; **23**:1666–79.
76. Tianyu W, Tang Y, Sun Q. *et al.* Molecular joint representation learning via multi-modal information of smiles and graphs.

IEEE/ACM Trans Comput Biol Bioinform 2023; **20**:3044–55. https://doi.org/10.1109/TCBB.2023.3253862
77. Gong X, Liu M, Liu Q. *et al.* MDFCL: multimodal data fusion-based graph contrastive learning framework for molecular property prediction. *Pattern Recognit* 2025; **163**:111463. https://doi.org/10.1016/j.patcog.2025.111463
78. Xiao J, Jianfeng L, Li X. Davies bouldin index based hierarchical initialization k-means. *Intell Data Anal* 2017; **21**:1327–38. https://doi.org/10.3233/IDA-163129
79. Kipf TN. Semi-supervised classification with graph convolutional networks[J]. arXiv preprint arXiv:1609.02907, 2016.
80. Veličković P, Cucurull G, Casanova A. *et al.* Graph attention networks[J]. arXiv preprint arXiv:1710.10903, 2017.
81. Baek J, Kang M, Hwang SJ. Accurate learning of graph representations with graph multiset pooling[J]. arXiv preprint arXiv:2102.11533, 2021.
82. Devlin J, Chang MW, Lee K. *et al.* Bert: Pre-training of deep bidirectional transformers for language understanding[C]. *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies*, volume 1 (long and short papers). 2019, 4171–86.

© The Author(s) 2025. Published by Oxford University Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.
Briefings in Bioinformatics, 2025, **26(6)**, bbaf629
https://doi.org/10.1093/bib/bbaf629
Problem Solving Protocol