GENERAL-PURPOSE PRE-TRAINED MODEL TOWARDS CROSS-DOMAIN MOLECULE LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Self-supervised pre-training on biomolecules has achieved remarkable success in various biochemical applications, such as drug discovery and protein design. However, in most approaches, the learning model is primarily constructed based on the characteristics of either small molecules or proteins, without exploring their potential binding interactions – an essential cross-domain relationship crucial for driving numerous biological processes. In this paper, inspired by the success of multimodal learning, we fill this gap by proposing a general-purpose foundation model named **BIT** (an abbreviation for **B**iomolecular **I**nteraction **T**ransformer), which is capable of encoding a range of biochemical entities, including small molecules, proteins, and protein-ligand complexes, as well as various data formats, encompassing both 2D and 3D structures, all within a shared Transformer backbone, via multiple unified self-supervised atom-level *denoising* tasks. We introduce *Mixture-of-Domain-Experts* (MoDE) to handle the biomolecules from diverse chemical domains and incorporate separate structural channels to capture positional dependencies in the molecular structures. The proposed MoDE allows BIT to enable both deep fusion and domain-specific encoding and learn cross-domain relationships on protein-ligand complexes with 3D cocrystal structures. Experimental results demonstrate that BIT achieves exceptional performance in both protein-ligand binding and molecular learning downstream tasks, including binding affinity prediction, virtual screening, and molecular property prediction.

1 INTRODUCTION

In the past few years, self-supervised pre-training of the foundation model has witnessed remarkable success in natural language processing (Devlin et al., 2018; Brown et al., 2020) and computer vision (Chen et al., 2020; He et al., 2022). Recently, pre-training on biomolecules has attracted growing attention. By fine-tuning the large-scale pre-trained model, one can significantly improve the performance on diverse biological downstream tasks, such as molecular property prediction (Rong et al., 2020), protein structure prediction (Lin et al., 2023) and protein design (Madani et al., 2023). Thus, substantial efforts have been devoted to biomolecule pre-training to leverage the potential inherent in the large-scale unlabeled molecule corpus, including molecular graphs and protein sequences (Hu et al., 2020; Rives et al., 2021). However, the majority of existing approaches are tailored to a single data domain, focusing on either small molecules or proteins. This restricts the pre-trained model from capturing molecular interactions across distinct chemical domains. Consequently, it limits the learning performance in the downstream tasks that highly depend on this information, such as structure-based binding affinity prediction and virtual screening (Ain et al., 2015).

The interactions between proteins and small molecules, known as ligands, play a pivotal role in orchestrating molecular-level biological processes (Tomasi & Persico, 1994). Understanding the underlying principles of these interactions is crucial in scientific fields. Besides, it is essential for a range of biomedical applications, particularly in structure-based drug discovery (SBDD) (Anderson, 2003; Vamathevan et al., 2019). To be more precise, by considering the 3D geometry of the binding pocket on the target protein, medicinal chemists strive to identify prospective drug candidates capable of modulating the function of the target. Recent studies have shown the significant potential of deep learning methods in modeling molecular interactions and facilitating the SBDD process (Li et al., 2021b; Luo et al., 2021; Corso et al., 2022). However, given the vastness of the chemical space

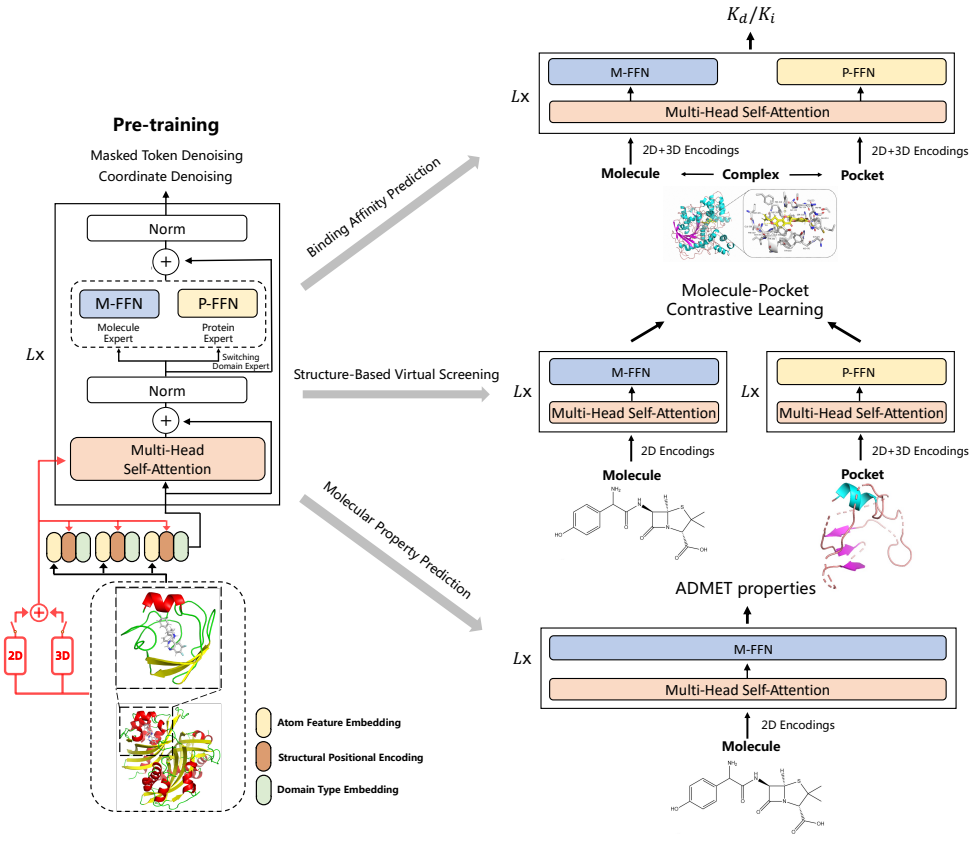

Figure 1: Overview of BIT pre-training and fine-tuning. We perform denoising tasks on protein-ligand complexes with a general-purpose foundation model named BIT. During fine-tuning, BIT can be used as (a) fusion encoder for binding affinity prediction; (b) dual encoder for virtual screening; (c) molecule encoder for molecular property prediction.

and the scarcity of task-specific labeled data, building a highly accurate prediction model for such problems remains a significant challenge.

More recently, preliminary attempts have been dedicated to pre-training a model that is explicitly designed to capture cross-domain dependencies between protein pockets and ligands, as exemplified by CoSP (Gao et al., 2022). Specifically, it distinguishes the two different domains as independent signals and adopts a dual-encoder architecture similar to CLIP (Radford et al., 2021) to encode pockets and ligands separately, then employs contrastive learning (Oord et al., 2018; Chuang et al., 2020) to learn a shared embedding space in which bindable pockets and ligands are pulled closer, while unbindable pocket-ligand pairs are pushed apart. However, the molecular interaction is merely inferred through a shallow interaction module, which involves the dot product of the pocket and ligand feature vectors, without considering inter-molecular connection information. As a result, CoSP remains unsatisfactory for addressing complex protein-ligand binding tasks. To effectively integrate both intra-molecular and inter-molecular interactions from protein-ligand complexes, a more intricate alignment between proteins and ligands is required.

Considering the aforementioned issues, it is straightforward to adopt multimodal learning (Xu et al., 2023b) to leverage all essential information available across diverse chemical domains, encompassing various data formats. The Transformer (Vaswani et al., 2017) is chosen as a preferred backbone as its variants have demonstrated effectiveness in modeling text (Devlin et al., 2018), images (Dosovitskiy et al., 2021), graphs (Rampášek et al., 2022), and molecule (Ying et al., 2021). Furthermore, it can handle multiple modalities in a unified manner. For example, Transformer-based multimodal models (Kim et al., 2021; Wang et al., 2023a) have significantly advanced vision-language understanding and generation ability, resulting in improved performance across various vision-language tasks.

In this work, we present a general-purpose pre-trained model within the *protein-ligand pre-training* paradigm. This model encodes molecules across a wide range of biochemical domains, including small molecules, proteins, and protein-ligand complexes, as well as diverse data formats, encompassing both 2D and 3D structures, all within a unified Transformer framework, referred to as **B**iomolecular **I**nteraction **T**ransformer (**BIT**). We construct our model upon Transformer-M (Luo et al., 2022), a model renowned for its flexibility and effectiveness in handling both 2D and 3D structural data. Then we enhance it to capture multi-domain specificity and inter-domain relationships by incorporating *Mixture-of-Domain-Experts* (MoDE). In each Transformer block, we replace the feed-forward network with two distinct domain experts, namely the molecule expert and the protein expert, while retaining a shared self-attention module across domains to facilitate alignment between different domains. In BIT, each input atom token is routed to its respective domain expert, enabling BIT to function as either a fusion encoder to model molecular interactions in protein-ligand complexes or as a dual encoder to separately encode small molecules and proteins.

To learn more precise cross-domain representations, we pre-train BIT on both protein-ligand complexes with 3D cocrystal structures (Wei et al., 2023) and large-scale small molecules with 3D equilibrium structures (Hu et al., 2021) in a unified manner via denoising tasks for both continuous atom coordinates and categorical atom types. Finally, we demonstrate the superiority of the proposed BIT through extensive experiments across various downstream tasks, including both protein-ligand binding and molecular learning. When employed as a fusion encoder, BIT consistently outperforms specialized baselines by a decent margin in binding affinity prediction. When used as a dual encoder, BIT still achieves state-of-the-art performance while offering significantly faster inference speed in virtual screening. Moreover, BIT surpasses related state-of-the-art pre-trained models in a series of molecular property prediction tasks.

The main contributions of this work are summarized as follows:

- We present BIT, a general-purpose foundation model designed to encode a range of biochemical entities, including small molecules, proteins, and protein-ligand complexes, across various data formats, encompassing both 2D and 3D structures, all within a unified Transformer backbone.
- We pre-train BIT on protein-ligand complexes with 3D cocrystal structures, alongside large-scale small molecule ligands with 3D equilibrium structures, to learn cross-domain biomolecule relationships using cross-domain attention.
- Experiments verify that BIT achieves exceptional performance in downstream tasks, including both protein-ligand binding and molecular learning, after fine-tuning.

2 Related work

2.1 Molecular representation learning

Learning meaningful and effective molecular representations is fundamental to AI-driven drug discovery. Self-supervised pre-training serves as a powerful tool in this area, thanks to the availability of the abundance of molecule data. Recently, several self-supervised pre-training models have been proposed separately for small molecules or proteins.

Pre-training on small molecules: Initially, researchers employ sequence-based pre-training strategies on string-based molecular data such as SMILES (Weininger, 1988). Representative works include SMILES-BERT (Wang et al., 2019) and ChemBERTa (Chithrananda et al., 2020). As molecular graphs can provide richer 2D structural information, more efforts (Hu et al., 2020; Rong et al., 2020; Wang et al., 2022a) have focused on pre-training graph neural networks (Xu et al., 2019) or Transformers (Vaswani et al., 2017) on molecular graphs. Moreover, there are recent studies aiming to explore pre-training on 3D molecular structures to improve model performance in predicting molecular properties using 3D structural data (Zaidi et al., 2022; Zhou et al., 2023; Feng et al., 2023).

Pre-training on proteins: Learning effective protein representations is also of great importance, such as for protein understanding and generation. Protein language models have achieved remarkable success in capturing biological co-evolutionary information from millions of diverse protein sequences (Elnaggar et al., 2021; Lin et al., 2023), or families of evolutionarily related sequences (Rao et al., 2021). Beyond these sequence-based approaches, there is a growing interest in exploring pre-training techniques for protein structures (Zhou et al., 2023; Zhang et al., 2023).

For more comprehensive reviews, we refer the reader to Xia et al. (2023b); Ferruz & Höcker (2022). While most prior work constructed models based on the characteristics of either small molecules or proteins, our work aims to enhance molecular representation learning by incorporating additional cross-domain relationships learned from biologically relevant protein-ligand complexes.

2.2 MULTIMODAL REPRESENTATION LEARNING

In recent years, multimodal representation learning has gained significant attention and has been extensively studied to enhance understanding across various areas, including image analysis (Radford et al., 2021), video processing (Sun et al., 2019), and speech recognition (Ao et al., 2022), often by incorporating additional textual information. Among these multimodal learning applications, Transformer (Vaswani et al., 2017; Dosovitskiy et al., 2021) has become a critical building block, thanks to its outstanding performance in learning from the corresponding monomodal data and its flexibility in aligning and integrating information across multimodal data sources. There are three main types of architectures to cater to different multimodal learning requirements: dual encoder (Radford et al., 2021; Jia et al., 2021) for efficient retrieval, fusion encoder (Kim et al., 2021; Li et al., 2021a) for deep understanding, and encoder-decoder architectures (Wang et al., 2022b) for generation. Some research (Li et al., 2022; Bao et al., 2022; Wang et al., 2023a) have explored effective ways to integrate the strengths of these architectures. Recently, multimodal learning has also found applications in the biomedical field. There have been early attempts to enhance molecular representation learning by leveraging the correspondence and consistency between 2D topological structures and 3D geometric views (Liu et al., 2022; Stärk et al., 2022; Liu et al., 2023) or incorporating biomedical text (Liu et al., 2022; Xu et al., 2023a).

2.3 STRUCTURE-BASED DRUG DISCOVERY

Structure-based drug discovery (SBDD) refers to a systematic scientific approach to design and develop new drugs by leveraging the detailed physical structure of the binding protein or molecular target. It involves analyzing the structure of the target, understanding its function, and designing molecules capable of interacting with the target in a specific and favorable manner to regulate its activity. To complement labor-intensive traditional methods, geometric deep learning algorithms (Atz et al., 2021) have recently been proposed to improve the efficiency and performance of various stages in the SBDD process, including binding site identification (Sverrisson et al., 2021), affinity prediction (Li et al., 2021b), virtual screening (Torng & Altman, 2019), *de novo* molecule generation (Luo et al., 2021), etc. Our proposed versatile model will further streamline and enhance this process.

3 METHODS

In this section, we present the details of BIT, a general-purpose pre-trained model designed to encode molecules across various biochemical domains, including small molecules, proteins, and protein-ligand complexes, in different data formats, including 2D and 3D structures. BIT can be fine-tuned as either a fusion encoder to model intricate molecular interactions within protein-ligand complexes for precise binding affinity prediction, or as a dual encoder to enable efficient virtual screening.

3.1 INPUT REPRESENTATIONS

In biochemical applications, data are collected in the form of molecules represented at different levels of granularity, such as atoms, residues, and nucleobases. However, all molecules can be uniformly represented as sets of atoms held together by attractive or repulsive forces. To more effectively capture and transfer atom-level knowledge across different domains, we propose to share atom embeddings and incorporate domain embeddings to distinguish between small molecules and proteins.

Both small molecule, denoted as $\mathcal{M}$, and protein, denoted as $\mathcal{P}$, can be represented as a geometric graphs of atoms $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. Here $\mathcal{V} = (\boldsymbol{X}, \vec{R})$ includes all atoms and $\mathcal{E}$ includes all chemical bonds. In a molecule consisting of n atoms, $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ denotes a set of atom feature vectors, $\vec{R} \in \mathbb{R}^{n \times 3}$ denotes a set of atom Cartesian coordinates, and $e_{ij} \in \mathcal{E}$ denotes the feature vector of the edge between atoms i and j if the edge exists. The molecule and protein input representations are computed via summing atom feature embeddings $\boldsymbol{X}$, structural positional encodings $\Psi \in \mathbb{R}^{n \times d}$ (Ying et al.,

2021; Shi et al., 2022), and the corresponding domain-type embedding vectors $\boldsymbol{m}_{\text{type}}, \boldsymbol{p}_{\text{type}} \in \mathbb{R}^d$. Following Ying et al. (2021), we introduce special virtual nodes [M_VNode] for small molecules and [P_VNode] for proteins, and make connection between virtual node and each atom node individually.

It is noteworthy that we only use the binding pocket as the model input rather than the entire protein primarily for the following two reasons: (1) the binding pocket is the paramount region of protein-ligand interaction, experiencing the most significant spatial alterations during the binding process and providing sufficient insight into molecular interactions; (2) the binding pocket contains significantly fewer atoms than the entire protein, leading to lower computational costs and faster training speeds.

Given a protein-ligand complex $< \mathcal{M}, \mathcal{P} >$ with 3D cocrystal structures, we first identify the binding pocket as the protein atoms located within a minimum distance of 5 Å from the ligand, as suggested in Muegge & Martin (1999). Then we input the extracted pocket-ligand complex into BIT to learn contextualized representations.

3.2 BACKBONE

Recently, several studies have extended the Transformers to model molecules (Rong et al., 2020; Ying et al., 2021; Rampášek et al., 2022; Luo et al., 2022). The vanilla Transformer architecture comprises stacked Transformer blocks (Vaswani et al., 2017). Each Transformer block consists of two components: a multi-head self-attention (MSA) layer followed by a feed-forward network (FFN). Layer normalization (LN) (Ba et al., 2016) is applied after both the MSA and FFN. Let $\boldsymbol{H}_{l-1}$ denotes the input, the l-th Transformer block works as follows:

$$\boldsymbol{H}'_l = \text{LN}(\text{MSA}(\boldsymbol{H}_{l-1}) + \boldsymbol{H}_{l-1}) \tag{1}$$

$$\boldsymbol{H}_l = \text{LN}(\text{FFN}(\boldsymbol{H}'_l) + \boldsymbol{H}'_l) \tag{2}$$

For our general-purpose modeling, we start with Transformer-M (Luo et al., 2022), a model known for its versatility and effectiveness in handling both 2D or 3D molecule data. To provide a comprehensive overview, we briefly introduce the core concept of Transformer-M here and recommend that readers refer to Luo et al. (2022) for more technical details. Transformer-M introduces two separate channels to encode 2D and 3D structural information and integrate them into the MSA module as bias terms. The modified attention matrix $\boldsymbol{A}$ is calculated as:

$$\boldsymbol{A}(\boldsymbol{H}) = \text{softmax}\left(\frac{\boldsymbol{H}\boldsymbol{W}_Q(\boldsymbol{H}\boldsymbol{W}_K)^\top}{\sqrt{d_K}} + \underbrace{\Phi^{\text{SPD}} + \Phi^{\text{Edge}}}_{\text{2D pair-wise channel}} + \underbrace{\Phi^{\text{3D Distance}}}_{\text{3D pair-wise channel}}\right) \tag{3}$$

where $\boldsymbol{W}_Q, \boldsymbol{W}_K \in \mathbb{R}^{d \times d_K}$ are learnable weight matrices, the 2D terms (Φ^{SPD} and Φ^{Edge}) and the 3D term ($\Phi^{\text{3D Distance}}$) originate from Ying et al. (2021) and Shi et al. (2022) respectively. To simplify the illustration, we omit the attention head index h and layer index l. When molecules are associated with specific 2D or 3D structural information, the corresponding channel will be activated, while the other will be disabled. In combination with the dropout-like 2D-3D joint training strategy (Luo et al., 2022), where the format of structural information for each data instance is randomly selected, Transformer-M learns to identify chemical knowledge from different data formats and generates meaningful semantic representations for each data format.

To further encode molecules across biochemical domains and learn cross-domain molecular representations enriched with molecular interaction knowledge, we propose to extend Transformer-M with a Mixture-of-Domain-Experts (MoDE) mechanism, employing specialized expert networks for different domains. As shown in Figure 1, each Transformer block in BIT consists of a shared MSA module and two FFNs, presenting domain experts, namely the molecule expert and the protein expert. In contrast to conventional mixture-of-experts layer (Shazeer et al., 2017; Fedus et al., 2022), which routes input tokens by a trainable gating network, we directly assign an expert to process each atom token based on its molecule data domain. Sharing the MSA module encourages the model to align protein and ligand, while employing MoDE in place of the FFN encourages the model to capture domain-specific knowledge. The Transformer block of BIT can be abstractly summarized as follows:

$$\boldsymbol{H}'_l = \text{LN}(\text{MSA-M}(\boldsymbol{H}_{l-1}) + \boldsymbol{H}_{l-1}) \tag{4}$$

$$\boldsymbol{H}_l = \text{LN}(\text{MoDE-FFN}(\boldsymbol{H}'_l) + \boldsymbol{H}'_l) \tag{5}$$

where MSA-M denotes the variant of MSA used in Transformer-M.

Thanks to MoDE, BIT decouples the encoding process across different domains. As a result, BIT can be fine-tuned to function as either a fusion encoder or a dual encoder, depending on the specific formulation of various downstream protein-ligand binding tasks. Further discussion on this aspect is presented in Section 3.4.

Remarks. Our proposed MoDE seamlessly integrates with Transformer-M. Both of them employ shared self-attention modules for unified modeling and use distinct parameters to capture the data specificity among different inputs. The main difference is that MoDE captures instance-level domain specificity with separate expert networks, while Transformer-M locates pairwise structural specificity in the MSA via separate bias terms. In summary, our extension is straightforward yet highly effective.

3.3 PRE-TRAINING BIT

We pre-train BIT on both protein-ligand complex and small molecule datasets. We use the Q-BioLiP database (Wei et al., 2023) as the complex corpus. It contains approximately 1.0 million biologically relevant interactions associated with 3D cocrystal structures. This dataset is sourced from the Protein Data Bank (Berman et al., 2000) through manual process (Yang et al., 2012). To prevent potential overfitting to a limited portion of the chemical space represented by ligands in the Q-BioLiP dataset, we additionally pre-train BIT on large-scale small-molecule-only corpus. For this purpose, we incorporate the PCQM4Mv2 dataset (Nakata & Shimazaki, 2017), which has been widely used for 3D molecular pre-training (Zaidi et al., 2022; Wang et al., 2023b).

To ensure the scalability of the pre-training process, we employ a unified corrupt-then-recover objective to pre-train BIT. During this pre-training approach, we randomly corrupt the continuous atom coordinates and the categorical atom types of single-domain molecules and ligands from protein-ligand complexes, and guide BIT to restore the original states. The detailed explanations of the two denoising tasks are provided below.

Coordinate denoising aims to learn meaningful representations that capture the inter-atomic interactions within the molecular structure. It has been demonstrated to be effective in improving 3D molecular property prediction (Zaidi et al., 2022). Theoretically, this objective can be interpreted as learning an approximate molecular force field from equilibrium structures (Zaidi et al., 2022). Thus, we can extend coordinate denoising to protein-ligand complexes, as the experimentally-determined cocrystal structures of the complexes typically represent equilibrium conformations and correspond to local energy minima. To further capture the inter-molecular interactions, we encourage the model to restore the corrupted ligand pose based on the information from both the ligand and pocket.

Formally, let $\vec{R} = \{\vec{r}_1, \vec{r}_2, ..., \vec{r}_n\}, \vec{r}_i \in \mathbb{R}^3$ denote the binding pose of a bound ligand. We perturb it by adding independent and identically distributed ($i.i.d.$) Gaussian noise to its atomic coordinates $\vec{r}_i$. The resulting noisy atom positions are denoted as $\hat{R} = \{\vec{r}_1 + \sigma\vec{\epsilon}_1, \vec{r}_2 + \sigma\vec{\epsilon}_2, ..., \vec{r}_n + \sigma\vec{\epsilon}_n\}$, where $\vec{\epsilon}_i \sim \mathcal{N}(\vec{0}, \boldsymbol{I})$ and σ is a hyperparameter controlling the noise scale. The model is trained to predict the noise from the noisy input. The output of the last Transformer block is then fed into an SE(3) equivariant prediction head (Shi et al., 2022), driven by the denoising loss $\mathcal{L}_{pos} = \frac{1}{|\mathcal{V}|}\sum_{i\in V} \|\hat{\vec{\epsilon}}_i - \vec{\epsilon}_i\|^2$.

Masked token denoising aims to learn fundamental physicochemical information contained within the molecules or complexes by modeling the dependency between their atoms. This task is similar to the masked language modeling (MLM) task used in BERT (Devlin et al., 2018) and has achieved remarkable performance in molecular pre-training (Hu et al., 2020). As discussed in Austin et al. (2021), MLM can be interpreted as a categorical denoising process. Given an input molecule, we randomly mask 15% of its atoms and predict each masked atom based on its contextualized representation extracted by BIT. The cross-entropy prediction loss for this task is denoted as $\mathcal{L}_{atom}$.

3.4 FINE-TUNING BIT ON DOWNSTREAM TASKS

As BIT is designed to be a general-purpose cross-domain pre-train model, it is straightforward to supervised fine-tune it with task-specific data to adapt to various protein-ligand binding tasks.

Protein-ligand binding affinity prediction. As aforementioned, our model can serve as a fusion encoder to model the molecular interactions between proteins and ligands. Therefore, we extract the

final encoding vector from the special token [M_VNode] as the representation of the protein-ligand complexes and feed it to a task-specific prediction head to make the final prediction.

Structure-based virtual screening. We formulate large-scale virtual screening as a pocket-to-ligand retrieval task. In this task, our model is used as a dual encoder to encode both 3D protein pockets and 2D ligands to vectors of equal length. In fine-tuning, the pre-trained model is further optimized on task-specific data using contrastive learning. During inference, we compute representations of the target pocket and all candidate ligands, and then obtain pocket-to-ligand similarity scores of all possible pocket-ligand pairs using dot products. Hits are identified as ligands that exhibit a high level of similarity to the target pocket. This approach allows for much faster inference speeds than fusion encoder-based methods, which require preliminary molecular docking.

4 EXPERIMENTS

In this section, we pre-train BIT and extensively evaluate BIT on well-established public benchmarks, including both protein-ligand binding tasks and molecular learning tasks.

4.1 PRE-TRAINING SETUPS

4.1.1 DATASETS

We pre-train BIT on both protein-ligand complex data and small molecule data. For complex data, we use the Q-BioLiP database (Wei et al., 2023), which contains 967,085 biological relevant interactions associated with 3D cocrystal structures as of June 14th, 2023. Q-BioLiP is an updated version of the original BioLiP database (Yang et al., 2012), where protein-ligand interactions are based on the quaternary structure rather than the single-chain monomer structure. This alteration provides higher-quality interactions for analyzing the binding mode. Since our primary focus is on regular ligands, i.e., small molecules, we filter out complexes containing metal ions and DNA/RNA ligands. For small molecule data, we utilize the PCQM4Mv2 dataset from the OGB Large-Scale Challenge (Hu et al., 2021), which has 3.4M organic molecules. These molecules are characterized by their 3D structures at equilibrium, calculated using density functional theory (DFT).

4.1.2 TRAINING SETTINGS

Our model adopts the same network configuration as Transformer-M (Luo et al., 2022). We employ a 12-layer Transformer with a hidden size of 768 and 32 attention heads. We use AdamW optimizer (Loshchilov & Hutter, 2018) with hyper-parameter ϵ set to 1e-8 and (β_1, β_2) set to (0.9,0.999). The gradient clip norm is set to 5. The peak learning rate is set to 2e-4, and we employ a 12k-step warm-up stage followed by a linear decay scheduler. The total training steps are 200k. Each batch contains 1024 samples, including 512 small molecules and 512 pocket-ligand complexes. We adopt the 2D-3D joint training strategy proposed in (Luo et al., 2022). In the coordinate denoising objective, σ is set to 0.2. All models are trained on 64 NVIDIA Tesla V100 GPUs for approximately 2 days.

4.2 PROTEIN-LIGAND BINDING TASKS

4.2.1 BINDING AFFINITY PREDICTION

In this task, the pre-trained model is fine-tuned to predict binding affinities pK_a (or $-\log K_d$, $-\log K_i$) for protein-ligand complexes. Following previous studies (Li et al., 2021b; Luo et al., 2022), we perform fine-tuning experiments using the PDBbind v2016 dataset (Wang et al., 2004; 2005). The PDBbind dataset consists of three subsets: the general set, which includes 13,283 protein-ligand complexes; the refined set, comprising 4,057 complexes selected from the general set for higher data quality, and the core set, consisting of 285 complexes chosen for the highest data quality (Su et al., 2018). We fine-tune the pre-trained BIT using the refined set. To prevent data leakage, we remove the data instances in the core set from the refined set. We evaluate the prediction performance using metrics such as Pearson's correlation coefficient (R), Mean Absolute Error (MAE), Root-Mean Squared Error (RMSE), and Standard Deviation (SD) (Su et al., 2018).

We compare BIT with DMPNN (Yang et al., 2019), MAT (Maziarka et al., 2020), DimeNet (Gasteiger et al., 2020), CMPNN (Song et al., 2020), SIGN (Li et al., 2021b), MBP (Yan et al., 2023), and

Table 1: Binding affinity prediction results on the PDBbind core set.

Method	R ↑	MAE ↓	RMSE ↓	SD ↓
DMPNN	$0.729_{\pm 0.006}$	$1.188_{\pm 0.009}$	$1.493_{\pm 0.016}$	$1.489_{\pm 0.014}$
MAT	$0.747_{\pm 0.013}$	$1.154_{\pm 0.037}$	$1.457_{\pm 0.037}$	$1.445_{\pm 0.033}$
DimeNet	$0.752_{\pm 0.010}$	$1.138_{\pm 0.026}$	$1.453_{\pm 0.027}$	$1.434_{\pm 0.023}$
CMPNN	$0.765_{\pm 0.009}$	$1.117_{\pm 0.031}$	$1.408_{\pm 0.028}$	$1.399_{\pm 0.025}$
SIGN	$0.797_{\pm 0.012}$	$1.027_{\pm 0.025}$	$1.316_{\pm 0.031}$	$1.312_{\pm 0.035}$
MBP	$0.825_{\pm 0.008}$	$0.999_{\pm 0.024}$	$1.263_{\pm 0.023}$	$1.229_{\pm 0.026}$
Transformer-M	$0.830_{\pm 0.011}$	$0.940_{\pm 0.006}$	$1.232_{\pm 0.013}$	$1.207_{\pm 0.007}$
BIT	$\mathbf{0.842}_{\pm 0.002}$	$\mathbf{0.927}_{\pm 0.008}$	$\mathbf{1.179}_{\pm 0.007}$	$\mathbf{1.173}_{\pm 0.006}$

Table 2: Virtual screening results on the DUD-E dataset.

Method	AUC ↑	$RE_{0.5\%}$ ↑	$RE_{1.0\%}$ ↑	$RE_{2.0\%}$ ↑	$RE_{5.0\%}$ ↑
NNScore	58.4	4.17	2.98	2.46	1.89
RF-Score	62.2	5.63	4.27	3.50	2.68
Vina	71.6	9.14	7.32	5.88	4.44
3DCNN	86.8	42.56	29.65	19.36	10.71
Graph CNN	88.6	44.41	29.75	19.41	10.74
CoSP	90.1	51.05	35.98	23.68	12.21
DrugVQA	97.2	88.17	58.71	35.06	17.39
BIT	**98.4**	**141.63**	**78.31**	**42.43**	**18.44**

Transformer-M (Luo et al., 2022). We report the official results of baselines from Li et al. (2021b); Luo et al. (2022); Yan et al. (2023). As presented in Table 1, BIT consistently outperforms pre-training baselines and other approaches tailored for binding affinity prediction across all evaluation metrics, demonstrating the effectiveness of BIT in capturing intricate molecular interactions present in complexes.

4.2.2 Structure-Based Virtual Screening

Structure-based virtual screening of potential drug-like molecules against a protein target of interest, as outlined by Lionta et al. (Lionta et al., 2014) is a critical goal in structure-based drug discovery. This task is to identify the molecules with the highest likelihood of binding to protein pockets with known 3D structures. We choose the widely-used DUD-E dataset (Mysinger et al., 2012) for our model evaluation, following previous study (Gao et al., 2022). The DUD-E dataset comprises 102 targets across different protein families. Each target, on average, is assigned 224 binding compounds and over 10,000 decoys. These decoys are physically similar to the active compounds but differ in terms of their topology. We adopt a four-fold cross-validation strategy and use the same data split approach outlined in GraphCNN (Torng & Altman, 2019). In our data splits, we ensure that no two folds contain targets with greater than 75% sequence identity. We provide results in terms of the AUC-ROC and ROC enrichment (RE) scores. The RE score measures early enrichment and is calculated as the ratio of the true positive rate (TPR) to the false positive rate (FPR) at a given FPR threshold. Here, we report the RE scores at 0.5%, 1.0%, 2.0%, and 5.0% FPR thresholds.

Since most of the protein-ligand pairs of interest do not have experimentally solved cocrystal structures, conventional affinity prediction models that rely on this information must be complemented with molecular docking software, such as AutoDock (Trott & Olson, 2010). However, this integration often leads to significant computational expenses, particularly in large-scale virtual screening tasks. By framing virtual screening as a pocket-to-ligand retrieval task, BIT can be adopted as a dual encoder. We encode 3D protein pockets and 2D molecular graphs separately to obtain their representations in a shared subspace and compute their similarity scores by the dot product. During fine-tuning, BIT is optimized using the contrastive loss function InfoNCE (Oord et al., 2018), with 64 randomly sampled decoys per active compound.

In this task, we compare BIT with NNScore (Durrant & McCammon, 2010), RF-Score (Ballester & Mitchell, 2010), Vina (Trott & Olson, 2010), 3DCNN (Ragoza et al., 2017), Graph CNN (Torng & Altman, 2019), CoSP (Gao et al., 2022), and DrugVQA (Zheng et al., 2020). As presented in Table 2, BIT achieves superior performance compared to the baselines. Furthermore, BIT is not required to jointly encode all potential pocket-ligand pairs and can store the pre-computed representations of pockets and ligands. This enables it to achieve high screening efficiency without compromising learning precision. In the empirical study, we can compute representations for one billion molecules in an ultra-large-scale screening library (e.g., ZINC (Irwin & Shoichet, 2005)) in just under 2 days using a single NVIDIA V100 GPU.

4.3 Molecular Property Prediction

In addition to the protein-ligand binding task, we also assess the capabilities of BIT in the molecular property prediction task, where BIT is used as an encoder for small molecules. In this task, we aim to predict the absorption, distribution, metabolism, excretion, and toxicity properties of molecules. We consider eight binary classification datasets from the MoleculeNet benchmark (Wu et al., 2018).

Table 3: Molecular property prediction results (with 2D topology only) on the MoleculeNet benchmark. The best and second best results are marked bold and bold, respectively.

Methods	BBBP ↑	Tox21 ↑	ToxCast ↑	SIDER ↑	ClinTox ↑	MUV ↑	HIV ↑	BACE ↑	Avg ↑
AttrMask (Hu et al., 2020)	65.0±2.36	74.8±0.25	62.9±0.11	61.2±0.12	**87.7±1.19**	73.4±2.02	76.8±0.53	79.7±0.33	72.68
ContextPred (Hu et al., 2020)	65.7±0.62	74.2±0.06	62.5±0.31	62.2±0.59	77.2±0.88	75.3±1.57	77.1±0.86	76.0±2.08	71.28
GraphCL (You et al., 2020)	69.7±0.67	73.9±0.66	62.4±0.57	60.5±0.88	76.0±2.65	69.8±2.66	78.5±1.22	75.4±1.44	70.78
InfoGraph (Sun et al., 2020)	67.5±0.11	73.2±0.43	63.7±0.50	59.9±0.30	76.5±1.07	74.1±0.74	75.1±0.99	77.8±0.88	70.96
GROVER (Rong et al., 2020)	70.0±0.10	74.3±0.10	65.4±0.40	64.8±0.60	81.2±3.00	67.3±1.80	62.5±0.90	82.6±0.70	71.01
MolCLR (Wang et al., 2022a)	66.6±1.89	73.0±0.16	62.9±0.38	57.5±1.77	86.1±0.95	72.5±2.38	76.2±1.51	71.5±3.17	70.79
GraphMAE (Hou et al., 2022)	72.0±0.60	75.5±0.60	64.1±0.30	60.3±1.10	82.3±1.20	76.3±2.40	77.2±1.00	83.1±0.90	73.85
Mole-BERT (Xia et al., 2023a)	71.9±1.60	76.8±0.50	64.3±0.20	62.8±1.10	78.9±3.00	78.6±1.80	78.2±0.80	80.8±1.40	74.04
3D InfoMax (Stärk et al., 2022)	69.1±1.07	74.5±0.74	64.4±0.88	60.6±0.78	79.9±3.49	74.4±2.45	76.1±1.33	79.7±1.54	72.34
GraphMVP (Liu et al., 2021)	72.4±1.60	74.4±0.20	63.1±0.40	63.9±1.20	77.5±4.20	75.0±1.00	77.0±1.20	81.2±0.90	73.07
MoleculeSDE (Liu et al., 2023)	71.8±0.76	76.8±0.34	65.0±0.26	60.8±0.39	87.0±0.53	**80.9±0.37**	78.8±0.92	79.5±2.17	75.07
MoleBLEND (Yu et al., 2023)	**73.0±0.81**	**77.8±0.89**	**66.1±0.03**	**64.9±0.35**	87.6±0.75	77.2±2.38	**79.0±0.89**	**83.7±1.46**	**76.16**
BIT	**74.3±0.81**	**78.1±0.91**	**66.4±0.18**	**64.8±0.47**	**91.3±1.21**	**79.4±0.87**	**80.2±0.67**	**84.5±0.85**	**77.38**

Table 4: Ablation studies of MoDE and pre-training tasks.

	Pre-Training Tasks		Backbone	Property		Binding	
	Token	Coordinate	MoDE	HIV ↑	Tox21 ↑	PDBbind (MAE) ↓	DUD-E (AUC) ↑
[1]	✗	✗	✓	70.9	75.1	1.114	95.7
[2]	✓	✗	✓	78.5	76.1	1.016	97.1
[3]	✗	✓	✓	78.2	77.6	0.939	96.5
[4]	✓	✓	✗	78.3	76.9	0.977	98.0
[5]	✓	✓	✓	**80.2**	**78.1**	**0.927**	**98.4**

Following previous studies (Hu et al., 2020), we employ scaffold splitting to divide the dataset into training, validation, and test sets in an 8:1:1 ratio. We use the ROC-AUC as the evaluation metric and report the mean and standard deviation of the results obtained from 3 random seed runs. We compare BIT against the most representative molecular graph-based as well as multimodal pre-trained models. Detailed descriptions of the baselines are presented in Appendix B.3.

The performance of BIT, compared to competitive baselines, is summarized in Table 3. We observe that BIT outperforms the baselines on 6 out of 8 tasks, and achieves an overall relative improvement of 1.6% in terms of average ROC-AUC compared to the previous state-of-the-art result.

4.4 Ablation Studies

MoDE. We conduct ablation experiments to investigate the impact of MoDE. As presented in Table 4, the integration of MoDE significantly boosts performance across various tasks, particularly in the binding affinity prediction task (PDBbind), where it is essential to encode both ligands and proteins concurrently while capturing the fine-grained inter-molecular interactions. Such enhancement is in line with our motivation to introduce MoDE.

pre-training tasks. We also perform ablation studies to analyze the contribution of different pre-training tasks, and the results are presented in Table 4. Eliminating either pre-training objective leads to pronounced declines in performance. We observe that coordinate denoising is indispensable for 3D representations, whereas masked token denoising is paramount for 2D representations. These results indicate that our unified pre-training is crucial and yields positive outcomes.

5 Conclusion and Future Work

In this work, we take further strides towards general-purpose molecular modeling. We introduce BIT, a pre-trained foundation model, which is designed to encode molecules across various biochemical domains, including small molecules, proteins, and protein-ligand complexes, in different data formats, including 2D and 3D structures. Experimental results demonstrate that BIT excels across a broad spectrum of protein-ligand binding and molecular learning tasks. In our **future work**, we plan to work on fine-tuning BIT for structure-based molecular generation tasks, such as target protein binding (Luo et al., 2021) and molecular docking (Corso et al., 2022). We are also working on gathering more diverse real-world and synthetic protein-ligand complexes to facilitate the training of larger models.

REFERENCES

Qurrat Ul Ain, Antoniya Aleksandrova, Florian D Roessler, and Pedro J Ballester. Machine-learning scoring functions to improve structure-based binding affinity prediction and virtual screening. Wiley Interdisciplinary Reviews: Computational Molecular Science, 5(6):405–424, 2015.

Amy C Anderson. The process of structure-based drug design. Chemistry & biology, 10(9):787–797, 2003.

Junyi Ao, Rui Wang, Long Zhou, Chengyi Wang, Shuo Ren, Yu Wu, Shujie Liu, Tom Ko, Qing Li, Yu Zhang, et al. Speecht5: Unified-modal encoder-decoder pre-training for spoken language processing. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 5723–5738, 2022.

Kenneth Atz, Francesca Grisoni, and Gisbert Schneider. Geometric deep learning on molecular representations. Nature Machine Intelligence, 3(12):1023–1032, 2021.

Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne Van Den Berg. Structured denoising diffusion models in discrete state-spaces. Advances in Neural Information Processing Systems, 34:17981–17993, 2021.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. arXiv preprint arXiv:1607.06450, 2016.

Pedro J Ballester and John BO Mitchell. A machine learning approach to predicting protein–ligand binding affinity with applications to molecular docking. Bioinformatics, 26(9):1169–1175, 2010.

Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. Advances in Neural Information Processing Systems, 35:32897–32912, 2022.

Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The protein data bank. Nucleic acids research, 28(1): 235–242, 2000.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. Advances in neural information processing systems, 33:1877–1901, 2020.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In International conference on machine learning, pp. 1597–1607. PMLR, 2020.

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

Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. Debiased contrastive learning. Advances in neural information processing systems, 33:8765–8775, 2020.

Gabriele Corso, Hannes Stärk, Bowen Jing, Regina Barzilay, and Tommi S Jaakkola. Diffdock: Diffusion steps, twists, and turns for molecular docking. In The Eleventh International Conference on Learning Representations, 2022.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In International Conference on Learning Representations, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Jacob D Durrant and J Andrew McCammon. Nnscore: a neural-network-based scoring function for the characterization of protein- ligand complexes. Journal of chemical information and modeling, 50(10):1865–1871, 2010.

Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, et al. Prottrans: Toward understanding the language of life through self-supervised learning. IEEE transactions on pattern analysis and machine intelligence, 44(10):7112–7127, 2021.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. The Journal of Machine Learning Research, 23(1): 5232–5270, 2022.

Shikun Feng, Yuyan Ni, Yanyan Lan, Zhi-Ming Ma, and Wei-Ying Ma. Fractional denoising for 3d molecular pre-training. In International Conference on Machine Learning, pp. 9938–9961. PMLR, 2023.

Noelia Ferruz and Birte Höcker. Controllable protein design with language models. Nature Machine Intelligence, 4(6):521–532, 2022.

Zhangyang Gao, Cheng Tan, Lirong Wu, and Stan Z Li. Cosp: Co-supervised pretraining of pocket and ligand. arXiv preprint arXiv:2206.12241, 2022.

Johannes Gasteiger, Janek Groß, and Stephan Günnemann. Directional message passing for molecular graphs. arXiv preprint arXiv:2003.03123, 2020.

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.

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, pp. 594–604, 2022.

W Hu, B Liu, J Gomes, M Zitnik, P Liang, V Pande, and J Leskovec. Strategies for pre-training graph neural networks. In International Conference on Learning Representations, 2020.

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.

John J Irwin and Brian K Shoichet. Zinc- a free database of commercially available compounds for virtual screening. Journal of chemical information and modeling, 45(1):177–182, 2005.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In International conference on machine learning, pp. 4904–4916. PMLR, 2021.

Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In International Conference on Machine Learning, pp. 5583–5594. PMLR, 2021.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. Advances in neural information processing systems, 34:9694–9705, 2021a.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In International Conference on Machine Learning, pp. 12888–12900. PMLR, 2022.

Shuangli Li, Jingbo Zhou, Tong Xu, Liang Huang, Fan Wang, Haoyi Xiong, Weili Huang, Dejing Dou, and Hui Xiong. Structure-aware interactive graph neural networks for the prediction of protein-ligand binding affinity. In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, pp. 975–985, 2021b.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science, 379(6637):1123–1130, 2023.

Evanthia Lionta, George Spyrou, Demetrios K Vassilatis, and Zoe Cournia. Structure-based virtual screening for drug discovery: principles, applications and recent advances. Current topics in medicinal chemistry, 14(16):1923–1938, 2014.

Shengchao Liu, Hanchen Wang, Weiyang Liu, Joan Lasenby, Hongyu Guo, and Jian Tang. Pre-training molecular graph representation with 3d geometry. arXiv preprint arXiv:2110.07728, 2021.

Shengchao Liu, Weili Nie, Chengpeng Wang, Jiarui Lu, Zhuoran Qiao, Ling Liu, Jian Tang, Chaowei Xiao, and Anima Anandkumar. Multi-modal molecule structure-text model for text-based retrieval and editing. arXiv preprint arXiv:2212.10789, 2022.

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, 2023.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In International Conference on Learning Representations, 2018.

Shengjie Luo, Tianlang Chen, Yixian Xu, Shuxin Zheng, Tie-Yan Liu, Liwei Wang, and Di He. One transformer can understand both 2d & 3d molecular data. arXiv preprint arXiv:2210.01765, 2022.

Shitong Luo, Jiaqi Guan, Jianzhu Ma, and Jian Peng. A 3d generative model for structure-based drug design. Advances in Neural Information Processing Systems, 34:6229–6239, 2021.

Ali Madani, Ben Krause, Eric R Greene, Subu Subramanian, Benjamin P Mohr, James M Holton, Jose Luis Olmos Jr, Caiming Xiong, Zachary Z Sun, Richard Socher, et al. Large language models generate functional protein sequences across diverse families. Nature Biotechnology, pp. 1–8, 2023.

Łukasz Maziarka, Tomasz Danel, Sławomir Mucha, Krzysztof Rataj, Jacek Tabor, and Stanisław Jastrzębski. Molecule attention transformer. arXiv preprint arXiv:2002.08264, 2020.

Ingo Muegge and Yvonne C Martin. A general and fast scoring function for protein- ligand interactions: a simplified potential approach. Journal of medicinal chemistry, 42(5):791–804, 1999.

Michael M Mysinger, Michael Carchia, John J Irwin, and Brian K Shoichet. Directory of useful decoys, enhanced (dud-e): better ligands and decoys for better benchmarking. Journal of medicinal chemistry, 55(14):6582–6594, 2012.

Maho Nakata and Tomomi Shimazaki. Pubchemqc project: a large-scale first-principles electronic structure database for data-driven chemistry. Journal of chemical information and modeling, 57 (6):1300–1308, 2017.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748, 2018.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In International conference on machine learning, pp. 8748–8763. PMLR, 2021.

Matthew Ragoza, Joshua Hochuli, Elisa Idrobo, Jocelyn Sunseri, and David Ryan Koes. Protein–ligand scoring with convolutional neural networks. Journal of chemical information and modeling, 57(4):942–957, 2017.

Ladislav Rampášek, Michael Galkin, Vijay Prakash Dwivedi, Anh Tuan Luu, Guy Wolf, and Dominique Beaini. Recipe for a general, powerful, scalable graph transformer. Advances in Neural Information Processing Systems, 35:14501–14515, 2022.

Roshan M Rao, Jason Liu, Robert Verkuil, Joshua Meier, John Canny, Pieter Abbeel, Tom Sercu, and Alexander Rives. Msa transformer. In International Conference on Machine Learning, pp. 8844–8856. PMLR, 2021.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proceedings of the National Academy of Sciences, 118(15):e2016239118, 2021.

Yu Rong, Yatao Bian, Tingyang Xu, Weiyang Xie, Ying Wei, Wenbing Huang, and Junzhou Huang. Self-supervised graph transformer on large-scale molecular data. Advances in Neural Information Processing Systems, 33:12559–12571, 2020.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. arXiv preprint arXiv:1701.06538, 2017.

Yu Shi, Shuxin Zheng, Guolin Ke, Yifei Shen, Jiacheng You, Jiyan He, Shengjie Luo, Chang Liu, Di He, and Tie-Yan Liu. Benchmarking graphormer on large-scale molecular modeling datasets. arXiv preprint arXiv:2203.04810, 2022.

Ying Song, Shuangjia Zheng, Zhangming Niu, Zhang-Hua Fu, Yutong Lu, and Yuedong Yang. Communicative representation learning on attributed molecular graphs. In IJCAI, volume 2020, pp. 2831–2838, 2020.

Hannes Stärk, Dominique Beaini, Gabriele Corso, Prudencio Tossou, Christian Dallago, Stephan Günnemann, and Pietro Liò. 3d infomax improves gnns for molecular property prediction. In International Conference on Machine Learning, pp. 20479–20502. PMLR, 2022.

Minyi Su, Qifan Yang, Yu Du, Guoqin Feng, Zhihai Liu, Yan Li, and Renxiao Wang. Comparative assessment of scoring functions: the casf-2016 update. Journal of chemical information and modeling, 59(2):895–913, 2018.

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In Proceedings of the IEEE/CVF international conference on computer vision, pp. 7464–7473, 2019.

Fan-Yun Sun, Jordon Hoffman, Vikas Verma, and Jian Tang. Infograph: Unsupervised and semi-supervised graph-level representation learning via mutual information maximization. In International Conference on Learning Representations. OpenReview. net, 2020.

Freyr Sverrisson, Jean Feydy, Bruno E Correia, and Michael M Bronstein. Fast end-to-end learning on protein surfaces. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15272–15281, 2021.

Jacopo Tomasi and Maurizio Persico. Molecular interactions in solution: an overview of methods based on continuous distributions of the solvent. Chemical Reviews, 94(7):2027–2094, 1994.

Wen Torng and Russ B Altman. Graph convolutional neural networks for predicting drug-target interactions. Journal of chemical information and modeling, 59(10):4131–4149, 2019.

Oleg Trott and Arthur J Olson. Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. Journal of computational chemistry, 31(2):455–461, 2010.

Jessica Vamathevan, Dominic Clark, Paul Czodrowski, Ian Dunham, Edgardo Ferran, George Lee, Bin Li, Anant Madabhushi, Parantu Shah, Michaela Spitzer, et al. Applications of machine learning in drug discovery and development. Nature reviews Drug discovery, 18(6):463–477, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. Advances in neural information processing systems, 30, 2017.

Renxiao Wang, Xueliang Fang, Yipin Lu, and Shaomeng Wang. The pdbbind database: Collection of binding affinities for protein- ligand complexes with known three-dimensional structures. Journal of medicinal chemistry, 47(12):2977–2980, 2004.

Renxiao Wang, Xueliang Fang, Yipin Lu, Chao-Yie Yang, and Shaomeng Wang. The pdbbind database: methodologies and updates. Journal of medicinal chemistry, 48(12):4111–4119, 2005.

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.

Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for vision and vision-language tasks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 19175–19186, 2023a.

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, 2022a.

Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. SimVLM: Simple visual language model pretraining with weak supervision. In International Conference on Learning Representations, 2022b. URL https://openreview.net/forum?id=GUrhfTuf_3.

Hong Wei, Wenkai Wang, Zhenling Peng, and Jianyi Yang. Biolip2: a database for biological unit-based protein-ligand interactions. bioRxiv, pp. 2023–06, 2023.

David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. Journal of chemical information and computer sciences, 28(1):31–36, 1988.

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. In The Eleventh International Conference on Learning Representations, 2023a. URL https://openreview.net/forum?id=jevY-DtiZTR.

Jun Xia, Yanqiao Zhu, Yuanqi Du, Yue Liu, and Stan Z Li. A systematic survey of chemical pre-trained models. IJCAI, 2023b.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In International Conference on Learning Representations, 2019. URL https://openreview.net/forum?id=ryGs6iA5Km.

Minghao Xu, Xinyu Yuan, Santiago Miret, and Jian Tang. Protst: Multi-modality learning of protein sequences and biomedical texts. arXiv preprint arXiv:2301.12040, 2023a.

Peng Xu, Xiatian Zhu, and David A Clifton. Multimodal learning with transformers: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023b.

Jiaxian Yan, Zhaofeng Ye, Ziyi Yang, Chengqiang Lu, Shengyu Zhang, Qi Liu, and Jiezhong Qiu. Multi-task bioassay pre-training for protein-ligand binding affinity prediction. arXiv preprint arXiv:2306.04886, 2023.

Jianyi Yang, Ambrish Roy, and Yang Zhang. Biolip: a semi-manually curated database for biologically relevant ligand–protein interactions. Nucleic acids research, 41(D1):D1096–D1103, 2012.

Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, et al. Analyzing learned molecular representations for property prediction. Journal of chemical information and modeling, 59(8): 3370–3388, 2019.

Chengxuan Ying, Tianle Cai, Shengjie Luo, Shuxin Zheng, Guolin Ke, Di He, Yanming Shen, and Tie-Yan Liu. Do transformers really perform badly for graph representation? Advances in Neural Information Processing Systems, 34:28877–28888, 2021.

Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. Graph contrastive learning with augmentations. Advances in neural information processing systems, 33: 5812–5823, 2020.

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

Sheheryar Zaidi, Michael Schaarschmidt, James Martens, Hyunjik Kim, Yee Whye Teh, Alvaro Sanchez-Gonzalez, Peter Battaglia, Razvan Pascanu, and Jonathan Godwin. Pre-training via denoising for molecular property prediction. arXiv preprint arXiv:2206.00133, 2022.

Zuobai Zhang, Minghao Xu, Arian Rokkum Jamasb, Vijil Chenthamarakshan, Aurelie Lozano, Payel Das, and Jian Tang. Protein representation learning by geometric structure pretraining. In The Eleventh International Conference on Learning Representations, 2023. URL https://openreview.net/forum?id=to3qCB3tOh9.

Shuangjia Zheng, Yongjian Li, Sheng Chen, Jun Xu, and Yuedong Yang. Predicting drug–protein interaction using quasi-visual question answering system. Nature Machine Intelligence, 2(2): 134–140, 2020.

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. In The Eleventh International Conference on Learning Representations, 2023. URL https://openreview.net/forum?id=6K2RM6wVqKu.

A IMPLEMENTATION DETAILS OF TRANSFORMER-M

A.1 PREDICTION HEAD FOR POSITION OUTPUT.

We use the SE(3) equivariant prediction head proposed in Shi et al. (2022):

$$\hat{\vec{\epsilon}}_i^k = (\sum_{v_j \in V} a_{ij} \Delta_{ij}^k \boldsymbol{X}_j^{(L)} \boldsymbol{W}_N^1) \boldsymbol{W}_N^2, \quad k = 0, 1, 2 \tag{6}$$

where $\boldsymbol{X}_j^{(L)}$ is the output of the last Transformer block, a_{ij} is the attention score between atom i and j calculated by Eqn.3, Δ_{ij}^k is the k-th element of the directional vector $\frac{\vec{r}_i - \vec{r}_j}{\|\vec{r}_i - \vec{r}_j\|}$ between atom i and j, and $\boldsymbol{W}_N^1 \in \mathbb{R}^{d \times d}$, $\boldsymbol{W}_N^2 \in \mathbb{R}^{d \times 1}$ are learnable weight matrices.

B EXPERIMENTAL DETAILS

B.1 PDBBIND

Evaluation Metrics. Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and Pearson correlation coefficient (R) are defined as:

$$RMSE = \sqrt{\frac{1}{|\mathcal{D}|} \sum_{i=1}^{|\mathcal{D}|} (\hat{y}_i - y_i)^2}, \; MAE = \frac{1}{|\mathcal{D}|} \sum_{i=1}^{|\mathcal{D}|} |\hat{y}_i - y_i| \tag{7}$$

$$R = \frac{\sum_{i=1}^{|\mathcal{D}|} (\hat{y}_i - \bar{\hat{y}})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{|\mathcal{D}|} (\hat{y}_i - \bar{\hat{y}})^2 (y_i - \bar{y})^2}} \tag{8}$$

$\hat{y}_i$ and y_i respectively represent the predicted and experimental binding affinity of the i-th complex in dataset $\mathcal{D}$. The standard deviation (SD) is defined as follows:

$$SD = \sqrt{\frac{1}{|\mathcal{D}| - 1} \sum_{i=1}^{|\mathcal{D}|} [y_i - (a + b\hat{y}_i)]^2} \tag{9}$$

where a and b are the intercept and the slope of the regression line, respectively.

Settings. We fine-tune the pre-trained BIT on the PDBbind dataset. We use AdamW (Loshchilov & Hutter, 2018) as the optimizer and set its hyperparameter ϵ to 1e-8 and (β_1, β_2) to (0.9,0.999). The gradient clip norm is set to 5.0. The peak learning rate is set to 1e-5. The total number of epochs is set to 120. The ratio of the warm-up steps to the total steps is set to 0.06. The batch size is set to 32. The dropout ratios for the input embeddings, attention matrices, and hidden representations are set to 0.0, 0.1, and 0.0 respectively. The weight decay is set to 0.0.

B.2 DUD-E

Settings. We fine-tune the pre-trained BIT on the DUD-E dataset. We use AdamW (Loshchilov & Hutter, 2018) as the optimizer and set its hyperparameter ϵ to 1e-8 and (β_1, β_2) to (0.9,0.999). The gradient clip norm is set to 5.0. The peak learning rate is set to 2e-4. The total number of epochs is set to 10. The ratio of the warm-up steps to the total steps is set to 0.06. The batch size is set to 16. The dropout ratios for the input embeddings, attention matrices, and hidden representations are set to 0.0, 0.1, and 0.0 respectively. The weight decay is set to 0.0.

B.3 MOLECULENET

Dataset. The details of the 8 datasets used in this work are described below.

- BBBP: Blood-brain barrier penetration (BBBP) contains the ability of small molecules to penetrate the blood-brain barrier.

- Tox21: The dataset contains toxicity measurements of 8k molecules for 12 targets.
- ToxCast: This dataset is derived from toxicology data from in vitro high-throughput screening and contains toxicity measurements for 8k molecules against 617 targets.
- SIDER: The Side Effect Resource (SIDER) contains side effects of drugs on 27 system organs. These drugs are not only small molecules but also some peptides with molecular weights over 1000.
- ClinTox: This dataset contains the toxicity of the drug in clinical trials and the status of the drug for FDA approval.
- MUV: Maximum Unbiased Validation (MUV) is another subset of PubChem BioAssay, containing 90k molecules and 17 bioassays.
- HIV: This dataset contains 40k compounds with the ability to inhibit HIV replication.
- BACE: This dataset contains the results of small molecules as inhibitors of binding to human β-secretase 1 (BACE-1).

Baselines. We compare BIT against both molecular graph-based pre-trained models, including AttrMask (Hu et al., 2020), GraphCL (You et al., 2020), InfoGraph (Sun et al., 2020), GROVER (Rong et al., 2020), MolCLR (Wang et al., 2022a), GraphMAE (Hou et al., 2022) and Mole-BERT (Xia et al., 2023a), as well as multimodal pre-trained models, including 3D infoMax (Stärk et al., 2022), GraphMVP (Liu et al., 2021), MoleculeSDE (Liu et al., 2023), and MoleBLEND (Yu et al., 2023).

Settings. We use a grid search to find the best combination of hyperparameters for the molecular property prediction task. The specific search space is shown in Table 5. In all experiments, we choose the checkpoint with the lowest validation loss, and report the results on the test set run by that checkpoint.

Table 5: Search space for the MoleculeNet benchmark.

Hyperparameter	Search space
Learning rate	[2e-5, 5e-5, 1e-4, 2e-4]
Batch size	[32, 64, 128, 256]
Warmup ratio	[0, 0.06]