LEARNING THE NEIGHBORHOOD: CONTRAST-FREE MULTIMODAL SELF-SUPERVISED MOLECULAR GRAPH PRETRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

High-quality molecular representations are essential for property prediction and molecular design, yet large labeled datasets remain scarce. While self-supervised pretraining on molecular graphs has shown promise, many existing approaches either depend on hand-crafted augmentations or complex generative objectives, and often rely solely on 2D topology, leaving valuable 3D structural information underutilized. To address this gap, we introduce C-FREE (***C**ontrast-**F**ree **R**epresentation learning on **E**go-n**e**ts*), a simple framework that integrates 2D graphs with ensembles of 3D conformers. C-FREE learns molecular representations by predicting subgraph embeddings from their complementary neighborhoods in the latent space, using fixed-radius ego-nets as modeling units across different conformers. This design allows us to integrate both geometric and topological information within a hybrid Graph Neural Network (GNN)-Transformer backbone, without negatives, positional encodings, or expensive pre-processing. Pretraining on the GEOM dataset, which provides rich 3D conformational diversity, C-FREE achieves state-of-the-art results on MoleculeNet, surpassing contrastive, generative, and other multimodal self-supervised methods. Fine-tuning across datasets with diverse sizes and molecule types further demonstrates that pretraining transfers effectively to new chemical domains, highlighting the importance of 3D-informed molecular representations. We will make our code and checkpoints publicly available for the final version of the paper.

1 INTRODUCTION

High-quality molecular representations are critical for predicting properties, interpreting chemical behavior, and accelerating compound discovery (Wigh et al., 2022; Elton et al., 2019). Many existing approaches, however, rely on a single modality, such as SMILES strings (Hirohara et al., 2018; Wang et al., 2019), 2D graph structures (Gilmer et al., 2017; Kipf & Welling, 2017; Xu et al., 2019), or 3D conformations (Schütt et al., 2017; Gasteiger et al., 2021). While effective, each of these methods captures only part of the molecular information and overlooks complementary aspects available in other modalities (Liu et al., 2023b). Beyond modality limitations, these models often require large, curated datasets, which restricts their use in low-data settings.

Because such curated datasets are often unavailable, especially in low-data regimes, self-supervised learning (SSL) provides a promising alternative to fully supervised training. Recent advances in vision and language modeling (Devlin et al., 2019; Chen et al., 2020; Grill et al., 2020a; Caron et al., 2021; He et al., 2022; LeCun & Courant, 2022) have motivated similar methods for molecular graphs, especially approaches that aim to combine structural and geometric information. While these approaches have advanced molecular representation learning, each comes with trade-offs: contrastive methods hinge on carefully chosen negative samples (You et al., 2021b;a), generative methods often require discrete reconstruction in the input graph space with graph tokenization (Hu et al., 2020b), and latent-predictive methods can depend on augmentations or expensive procedures such as clustering (Skenderi et al., 2025). These challenges motivate simpler predictive frameworks that combine 2D topology and 3D conformations without relying on negatives or input-space reconstruction, and that work reliably across both high- and low-data settings.

Current work. Our self-supervised framework C-FREE (Contrast-Free Representation learning on Ego-nets) adopts a predictive learning strategy with subgraphs as the basic modeling unit. It is

motivated by three goals: (i) *avoiding computationally intensive or ambiguous design choices*, including expensive augmentations, heavy subgraph-construction procedures, and complex negative-sampling schemes. For example, clustering-based subgraph algorithms such as METIS (Skenderi et al., 2025) can be costly, and defining suitable augmentations or negatives is often non-trivial, since molecules with nearly identical structures (e.g., chiral isomers) may still have very different properties; (ii) *leveraging the success of subgraph-based methods in 2D graph supervised learning* (Bevilacqua et al., 2022; Wollschläger et al., 2024), which suggest that aggregating information from substructures can yield richer graph-level representations, and (iii) *harnessing the benefits of multi-modal architectures in the supervised setting* (Zhu et al., 2024; Nguyen et al., 2024; Manolache et al., 2024). Since many molecular properties depend on multiple conformations and their probabilities (Cao et al., 2022), using multiple high-probability conformers alongside 2D topology helps capture this variability and improves predictive performance. Building on JEPA (Assran et al., 2023) and Equivariant Subgraph Aggregation Networks (ESAN) (Bevilacqua et al., 2022), our method segments graphs into disjoint subgraphs, similar to image patches or language tokens, and learns to align each subgraph with its context in latent space. Unlike GraphJEPA (Skenderi et al., 2025) and I-JEPA (Assran et al., 2023), it avoids positional encodings, hierarchical objectives, and costly clustering, and instead leverages the inductive bias of 2D and 3D encoders together with subgraph-based pre-training to learn rich embeddings. Our contributions are as follows:

1. **A new multi-modal pretraining task for molecular graphs.** We introduce a broadly applicable predictive objective based on k-EgoNet subgraphs, avoiding costly hand-crafted augmentations and utilizing both 2D and 3D views of the molecule.
2. **Robust performance in both multimodal and 2D-only settings.** Our framework leverages 2D topology together with multiple 3D conformations when available, but also performs strongly in purely 2D settings where conformers are absent.
3. **A simple and effective training scheme.** We adopt non-contrastive predictive learning, avoiding the pre-train/fine-tune mismatch and removing the need for negative samples or augmentations. Moreover, when fine-tuning, our framework simulates ESAN (Bevilacqua et al., 2022) and is provably more expressive than 1-WL (Weisfeiler & Leman, 1968)
4. **State-of-the-art results.** Our approach matches or surpasses other self-supervised models under both linear-probe evaluation, where the backbone is frozen, and full fine-tuning. It achieves the best average performance on MoleculeNet (Wu et al., 2018) and shows strong transfer to novel multimodal molecular benchmarks such as MARCEL (Zhu et al., 2024).

2 Related Work

Existing approaches to graph self-supervised learning can be grouped into three main categories: *contrastive learning*, *generative pre-training*, and *latent representation learning*. Each of these has also been extended to molecular graphs, with varying degrees of multimodal integration.

Contrastive learning aligns representations of similar instances while pushing apart dissimilar ones and has become central to graph representation learning. GraphCL (You et al., 2021b) and JOAO (You et al., 2021a) pioneered this idea through graph augmentations, while InfoGraph (Sun et al., 2020) maximized mutual information across views. Extensions to molecules incorporate 3D information: GraphMVP (Liu et al., 2022a) aligns 2D topology and 3D conformations with generative objectives, MoleculeSDE (Liu et al., 2023a) introduces symmetry-aware stochastic differential equations, and 3D InfoMax (Stärk et al., 2022) encodes 3D from 2D via mutual information. While effective (Wang et al., 2023a), these methods depend on negative samples and large batches (You et al., 2021b), a limitation exacerbated by irregular graph structures.

Generative pre-training forms the second category of self-supervised learning, where models reconstructs masked or missing components of a graph from its surrounding context. Early works such as AttrMask (Hu et al., 2020a), ContextPred (Hu et al., 2020a), and EdgePred (Hamilton et al., 2017) predict attributes or edges, while GPT-GNN (Hu et al., 2020b) and GROVER (Rong et al., 2020) extend to autoregressive reconstruction and chemically informed motif prediction. More recent methods incorporate multimodal and geometric signals: unified cross-modal generation of 2D/3D (Zhu et al., 2022), modality integration via MoleBlend (Yu et al., 2023), and geometry-aware prediction in 3D PGT (Wang et al., 2023b). Despite their promise, generative methods must reconstruct both

discrete graph structure and continuous features, and autoregressive variants are further complicated by the lack of a natural ordering over graph nodes.

Latent representation learning forms the third category of self-supervised methods. Instead of reconstructing raw graph structures or features, these approaches predict target embeddings directly in latent space, yielding compact, denoised, and often across multimodal representations. BGRL (Thakoor et al., 2023) employs a bootstrapped online–target encoder scheme under augmentations, while LaGraph (Xie et al., 2022) frames the task as latent graph prediction, optimizing an upper bound with context-aware regularization on masked nodes. While latent prediction methods avoid the costly generation of negatives, they remain sensitive to augmentation quality and model update stability, and are prone to representation collapse (Assran et al., 2023; Grill et al., 2020b).

Within latent representation learning, GraphJEPA (Skenderi et al., 2025) extends the Joint Embedding Predictive Architecture (JEPA) (Assran et al., 2023) to graphs by masking METIS-generated clusters and predicting them as patch-like substructures, while also encoding hierarchical information via hyperbolic subgraph coordinates. While effective, this approach incurs significant computational cost and depends on auxiliary components —such as clustering, hierarchical encodings, positional embeddings— that add complexity without being clearly essential for representation learning.

Another direction explores large-scale supervised pretraining on massive labeled molecular datasets, aiming to transfer knowledge to downstream tasks (Beaini et al., 2024). While effective in some cases, this approach still depends on labeled data and may be domain-specific, since source and target distributions can differ. In contrast, self-supervised methods avoid this reliance on labels and can transfer more flexibly. Importantly, the two strategies are complementary: self-supervised pretraining can provide strong initializations that are later fine-tuned on labeled data.

To the best of our knowledge, our work is the first non-contrastive, non-generative predictive framework for multimodal molecular representation learning. We now turn to its design.

3 Contrast-Free Multimodal Self-Supervised Pretraining

In the following, we outline our proposed training pipeline, illustrated in Fig. 2. Unlike most generative methods (Hu et al., 2020a; Hamilton et al., 2017), we apply our training objective fully in the latent space, without reconstructing the original features of the masked components.

The core principle of our approach is to learn representations by aligning the embedding of a target view with that of a related context view. Specifically, we represent each molecule as a 2D graph $G = (V, E)$, where V is the set of nodes (atoms) and E the set of edges (covalent bonds). For each atom $v \in V$, we include 3D coordinates $r_v \in \mathbb{R}^3$, taken from multiple conformers of the molecule. Using these graph and geometric features, we construct complementary context–target views that serve as inputs for our contrast-free pretraining scheme. Each 2D and 3D view is encoded independently, and the resulting embeddings are concatenated into a single multimodal sequence processed by a transformer (Vaswani et al., 2017). We then align the embedding of the target subgraph with that of its associated context subgraph. This design is loosely inspired by ESAN (Bevilacqua et al., 2022), but adopts a simplified variant: we use fixed-radius ego-nets to obtain complementary views during pretraining, and at fine-tuning we evaluate both linear probing on whole-graph embeddings and an aggregation of subgraph embeddings using DeepSets (Zaheer et al., 2017).

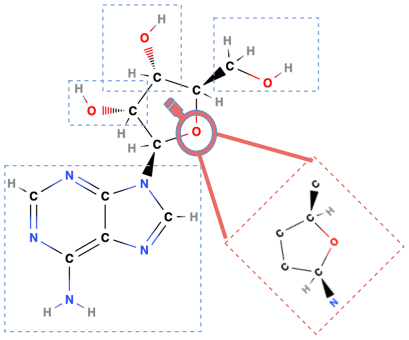

Figure 1: To generate subgraphs, we sample a random node from the original graph (here, the oxygen atom) and extract its 2-EgoNet as the context subgraph (outlined by red square). The remaining components (outlined by blue squares) constitute the target subgraph.

Context–Target View Generation. We generate complementary 2D views by sampling k-EgoNets, where the k-hop neighborhood of a node defines one subgraph and the remaining nodes and edges define its complement (see Fig. 1). The 3D coordinates are added to generate the corresponding 3D views for all the conformers. Either view can serve as the target while the other acts as context, and their roles are alternated during training to avoid prediction bias. We adopt fixed-radius

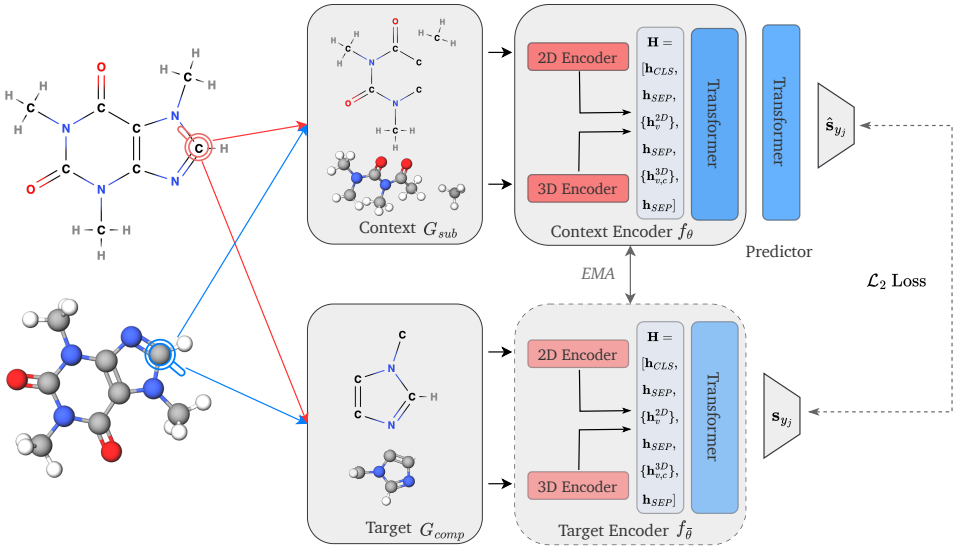

Figure 2: From each molecular graph, we sample a random node and extract its k-EgoNet (Bevilacqua et al., 2022) with $k \in \{3, 4\}$ to form complementary context and target subgraphs. Both 2D and 3D views are encoded with a GINE and a SchNet, concatenated, and passed through a transformer; the context embedding is further processed by a predictor to estimate the target. Training minimizes the mean squared $\mathcal{L}_2$ loss between predicted and encoded targets, with the target encoder updated as an exponential moving average (EMA) of the context encoder (Grill et al., 2020b; Assran et al., 2023). For clarity, only one 3D conformation is shown, though in practice we use three. During fine-tuning, the target encoder serves as the pretrained backbone for graph embeddings, with lightweight task-specific heads added. See Appendix Fig. 5 for an illustration.

neighborhoods with $k \in \{3, 4\}$, analogous to fixed-size patches in vision-based methods (Assran et al., 2023). Although graphs vary in size and structure, this ensures that each subgraph captures a comparable amount of local information. To further diversify training, we sample multiple nodes $v_1, v_2, \ldots, v_n$ per molecule and construct their corresponding k-EgoNets $E(v_1), E(v_2), \ldots, E(v_n)$, yielding multiple complementary context–target pairs without increasing dataset size.

Context Encoder. We aim to learn subgraph representations that generalize to whole-molecule embeddings. Following the architecture proposed in Manolache et al. (2024), we use a message-passing neural network (MPNN) with GINE (Hu et al., 2020a; Xu et al., 2019) as the 2D encoder, and SchNet (Schütt et al., 2017) as the 3D encoder used to process multiple conformers. From GINE, we obtain node-level embeddings $\{\mathbf{h}_v^{2D}\}$ for all atoms in the subgraph by averaging their intermediate representations across layers. From SchNet, we extract node-level embeddings $\{\mathbf{h}_{v,c}^{3D}\}$ for each conformer c, preserving per-atom detail across conformations. To build the multimodal sequence, we prepend a learnable classification token $\mathbf{h}_{CLS}$ and insert a learnable separation token $\mathbf{h}_{SEP}$ between the 2D and 3D components, resulting in the following multi-modal sequence:

$$\mathbf{H} = [\mathbf{h}_{CLS}, \mathbf{h}_{SEP}, \{\mathbf{h}_v^{2D}\}, \mathbf{h}_{SEP}, \{\mathbf{h}_{v,c}^{3D}\}, \mathbf{h}_{SEP}]$$

To distinguish between modalities, we add learnable modality embeddings that mark whether a token comes from the 2D or 3D graph. The full sequence is then passed through a Transformer with multiple self-attention layers to capture global dependencies both within and across modalities.

Predictor Network. The predictor takes the multimodal embedding of the context subgraph, given by the $\mathbf{h}_{CLS}^{out}$ embedding from the context encoder, which fuses outputs from the 2D GINE and 3D SchNet. The predictor is a lightweight transformer followed by an MLP, which maps the context embedding to the representation of the complementary subgraph. Since the upstream modality encoders already capture spatial and relational information, we do not add explicit positional encodings, unlike image-based JEPA (Assran et al., 2023) and 2D GraphJEPA (Skenderi et al., 2025).

Target Encoder. The target subgraph $f_{\bar{\theta}}$ is encoded by a separate instance of the context encoder. Maintaining two distinct networks stabilizes training and mitigates representation collapse, a strategy widely adopted in self-predictive frameworks such as BYOL (Grill et al., 2020b), I-JEPA (Ass-

ran et al., 2023), and BGRL (Thakoor et al., 2023). The target encoder's weights are updated via an exponential moving average (EMA) of the context encoder's parameters:

$$\bar{\theta}^{(t)} = \tau\, \bar{\theta}^{(t-1)} + (1-\tau)\, \theta^{(t)},$$

where $\bar{\theta}^{(t)}$ are the exponentially moving averaged parameters at step t, $\theta^{(t)}$ are the current context encoder parameters, and $\tau \in [0, 1]$ is the decay rate controlling the contribution of past parameters.

Pretraining task. Each subgraph is represented by a single multimodal embedding, taken from the final classification token embedding $\mathbf{h}_{CLS}^{out}$. For the *context* subgraph, we feed the entire multimodal token sequence from the context encoder into the predictor transformer and take its output CLS token as the predicted embedding. For the *target* subgraph, we use the CLS token directly from the target encoder. The self-supervised objective minimizes the mean squared $\mathcal{L}_2$ distance between the predicted context embedding and the target embedding:

$$\frac{1}{M} \sum_{i=1}^{M} \sum_{j=1}^{k} \left\| \hat{\mathbf{s}}_{y_j} - \mathbf{s}_{y_j} \right\|^2,$$

where $\hat{\mathbf{s}}_{y_j}$ and $\mathbf{s}_{y_j}$ denote the predicted and target subgraph embeddings, M is the batch size, and k the number of sampled views (ego-nets and their complements). All views are treated as separate instances when computing the loss.

Fine-tuning. When fine-tuning, we use the target encoder as our pretrained backbone to generate graph embeddings, and add lightweight task-specific heads. We consider two types of task heads: (i) linear probing on whole-graph embeddings with a single linear layer (C-FREE$_{\text{LIN}}$) to evaluate representation quality, and (ii) aggregating k-EgoNet subgraph embeddings with DeepSets (Zaheer et al., 2017) (C-FREE$_{\text{DS}}$), showing that subgraph pretraining transfers both to whole-molecule prediction and to ESAN-style fine-tuning schemes. We find that C-FREE$_{\text{DS}}$ is especially beneficial in the 2D-only setting, while for multimodal inputs both heads perform similarly, likely because 3D information compensates for the lower expressiveness. Nevertheless, downstream convergence is faster with DeepSets, suggesting advantages in aligning pretraining and fine-tuning.

3.1 Invariance and Expressiveness

We make two theoretical observations. First, with the DeepSets head, C-FREE$_{\text{DS}}$ simulates ESAN (Bevilacqua et al., 2022) and is more expressive than 1-WL (Weisfeiler & Leman, 1968). An informal statement is given below; the full theorem and proof are in Appendix Section A.1.

(Informal) Lemma 1. *Under the assumptions from Theorem 2 of (Bevilacqua et al., 2022), C-FREE with a DeepSets task head is as expressive as ESAN, hence it is strictly more expressive than the* 1*-WL algorithm (Weisfeiler & Leman, 1968).*

Second, C-FREE preserves the invariances of its modality encoders. Prior work has shown that architectures of this form inherit invariance from their encoders (Manolache et al., 2024), and the same holds for our framework. For completeness, we include the lemma in Appendix Section A.1.

4 Empirical Evaluation

We evaluate our framework through four complementary sets of experiments:

(i) **Frozen backbone evaluation.** We assess representation quality by freezing the backbone and training linear probes (Section 4.1). On MoleculeNet (Wu et al., 2018), C-FREE outperforms contrastive and non-contrastive baselines and remains effective even with 2D-only inputs (Table 1). On Kraken (Gensch et al., 2022), pretrained backbones not only converge faster but also achieve lower error than random initialization, with even the 2D-only pretrained model surpassing a randomly initialized multimodal one (Fig. 3).

(ii) **Full fine-tuning.** We further evaluate end-to-end adaptability by updating the full backbone with task-specific heads. On MoleculeNet, this setup tests how well pretraining transfers to dataset-specific classification tasks (Section 4.2.1). On Kraken, we find that pretraining improves downstream regression performance over random initialization (Table 6). Finally, on the larger Drugs-75K dataset, we study label efficiency by fine-tuning on progressively larger subsets of labeled data (Section 4.2.3).

Table 1: Performance on MoleculeNet (Wu et al., 2018) with frozen backbones. **Non-CL** denotes non-contrastive and **CL** contrastive methods. We report **C-FREE$_{2D}$** (2D-only) and **C-FREE$_{MM}$** (multi-modal), each with linear probing on whole-molecule embeddings (**MOL**) or on subgraphs using subgraph aggregation with DeepSets (Zaheer et al., 2017) (**SUB**). Metric: ROC-AUC (↑). **Red** marks the best model and **Blue** the second best. **C-FREE** ranks first or second on 6 of 8 datasets, with **MM-MOL** best overall, while even the 2D-only variants of C-FREE outperform all baselines on average.

		MoleculeNet Datasets (Linear Probe)								
		BBBP (↑)	Tox21 (↑)	ToxCast (↑)	Sider (↑)	ClinTox (↑)	MUV (↑)	HIV (↑)	Bace (↑)	Avg (↑)
	Random Init.	$50.7_{\pm 2.5}$	$64.9_{\pm 0.5}$	$53.2_{\pm 0.3}$	$53.2_{\pm 1.1}$	$63.1_{\pm 2.3}$	$62.1_{\pm 1.3}$	$66.1_{\pm 0.7}$	$63.4_{\pm 1.8}$	59.60
CL	InfoGraph	$65.9_{\pm 0.6}$	$65.8_{\pm 0.7}$	$54.6_{\pm 0.1}$	$57.2_{\pm 1.0}$	$61.4_{\pm 4.8}$	$63.9_{\pm 1.9}$	$71.4_{\pm 0.6}$	$67.4_{\pm 4.9}$	63.44
CL	GROVER	$67.0_{\pm 0.3}$	$63.9_{\pm 0.3}$	$53.6_{\pm 0.4}$	**$59.9_{\pm 1.7}$**	$65.0_{\pm 6.4}$	$62.7_{\pm 1.4}$	$67.8_{\pm 1.0}$	$69.0_{\pm 4.7}$	63.62
CL	GraphCL	$64.7_{\pm 1.7}$	$69.1_{\pm 0.5}$	$56.2_{\pm 0.2}$	**$59.5_{\pm 0.9}$**	$60.8_{\pm 3.0}$	$60.6_{\pm 1.8}$	$72.5_{\pm 1.4}$	**$77.0_{\pm 1.7}$**	65.04
CL	JOAO	$66.1_{\pm 0.8}$	$68.1_{\pm 0.2}$	$55.1_{\pm 0.4}$	$58.3_{\pm 0.3}$	$65.3_{\pm 6.1}$	$62.4_{\pm 1.2}$	**$73.8_{\pm 1.2}$**	$71.1_{\pm 0.8}$	65.05
Non-CL	EdgePred	$54.2_{\pm 1.0}$	$66.2_{\pm 0.2}$	$54.4_{\pm 0.1}$	$56.1_{\pm 0.1}$	$65.4_{\pm 5.0}$	$59.5_{\pm 0.9}$	**$73.6_{\pm 0.4}$**	$71.4_{\pm 1.2}$	62.59
Non-CL	AttrMask	$62.7_{\pm 2.7}$	$65.7_{\pm 0.8}$	$56.1_{\pm 0.2}$	$58.3_{\pm 1.5}$	$61.9_{\pm 6.4}$	$60.9_{\pm 1.8}$	$65.5_{\pm 1.4}$	$64.8_{\pm 2.6}$	61.99
Non-CL	GPT-GNN	$62.0_{\pm 0.9}$	$64.9_{\pm 0.7}$	$55.4_{\pm 0.2}$	$55.3_{\pm 0.8}$	$55.0_{\pm 5.1}$	$61.2_{\pm 1.5}$	$71.2_{\pm 1.5}$	$61.0_{\pm 1.2}$	60.74
Non-CL	Cont. Pred	$55.5_{\pm 2.0}$	$67.9_{\pm 0.7}$	$54.0_{\pm 0.3}$	$57.1_{\pm 0.5}$	$67.4_{\pm 4.3}$	$60.5_{\pm 0.9}$	$66.2_{\pm 1.5}$	$54.4_{\pm 3.2}$	60.36
Non-CL	**C-FREE$_{2D\text{-}MOL}$**	$60.5_{\pm 1.7}$	$76.1_{\pm 0.2}$	$62.7_{\pm 0.4}$	$59.0_{\pm 0.6}$	$62.7_{\pm 1.0}$	$67.6_{\pm 0.5}$	$68.7_{\pm 0.4}$	**$75.8_{\pm 0.9}$**	66.63
Non-CL	**C-FREE$_{2D\text{-}SUB}$**	$64.2_{\pm 3.8}$	**$76.7_{\pm 0.6}$**	$63.9_{\pm 0.3}$	$58.0_{\pm 0.7}$	$71.4_{\pm 3.7}$	$64.6_{\pm 3.1}$	$65.5_{\pm 0.6}$	$73.9_{\pm 0.7}$	67.27
Non-CL	**C-FREE$_{MM\text{-}MOL}$**	**$69.8_{\pm 2.6}$**	**$79.9_{\pm 1.1}$**	**$65.8_{\pm 0.7}$**	$58.5_{\pm 2.5}$	**$69.9_{\pm 1.9}$**	**$76.6_{\pm 2.8}$**	$72.8_{\pm 0.7}$	$75.3_{\pm 1.1}$	**71.07**
Non-CL	**C-FREE$_{MM\text{-}SUB}$**	**$73.8_{\pm 2.1}$**	**$76.7_{\pm 0.7}$**	**$66.8_{\pm 0.2}$**	$56.4_{\pm 1.5}$	**$75.7_{\pm 2.2}$**	**$70.6_{\pm 1.0}$**	$71.9_{\pm 1.5}$	$75.5_{\pm 1.9}$	**70.92**

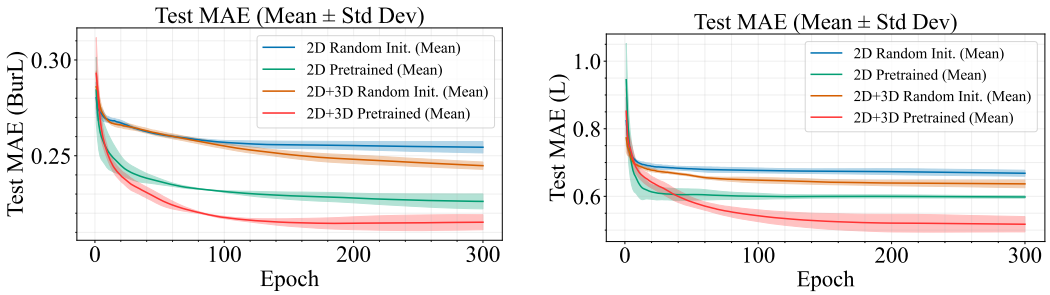

Figure 3: Test MAE on the Kraken regression tasks (Sterimol BurL and Sterimol L) with frozen backbones. GEOM-pretrained models consistently outperform random initialization for both 2D-only and multimodal variants. Pretrained models begin with lower error and converge faster, while randomly initialized models fail to match performance even after 300 epochs. Incorporating the 3D modality yields further gains over 2D-only backbones, with pretraining amplifying this advantage. Curves show the mean over 3 runs, with shaded regions indicating the standard deviation.

(iii) **Ablations.** We conduct two ablations: (i) removing each modality to assess its contribution (Section 4.2.4), and (ii) removing the predictor network to evaluate its impact on representation quality and training stability (Section 4.3).

(iv) **Theory alignment.** We verify whether the empirical expressiveness aligns with the theoretical result from Lemma 1 (Section 4.4).

Implementation details for pretraining and evaluation are provided in Section A.5 in the Appendix.

4.1 Comparison with Frozen Backbones

For the first experiments, we compare our framework against state-of-the-art contrastive and non-contrastive self-supervised methods on molecular property classification tasks. Following Wang et al. (2023a), we pre-train two backbones on 0.33M molecules from GEOM (Axelrod & Gomez-Bombarelli, 2022) and evaluate them on MoleculeNet (Wu et al., 2018). One backbone has 4M parameters and uses 2D inputs, while the other has 9.1M parameters and incorporates both 2D graphs and 3D conformer ensembles available in GEOM, with three additional conformers generated using RDKit (Landrum, 2016) at fine-tuning. This setup enables fair comparison to 2D-only baselines while also testing the benefit of 3D information. Performance is reported as mean ROC-AUC over three scaffold splits, with frozen checkpoints chosen by best self-supervised loss and linear probes selected by downstream validation loss. We evaluate two strategies: linear probing on whole-graph embeddings and on subgraph embeddings aggregated with DeepSets (Zaheer et al., 2017).

Table 2: Performance on MoleculeNet (Wu et al., 2018) with full end-to-end fine-tuning. For this experiment, the downstream model receives the entire molecule as input, and we limit evaluation to the multi-modal variant. **Non-CL** denotes non-contrastive, **CL** contrastive, and **Multi** multi-modal methods. Our model is reported as **C-FREE$_{\text{MM-MOL}}$**, with the multi-modal variant using both modalities. The evaluation metric is ROC-AUC (↑). **Red** highlights the best results and **Blue** the second best. **C-FREE** achieves the best results on 4 out of 8 datasets and ranks first overall, outperforming both multi-modal baselines.

		MoleculeNet Datasets (Full Fine-Tuning)								
		BBBP (↑)	Tox21 (↑)	ToxCast (↑)	Sider (↑)	ClinTox (↑)	MUV (↑)	HIV (↑)	Bace (↑)	Avg (↑)
CL	InfoGraph	67.5$_{\pm 0.1}$	73.2$_{\pm 0.4}$	63.7$_{\pm 0.5}$	59.9$_{\pm 0.3}$	76.5$_{\pm 1.0}$	74.1$_{\pm 0.7}$	75.1$_{\pm 0.9}$	77.8$_{\pm 0.8}$	70.98
	GROVER	70.0$_{\pm 0.1}$	74.3$_{\pm 0.1}$	65.4$_{\pm 0.4}$	64.8$_{\pm 0.6}$	81.2$_{\pm 3.0}$	67.3$_{\pm 1.8}$	62.5$_{\pm 0.9}$	82.6$_{\pm 0.7}$	71.01
	GraphCL	69.7$_{\pm 0.6}$	73.9$_{\pm 0.6}$	62.4$_{\pm 0.5}$	60.5$_{\pm 0.8}$	76.0$_{\pm 2.6}$	69.8$_{\pm 2.6}$	78.5$_{\pm 1.2}$	75.4$_{\pm 1.4}$	70.78
	MolCLR	66.6$_{\pm 1.8}$	73.0$_{\pm 0.1}$	62.9$_{\pm 0.3}$	57.5$_{\pm 1.7}$	86.1$_{\pm 0.9}$	72.5$_{\pm 2.3}$	76.2$_{\pm 1.5}$	71.5$_{\pm 3.1}$	70.79
	GraphLoG	72.5$_{\pm 0.8}$	75.7$_{\pm 0.5}$	63.5$_{\pm 0.7}$	61.2$_{\pm 1.1}$	76.7$_{\pm 3.3}$	76.0$_{\pm 1.1}$	77.8$_{\pm 0.8}$	83.5$_{\pm 1.2}$	73.40
Non-CL	AttrMask	65.0$_{\pm 2.3}$	74.8$_{\pm 0.2}$	62.9$_{\pm 0.1}$	61.2$_{\pm 0.1}$	87.7$_{\pm 1.1}$	73.4$_{\pm 2.0}$	76.8$_{\pm 0.5}$	79.7$_{\pm 0.3}$	72.68
	ContextPred	65.7$_{\pm 0.6}$	74.2$_{\pm 0.0}$	62.5$_{\pm 0.3}$	62.2$_{\pm 0.5}$	77.2$_{\pm 0.8}$	75.3$_{\pm 1.5}$	77.1$_{\pm 0.8}$	76.0$_{\pm 2.0}$	71.28
	GraphMAE	72.0$_{\pm 0.6}$	75.5$_{\pm 0.6}$	64.1$_{\pm 0.3}$	60.3$_{\pm 1.1}$	82.3$_{\pm 1.2}$	76.3$_{\pm 2.4}$	77.2$_{\pm 1.0}$	83.1$_{\pm 0.9}$	73.85
	MGSSL	69.7$_{\pm 0.9}$	76.5$_{\pm 0.3}$	64.1$_{\pm 0.7}$	61.8$_{\pm 0.8}$	80.7$_{\pm 2.1}$	78.7$_{\pm 1.5}$	78.8$_{\pm 1.2}$	79.1$_{\pm 0.9}$	73.70
	Mole-BERT	71.9$_{\pm 1.6}$	76.8$_{\pm 0.5}$	64.3$_{\pm 0.2}$	62.8$_{\pm 1.1}$	78.9$_{\pm 3.0}$	78.6$_{\pm 1.8}$	78.2$_{\pm 0.8}$	80.8$_{\pm 1.4}$	74.04
Multi	GraphMVP	68.5$_{\pm 0.2}$	74.5$_{\pm 0.4}$	62.7$_{\pm 0.1}$	62.3$_{\pm 1.6}$	79.0$_{\pm 2.5}$	75.0$_{\pm 1.4}$	74.8$_{\pm 1.4}$	76.8$_{\pm 1.1}$	71.69
	3D InfoMax	69.1$_{\pm 1.0}$	74.5$_{\pm 0.7}$	64.4$_{\pm 0.8}$	60.6$_{\pm 0.7}$	79.9$_{\pm 3.4}$	74.4$_{\pm 2.4}$	76.1$_{\pm 1.3}$	79.7$_{\pm 1.5}$	72.34
	MoleculeSDE	71.8$_{\pm 0.7}$	76.8$_{\pm 0.3}$	65.0$_{\pm 0.2}$	60.8$_{\pm 0.3}$	87.0$_{\pm 0.5}$	80.9$_{\pm 0.3}$	78.8$_{\pm 0.9}$	79.5$_{\pm 2.1}$	75.07
	MoleBlend	73.0$_{\pm 0.8}$	77.8$_{\pm 0.8}$	66.1$_{\pm 0.0}$	64.9$_{\pm 0.3}$	87.6$_{\pm 0.7}$	77.2$_{\pm 2.3}$	79.0$_{\pm 0.8}$	83.7$_{\pm 1.4}$	76.16
	VideoMol	70.7$_{\pm 1.5}$	78.8$_{\pm 0.4}$	66.7$_{\pm 0.5}$	**66.3$_{\pm 0.9}$**	–	–	79.4$_{\pm 0.5}$	82.4$_{\pm 0.9}$	–
Multi-FM	ChemBERTa-1	64.3	72.8	–	–	73.3	–	62.2	–	–
	ChemBERTa-2	72.8	–	–	–	56.3	–	–	79.9	–
	MolFM	72.9$_{\pm 0.1}$	77.2$_{\pm 0.7}$	64.4$_{\pm 0.2}$	64.2$_{\pm 0.9}$	79.7$_{\pm 1.6}$	76.0$_{\pm 0.8}$	78.8$_{\pm 1.1}$	83.9$_{\pm 1.1}$	74.63
	SCAGE	73.4$_{\pm 1.1}$	79.4$_{\pm 1.2}$	69.3$_{\pm 0.5}$	66.0$_{\pm 1.2}$	**92.7$_{\pm 0.9}$**	–	–	85.4$_{\pm 1.2}$	–
	GEM	72.4$_{\pm 0.4}$	78.1$_{\pm 0.1}$	69.2$_{\pm 0.4}$	**67.2$_{\pm 0.4}$**	90.1$_{\pm 1.3}$	81.7$_{\pm 0.5}$	80.6$_{\pm 0.9}$	**85.6$_{\pm 1.1}$**	78.11
	UniMol	72.9$_{\pm 0.6}$	79.6$_{\pm 0.5}$	69.6$_{\pm 0.1}$	65.9$_{\pm 1.3}$	**91.9$_{\pm 1.8}$**	**82.1$_{\pm 1.3}$**	**80.8$_{\pm 0.3}$**	**85.7$_{\pm 0.2}$**	**78.56**
	C-FREE$_{\text{Sch-1C}}$	**88.6$_{\pm 1.8}$**	**84.7$_{\pm 0.8}$**	71.3$_{\pm 1.6}$	61.8$_{\pm 1.7}$	73.1$_{\pm 1.7}$	75.9$_{\pm 1.3}$	78.9$_{\pm 1.8}$	78.8$_{\pm 2.1}$	76.63
	C-FREE$_{\text{Sch-3C}}$	78.9$_{\pm 1.1}$	84.2$_{\pm 0.4}$	**71.7$_{\pm 0.9}$**	62.5$_{\pm 1.9}$	83.7$_{\pm 2.9}$	**82.5$_{\pm 0.1}$**	77.9$_{\pm 1.2}$	78.6$_{\pm 1.1}$	77.50
	C-FREE$_{\text{PaiNN-3C}}$	**88.9$_{\pm 0.7}$**	**86.2$_{\pm 0.2}$**	**71.6$_{\pm 0.9}$**	64.8$_{\pm 0.9}$	87.6$_{\pm 0.9}$	75.7$_{\pm 1.2}$	**81.4$_{\pm 4.7}$**	82.3$_{\pm 0.7}$	**79.81**

As shown in Table 1, our framework achieves the best average performance, outperforming baselines on 6 of 8 tasks. In the 2D-only setting, DeepSets aggregation provides clear gains and yields the best average result, with further improvements from adding 3D information. We also see strong gains on multi-task datasets such as Tox21, ToxCast, and MUV, suggesting that our model captures more generalizable features. While DeepSets helps in 2D-only, its effect is minimal in the multimodal case, likely because 3D inputs already encode rich structural detail.

We evaluate transfer to molecular property regression on Kraken (Gensch et al., 2022), which contains 1,552 ligands labeled with four 3D descriptors (Sterimol B5, Sterimol L, buried Sterimol B5, buried Sterimol L). Since Kraken molecules are disjoint from GEOM, it provides a strong test of generalization. As shown in Fig. 3, GEOM-pretrained backbones start with lower error, converge faster, and consistently outperform random initialization, with even the 2D-only pretrained model surpassing a randomly initialized multimodal one. Adding 3D information yields further gains, amplified by pretraining. Results for BurB5 and B5 are included in Appendix Section A.6.

4.2 Full Fine-tuning

Beyond evaluating frozen representations, we next assess the adaptability of our pretrained models through full end-to-end fine-tuning. In this setting, the entire backbone is updated jointly with a task-specific head, enabling us to test how pretraining improves convergence and downstream performance. We consider two scenarios: (i) property classification on MoleculeNet, where we compare against both contrastive, non-contrastive and multimodal self-supervised baselines, and (ii) property regression on the MARCEL benchmark, where we evaluate transferability to the Kraken dataset and study label efficiency on the larger Drugs-75K dataset.

4.2.1 Full Fine-tuning for Property Prediction

Building on the frozen backbone results, we next evaluate the adaptability of our models through end-to-end fine-tuning. We compare against the baselines reported in Yu et al. (2023), reproducing their evaluation protocol by attaching a linear classifier head and fine-tuning the entire backbone on each MoleculeNet dataset. We report results for our backbones pre-trained with an ensemble of conformers: one using SchNet (Schütt et al., 2017) as the 3D encoder (**C-FREE$_{\text{Sch-3C}}$**) and one using PaiNN (Schütt et al., 2021) (**C-FREE$_{\text{PaiNN-3C}}$**). For fairness, we also include a variant pre-trained with a single conformer, **C-FREE$_{\text{Sch-1C}}$**. As in Table 1, all backbones are pre-trained on the GEOM

Table 3: Performance on QM9 (Ramakrishnan et al., 2014) with full end-to-end fine-tuning, following the same protocol as in Ji et al. (2024). In this setup, the full molecule is used as input to the downstream model. **C-FREE$_{\text{EGO}}$** and **C-FREE$_{\text{MURCKO}}$** denote the variant of C-FREE pre-trained with EgoNets and Murcko segments respectively. The evaluation metric is the Mean Absolute Error (MAE) (↓). **C-FREE** achieves the best results on 4 out of 6 datasets.

METHOD	MU (↓)	ALPHA (↓)	HOMO/LUMO/GAP (↓)	R^2 (↓)	C_v (↓)	ZPVE (↓)
GEM	$0.444_{\pm 1.5e^{-3}}$	$0.589_{\pm 4.2e^{-3}}$	$0.0067_{\pm 4e^{-5}}$	$25.67_{\pm 0.743}$	$0.237_{\pm 1.4e^{-3}}$	$0.0011_{\pm 2.0e^{-5}}$
UNIMOL (1)	$0.155_{\pm 1.5e^{-3}}$	$0.363_{\pm 9.0e^{-3}}$	$0.0043_{\pm 2e^{-5}}$	$4.805_{\pm 0.055}$	$0.183_{\pm 2.0e^{-3}}$	$0.0011_{\pm 3.0e^{-5}}$
UNIMOL2 310M	$0.092_{\pm 1.3e^{-3}}$	$0.315_{\pm 3.0e^{-3}}$	$0.0036_{\pm 1e^{-5}}$	$4.672_{\pm 0.245}$	$0.143_{\pm 2.0e^{-3}}$	$\mathbf{0.0005}_{\pm \mathbf{1.0e^{-5}}}$
UNIMOL2 1.1B	$0.089_{\pm 4.0e^{-4}}$	$0.305_{\pm 3.0e^{-3}}$	$\mathbf{0.0035}_{\pm \mathbf{1e^{-5}}}$	$4.265_{\pm 0.067}$	$0.144_{\pm 2.0e^{-3}}$	$\mathbf{0.0005}_{\pm \mathbf{1.0e^{-5}}}$
C-FREE$_{\text{EGO}}$	$\mathbf{0.064}_{\pm \mathbf{1.7e^{-4}}}$	$\mathbf{0.116}_{\pm \mathbf{4.9e^{-4}}}$	$0.00487_{\pm 7e^{-4}}$	$1.155_{\pm 0.010}$	$\mathbf{0.049}_{\pm \mathbf{5.0e^{-4}}}$	$0.0061_{\pm 4.1e^{-4}}$
C-FREE$_{\text{MURCKO}}$	$0.077_{\pm 2.6e^{-3}}$	$0.124_{\pm 9.0e^{-3}}$	$0.00491_{\pm 5e^{-4}}$	$\mathbf{1.135}_{\pm 0.025}$	$0.052_{\pm 2.5e^{-3}}$	$0.0040_{\pm 2.0e^{-4}}$

dataset (Axelrod & Gomez-Bombarelli, 2022), which contains approximately 330K molecules. This contrasts with the contrastive, non-contrastive, and multi-modal baselines, which are trained on the substantially larger PCQM4Mv2 dataset from the OGB Large-Scale Challenge (Hu et al., 2021) (3M molecules), potentially placing our approach at a disadvantage.

We additionally include a suite of recent multi-modal foundation models—ChemBERTa (v1 and v2), MolFM, SCAGE, GEM, and UniMol—and compare with their published results on the same downstream tasks. Note that these models are trained on substantially larger datasets: UniMol on 19M molecules aggregated from multiple sources, GEOM on ZINC-20M, while MolFM, SCAGE, and ChemBERTa rely on the PubChem dataset comprising 77M molecules. This setting tests how well the pretrained representations adjust to dataset-specific distributions and whether the multi-modal backbone provides additional benefits. Performance is reported as mean ROC-AUC over three scaffold splits. A rundown and explanation of each baseline is found in Appendix Section A.7.

As shown in Table 2, our method achieves notable gains on 5 of 8 datasets and delivers the best overall average when using the PaiNN variant. The strongest competitor, UniMol (Zhou et al., 2023), also leverages multimodal inputs but was pretrained at a larger scale, namely on 19M molecules with roughly 209 million molecular conformations. Despite this large difference in pretraining scale, our approach matches or surpasses UniMol on several MoleculeNet tasks.

To further compare against larger-scale models, we conduct additional experiments on QM9 (Ramakrishnan et al., 2014), ZINC (Gómez-Bombarelli et al., 2018), and SPICE (Eastman et al., 2023). For QM9, we follow the setup in (Ji et al., 2024), restricting C-FREE to its 3D encoder and comparing it with the Uni-Mol (Ji et al., 2024) family, whose largest model contains 1.1 billion parameters. Despite its substantially smaller size, our model outperforms Uni-Mol2 on four of the six prediction targets; the exceptions being HOMO/LUMO/GAP and ZPVE tasks, where Uni-Mol2 retains an advantage. On ZINC, we compare our approach with GraphJEPA and find that it achieves superior performance. Finally, on SPICE we compare against the supervised TensorNet (Simeon & de Fabritiis, 2023) and the Equivariant Transformer (Eastman et al., 2023), and observe significant improvements over both. Further experimental details are provided in Appendix Section A.6.

4.2.2 COMPARISON WITH CHEMICALLY-INFORMED SUBGRAPH PRE-TRAINING

We perform a comparison of our pre-trained backbone to a variant trained using Murcko-based fragments (Bemis & Murcko) on the QM9 (Ramakrishnan et al., 2014) dataset in Table 3. Murcko scaffolds capture the core ring systems of molecules and are commonly used to define chemically meaningful fragments. In this variant, the Murcko scaffold of each molecule serves as the target subgraph, while the remaining atoms and edges form the context, allowing us to assess the effect of using chemically informed fragments. We observe gains on the R^2 and ZPVE targets, whereas the EgoNet variant performs best on the remaining four tasks. Overall, these results suggest that our EgoNet-based approach already captures useful structural patterns without requiring hand-crafted chemical partitions, while remaining competitive across tasks.

4.2.3 LABEL-EFFICIENT FINE-TUNING

Building on the observation that fine-tuning outperforms supervised training, we next study label efficiency on the Drugs-75K dataset (Zhu et al., 2024), a GEOM-Drugs subset with 75.099 molecules and at least 5 rotatable bonds. For each molecule, Auto3D (Liu et al., 2022b) generates conformer ensembles, and three DFT-based reactivity descriptors serve as targets: ionization potential (IP), electron affinity (EA), and electronegativity (χ).

Table 4: Fine-tuning on the Drugs-75K dataset (Zhu et al., 2024) with limited labeled data. We compare our pretrained backbone **(FFT)** against a model trained from scratch **(RND)** using 1%, 10%, 50%, and 100% of the labels. Pretraining offers clear gains in low-label regimes, while performance is comparable when using the full dataset.

		IP ↓	EA ↓	χ ↓
1%	RND	$0.638_{\pm 0.001}$	$0.613_{\pm 0.002}$	$0.334_{\pm 0.001}$
	FFT	$\mathbf{0.608}_{\pm 0.001}$	$\mathbf{0.583}_{\pm 0.001}$	$\mathbf{0.317}_{\pm 0.001}$
10%	RND	$0.561_{\pm 0.002}$	$0.526_{\pm 0.001}$	$0.277_{\pm 0.001}$
	FFT	$\mathbf{0.520}_{\pm 0.005}$	$\mathbf{0.494}_{\pm 0.002}$	$\mathbf{0.267}_{\pm 0.001}$
50%	RND	$0.457_{\pm 0.002}$	$0.433_{\pm 0.002}$	$0.233_{\pm 0.001}$
	FFT	$\mathbf{0.454}_{\pm 0.001}$	$\mathbf{0.421}_{\pm 0.002}$	$\mathbf{0.230}_{\pm 0.001}$
100%	RND	$\mathbf{0.419}_{\pm 0.005}$	$0.403_{\pm 0.002}$	$\mathbf{0.211}_{\pm 0.003}$
	FFT	$\mathbf{0.419}_{\pm 0.002}$	$\mathbf{0.395}_{\pm 0.003}$	$0.213_{\pm 0.001}$

Table 5: Expressiveness results on EXP (Abboud et al., 2021), where 1-WL GNNs cannot surpass random guessing. Even the 1-EgoNet variant of C-FREE approaches the theoretical upper bound, while 2- and 3-EgoNet variants achieve the highest accuracy, outperforming GINE (Hu et al., 2020a) and GraphJEPA (Assran et al., 2023). Unlike GraphJEPA, our method avoids costly METIS clustering (Karypis & Kumar, 1998).

Method	Accuracy (↑)
GINE	$50.69_{\pm 1.39}$
GraphJEPA	$98.77_{\pm 0.99}$
C-FREE (1-Ego)	$96.03_{\pm 1.22}$
C-FREE (2-Ego)	$\mathbf{99.33}_{\pm 0.18}$
C-FREE (3-Ego)	$99.08_{\pm 0.20}$

Table 6: Modality ablation on the Kraken dataset (MAE ↓). Using the same pretrained backbone, we feed only the 2D sequence for the 2D variant and only the 3D sequence for the 3D variant. Pretraining consistently outperforms training from scratch. The multimodal model delivers the strongest performance, with 3D-only close behind, indicating that 3D features have greater impact than 2D.

	Method	B5 ↓	L ↓	BurB5 ↓	BurL ↓
2D	Random Init.	$0.297_{\pm 0.006}$	$0.396_{\pm 0.026}$	$0.205_{\pm 0.006}$	$0.152_{\pm 0.006}$
	Fine-tuned	$0.276_{\pm 0.012}$	$0.340_{\pm 0.028}$	$0.176_{\pm 0.002}$	$0.146_{\pm 0.005}$
3D	Random Init.	$0.197_{\pm 0.006}$	$0.345_{\pm 0.011}$	$0.162_{\pm 0.006}$	$0.135_{\pm 0.009}$
	Fine-tuned	$\mathbf{0.194}_{\pm 0.003}$	$0.329_{\pm 0.002}$	$\mathbf{0.134}_{\pm 0.005}$	$0.131_{\pm 0.004}$
MM	Random Init.	$0.203_{\pm 0.008}$	$0.378_{\pm 0.003}$	$0.161_{\pm 0.002}$	$0.142_{\pm 0.001}$
	Fine-tuned	$\mathbf{0.193}_{\pm 0.017}$	$\mathbf{0.306}_{\pm 0.011}$	$\mathbf{0.134}_{\pm 0.009}$	$\mathbf{0.126}_{\pm 0.004}$

We fine-tune our pretrained backbone using 1%, 10%, 50%, and 100% of the available data and compare against a model trained fully supervised from random initialization. With the full dataset, performance is broadly comparable, which is reasonable given the scale of the dataset. More importantly, in low-data regimes, initializing from self-supervised pretraining provides clear gains, consistently outperforming training from scratch (Table 13).

4.2.4 Modality Ablation

After establishing the benefits of our multimodal backbone, we next analyze the contribution of each modality through targeted ablations. Starting from the same GEOM-pretrained backbone, we fine-tune on Kraken while feeding either only the 2D encoder sequence, only the 3D sequence, or both. This setup keeps the architecture and pretraining signal fixed, isolating the effect of each modality and mimicking transfer scenarios where only 2D or 3D data is available. Table 6 shows that pretraining consistently improves over random initialization across all settings, confirming that useful information is transferred even when restricted to a single modality. Among unimodal variants, the 3D-only backbone performs best, suggesting that geometric information has greater impact than 2D topology alone. Combining both modalities achieves the strongest results overall, reinforcing the view that 2D and 3D provide complementary signals.

4.3 Ablation on Predictor Types

We hypothesize that our model's strong performance stems from the predictor network, which serves as a guiding signal to refine the representations produced by the encoder. To test this, we pretrain on GEOM following the same setup previously described and perform an ablation with three predictor variants: none, a linear predictor, and a transformer. We then evaluate downstream performance using either a linear probe or full fine-tuning. To keep the study computationally lightweight, we restrict these experiments to the smaller-sized 2D-only pretrained backbone. Since the predictor design is identical in both the 2D-only and multimodal variants, this enables a focused analysis of the predictor's contribution without incurring excessive computational cost.

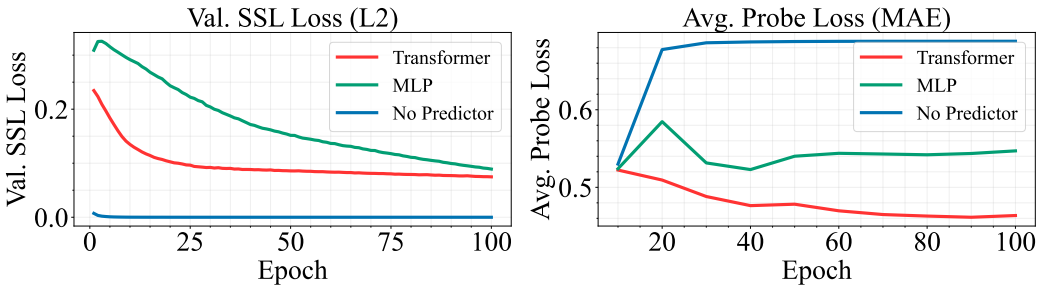

Figure 4: Predictor ablation. **Left**: SSL validation loss on GEOM during pretraining. **Right**: Average linear-probe MAE (↓) on Kraken with frozen backbones. Without a predictor, training collapses (loss ∼0) and probes perform worst; an MLP predictor helps but underperforms, while a Transformer predictor achieves the best results.

As shown in Appendix Table 13, removing the predictor leads to poor downstream performance across regression tasks, with the self-supervised loss collapsing to zero (Fig. 4). While the EMA target encoder stabilizes training (Tarvainen & Valpola, 2017; Grill et al., 2020b; Assran et al., 2023), it is insufficient on its own: without an asymmetric architecture, the model collapses to trivial solutions. Adding a predictor breaks this symmetry and prevents collapse (Richemond et al., 2020), consistent with theory showing that a trainable prediction head enables richer representations (Wen & Li, 2022). Even a simple MLP improves results, but the transformer predictor performs best, likely because it operates at the node level before pooling, yielding more informative graph-level embeddings. We therefore adopt it as the default. Further analysis of the mechanisms that prevent collapse in latent methods can be found in the Appendix in Section A.2.

4.4 Theory Alignment

Finally, we validate our theoretical findings with an experiment on the EXP dataset, designed by Abboud et al. (2021) such that any 1-WL GNN cannot exceed random guessing. We train a smaller version of our encoder and compare it against GINE (Hu et al., 2020a) and GraphJEPA (Assran et al., 2023). Results are averaged over three runs with resampled EgoNets. As shown in Table 5, even the 1-EgoNet variant of C-FREE approaches the theoretical upper bound, while 2- and 3-EgoNet variants achieve the highest accuracy, outperforming both GINE and GraphJEPA, the latter relying on the costly METIS (Karypis & Kumar, 1998) algorithm. Nevertheless, the practical significance of expressiveness for molecular learning is limited, as recent work (Pellizzoni et al., 2025) shows that 1-WL already distinguishes most samples in molecular datasets nearly perfectly. Although less decisive for molecules, expressiveness may be more important in other domains, where the flexibility of our framework could be advantageous.

5 Conclusions and Future Work

We introduced C-FREE, a contrast-free multimodal self-supervised framework for molecular representation learning. Its core idea is to align embeddings of complementary subgraphs, enabling predictive pretraining without negatives, positional encodings, or costly clustering. By combining 2D molecular graphs and 3D conformer ensembles, C-FREE achieves state-of-the-art results on classification (MoleculeNet) and regression (Kraken, Drugs-75K), with clear gains in low-label regimes. Ablations show that the predictor is critical to avoid collapse and that 2D and 3D provide complementary signals, with 3D offering stronger performance but 2D remaining competitive. Finally, experiments on the EXP benchmark confirm that C-FREE is more expressive than 1-WL.

Looking ahead, several extensions are promising. Our largest model has only 9.1M parameters and was pretrained on 0.33M molecules; scaling to larger architectures and datasets, as the 100M-molecule collections in (Beaini et al., 2024), could unlock further gains. Another extension is combining self-supervised pretraining with an additional supervised stage on such large-scale data before transferring to downstream tasks. Extending to new modalities, such as SMILES strings, is also appealing, though it may require modality-specific masking. Finally, more chemically informed subgraph selection or alternative alignment objectives could further enrich representations.

6 Reproducibility Statement

We have taken several steps to ensure reproducibility of our work. A full description of the C-FREE framework, including architecture choices, training setup, and pretraining objectives, is provided in Section 3. Implementation details for all experiments, including hyperparameters, optimizers, and fine-tuning protocols, are included in Section A.5. Theoretical claims, such as expressiveness and invariance, are clearly stated in Section 3.1 with proofs in Appendix Section A.1. Datasets used in this work (GEOM, MoleculeNet, Kraken, Drugs-75K, and EXP) are publicly available, and all preprocessing steps are described in Section 4 and Section A.5. To further facilitate reproducibility, we will release our code, pretrained checkpoints, and data processing scripts in the final version of the paper.

References

Ralph Abboud, İsmail İlkan Ceylan, Martin Grohe, and Thomas Lukasiewicz. The surprising power of graph neural networks with random node initialization, 2021. URL https://arxiv.org/abs/2010.01179.

Walid Ahmad, Elana Simon, Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. Chemberta-2: Towards chemical foundation models. *arXiv preprint arXiv:2209.01712*, 2022.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-Supervised Learning From Images With a Joint-Embedding Predictive Architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15619–15629, 2023.

Simon Axelrod and Rafael Gomez-Bombarelli. Geom, energy-annotated molecular conformations for property prediction and molecular generation. *Scientific Data*, 9(1):185, 2022.

Peizhen Bai, Xianyuan Liu, and Haiping Lu. Geometry-aware line graph transformer pre-training for molecular property prediction, 2023. URL https://arxiv.org/abs/2309.00483.

Dominique Beaini, Shenyang Huang, Joao Alex Cunha, Zhiyi Li, Gabriela Moisescu-Pareja, Oleksandr Dymov, Samuel Maddrell-Mander, Callum McLean, Frederik Wenkel, Luis Müller, Jama Hussein Mohamud, Ali Parviz, Michael Craig, Michał Koziarski, Jiarui Lu, Zhaocheng Zhu, Cristian Gabellini, Kerstin Klaser, Josef Dean, Cas Wognum, Maciej Sypetkowski, Guillaume Rabusseau, Reihaneh Rabbany, Jian Tang, Christopher Morris, Mirco Ravanelli, Guy Wolf, Prudencio Tossou, Hadrien Mary, Therence Bois, Andrew W Fitzgibbon, Blazej Banaszewski, Chad Martin, and Dominic Masters. Towards foundational models for molecular learning on large-scale multi-task datasets. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Zc2aIcucwc.

G. W. Bemis and M. A. Murcko. The properties of known drugs. 1. molecular frameworks. *Journal of Medicinal Chemistry*.

Beatrice Bevilacqua, Fabrizio Frasca, Derek Lim, Balasubramaniam Srinivasan, Chen Cai, Gopinath Balamurugan, Michael M. Bronstein, and Haggai Maron. Equivariant Subgraph Aggregation Networks, March 2022.

Longxing Cao, Brian Coventry, Inna Goreshnik, Buwei Huang, William Sheffler, Joon Sung Park, Kevin M Jude, Iva Marković, Rameshwar U Kadam, Koen HG Verschueren, et al. Design of protein-binding proteins from the target structure alone. *Nature*, 605(7910), 2022.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jegou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9630–9640, 2021. doi: 10.1109/ICCV48922.2021.00951.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Guide Proceedings*, volume 119, pp. 1597–1607. JMLR.org, July 2020. doi: 10.5555/3524938.3525087.

Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. Chemberta: Large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.09885*, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *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)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423.

Peter Eastman, Pavan Kumar Behara, David L Dotson, Raimondas Galvelis, John E Herr, Josh T Horton, Yuezhi Mao, John D Chodera, Benjamin P Pritchard, Yuanqing Wang, et al. Spice, a dataset of drug-like molecules and peptides for training machine learning potentials. *Scientific Data*, 10(1):11, 2023.

Daniel C. Elton, Zois Boukouvalas, Mark D. Fuge, and Peter W. Chung. Deep learning for molecular design—a review of the state of the art. *Mol. Syst. Des. Eng.*, 4(4):828–849, August 2019. ISSN 2058-9689. doi: 10.1039/C9ME00039A.

Jiayu Fang et al. Gem: A generalizable molecular representation via large-scale graph foundation models. *ICLR*, 2024.

Shikun Feng, Yuyan Ni, Minghao Li, Yanwen Huang, Zhi-Ming Ma, Wei-Ying Ma, and Yanyan Lan. UniCorn: A Unified Contrastive Learning Approach for Multi-view Molecular Representation Learning. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 13256–13277. PMLR, July 2024.

Matthias Fey and Jan Eric Lenssen. Fast graph representation learning with pytorch geometric. *arXiv preprint arXiv:1903.02428*, 2019.

Johannes Gasteiger, Florian Becker, and Stephan Günnemann. Gemnet: Universal directional graph neural networks for molecules. *Advances in Neural Information Processing Systems*, 34:6790–6802, 2021.

Tobias Gensch, Gabriel dos Passos Gomes, Pascal Friederich, Ellyn Peters, Théophile Gaudin, Robert Pollice, Kjell Jorner, AkshatKumar Nigam, Michael Lindner-D'Addario, Matthew S Sigman, et al. A comprehensive discovery platform for organophosphorus ligands for catalysis. *Journal of the American Chemical Society*, 144(3):1205–1217, 2022.

Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural Message Passing for Quantum Chemistry. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 1263–1272. PMLR, July 2017.

Rafael Gómez-Bombarelli, Jennifer N. Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D. Hirzel, Ryan P. Adams, and Alán Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Central Science*, 4(2):268–276, February 2018. ISSN 2374-7943, 2374-7951. doi: 10.1021/acscentsci.7b00572.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent - a new approach to self-supervised learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 21271–21284. Curran Associates, Inc., 2020a. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/f3ada80d5c4ee70142b17b8192b2958e-Paper.pdf.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap Your Own Latent - A New Approach to Self-Supervised Learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 21271–21284. Curran Associates, Inc., 2020b.

Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive Representation Learning on Large Graphs. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked Autoencoders Are Scalable Vision Learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022.

Maya Hirohara, Yutaka Saito, Yuki Koda, Kengo Sato, and Yasubumi Sakakibara. Convolutional neural network based on smiles representation of compounds for detecting chemical motif. *BMC bioinformatics*, 19(Suppl 19):526, 2018.

Zhenyu Hou, Xiao Liu, Yukuo Cen, Yuxiao Dong, Hongxia Yang, Chunjie Wang, and Jie Tang. GraphMAE: Self-Supervised Masked Graph Autoencoders. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, pp. 594–604, New York, NY, USA, August 2022. Association for Computing Machinery. ISBN 978-1-4503-9385-0. doi: 10.1145/3534678.3539321.

Weihua Hu, Bowen Liu, Joseph Gomes, Marinka Zitnik, Percy Liang, Vijay Pande, and Jure Leskovec. Strategies for Pre-training Graph Neural Networks, February 2020a.

Weihua Hu, Matthias Fey, Hongyu Ren, Maho Nakata, Yuxiao Dong, and Jure Leskovec. Ogb-lsc: A large-scale challenge for machine learning on graphs. *arXiv preprint arXiv:2103.09430*, 2021.

Ziniu Hu, Yuxiao Dong, Kuansan Wang, Kai-Wei Chang, and Yizhou Sun. Gpt-gnn: Generative pre-training of graph neural networks. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 1857–1867, 2020b.

Xiaohong Ji, Zhen Wang, Zhifeng Gao, Hang Zheng, Linfeng Zhang, Guolin Ke, and Weinan E. Uni-mol2: Exploring molecular pretraining model at scale, 2024. URL `https://arxiv.org/abs/2406.14969`.

George Karypis and Vipin Kumar. A fast and high quality multilevel scheme for partitioning irregular graphs. *SIAM Journal on scientific Computing*, 20(1):359–392, 1998.

Thomas N. Kipf and Max Welling. Semi-Supervised Classification with Graph Convolutional Networks, February 2017.

Taojie Kuang, Yiming Ren, and Zhixiang Ren. 3D-Mol: A Novel Contrastive Learning Framework for Molecular Property Prediction with 3D Information, June 2024.

G Landrum. Rdkit: Open-source cheminformatics `http://www.rdkit.org`. 2016.

Yann LeCun and Courant. A path towards autonomous machine intelligence. 2022. URL `https://api.semanticscholar.org/CorpusID:251881108`.

Namkyeong Lee, Junseok Lee, and Chanyoung Park. Augmentation-Free Self-Supervised Learning on Graphs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(7):7372–7380, June 2022. ISSN 2374-3468. doi: 10.1609/aaai.v36i7.20700.

Shengchao Liu, Hanchen Wang, Weiyang Liu, Joan Lasenby, Hongyu Guo, and Jian Tang. Pre-training Molecular Graph Representation with 3D Geometry, May 2022a.

Shengchao Liu, Weitao Du, Zhi-Ming Ma, Hongyu Guo, and Jian Tang. A group symmetric stochastic differential equation model for molecule multi-modal pretraining. In *International Conference on Machine Learning*, pp. 21497–21526. PMLR, 2023a.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pp. 38–55. Springer, 2024.

Zhen Liu, Tetiana Zubatiuk, Adrian Roitberg, and Olexandr Isayev. Auto3d: Automatic generation of the low-energy 3d structures with ani neural network potentials. *Journal of Chemical Information and Modeling*, 62(22), 2022b. PMID: 36112860.

Zhen Liu, Yurii S Moroz, and Olexandr Isayev. The challenge of balancing model sensitivity and robustness in predicting yields: a benchmarking study of amide coupling reactions. *Chemical Science*, 14(39):10835–10846, 2023b.

Yizhen Luo, Kai Yang, Massimo Hong, Xing Yi Liu, and Zaiqing Nie. Molfm: A multimodal molecular foundation model. *arXiv preprint arXiv:2307.09484*, 2023.

Kha-Dinh Luong and Ambuj K. Singh. Fragment-based Pretraining and Finetuning on Molecular Graphs. *Advances in Neural Information Processing Systems*, 36:17584–17601, December 2023.

Andrei Manolache, Dragos Tantaru, and Mathias Niepert. MolMix: A Simple Yet Effective Baseline for Multimodal Molecular Representation Learning. In *Advances in Neural Information Processing Systems (NeurIPS), Machine Learning for Structural Biology Workshop*, 2024.

Duy M. H. Nguyen, Nina Lukashina, Tai Nguyen, An T. Le, TrungTin Nguyen, Nhat Ho, Jan Peters, Daniel Sonntag, Viktor Zaverkin, and Mathias Niepert. Structure-aware e(3)-invariant molecular conformer aggregation networks. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Paolo Pellizzoni, Till Hendrik Schulz, and Karsten Borgwardt. Graph neural networks can (often) count substructures. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=sZQRUrvLn4.

Raghunathan Ramakrishnan, Pavlo O Dral, Matthias Rupp, and O Anatole Von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific data*, 1(1):1–7, 2014.

Pierre H. Richemond, Jean-Bastien Grill, Florent Altché, Corentin Tallec, Florian Strub, Andrew Brock, Samuel Smith, Soham De, Razvan Pascanu, Bilal Piot, and Michal Valko. Byol works even without batch statistics, 2020. URL https://arxiv.org/abs/2010.10241.

Yu Rong, Yatao Bian, Tingyang Xu, Weiyang Xie, Ying WEI, Wenbing Huang, and Junzhou Huang. Self-Supervised Graph Transformer on Large-Scale Molecular Data. In *Advances in Neural Information Processing Systems*, volume 33, pp. 12559–12571. Curran Associates, Inc., 2020.

Kristof T. Schütt, Pieter-Jan Kindermans, Huziel E. Sauceda, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert Müller. SchNet: A continuous-filter convolutional neural network for modeling quantum interactions, December 2017.

Kristof T. Schütt, Oliver T. Unke, and Michael Gastegger. Equivariant message passing for the prediction of tensorial properties and molecular spectra, 2021. URL https://arxiv.org/abs/2102.03150.

Guillem Simeon and Gianni De Fabritiis. Tensornet: Cartesian tensor representations for efficient learning of molecular potentials. *Advances in Neural Information Processing Systems*, 36:37334–37353, 2023.

Guillem Simeon and Gianni de Fabritiis. TensorNet: Cartesian tensor representations for efficient learning of molecular potentials, October 2023.

Geri Skenderi, Hang Li, Jiliang Tang, and Marco Cristani. Graph-level Representation Learning with Joint-Embedding Predictive Architectures, January 2025.

Hannes Stärk, Dominique Beaini, Gabriele Corso, Prudencio Tossou, Christian Dallago, Stephan Günnemann, and Pietro Lió. 3D Infomax improves GNNs for Molecular Property Prediction. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 20479–20502. PMLR, June 2022.

Fan-Yun Sun, Jordan Hoffmann, Vikas Verma, and Jian Tang. InfoGraph: Unsupervised and Semi-supervised Graph-Level Representation Learning via Mutual Information Maximization, January 2020.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/68053af2923e00204c3ca7c6a3150cf7-Paper.pdf.

Saisai Teng et al. A self-conformation-aware pre-training framework for molecular property prediction with substructure interpretability. *arXiv (or conference / journal) — if available*, 2025. preprint.

Shantanu Thakoor, Corentin Tallec, Mohammad Gheshlaghi Azar, Mehdi Azabou, Eva L. Dyer, Rémi Munos, Petar Veličković, and Michal Valko. Large-Scale Representation Learning on Graphs via Bootstrapping, February 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Hanchen Wang, Jean Kaddour, Shengchao Liu, Jian Tang, Joan Lasenby, and Qi Liu. Evaluating self-supervised learning for molecular graph embeddings. *Advances in Neural Information Processing Systems*, 36:68028–68060, 2023a.

Sheng Wang, Yuzhi Guo, Yuhong Wang, Hongmao Sun, and Junzhou Huang. Smiles-bert: large scale unsupervised pre-training for molecular property prediction. In *Proceedings of the 10th ACM international conference on bioinformatics, computational biology and health informatics*, pp. 429–436, 2019.

Xu Wang, Huan Zhao, Wei-wei Tu, and Quanming Yao. Automated 3d pre-training for molecular property prediction. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2419–2430, 2023b.

Yuyang Wang, Jianren Wang, Zhonglin Cao, and Amir Barati Farimani. Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence*, 4(3):279–287, 2022.

Boris Weisfeiler and Andrei Leman. The reduction of a graph to canonical form and the algebra which appears therein. *nti, Series*, 2(9):12–16, 1968.

Zixin Wen and Yuanzhi Li. The mechanism of prediction head in non-contrastive self-supervised learning. *Advances in Neural Information Processing Systems*, 35:24794–24809, 2022.

Daniel S. Wigh, Jonathan M. Goodman, and Alexei A. Lapkin. A review of molecular representation in the age of machine learning. *WIREs Comput. Mol. Sci.*, 12(5):e1603, September 2022. ISSN 1759-0876. doi: 10.1002/wcms.1603.

Tom Wollschläger, Niklas Kemper, Leon Hetzel, Johanna Sommer, and Stephan Günnemann. Expressivity and generalization: Fragment-biases for molecular gnns. In *International Conference on Machine Learning*, 2024.

Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530, 2018.

Jun Xia, Chengshuai Zhao, Bozhen Hu, Zhangyang Gao, Cheng Tan, Yue Liu, Siyuan Li, and Stan Z. Li. Mole-BERT: Rethinking Pre-training Graph Neural Networks for Molecules, April 2023.

Yaochen Xie, Zhao Xu, and Shuiwang Ji. Self-Supervised Representation Learning via Latent Graph Prediction. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 24460–24477. PMLR, June 2022.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How Powerful are Graph Neural Networks?, February 2019.

Minghao Xu, Hang Wang, Bingbing Ni, Hongyu Guo, and Jian Tang. Self-supervised graph-level representation learning with local and global structure. In *International conference on machine learning*, pp. 11548–11558. PMLR, 2021.

Yuning You, Tianlong Chen, Yang Shen, and Zhangyang Wang. Graph contrastive learning automated. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12121–12132. PMLR, 18–24 Jul 2021a. URL https://proceedings.mlr.press/v139/you21a.html.

Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. Graph Contrastive Learning with Augmentations, April 2021b.

Qiying Yu, Yudi Zhang, Yuyan Ni, Shikun Feng, Yanyan Lan, Hao Zhou, and Jingjing Liu. Multimodal molecular pretraining via modality blending. *arXiv preprint arXiv:2307.06235*, 2023.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. Deep sets. *Advances in neural information processing systems*, 30, 2017.

Hengrui Zhang, Qitian Wu, Junchi Yan, David Wipf, and Philip S Yu. From Canonical Correlation Analysis to Self-supervised Graph Neural Networks. In *Advances in Neural Information Processing Systems*, volume 34, pp. 76–89. Curran Associates, Inc., 2021.

ZAIXI ZHANG, Qi Liu, Hao Wang, Chengqiang Lu, and Chee-Kong Lee. Motif-based Graph Self-Supervised Learning for Molecular Property Prediction. In *Advances in Neural Information Processing Systems*, volume 34, pp. 15870–15882. Curran Associates, Inc., 2021.

Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. Uni-mol: A universal 3d molecular representation learning framework. 2023.

Jinhua Zhu, Yingce Xia, Lijun Wu, Shufang Xie, Tao Qin, Wengang Zhou, Houqiang Li, and Tie-Yan Liu. Unified 2d and 3d pre-training of molecular representations. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 2626–2636, 2022.

Yanqiao Zhu, Jeehyun Hwang, Keir Adams, Zhen Liu, Bozhao Nan, Brock Stenfors, Yuanqi Du, Jatin Chauhan, Olaf Wiest, Olexandr Isayev, Connor W. Coley, Yizhou Sun, and Wei Wang. Learning Over Molecular Conformer Ensembles: Datasets and Benchmarks, July 2024.

LLM USAGE STATEMENT

During the course of this work, AI-assisted tools were employed to support specific aspects of the research and writing process:

- **Literature Search:** LLMs were used to identify potentially relevant papers. All references included in this work were thoroughly read, analyzed, and vetted by the authors to ensure their accuracy and relevance.
- **Summarization and Writing Support:** LLMs aided in editing, reformulating, and refining text for clarity, conciseness, and academic style. All AI-generated content was reviewed, adapted, and integrated by the authors to maintain coherence, accuracy, and contextual relevance.

All ideas, critical analyses, interpretations, and conclusions presented in this work are solely those of the authors.

A SUPPLEMENTARY MATERIALS

A.1 EXPRESSIVENESS AND INVARIANCE

We preserve the invariance properties of the encoders used in our framework. In particular, we restate the result of Manolache et al. (2024) for our setting, restricted to 2D graphs and 3D conformers:

Lemma 1. *Let G be a 2D molecular graph and $\{c_1, \ldots, c_k\}$ a set of k 3D conformers for the same molecule. Let $\hat{y} = f_\theta(G, \{c_1, \ldots, c_k\})$ be the encoder output as defined in Section 3. Suppose the 3D encoder is invariant to the actions of a group $\mathcal{G}$. Then f_θ is also invariant to any $T_1, \ldots, T_k \in \mathcal{G}$, i.e.,*

$$f_\theta(G, \{T_1 c_1, \ldots, T_k c_k\}) = f_\theta(G, \{c_1, \ldots, c_k\}).$$

The proof of this result is given in Manolache et al. (2024); it applies directly in our case after removing the SMILES modality.

Next, we provide the proof for Lemma 1 in the main paper.

Lemma 2. *Let C-FREE$_{DS}$ be a model as defined in Section 3 with a subgraph encoder f_θ consisting of a 1-WL MPNN (e.g., GIN/GINE) followed by a Transformer without positional encodings, and a DeepSets task head DS. For any k-EgoNet policy with $k \geq 1$ under the assumptions of Theorem 2 from (Bevilacqua et al., 2022), C-FREE$_{DS}$ is as expressive as ESAN (Bevilacqua et al., 2022) with an EGO policy, therefore it is as most as expressive as DS-WL and strictly more expressive than 1-WL.*

Proof. Fix a k-EGO policy $\pi = \text{EGO}_k$ with $k \geq 1$ and let $S_\pi(G)$ be the multiset of k-ego-nets with their complements (edge-covering in the ESAN sense). Define:

$$f_{C\text{-}FREE}(G) = DS\big(\{f_\theta(S) : S \in S_\pi(G)\}\big),$$

Due to Lemma 1, we have that f_θ maintains the permutation equivariance of the MPNN; moreover, since there exists a parametrization of the Transformer that can approximate the identity map arbitrarily well, the Transformer does not lower the expressive power of the MPNN. We therefore have that f_θ is as powerful as 1-WL.

Since we have a DeepSets encoder DS and an edge-covering k-EGO policy, we can use the same proof argument as in Theorem 2 from (Bevilacqua et al., 2022), i.e. we apply f_θ to each $S \in S_\pi(G)$ and then aggregate the multisets with DS, therefore $f_{C\text{-}FREE}$ simulates ESAN, and is at most as expressive as *DS-WL* and strictly more expressive than 1-WL.

□

A.2 Prevention of Collapse in Latent Learning

While the target encoder is updated via Exponential Moving Average to stabilize training (Tarvainen & Valpola, 2017; Grill et al., 2020b; Assran et al., 2023), this alone is insufficient: without an asymmetric architecture, the network converges to a trivial solution. Introducing a predictor breaks this symmetry and is hypothesized to empirically prevent collapse (Richemond et al., 2020). Even a simple two-layer MLP improves performance, while a transformer-based predictor yields the strongest results. This is further supported by the theoretical work of Wen & Li (2022), which shows that, in a simplified two-feature setting, a trainable prediction head avoids dimensional collapse by leveraging two mechanisms: (i) the *substitution effect*, where strong features learned by some neurons substitute for others, and (ii) the *acceleration effect*, where such substitutions accelerate the learning of weaker features, ultimately enabling the network to capture a more diverse set of representations. Therefore, when extended to settings with more features and a higher-capacity predictor—such as the transformer used here—these mechanisms are expected to further enhance performance.

A.3 Backbone Parameter Efficiency

Table 7 compares the number of trainable parameters across different SSL backbones. For our method, we report both the total parameters and the encoder-only count used during downstream evaluation. This distinction arises because only the target encoder is retained as the backbone, during downstream tasks, while the predictor is discarded. As a result, nearly half of the parameters are removed at this stage, allowing our method to remain competitive without increasing the parameter load for downstream evaluation, further underscoring its efficiency.

Table 7: Computational efficiency of different SSL methods from Wang et al. (2023a), showing the number of trainable parameters for each backbone. We report both the total parameters of our backbone and those of the encoder alone, since only the latter is used for downstream evaluation. By discarding nearly half of the backbone parameters in this stage, our approach remains competitive without increasing the parameter count for downstream tasks, further highlighting its efficiency.

METHOD	#PARAMETERS (MILLION)
EDGEPRED	7.46
ATTRMASK	7.61
GPT-GNN	7.61
INFOGRAPH	7.82
GROVER	7.57
CONT.PRED	12.00
GRAPHCL	8.19
JOAO	8.19
GRAPHMVP	15.84
C-FREE$_{2D}$ (FULL)	8.09
C-FREE$_{2D}$ (ENCODER)	4.67
C-FREE$_{MM}$ (FULL)	14.65
C-FREE$_{MM}$ (ENCODER)	9.12

Table 8: Average runtime (in milliseconds) for generating a single subgraph on the GEOM dataset, comparing METIS partitions with $n \in \{16, 32\}$ patches and k-EgoNets with $k \in \{3, 4\}$.

	METHOD	AVG. RUNTIME (MS)
METIS	N = 32	1.123
	N = 16	1.031
K-EGONETS	K = 3	0.171
	K = 4	0.185

A.4 COMPUTATIONAL COMPLEXITY ANALYSIS

For generating the subgraphs used as input units in our pre-training scheme, we employ k-EgoNets with fixed radii $k \in \{3, 4\}$. We extract k-hop neighborhoods using PyTorch Geometric's (Fey & Lenssen, 2019) `k_hop_subgraph` function, which performs a breadth-first search (BFS) from each node and collects all nodes reachable within k hops. For constant k, the BFS cost is bounded by the number of explored edges, yielding a worst-case complexity of $O(|E|)$. When repeated for all k radii, this results in $O(k \cdot |E|)$, where $|E|$ denotes the total number of edges. In practice, the number of explored edges is proportional to the average degree d, giving a total cost of $O(k \cdot d \cdot |V|)$, where $|V|$ is the number of nodes. Since molecular graphs are sparse, neighborhood growth is modest: an analysis of the GEOM dataset used for pre-training shows that the average degree is only $d = 2.1$. As a result, k-hop neighborhoods remain small, and `k_hop_subgraph` is computationally efficient, scaling linearly with the number of nodes $O(|V|)$ in practice.

In comparison, the METIS algorithm (Karypis & Kumar, 1998) used in GraphJEPA (Skenderi et al., 2025) employs a multilevel graph partitioning approach. While the algorithm is often cited with an overall complexity of $O(|E| \cdot \log |V|)$ in practice, it consists of three main phases: (1) a coarsening phase that uses heavy-edge matching to successively reduce the graph size, (2) an initial partitioning phase that partitions the smallest coarsened graph (with negligible complexity due to its small size), and (3) an uncoarsening/refinement phase that projects the partition back to the original graph while refining it at each level. The coarsening and refinement phases dominate the computational cost, each contributing $O(|E| \cdot \log |V|)$ complexity. However, since molecular graphs are sparse, similar to above, this results in a total complexity of $O(d \cdot |V| \cdot \log |V|)$, which is theoretically higher than that of fixed-radius EgoNets.

To further validate this, we ran timed experiments comparing the generation of k-EgoNet subgraphs with the generation of METIS partitions on the GEOM dataset. Section A.4 reports the average runtime of each method, computed over all graphs in the dataset.

Table 9: Performance on SPICE (Eastman et al., 2023)with full end-to-end fine-tuning. The numbers are taken from the TensorNet (Simeon & De Fabritiis, 2023) paper. C-FREE outperforms both TensorNet and the Equivarient Transformer, which were trained under a fully supervised setting.

METHOD	ENERGIES (MEV) (↓)
TENSORNET	25.0
EQU. TRANSFORMER	31.2
OURS	**22.8**

Table 10: Performance on ZINC (Gómez-Bombarelli et al., 2018)with full end-to-end fine-tuning. C-FREE outperforms GraphJEPA despite not relying on positional encoding.

MODEL	ZINC (↓)
GRAPHJEPA	$0.434_{\pm 0.014}$
C-FREE (OURS)	$\mathbf{0.204}_{\pm 0.015}$

A.5 ADDITIONAL DETAILS ON EMPIRICAL EVALUATION

For the multi-modal variant, the context encoder consists of three components: a 6-layer GINE (Xu et al., 2019) with hidden dimension 128, a SchNet with hidden dimension 128, 6 interaction steps, and a cutoff of 10, and 6 Transformer layers with 8 heads and hidden dimension 512. For the pre-trained 2D variant, we use the same GINE configuration and the same number of Transformer layers and heads, but reduce the hidden dimension to 387. In both variants, the predictor is implemented as 4 Transformer layers with 4 heads each. The parameters are updated via backpropagation using the Adam optimizer, while the target encoder is updated through an exponential moving average (EMA) schedule, with the decay rate τ_t gradually increasing from 0.995 to 1.0 over the course of training.

Since the EMA decay reaches $\tau_t = 1$ in the final epoch, the context and target encoders converge to identical parameters. Nevertheless, we follow Assran et al. (2023) and report results using the target encoder.

For the choice of the scheduler we opt for a cosine scheduler without warmup. We notice that using a very small learning rate prevented convergence, while a moderate learning rate caused an early loss drop followed by stagnating representations. Adding a warmup phase allows the model to adapt gradually before the cosine decay, improving stability and representation learning. Thus we begin with a learning rate of 2×10^{-6} over 30 epochs warmup up to 5×10^{-5} and a patience of 50 epochs. For the batch size we use 256 and a weight decay of 0.04 and train for 200 epochs.

All experiments were performed on a mix of Nvidia A100/RTX 4090 GPUs and AMD EPYC 7713/Intel Xeon W-2225 CPUs for both the pre-training and downstream experiments. All experiments consumed a total of approximately 500 GPU hours, with the longest compute being consumed on the pre-training backbone run on the GEOM dataset with a total of 25 hours.

A.6 ADDITIONAL DETAILS AND RESULTS

On the ZINC benchmark (Table 10), our method shows a clear performance improvement over GraphJEPA. This suggests that the gains come from the combination of architectural choices and the overall training pipeline, with EgoNets contributing as one part of this broader design. On SPICE (Table 9), our method achieves results competitive with TensorNet and outperforms the equivariant transformer.

In Fig. 5, we illustrate the downstream pipeline. We retain the target encoder from pre-training and attach a prediction head. For the linear probe using the full molecule, we feed the entire molecule into the encoder, perform a single forward pass, and apply a linear head for prediction. For the subgraph-based variant, each subgraph is processed independently by the encoder; their representations are then aggregated using a DeepSets module, and the resulting aggregated embedding is passed through a linear head for the final prediction.

In Table 11, we summarize the tasks and dataset sizes of the MoleculeNet benchmarks. Likewise, in Table 12, we present an overview of the GEOM, Drugs-75K, and Kraken datasets, including the corresponding number of conformers.

In Fig. 6 we report the results on the Sterimol B5 and Sterimol BurB5 targets from the Kraken dataset following the experiment in Section 4.1.

In addition to the linear probe, we also report results for full fine-tuning on the Kraken dataset. For this, we use the same setup as in Section 4.1, with an MLP head, and fine-tune the model end-to-end.

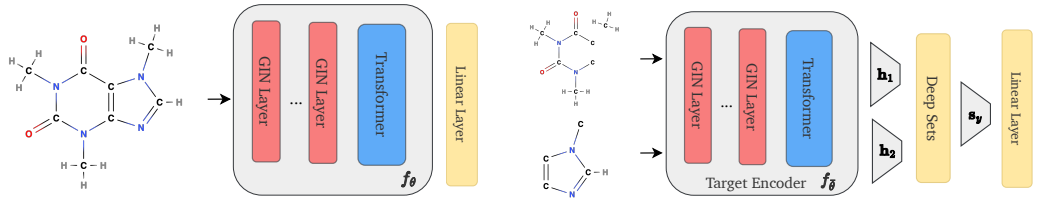

Figure 5: Fine-tuning strategies for C-FREE. Left: C-FREE$_{\text{MOL}}$: use the pretrained encoder to produce a whole-molecule embedding and attach a lightweight task head (e.g., linear probe or small MLP) for downstream prediction. Right: C-FREE$_{\text{DS}}$: aggregate multiple k-EgoNet subgraph embeddings with a DeepSets aggregator and apply a lightweight task head to the aggregated representation. For evaluation we typically use a linear probe on a frozen backbone to assess representation quality; for downstream tasks (regression, classification) the probe can be replaced by a task-specific head and the backbone optionally fine-tuned.

Table 11: Overview of tasks and sizes for the MoleculeNet datasets.

	BBBP	TOX21	TOXCAST	SIDER	CLINTOX	MUV	HIV	BACE
# MOLECULES	2,039	7,831	8,575	1,427	1,478	93,087	41,127	1,513
# TASKS	1	12	617	27	2	17	1	1

Table 12: Overview of tasks and sizes for the GEOM, Drugs-75K and Kraken datasets.

	GEOM	DRUGS-75K	KRAKEN
# MOLECULES	304,466	75,099	1,552
# CONFORMERS	25M	558,002	21,287
# TASKS	-	3	4

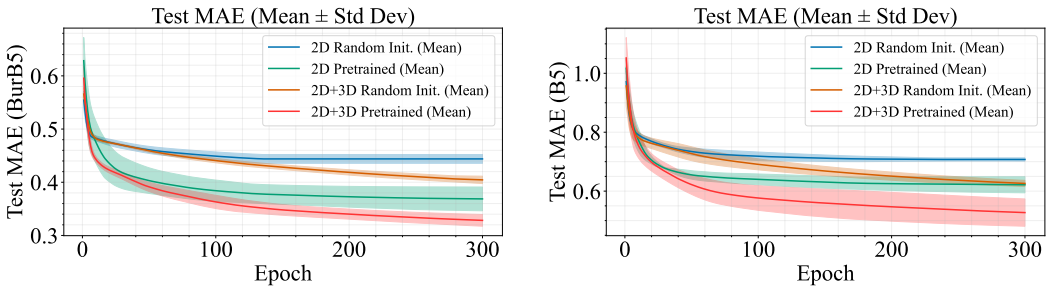

Figure 6: Test MAE on the Kraken regression tasks (Sterimol BurL and Sterimol L) with frozen backbones. GEOM-pretrained models consistently outperform random initialization for both 2D-only and multimodal variants. Pretrained models begin with lower error and converge faster, while randomly initialized models fail to match performance even after 300 epochs. Incorporating the 3D modality yields further gains over 2D-only backbones, with pretraining amplifying this advantage. Curves show the mean over 3 runs, with shaded regions indicating the standard deviation.

In Fig. 7, we show results for the 2D-only variant to highlight the initial performance gap, where the pre-trained backbone converges faster than the randomly initialized one.

Finally, in Table 13, we report the explicit numerical results of the probe presented in Section 4.3, evaluated after the full convergence of the model. Additionally, we include the results from end-to-end fine-tuning using the same experimental setup, providing a complete view of how the model performs when the entire backbone is updated. These fine-tuning results further confirm our observation that the predictor transformer consistently outperforms the MLP predictor, while omitting the predictor altogether leads to substantially worse performance, highlighting the importance of the predictor design in our framework.

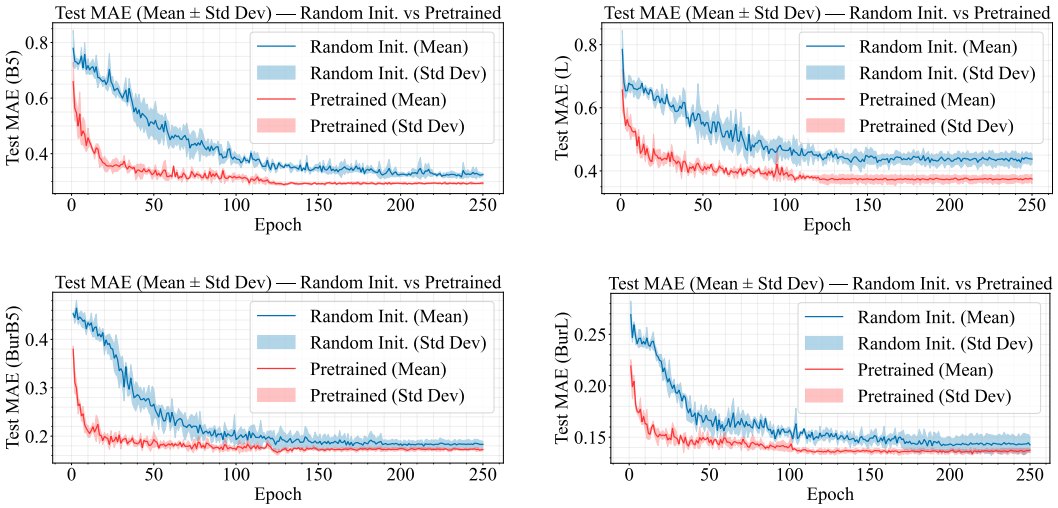

Figure 7: Test MAE on Kraken regression tasks (Sterimol L,B5, BurB5 and BurL) comparing random initialization and 2D-only GEOM-pretrained models. Pretrained models start with lower error and converge faster, while randomly initialized models fail to match their performance even after 250 epochs. Curves show the mean over 3 runs, with shaded regions indicating the standard deviation.

Table 13: Ablation study on the Kraken dataset (MAE ↓). We keep the encoder fixed and compare three predictors: (1) none, (2) a 2-layer MLP, and (3) a transformer. The transformer consistently achieves the best performance. The gap is especially pronounced in the linear probe (LIN. P.) setting, where the quality of the learned representations matters most. Even with full fine-tuning (FT), the no-predictor and MLP variants fail to match the transformer predictor.

	METHOD	B5 ↓	L ↓	BURB5 ↓	BURL ↓
FT	NONE	$0.381_{\pm 0.023}$	$0.494_{\pm 0.020}$	$0.202_{\pm 0.009}$	$0.157_{\pm 0.004}$
	2-LAYERS MLP	$0.315_{\pm 0.017}$	$0.396_{\pm 0.018}$	$0.185_{\pm 0.009}$	$\mathbf{0.144}_{\pm 0.004}$
	TRANSFORMER	$\mathbf{0.292}_{\pm 0.006}$	$\mathbf{0.380}_{\pm 0.023}$	$\mathbf{0.180}_{\pm 0.014}$	$0.146_{\pm 0.004}$
LIN. P.	NONE	$1.065_{\pm 0.001}$	$0.814_{\pm 0.001}$	$0.624_{\pm 0.001}$	$0.296_{\pm 0.001}$
	2-LAYERS MLP	$0.817_{\pm 0.002}$	$0.687_{\pm 0.008}$	$0.514_{\pm 0.001}$	$0.266_{\pm 0.001}$
	TRANSFORMER	$\mathbf{0.588}_{\pm 0.004}$	$\mathbf{0.554}_{\pm 0.007}$	$\mathbf{0.347}_{\pm 0.003}$	$\mathbf{0.202}_{\pm 0.008}$

A.7 ADDITIONAL RELATED WORK

Contrast-Free In this paper, our use of the term "contrast-free" follows the standard terminology in the representation learning literature, where methods such as BYOL (Grill et al., 2020b), DINO (Liu et al., 2024), and JEPA (Assran et al., 2023) are described as contrast-free because they avoid explicit contrastive objectives, large negative sets, and the large batch requirements of InfoNCE-style losses. We use the term in the same sense here: our model does not rely on negative pairs, large batch sizes, or explicit contrastive objectives, which helps avoid the computational overhead and instability often seen in contrastive training.

Contrastive Learning UniCorn (Feng et al., 2024) presents a unified contrastive learning framework that integrates multiple molecular views and existing methods into a single pre-training approach. 3D-Mol (Kuang et al., 2024) leverages 3D conformational information by constructing hierarchical graphs and applying contrastive learning to differentiate molecular conformations. GraphFP (Luong & Singh, 2023) captures higher-level connectivity through fragments—representations of molecular substructures—aligning fragment embeddings with their corresponding graph regions to enable multi-resolution structural learning. MolCLR (Wang et al., 2022) employs three graph augmentations—atom masking, bond deletion, and subgraph removal—and uses contrastive learning to bring augmented views of the same molecule closer, and Galformer (Bai et al., 2023) applies dual-view contrastive learning. Finally, GraphLoG (Xu et al., 2021) captures

both local similarities and global semantic clusters in whole-graph representations using hierarchical prototypes trained via an online EM algorithm.

Generative Learning GraphMAE (Hou et al., 2022) adapts the Masked Autoencoder (MAE) from the vision domain to graphs, focusing on reconstructing node attribute features using a scaled cosine error. Similarly, MoleBERT (Xia et al., 2023) extends this idea with a VQ-VAE-based context-aware tokenizer that encodes atom attributes into a larger, chemically meaningful discrete vocabulary, enabling a Masked Atoms Modeling (MAM) task where GNNs predict masked atom codes rather than raw features. Finally, MGSSL (ZHANG et al., 2021) leverages a BRICS-based fragmentation method to extract molecular motifs and pre-trains GNNs to predict motif topology and labels, incorporating multi-level pre-training to capture both local and global graph information.

Latent Representation Learning CCA-SSG (Zhang et al., 2021) introduces an alignment objective based on Canonical Correlation Analysis, encouraging the latent features of two augmented views to be maximally correlated while remaining de-correlated across dimensions. Complementing these augmentation-based methods, AFGRL (Lee et al., 2022) proposes a more principled strategy for view generation by identifying structurally and semantically similar anchor nodes, mitigating the reliance on handcrafted augmentations.

Multi-modal Foundation Models Early SMILES-only models such as ChemBERTa-v1 (Chithrananda et al., 2020) and its improved successor ChemBERTa-v2 (Ahmad et al., 2022) employ transformer-based language modeling to provide strong molecular embeddings from sequence data alone. Beyond purely textual inputs, SCAGE (Teng et al., 2025) incorporates 3D conformational information through a self-conformation-aware graph transformer. UniMol-v1 (Zhou et al., 2023) further established 3D-informed pretraining by combining SE(3)-equivariant architectures with large-scale conformer data, while UniMol-v2 (Ji et al., 2024) scaled this approach to the billion-parameter regime with substantially expanded datasets. More recently, MolFM (Luo et al., 2023) introduced a multimodal formulation integrating molecular graphs, biomedical text, and knowledge-graph information into a unified embedding space. GEM (Graph Foundation Model) (Fang et al., 2024) and its updated variant extend this multimodal perspective by combining structural, physical, and textual molecular descriptors at scale.