TOWARDS ALL-ATOM FOUNDATION MODELS FOR BIOMOLECULAR BINDING AFFINITY PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Biomolecular interactions play a critical role in biological processes. While recent breakthroughs like AlphaFold 3 have enabled accurate modeling of biomolecular complex structures, predicting binding affinity remains challenging mainly due to limited high-quality data. Recent methods are often specialized for specific types of biomolecular interactions, limiting their generalizability. In this work, we adapt AlphaFold 3 for representation learning to predict binding affinity, a non-trivial task that requires shifting from generative structure prediction to encoding observed geometry, simplifying the heavily conditioned trunk module, and designing a framework to jointly capture sequence and structural information. To address these challenges, we introduce the **Atom-level Diffusion Transformer** (**ADiT**), which takes sequence and structure as inputs, employs a unified tokenization scheme, integrates diffusion transformers, and removes dependencies on multiple sequence alignments and templates. We pre-train three ADiT variants on the PDB dataset with a denoising objective and evaluate them across protein-ligand, drug-target, protein-protein, and antibody-antigen interactions. The model achieves state-of-the-art or competitive performance across benchmarks, scales effectively with model size, and successfully identifies wet-lab validated affinity-enhancing antibody mutations, establishing a generalizable framework for biomolecular interactions. We plan to release the code upon acceptance.

1 INTRODUCTION

Biomolecular interactions, including protein-protein and protein-ligand interactions, are pivotal to a wide range of biological processes, such as cellular functions (Busch et al., 2025) and therapeutic applications (Pacholarz et al., 2012; Cheng et al., 2021). Recent breakthroughs, exemplified by AlphaFold 3 and related methods (Abramson et al., 2024; Krishna et al., 2024; Chai Discovery, 2024; Chen et al., 2025; Wohlwend et al., 2024), have enabled accurate structure modeling of complex biological proteins. For example, AlphaFold 3 demonstrates remarkable accuracy in predicting protein complex structures from sequences, effectively resolving both the spatial arrangement and interaction mechanisms of protein-protein and protein-ligand interactions. Despite recent research progress, structure prediction remains an intermediate step, with the ultimate goal being the design of functional proteins with strong binding affinities for specific targets.

Binding affinity prediction remains difficult due to the scarcity of reliable experimental affinity labels. Early methods relied exclusively on sequences as input (Sun et al., 2017; Rao et al., 2019b), failing to fully leverage recent advancements in structure prediction (Abramson et al., 2024), even though all-atom structures are fundamental to understanding binding mechanisms and determining affinity. Recent approaches have incorporated structural information but are often specialized for specific types of biomolecular interactions. For instance, RDE-Network (Luo et al., 2023), DiffAffinity (Liu et al., 2024), and Prompt-DDG (Wu et al., 2024) target protein-protein binding affinity, while MGraph-DTA (Yang et al., 2022), HGNN-DTA (He et al., 2023), and ProFSA (Gao et al., 2023) focus on protein-ligand binding affinity. This narrow specialization limits the generalizability of these approaches and their applicability across diverse interaction types and data modalities.

Recent advancements in foundation models across various domains, such as BERT (Kenton & Toutanova, 2019) and the GPT series (Floridi & Chiriatti, 2020; Achiam et al., 2023) for language, the SAM series (Kirillov et al., 2023; Ravi et al., 2024) for vision, and CLIP (Radford et al., 2021)

for multi-modality, have inspired new directions for addressing these challenges. These models are first pre-trained on broad datasets using self-supervised learning objectives and then adapted to a wide range of downstream tasks (Bommasani et al., 2021). By reducing the need for domain-specific inductive biases and leveraging large datasets, these models learn powerful representations that significantly enhance generalization capabilities across diverse tasks. In biology, AlphaFold 3 exemplifies this paradigm by developing a universal model capable of predicting various types of biomolecular complex structures, demonstrating the power of generalization across interaction types.

Inspired by these successes, we adapt the AlphaFold 3 architecture for representation learning to design a structure-based unified foundation model for binding affinity prediction. While AlphaFold 3 is primarily a generative model for predicting biomolecular structures, adopting AlphaFold 3 for affinity prediction requires non-trivial modifications. We propose a novel strategy to adapt AlphaFold 3 as a representation learner, guided by three key insights: (i) when the model's goal shifts from *predicting geometry* to *encoding observed geometry*, the heavy conditioning (trunk) module becomes less critical and might allow for a more lightweight design; (ii) the transformer-based, atom and sequence-level architecture offers a general framework for jointly encoding both structural and sequential information; (iii) by large-scale pre-training on structure data, it can potentially mitigate data scarcity in affinity prediction and improve generalization across diverse interaction tasks.

Building on this framework, we present the **Atom-level Diffusion Transformer** (**ADiT**), a universal, all-atom structure foundation model designed to learn transferable representations across diverse biomolecular interactions. Our model accepts sequence and structure as inputs, employs a unified scheme to tokenize proteins and molecules, and generates atom-level, token-level, and complex-level representations. We integrate key components of diffusion transformers while simplifying the architecture by removing dependencies on multiple sequence alignments and templates. Following a pre-training and fine-tuning framework, we train three versions of our ADiT, namely ADiT-S, ADiT-M, and ADiT-L, on the PDB dataset. These pre-trained models are evaluated across four different types of interactions including protein-ligand, drug-target, protein-protein, and antibody-antigen, and demonstrate state-of-the-art or competitive performance across all benchmarks. Ablation studies highlight the importance of large-scale pre-training and atom-level modeling, while consistent performance gains with increasing model size align with established scaling trends in other domains.

Main contributions: *(i)* We present a general-purpose foundation model, ADiT, designed for binding affinity prediction across various types of biomolecular interactions. *(ii)* Extensive experiments demonstrate that ADiT, empowered by large-scale pre-training, achieves state-of-the-art or competitive performance across diverse, well-established benchmarks. *(iii)* We observe consistent performance improvements across various benchmarks as model size increases, reflecting scaling trends observed in other domains. *(iv)* A case study on antibodies shows that ADiT can successfully identify affinity-enhancing mutations, with predictions validated by wet-lab experiment results.

2 Related Work

Protein Representation Learning. Research in protein representation learning can be broadly categorized into two main paradigms: sequence-based methods and structure-based methods. Sequence-based approaches interpret protein sequences as a form of biological language, leveraging large-scale pre-training on sequence or alignment datasets to extract functional and evolutionary information (Rao et al., 2019b;a; Elnaggar et al., 2021; Rives et al., 2021; Meier et al., 2021; Wang et al., 2024b; Xu et al., 2023). In contrast, structure-based representation learning has gained significant attention due to the critical role of protein structures in determining function, as well as recent breakthroughs in structure prediction tools (Jumper et al., 2021). These methods explicitly encode protein structures using techniques such as 3D CNNs (Derevyanko et al., 2018), GNNs (Zhang et al., 2023c; Zhao et al., 2024; Fan et al., 2023), and transformers (Jiao et al., 2024). They employ various self-supervised learning strategies, including contrastive learning, self-prediction and diffusion denoising, to train structure encoders effectively (Zhang et al., 2023c; Chen et al., 2022; Zhang et al., 2023d). Additionally, some approaches use structure tokenizers to convert 3D protein structures into token sequences, enabling the application of sequence-based methods for structural representation learning (Su et al., 2024; Wang et al., 2024a; Heinzinger et al., 2024). However, many current approaches have two key limitations: (1) they primarily focus on coarse-grained, residue-level structures, and (2) they are restricted to proteins alone. In this work, we address these gaps by focusing on all-atom representation learning for general biomolecular complexes.

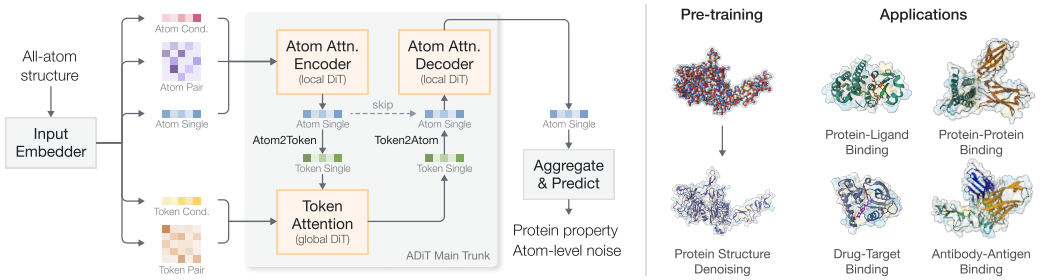


Figure 1: **Biomolecular complex representation learning with ADiT.** The model inputs all-atom structures, processes them through a multi-level attention trunk, and generates atom representations. It is pre-trained using a denoising objective and then fine-tuned for downstream application tasks.

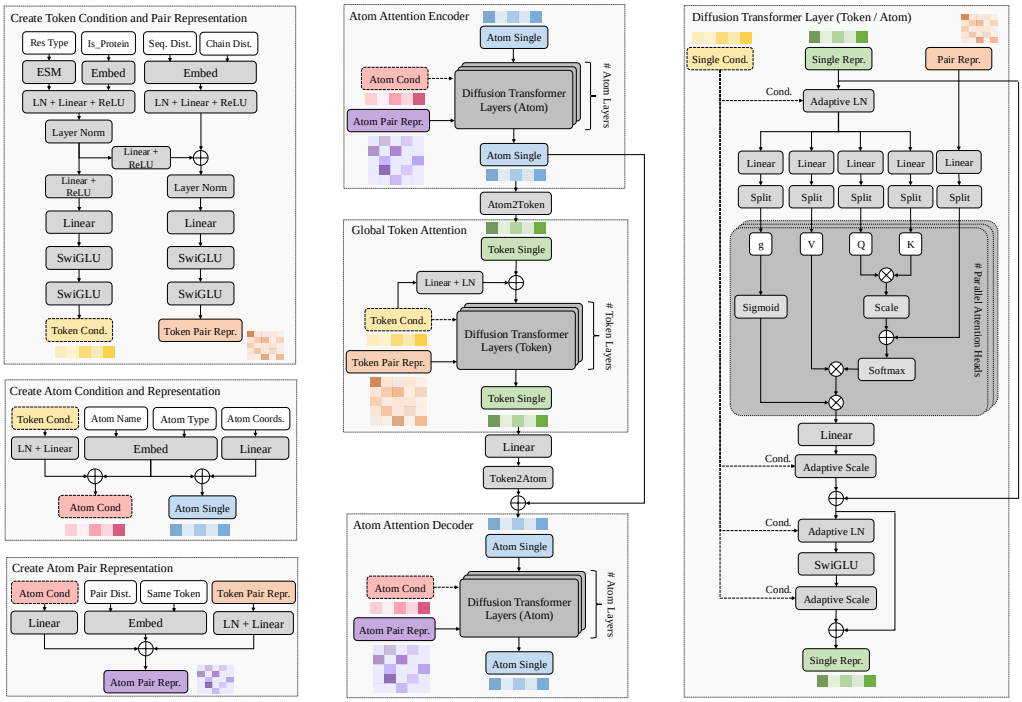
Biomolecular Complex Modeling. Recent advancements in molecular docking, both in predicting bound structures from unbound inputs (Ganea et al., 2022; Zhang et al., 2023a; Corso et al., 2023; Ketata et al., 2023) and in directly predicting bound structures from sequences (Abramson et al., 2024; Krishna et al., 2024; Chai Discovery, 2024; Chen et al., 2025; Wohlwend et al., 2024), have substantially advanced the field of biomolecular complex structure prediction. Despite these progress, predicting structures remains an intermediate step towards designing proteins with functional properties like strong binding affinity. Numerous studies have focused on predicting these binding affinities for protein-protein and protein-molecule interactions. For instance, prior work has explored energy-based methods (Delgado et al., 2019; Alford et al., 2017), evolution-based approaches (Meier et al., 2021; Rao et al., 2021; Notin et al., 2022), end-to-end learning models (Shan et al., 2022), and structure-informed pre-training methods (Yang et al., 2020; Luo et al., 2023; Wu et al., 2024; Mo et al., 2024; Liu et al., 2024) for protein-protein binding affinity. For protein-ligand binding, pre-training methods have become the dominant approach due to data scarcity (Karimi et al., 2019; Liu et al., 2023; Wu et al., 2022; Zhou et al., 2023; Gao et al., 2023). Protein-ligand affinity prediction is crucial in drug-target prediction (Yang et al., 2022; He et al., 2023; Kroll et al., 2023a;b), underscoring its importance in protein design. It is also common to benchmark against zero-shot baselines such as physics-based scoring functions (e.g., Rosetta (Barlow et al., 2018)), predicted mutational effects (e.g., NERE (Jin et al., 2023b) and DSMBind (Jin et al., 2023a)), and structural confidence metrics (e.g., AlphaFold 3's ipTM (Abramson et al., 2024)), although these typically fall far short of specialized fine-tuned models. Besides, Boltz-2 has been recently proposed for protein-ligand binding affinity prediction (Passaro et al., 2025). However, a unified framework for biomolecular binding affinity prediction remains elusive, with many studies still relying on sequence-based methods and overlooking breakthroughs like AlphaFold 3. In this work, we introduce a general foundation model for biomolecular complexes, leveraging structures to learn robust representations.

3 METHOD

In this section, we introduce the **A**tom-level **Di**ffusion **T**ransformer (**ADiT**), a unified foundation model for diverse biomolecular interactions. Our model employs a unified scheme to tokenize and featurize proteins and molecules (Sec. 3.2) and incorporates diffusion transformers to capture token- and atom-level interactions (Sec. 3.3). Following standard practices in modern foundation model development (Kenton & Toutanova, 2019; Achiam et al., 2023), we train ADiT with a pre-training and fine-tuning framework (Sec. 3.4). An overview of our method is shown in Figure 1.

3.1 ALPHAFOLD FOR BINDING AFFINITY PREDICTION

Given the sequence $A \in \{1, .., 20\}^L$ and the structure $\mathbf{x} \in \mathbb{R}^{L\times 3}$ of a biomolecular complex, our objective is to extract atom representations and use them to predict binding affinity through a downstream linear or muli-layer perceptron (MLP) prediction head. As AlphaFold 3 (Abramson et al., 2024) provides a highly expressive, transformer-based architecture that jointly encodes sequential and structural information at the atom level across diverse types of biomolecular interactions, it is a natural starting point for building general-purpose representation learner.



(a) Feature Initialization (b) Main Trunk (c) Diffusion Transformer Layer

Figure 2: **The Atom-level Diffusion Transformer (ADiT) architecture**. *(a)* Initialization of features for atom and token sequences. *(b)* Main trunk for hierarchical representation learning, comprising an atom attention encoder, global token attention, and an atom attention decoder. *(c)* Diffusion transformer blocks with multiple sparse attention update layers for atom-level modeling.

However, adapting AlphaFold 3 for representation learning is both non-trivial and essential. AlphaFold 3 was originally designed as a generative model for structure prediction, and prior work has shown that generative models often produce suboptimal representations when applied directly to downstream tasks such as protein function or affinity prediction (Hu et al., 2022). This limitation arises because generative objectives primarily optimize for reconstructing structures rather than capturing features that are informative for functional or interaction properties. Similar trends have been observed in other domains: for instance, generative models for images or text, while capable of producing high-quality outputs, do not automatically yield strong embeddings for tasks like classification, retrieval, or zero-shot generalization (Li et al., 2023; Achiam et al., 2023). Consequently, transforming AlphaFold 3 into ADiT, a model explicitly designed for representation learning, requires substantial modifications both its architecture and pre-training strategy.

Specifically, AlphaFold 3 is a conditional diffusion model that consists of two primary components: a trunk conditioning module and a diffusion module heavily conditioned on diffusion time step. The conditioning module is a multimodal encoder that integrates information from sequences, multiple sequence alignments (MSAs), and structural templates for structure prediction. However, in representation learning scenarios where complex structures are provided as input, the model's primary role shifts from inferring geometry to encoding the observed geometry. As a result, the reliance on heavy multimodal conditioning becomes less critical. Additionally, AlphaFold 3 depends heavily on MSAs and templates, which are not always available and may be suboptimal for learning generalizable representations. In contrast, the diffusion module employs a modern transformer architecture that is well-suited for scalability. It incorporates advanced techniques such as SwiGLU (Shazeer, 2020), gating mechanisms from AlphaFold 2 (Jumper et al., 2021), and a two-level architecture that alternates between atom-level and token-level representations. These features make the diffusion module particularly effective for updating representations based on both the geometry and the sequence. Based on these insights, we remove the dependencies on MSAs and templates, and introduce ADiT for representation learning across diverse biomolecular interactions and affinity prediction.

3.2 SIMPLIFIED FEATURIZATION SCHEME

We employ a generalized tokenization scheme, where each residue in proteins is represented as a single token, and each heavy atom in small molecules is encoded as a single token. We initialize the multi-level input features using a top-down approach. This involves first initializing features at the token level, followed by deriving atom-level features by integrating both atom-specific information and the propagated token-level features. Details of our feature embedder are illustrated in Fig. 2*(a)*.

Token Level Encoding. We first construct a token-level sequence embedder that encodes the chemical and biological information of complexes, producing representations for all tokens. These representations serve as the conditioning input for the subsequent diffusion transformer. Unlike AlphaFold 3, which relies on MSA and template inputs, introducing additional complexity and potential challenges when MSAs or templates are unavailable or of low quality, we use a pre-trained protein language model, ESM-2-650M (Lin et al., 2023), to incorporate evolutionary information. Specifically, the token conditioning representations consist of two components: the ESM sequence feature and the token type embedding. The ESM sequence feature is only computed for residue tokens, while for small molecule tokens, this feature is set to zero. The token type embedding encodes the origin of the token, distinguishing between protein and small molecule tokens.

Next, we construct pair representations by concatenating the corresponding token conditioning representations and integrate relative position encodings that include both relative sequence and chain distance. Different from AlphaFold 3, we avoid computationally intensive Pairformer blocks, but only use a concatenation and linear layer for embedding to save compute. We use $\boldsymbol{c}^{\text{token}}$ and $\boldsymbol{z}^{\text{token}}$ to denote the token conditioning and pair representations, respectively.

Atom Level Encoding. At the atomic level, we encode both chemical and evolutionary information, along with structural information for each atom. Note that the subsequent diffusion transformer will condition on the former (chemical and evolutionary information) while extracting representations from the latter (structural information). To achieve this, we distinguish between the single representation $\boldsymbol{s}^{\text{atom}}$, which includes structural information, and the conditions $\boldsymbol{c}^{\text{atom}}$, which exclude it.

To generate atom conditioning representations $\boldsymbol{c}^{\text{atom}}$, we embed the atom type and name (for protein atoms) and augment them with token conditioning representations. For single atom representations $\boldsymbol{s}^{\text{atom}}$, we encode the atom type, name, and coordinates. Then, for atom pair representations $\boldsymbol{z}^{\text{atom}}$, we combine atom conditioning representations, embed their Euclidean distance using RBF kernels, and encode whether they belong to the same token. These atom pair representations are then augmented with the corresponding token pair representations to capture both local and global interactions.

3.3 HIERARCHICAL REPRESENTATION LEARNING

To capture both token- and atom-level information, we adopt a hierarchical representation learning framework, utilizing Diffusion Transformer (DiT) modules (Peebles & Xie, 2023). The framework begins by updating atom representations through $N_{\text{block}}^{\text{atom}}$ DiT blocks, followed by an "Atom2Token" average pooling operation to generate token representations. These token representations are then refined using $N_{\text{block}}^{\text{token}}$ DiT blocks. Subsequently, a linear layer and a "Token2Atom" unpooling operation produce updated atom representations, which are combined with the previous atom representations via a skip connection. The atom representations are further refined through another $N_{\text{block}}^{\text{atom}}$ DiT blocks. An overview is illustrated in Figure 2*(b)* and outlined in Algorithm 1.

Diffusion Transformer Block. We adapt the diffusion transformer to model both interatomic and inter-token interactions for representation learning. The transformer comprises multiple stacked blocks, each containing an Adaptive LayerNorm function, a multi-head self-attention module, and a transition function, as detailed in Algorithm 2. We add skip connections to enhance training stability. The Adaptive LayerNorm function and transition function process both single and conditioning representations, while the multi-head self-attention module computes attention weights using single and pair representations. Formally, the multi-head self-attention module is defined as:

$$\boldsymbol{q}_i^h, \boldsymbol{k}_i^h, \boldsymbol{v}_i^h \leftarrow \text{Linear}_{q,k,v}(\boldsymbol{s}_i), \quad A_{ij}^h \leftarrow \text{softmax}_j\left(\frac{{\boldsymbol{q}_i^h}^\top \boldsymbol{k}_j^h}{\sqrt{d}} + \text{Linear}_b^h(\boldsymbol{z}_{ij}) + \beta_{ij}\right),$$

$$\boldsymbol{g}_i^h \leftarrow \text{sigmoid}(\text{Linear}_g(\boldsymbol{s}_i)), \quad \boldsymbol{s}_i \leftarrow \text{Linear}\left(\text{concat}_h\left(\boldsymbol{g}_i^h \odot \sum\nolimits_j A_{ij}^h \boldsymbol{v}_j^h\right)\right),$$

where h denotes the attention head index, and d is the dimension of the latent representation. The term β_{ij} controls whether interactions between pair (i, j) are modeled. For token-level interactions, all β_{ij} values are zero. For atom-level interactions, each group of 32 atoms attends to a nearby group of 128 atoms (based on sequence proximity). The DiT architecture is illustrated in Figure 2*(c)*.

Notably, ADiT utilizes only a single linear layer to embed all atom positions during the initialization of atom representations, without incorporating geometric biases such as locality or SE(3) invariance. This non-equivariant design contrasts with current trends that emphasize stronger domain-specific inductive biases, thereby facilitating the general usage on different data modalities. We approximate the required invariance through a combination of centroid-centering of the input coordinates and data augmentation via random rotations during pre-training. We hypothesize that this non-equivariant approach offers greater flexibility to capture crucial non-geometric features (e.g., electrostatic interactions and chemical semantics from RDKit and ESM-2 features) that govern binding thermodynamics, without being overly constrained by the strong geometric bias of equivariant layers. Furthermore, non-equivariant Transformers are architecturally simpler and highly scalable, enabling us to build the foundation model more efficiently than current equivariant frameworks.

3.4 Pre-training and Fine-tuning Framework

Inspired by recent foundation models, we adopt a two-stage training approach: first pre-training our model with a denoising objective, followed by fine-tuning it for specific downstream tasks.

Pre-training. Denoising pre-training has emerged as a widely used self-supervised learning approach for learning representations of 3D structures (Zaidi et al., 2023; Zhang et al., 2023b). The process involves adding Gaussian noise to each atom's coordinate as $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I})$, where σ represents the noise scale. The noisy structure is then input into ADiT, and a noise prediction head is applied to the resulting atom representations. As ADiT is a non-equivariant model, we incorporate random rotations of structures as data augmentation. Following (Zaidi et al., 2023), we use a fixed noise scale, treating it as a tunable hyperparameter. Preliminary experiments indicate that $\sigma = 0.5$Å yields the best performance. Although we also experimented with varying noise scales as in (Zhang et al., 2023d), no practical improvements were observed. Nonetheless, we hypothesize that a carefully curated training distribution of noise scales could enable the model to better capture both coarse- and fine-grained features. Due to computational constraints, we leave this exploration to our future work.

Fine-tuning. To adapt our pre-trained model for downstream tasks, we fine-tune it using a reduced learning rate. For predicting properties of the bound complex, we input the structure into the model to obtain output atom representations. These representations are then sequentially aggregated through "Atom2Token" average pooling and "Token2Complex" sum pooling, followed by a prediction head. In addition, our goal is to learn representations from clean samples, rather than to generate samples from noise as in standard diffusion models. We do not need to condition on the diffusion time step. Accordingly, during fine-tuning, the time step is always set to zero, corresponding to a clean sample.

Difference with AlphaFold 3. While ADiT leverages architectural components similar to those in AlphaFold 3, our model represents a significant re-engineering tailored for the task of affinity prediction, leading to substantial differences in architecture, input, and objective. The key distinctions are as follows: (1) **Goal:** AlphaFold 3 is fundamentally a generative model designed for structure prediction, while ADiT is a representation learner specialized for affinity prediction. (2) **Inputs:** We remove AlphaFold 3's reliance on computationally expensive MSA and template conditioning, which we replace with pre-trained protein features from ESM-2 and explicit RDKit features for small molecules. (3) **Architecture:** We simplify the backbone by removing the heavy trunk module and the computationally intensive Pairformer blocks, focusing instead on a more scalable stack of Diffusion Transformer blocks. (4) **Objective:** We adapt the core diffusion transformer module to a simpler denoising objective, rather than the generative structure prediction task. We believe these adaptations are important for the strong performance of ADiT on binding affinity prediction.

4 Experiments

The goal of this paper is to develop a *general-purpose* foundation model for predicting the binding affinity of various types of biomolecular complexes. In this section, we evaluate our model on four downstream tasks and compare its performance against recent strong baselines. The experimental

Table 1: **Protein-ligand binding affinity prediction results.** The best result is highlighted in **bold**, and the second-best result is underlined. Differences within 0.002 are considered negligible.

Model	Sequence Identity 30%			Sequence Identity 60%		
	RMSE↓	Pearson↑	Spear.↑	RMSE↓	Pearson↑	Spear.↑
B & B (Bepler & Berger, 2019)	1.985	0.165	0.152	1.891	0.249	0.275
DeepDTA (Öztürk et al., 2018)	1.866	0.472	0.471	1.762	0.666	0.663
DeepAffnity (Karimi et al., 2019)	1.893	0.415	0.426	–	–	–
MaSIF (Gainza et al., 2020)	1.484	0.467	0.455	1.426	0.709	0.701
Atom3D-GNN (Townshend et al., 2021)	1.601	0.545	0.533	1.408	0.743	0.743
IEConv (Hermosilla et al., 2021)	1.554	0.414	0.428	1.473	0.667	0.675
Holoprot (Somnath et al., 2021)	1.464	0.509	0.500	1.365	0.749	0.742
ProNet (Wang et al., 2023)	1.463	0.551	0.551	1.343	0.765	0.761
Uni-Mol (Zhou et al., 2023)	1.520	0.558	0.540	1.619	0.645	0.653
GeoSSL (Liu et al., 2023)	1.451	0.577	0.572	–	–	–
ProFSA (Gao et al., 2023)	1.377	0.628	0.620	1.377	0.764	0.762
GET (Kong et al., 2024)	1.327	0.620	0.611	1.510	0.675	0.688
GET-PS (Kong et al., 2024)	**1.309**	0.633	0.642	1.509	0.676	0.680
Protenix (Chen et al., 2025)	1.354	0.569	0.593	1.781	0.707	0.737
ADiT-S	1.337	0.626	0.618	1.413	0.740	0.740
ADiT-M	1.353	0.622	0.630	1.335	0.764	0.752
ADiT-L	**1.308**	**0.645**	**0.647**	**1.246**	**0.797**	**0.795**

results demonstrate that as a general-purpose foundation model, ADiT achieves state-of-the-art or competitive results across these tasks, even compared with specialized models.

4.1 Pre-training Setup

For pre-training, we curate datasets from the Protein Data Bank (PDB) (Berman et al., 2000), resulting in 433,297 protein single chains, 481,382 protein-protein interaction examples, and 427,947 protein-ligand interaction examples. Here the interaction example refers to the interaction between two specific chains that form chain-chain interfaces within a single PDB entry. Two chains are considered to interact if the minimum heavy-atom distance between them is less than 5 angstroms. To accelerate pre-training, we leverage the clustering results provided by the RCSB PDB, yielding a total of 150,009 clusters. Importantly, we do not use any function labels during pre-training to avoid potential data leakage and ensure a fair evaluation on downstream tasks.

To explore the scaling effects, we pre-trained three model variants: ADiT-S (12M params), ADiT-M (35M params), and ADiT-L (253M params). The largest variant, ADiT-L, matches the number of layers and hidden dimensions used in AlphaFold 3. The differences between these variants lie in the configurations of their atom and token transformer layers. For details, please refer to Appendix E.1.

4.2 Protein-Ligand Binding Affinity Prediction

Setup. For protein-ligand binding affinity, we adopt the setup from Atom3D (Townshend et al., 2021), where each sample consists of the bound structure of a protein-ligand complex, and our goal is to predict its binding affinity. The dataset includes two splits based on protein sequence similarity: one with a max. sequence identity 30% (LBA-30) and another with a max. sequence identity 60% (LBA-60). We use the Root Mean Square Error (RMSE), Pearson correlation coefficient, and Spearman's rank correlation coefficient as evaluation metrics. We include recent strong baselines, and additionally fine-tune Protenix (Chen et al., 2025), an open-source reproduction of AlphaFold3.

Results. The results in Table 1 demonstrate that all three ADiT variants achieve competitive performance compared to previous strong baselines. Among them, ADiT-L achieves state-of-the-art results across both data splits and all evaluation metrics. Specifically, on the LBA-30 benchmark, it outperforms the previous best method, GET, with a 1.9% increase in Pearson correlation and a 0.77% improvement in Spearman correlation. On LBA-60 benchmark, it surpasses prior best baselines ProNet and ProFSA, achieving a 7.2% reduction in RMSE, along with 4.2% and 4.3% increases in Pearson and Spearman correlations, respectively. Notably, even the smallest variant, ADiT-S, with only 12M parameters, outperforms most existing baselines. Furthermore, we observe that prior

Table 2: **Results for drug-target and protein-protein binding affinity prediction.** The best result is in **bold**, and the second-best is underlined. Differences within 0.002 are considered negligible.

Drug-Target			Protein-Protein				
Model	**MSE↓**	r_m^2↑	**Model**	**Pearson↑**	**Spear.↑**	**RMSE↓**	**MAE↓**
DeepDTA	0.261	0.630	ESM-IF	0.319	0.280	1.886	1.285
MT-DTI	0.245	0.665	End-to-End	0.637	0.488	1.619	1.176
GraphDTA	0.229	0.685	DDGPred	0.658	0.468	**1.499**	1.082
rzMLP	0.205	0.709	MIF-Network	0.652	0.513	1.593	1.146
EnsembleDLM	0.202	–	RDE-Network	0.644	0.558	1.579	1.112
FusionDTA	0.208	0.743	DiffAffinity	0.669	0.556	1.535	1.093
MgraphDTA	0.207	0.710	Prompt-DDG	0.677	**0.591**	1.520	**1.077**
NHGNN-DTA	**0.196**	0.744	Surface-VQMAE	0.648	0.561	1.588	1.127
ADiT-S	0.252	0.690	**ADiT-S**	0.660	0.524	1.597	1.132
ADiT-M	0.216	0.734	**ADiT-M**	0.683	0.539	1.559	1.098
ADiT-L	**0.198**	**0.751**	**ADiT-L**	**0.691**	0.560	1.540	1.088

methods exhibit limited generalization capabilities, performing well on either LBA-30 or LBA-60, but not both, *e.g.*, GET and ProNet. In contrast, our approach demonstrates consistent and robust performance across both dataset splits. Moreover, our results underscore the superiority of ADiT over Protenix across all evaluation metrics, showing that simply fine-tuning generative models for representation learning and downstream tasks is insufficient to unlock their full potential.

4.3 Drug-Target Binding Affinity Prediction

Setup. For the drug-target binding affinity task, we employ the widely used Davis dataset (Davis et al., 2011), consisting of 30,056 data points that capture binding affinities between 72 small-molecule drugs and 442 target protein kinases. Following (He et al., 2023), we perform a random split of the Davis dataset five times and report the average performance across these splits. We employ two evaluation metrics: Mean Squared Error (MSE) and the r_m^2 metric. Details in Appendix E.3.

Results. The results, summarized in the left part of Table 2, demonstrate that ADiT delivers competitive performance, with its best-performing variant, ADiT-L, achieving state-of-the-art results on both the MSE and r_m^2 metrics. Moreover, ADiT exhibits a clear scaling effect with respect to model size, suggesting that larger models have the potential to further improve performance.

4.4 Protein-Protein Binding Affinity Prediction

Setup. For the protein-protein binding affinity task, we adopt the SKEMPIv2 dataset (Jankauskaitė et al., 2019), the most widely used benchmark for $\Delta\Delta\mathcal{G}$ prediction. SKEMPIv2 is a comprehensive, annotated mutation dataset covering 348 protein complexes, containing 7,085 amino acid mutations along with corresponding changes in thermodynamic parameters and kinetic rate constants. Since SKEMPIv2 does not provide structures for the mutated complexes, we generate them using FoldX (Delgado et al., 2019). We evaluate the performance of ADiT using a split-by-complex threefold cross-validation on the SKEMPIv2 dataset following (Liu et al., 2024). The metrics include: (1) Pearson correlation coefficient, (2) Spearman's rank correlation coefficient, (3) root mean squared error (RMSE) and (4) mean absolute error (MAE). Details in Appendix E.4.

Results. The detailed results are presented in Table 2. Among all baselines, ADiT demonstrates strong and consistent performance. The best-performing variant, ADiT-L, achieves the highest Pearson correlation coefficient and the second-best Spearman's rank correlation coefficient. Although this is a ranking task, where metrics like RMSE and MAE are less critical, ADiT still performs comparably to the strongest baselines on these metrics, with only marginal differences.

4.5 Antibody-Antigen Binding Affinity Prediction

Setup. With the models trained on SKEMPIv2, we test their performance on the HER2 binders test set, following (Cai et al., 2024). The HER2 binders test set is collected from (Shanehsazzadeh et al., 2023) and contains high-quality binding affinity data, measured by surface plasmon resonance (SPR)

Table 4: **Ablation study results on SKEMPIv2 based on ADiT-M.**

NO.	Model	Pearsonr↑	Spearmanr↑	RMSE↓	MAE↓
0	**ADiT-M**	0.683	0.539	1.559	1.098
1	w/o pre-training	0.649	0.511	1.624	1.169
2	w/o all-atom info.	0.658	0.517	1.606	1.153
3	w/ larger size	0.691	0.560	1.540	1.088
4	w/ smaller size	0.660	0.524	1.597	1.132

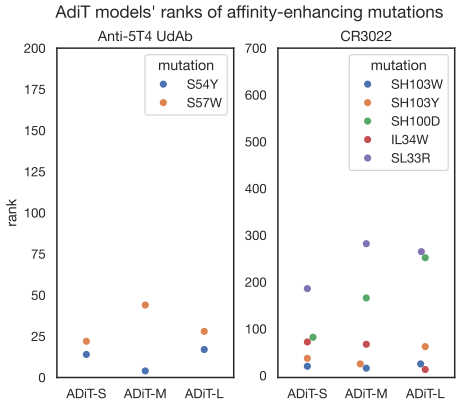


Figure 3: **Case Study on Antibody Affinity Maturation.** ADiT models' ranks of affinity-enhancing mutations for nanobody Anti-5T4 UdAb (780 mutations) and antibody CR3022 (1420 mutations) are shown. Lower rank is better.

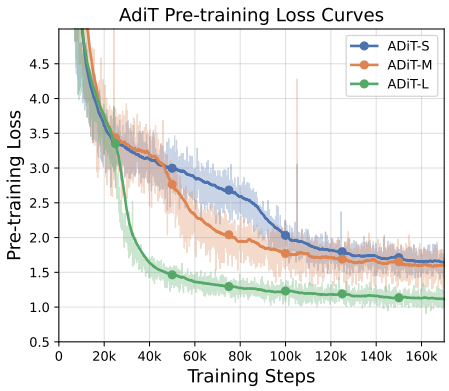


Figure 4: **Pre-training and Scaling Effects Across Model Sizes.** Pre-training loss curves for the small, medium, and large variants of ADiT. Larger models consistently achieve lower loss throughout the pre-training process.

on 419 HER2 binders with de novo designed CDR loops. The antibodies in the dataset are variants of Trastuzumab that have high edit distance (7.6 on average), making them potentially challenging for $\Delta\Delta\mathcal{G}$ bind predictors trained on low-edit-distance data. Here, we use Pearson and Spearman correlation as evaluation metrics to assess the performance of our ADiT. Details in Appendix E.5.

Results. Results are shown in Table 3. Our best-performing variant, ADiT-L, demonstrates superior performance, outperforming previous methods Bind-ddG (Shan et al., 2022) and GearBind (Cai et al., 2024). These findings underscore the robust generalization capability of ADiT to held-out antibody data.

Table 3: **Antibody-antigen binding affinity prediction results.** The best result is highlighted in **bold**, and the second-best result is underlined.

Model	Pearson↑	Spear.↑
Bind-ddG	0.387	0.388
GearBind	0.478	0.475
GearBind + P	0.515	0.517
FoldX	0.246	0.329
ADiT-S	<u>0.527</u>	<u>0.558</u>
ADiT-M	0.507	0.523
ADiT-L	**0.567**	**0.576**

Case Study. We evaluate ADiT's performance on real-world *in silico* affinity maturation using models trained on SKEMPIv2. Two starting antibodies with different formats and target antigens are tested (see Appendix E.5 for details). Figure 3 shows the rankings of seven wet-lab-validated affinity-enhancing mutations from (Cai et al., 2024)[1]. The results demonstrate that ADiT assigns better ranks to most beneficial mutations (e.g., S54Y and S57W for Anti-5T4 UdAb, and SH103W, SH103Y, and IL34W for CR3022), highlighting its potential as an effective tool for antibody optimization.

[1]Only a limited number of mutations were experimentally validated, suggesting that additional affinity-enhancing candidates likely exist.

4.6 Ablation Study

To better understand the key contributors to the performance of ADiT, we conduct an ablation study on the protein-protein binding affinity prediction task using the SKEMPIv2 dataset. We use ADiT-M as the baseline and evaluate the impact of several single-factor modifications: (1) removing pretraining and use randomly initialized parameters, (2) replacing all-atom (AA) structures with backbone-only structures in both the pretraining and finetuning stages, (3) scaling up the model size (35M) to ADiT-L (253M), and (4) scaling down to ADiT-S (12M).

The results are summarized in Table 4. (1) Comparing No.0 and No.1 highlights the importance of pretraining, which significantly contributes to performance gains (5.2% improvement in Pearson, 5.5% improvement in Spearman, 4% improvement in RMSE, 6% improvement in MAE). (2) The comparison between No.0 and No.2 demonstrates the value of all-atom structural modeling (3.8% improvement in Pearson, 4.3% improvement in Spearman, 3% improvement in RMSE, 4.8% improvement in MAE), which provides richer structural information and leads to improved representations. (3) Finally, the comparison among No.0, No.3, and No.4 reveals consistent performance improvements with increased model size, reflecting a clear scaling trend in line with observations from other domains. Figure 4 presents the pretraining loss curves for the small, medium, and large model variants. Larger models consistently achieve lower loss during pretraining, underscoring their enhanced ability to capture complex molecular patterns.

5 Conclusion

In this work, we take an important first step toward developing all-atom structure foundation models for biomolecular interactions. With large-scale pre-training, our unified foundation model achieves state-of-the-art or competitive performance across diverse tasks. Furthermore, we observe scaling effects in our model architecture that align with trends seen in other domains. Quantitative and qualitative analyses show the practical effectiveness of our approach. We further discuss the LLM usage in Appendix A, limitations in Appendix B, and broader impacts in Appendix C.

Reproducibility statement

We provide a clear and detailed description of the architecture in Sec. 3, along with the necessary pretraining and finetuning configurations in Sec. E. The pretraining dataset is constructed from publicly accessible PDB and clustered using publicly available clustering results. All finetuning datasets used in our experiments are also publicly accessible. Public access to codes will be released when not anonymized.

References

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, pp. 1–3, 2024.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Rebecca F Alford, Andrew Leaver-Fay, Jeliazko R Jeliazkov, Matthew J O'Meara, Frank P DiMaio, Hahnbeom Park, Maxim V Shapovalov, P Douglas Renfrew, Vikram K Mulligan, Kalli Kappel, et al. The rosetta all-atom energy function for macromolecular modeling and design. *Journal of chemical theory and computation*, 13(6):3031–3048, 2017.

Kyle A Barlow, Shane O Conchuir, Samuel Thompson, Pooja Suresh, James E Lucas, Markus Heinonen, and Tanja Kortemme. Flex ddg: Rosetta ensemble-based estimation of changes in protein–protein binding affinity upon mutation. *The Journal of Physical Chemistry B*, 122(21): 5389–5399, 2018.

Tristan Bepler and Bonnie Berger. Learning protein sequence embeddings using information from structure. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=SygLehCqtm.

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.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Hannah Busch, Muhammad Yasir Ateeque, Florian Taube, Thomas Wiegand, Björn Corzilius, and Georg Künze. Probing biomolecular interactions with paramagnetic nuclear magnetic resonance spectroscopy. *ChemBioChem*, pp. e202400903, 2025.

Huiyu Cai, Zuobai Zhang, Mingkai Wang, Bozitao Zhong, Quanxiao Li, Yuxuan Zhong, Yanling Wu, Tianlei Ying, and Jian Tang. Pretrainable geometric graph neural network for antibody affinity maturation. *Nature communications*, 15(1):7785, 2024.

Chai Discovery. Chai-1: Decoding the molecular interactions of life. *bioRxiv*, 2024. doi: 10.1101/2024.10.10.615955. URL https://www.biorxiv.org/content/early/2024/10/11/2024.10.10.615955.

Can (Sam) Chen, Jingbo Zhou, Fan Wang, Xue Liu, and Dejing Dou. Structure-aware protein self-supervised learning. *Bioinformatics*, 39, 2022.

Xinshi Chen, Yuxuan Zhang, Chan Lu, Wenzhi Ma, Jiaqi Guan, Chengyue Gong, Jincai Yang, Hanyu Zhang, Ke Zhang, Shenghao Wu, et al. Protenix-advancing structure prediction through a comprehensive alphafold3 reproduction. *bioRxiv*, pp. 2025–01, 2025.

Feixiong Cheng, Junfei Zhao, Yang Wang, Weiqiang Lu, Zehui Liu, Yadi Zhou, William R Martin, Ruisheng Wang, Jin Huang, Tong Hao, et al. Comprehensive characterization of protein–protein interactions perturbed by disease mutations. *Nature genetics*, 53(3):342–353, 2021.

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*, 2023. URL https://openreview.net/forum?id=kKF8_K-mBbS.

Mindy I Davis, Jeremy P Hunt, Sanna Herrgard, Pietro Ciceri, Lisa M Wodicka, Gabriel Pallares, Michael Hocker, Daniel K Treiber, and Patrick P Zarrinkar. Comprehensive analysis of kinase inhibitor selectivity. *Nature biotechnology*, 29(11):1046–1051, 2011.

Javier Delgado, Leandro G Radusky, Damiano Cianferoni, and Luis Serrano. Foldx 5.0: working with rna, small molecules and a new graphical interface. *Bioinformatics*, 35(20):4168–4169, 2019.

Georgy Derevyanko, Sergei Grudinin, Yoshua Bengio, and Guillaume Lamoureux. Deep convolutional networks for quality assessment of protein folds. *Bioinformatics*, 34(23):4046–4053, 2018.

Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Wang Yu, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, Debsindhu Bhowmik, and Burkhard Rost. Prottrans: Towards cracking the language of lifes code through self-supervised deep learning and high performance computing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2021. doi: 10.1109/TPAMI.2021.3095381.

Hehe Fan, Zhangyang Wang, Yi Yang, and Mohan Kankanhalli. Continuous-discrete convolution for geometry-sequence modeling in proteins. In *The Eleventh International Conference on Learning Representations*, 2023.

Luciano Floridi and Massimo Chiriatti. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30:681–694, 2020.

Pablo Gainza, Freyr Sverrisson, Frederico Monti, Emanuele Rodola, D Boscaini, MM Bronstein, and BE Correia. Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nature Methods*, 17(2):184–192, 2020.

Octavian-Eugen Ganea, Xinyuan Huang, Charlotte Bunne, Yatao Bian, Regina Barzilay, Tommi S. Jaakkola, and Andreas Krause. Independent SE(3)-equivariant models for end-to-end rigid protein docking. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=GQjaI9mLet.

Bowen Gao, Yinjun Jia, YuanLe Mo, Yuyan Ni, Wei-Ying Ma, Zhi-Ming Ma, and Yanyan Lan. Self-supervised pocket pretraining via protein fragment-surroundings alignment. In *The Twelfth International Conference on Learning Representations*, 2023.

Vladimir Gligorijević, P Douglas Renfrew, Tomasz Kosciolek, Julia Koehler Leman, Daniel Berenberg, Tommi Vatanen, Chris Chandler, Bryn C Taylor, Ian M Fisk, Hera Vlamakis, et al. Structure-based protein function prediction using graph convolutional networks. *Nature communications*, 12(1):3168, 2021.

Haohuai He, Guanxing Chen, and Calvin Yu-Chian Chen. Nhgnn-dta: a node-adaptive hybrid graph neural network for interpretable drug–target binding affinity prediction. *Bioinformatics*, 39(6): btad355, 2023.

Michael Heinzinger, Konstantin Weissenow, Joaquin Gomez Sanchez, Adrian Henkel, Milot Mirdita, Martin Steinegger, and Burkhard Rost. Bilingual language model for protein sequence and structure. *NAR Genomics and Bioinformatics*, 6(4):lqae150, 11 2024. ISSN 2631-9268. doi: 10.1093/nargab/lqae150. URL https://doi.org/10.1093/nargab/lqae150.

Pedro Hermosilla, Marco Schäfer, Matěj Lang, Gloria Fackelmann, Pere Pau Vázquez, Barbora Kozlíková, Michael Krone, Tobias Ritschel, and Timo Ropinski. Intrinsic-extrinsic convolution and pooling for learning on 3d protein structures. *International Conference on Learning Representations*, 2021.

Chloe Hsu, Robert Verkuil, Jason Liu, Zeming Lin, Brian Hie, Tom Sercu, Adam Lerer, and Alexander Rives. Learning inverse folding from millions of predicted structures. In *International conference on machine learning*, pp. 8946–8970. PMLR, 2022.

Mingyang Hu, Fajie Yuan, Kevin Yang, Fusong Ju, Jin Su, Hui Wang, Fei Yang, and Qiuyang Ding. Exploring evolution-aware &-free protein language models as protein function predictors. *Advances in Neural Information Processing Systems*, 35:38873–38884, 2022.

Justina Jankauskaitė, Brian Jiménez-García, Justas Dapkūnas, Juan Fernández-Recio, and Iain H Moal. Skempi 2.0: an updated benchmark of changes in protein–protein binding energy, kinetics and thermodynamics upon mutation. *Bioinformatics*, 35(3):462–469, 2019.

Rui Jiao, Xiangzhe Kong, Ziyang Yu, Wenbing Huang, and Yang Liu. Equivariant pretrained transformer for unified geometric learning on multi-domain 3d molecules. *arXiv preprint arXiv:2402.12714*, 2024.

Wengong Jin, Xun Chen, Amrita Vetticaden, Siranush Sarzikova, Raktima Raychowdhury, Caroline Uhler, and Nir Hacohen. Dsmbind: Se (3) denoising score matching for unsupervised binding energy prediction and nanobody design. *bioRxiv*, pp. 2023–12, 2023a.

Wengong Jin, Siranush Sarkizova, Xun Chen, Nir Hacohen, and Caroline Uhler. Unsupervised protein-ligand binding energy prediction via neural euler's rotation equation. *Advances in Neural Information Processing Systems*, 36:33514–33528, 2023b.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

Po-Yu Kao, Shu-Min Kao, Nan-Lan Huang, and Yen-Chu Lin. Toward drug-target interaction prediction via ensemble modeling and transfer learning. In *2021 IEEE international conference on bioinformatics and biomedicine (BIBM)*, pp. 2384–2391. IEEE, 2021.

Mostafa Karimi, Di Wu, Zhangyang Wang, and Yang Shen. Deepaffinity: interpretable deep learning of compound–protein affinity through unified recurrent and convolutional neural networks. *Bioinformatics*, 35(18):3329–3338, 2019.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1. Minneapolis, Minnesota, 2019.

Mohamed Amine Ketata, Cedrik Laue, Ruslan Mammadov, Hannes Stark, Menghua Wu, Gabriele Corso, Céline Marquet, Regina Barzilay, and Tommi S. Jaakkola. Diffdock-PP: Rigid protein-protein docking with diffusion models. In *ICLR 2023 - Machine Learning for Drug Discovery workshop*, 2023. URL https://openreview.net/forum?id=AM7WbQxuRS.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4026, 2023.

Xiangzhe Kong, Wenbing Huang, and Yang Liu. Generalist equivariant transformer towards 3d molecular interaction learning. In *ICML*. OpenReview.net, 2024.

Rohith Krishna, Jue Wang, Woody Ahern, Pascal Sturmfels, Preetham Venkatesh, Indrek Kalvet, Gyu Rie Lee, Felix S Morey-Burrows, Ivan Anishchenko, Ian R Humphreys, et al. Generalized biomolecular modeling and design with rosettafold all-atom. *Science*, 384(6693):eadl2528, 2024.

Alexander Kroll, Sahasra Ranjan, Martin KM Engqvist, and Martin J Lercher. A general model to predict small molecule substrates of enzymes based on machine and deep learning. *Nature communications*, 14(1):2787, 2023a.

Alexander Kroll, Sahasra Ranjan, and Martin J. Lercher. Drug-target interaction prediction using a multi-modal transformer network demonstrates high generalizability to unseen proteins. *bioRxiv*, 2023b. doi: 10.1101/2023.08.21.554147. URL https://www.biorxiv.org/content/early/2023/10/12/2023.08.21.554147.

Tianhong Li, Huiwen Chang, Shlok Mishra, Han Zhang, Dina Katabi, and Dilip Krishnan. Mage: Masked generative encoder to unify representation learning and image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2142–2152, 2023.

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.

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.

Shiwei Liu, Tian Zhu, Milong Ren, Chungong Yu, Dongbo Bu, and Haicang Zhang. Predicting mutational effects on protein-protein binding via a side-chain diffusion probabilistic model. *Advances in Neural Information Processing Systems*, 36, 2024.

Shitong Luo, Yufeng Su, Zuofan Wu, Chenpeng Su, Jian Peng, and Jianzhu Ma. Rotamer density estimator is an unsupervised learner of the effect of mutations on protein-protein interaction. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=_X9Yl1K2mD.

Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alexander Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021.

Yuanle Mo, Xin Hong, Bowen Gao, Yinjun Jia, and Yanyan Lan. Multi-level interaction modeling for protein mutational effect prediction. *arXiv preprint arXiv:2405.17802*, 2024.

Thin Nguyen, Hang Le, Thomas P Quinn, Tri Nguyen, Thuc Duy Le, and Svetha Venkatesh. Graphdta: predicting drug–target binding affinity with graph neural networks. *Bioinformatics*, 37 (8):1140–1147, 2021.

Pascal Notin, Mafalda Dias, Jonathan Frazer, Javier Marchena Hurtado, Aidan N Gomez, Debora Marks, and Yarin Gal. Tranception: protein fitness prediction with autoregressive transformers and inference-time retrieval. In *International Conference on Machine Learning*, pp. 16990–17017. PMLR, 2022.

Hakime Öztürk, Arzucan Özgür, and Elif Ozkirimli. Deepdta: deep drug–target binding affinity prediction. *Bioinformatics*, 34(17):i821–i829, 2018.

Kamila J Pacholarz, Rachel A Garlish, Richard J Taylor, and Perdita E Barran. Mass spectrometry based tools to investigate protein–ligand interactions for drug discovery. *Chemical Society Reviews*, 41(11):4335–4355, 2012.

Saro Passaro, Gabriele Corso, Jeremy Wohlwend, Mateo Reveiz, Stephan Thaler, Vignesh Ram Somnath, Noah Getz, Tally Portnoi, Julien Roy, Hannes Stark, et al. Boltz-2: Towards accurate and efficient binding affinity prediction. *BioRxiv*, 2025.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023.

Zongzhao Qiu, Qihong Jiao, Yuxiao Wang, Cheng Chen, Daming Zhu, and Xuefeng Cui. rzmlp-dta: gmlp network with rezero for sequence-based drug-target affinity prediction. In *2021 IEEE international conference on bioinformatics and biomedicine (BIBM)*, pp. 308–313. IEEE, 2021.

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.

Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Peter Chen, John Canny, Pieter Abbeel, and Yun Song. Evaluating protein transfer learning with tape. *Advances in neural information processing systems*, 32, 2019a.

Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Xi Chen, John Canny, Pieter Abbeel, and Yun S Song. Evaluating protein transfer learning with tape. In *Advances in Neural Information Processing Systems*, 2019b.

Roshan M Rao, Jason Liu, Robert Verkuil, Joshua Meier, John Canny, Pieter Abbeel, Tom Sercu, and Alexander Rives. Msa transformer. 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. 8844–8856. PMLR, 18–24 Jul 2021.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

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), 2021.

Sisi Shan, Shitong Luo, Ziqing Yang, Junxian Hong, Yufeng Su, Fan Ding, Lili Fu, Chenyu Li, Peng Chen, Jianzhu Ma, et al. Deep learning guided optimization of human antibody against sars-cov-2 variants with broad neutralization. *Proceedings of the National Academy of Sciences*, 119(11): e2122954119, 2022.

Amir Shanehsazzadeh, Sharrol Bachas, Matt McPartlon, George Kasun, John M Sutton, Andrea K Steiger, Richard Shuai, Christa Kohnert, Goran Rakocevic, Jahir M Gutierrez, et al. Unlocking de novo antibody design with generative artificial intelligence. *bioRxiv*, pp. 2023–01, 2023.

Noam M. Shazeer. Glu variants improve transformer. *arxiv*, abs/2002.05202, 2020. URL https://api.semanticscholar.org/CorpusID:211096588.

Bonggun Shin, Sungsoo Park, Keunsoo Kang, and Joyce C Ho. Self-attention based molecule representation for predicting drug-target interaction. In *Machine learning for healthcare conference*, pp. 230–248. PMLR, 2019.

Vignesh Ram Somnath, Charlotte Bunne, and Andreas Krause. Multi-scale representation learning on proteins. *Advances in Neural Information Processing Systems*, 34, 2021.

Jin Su, Chenchen Han, Yuyang Zhou, Junjie Shan, Xibin Zhou, and Fajie Yuan. Saprot: Protein language modeling with structure-aware vocabulary. In *International Conference on Learning Representations*, 2024.

Tanlin Sun, Bo Zhou, Luhua Lai, and Jianfeng Pei. Sequence-based prediction of protein protein interaction using a deep-learning algorithm. *BMC bioinformatics*, 18:1–8, 2017.

Raphael John Lamarre Townshend, Martin Vögele, Patricia Adriana Suriana, Alexander Derry, Alexander Powers, Yianni Laloudakis, Sidhika Balachandar, Bowen Jing, Brandon M. Anderson, Stephan Eismann, Risi Kondor, Russ Altman, and Ron O. Dror. ATOM3d: Tasks on molecules in three dimensions. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021. URL https://openreview.net/forum?id=FkDZLpK1Ml2.

Limei Wang, Haoran Liu, Yi Liu, Jerry Kurtin, and Shuiwang Ji. Learning hierarchical protein representations via complete 3d graph networks. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=9X-hgLDLYkQ.

Xinyou Wang, Zaixiang Zheng, Fei Ye, Dongyu Xue, Shujian Huang, and Quanquan Gu. Diffusion language models are versatile protein learners. In *International Conference on Machine Learning*, 2024a.

Xinyou Wang, Zaixiang Zheng, Fei Ye, Dongyu Xue, Shujian Huang, and Quanquan Gu. Dplm-2: A multimodal diffusion protein language model. *arXiv preprint arXiv:2410.13782*, 2024b.

Jeremy Wohlwend, Gabriele Corso, Saro Passaro, Mateo Reveiz, Ken Leidal, Wojtek Swiderski, Tally Portnoi, Itamar Chinn, Jacob Silterra, Tommi Jaakkola, and Regina Barzilay. Boltz-1: Democratizing biomolecular interaction modeling. *bioRxiv*, 2024. doi: 10.1101/2024.11.19.624167.

Fang Wu and Stan Z. Li. Surface-VQMAE: Vector-quantized masked auto-encoders on molecular surfaces. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=szxtVHOh0C.

Fang Wu, Siyuan Li, Lirong Wu, Dragomir Radev, Qiang Zhang, and Stan Z Li. Discovering the representation bottleneck of graph neural networks from multi-order interactions. 2022.

Lirong Wu, Yijun Tian, Haitao Lin, Yufei Huang, Siyuan Li, Nitesh V Chawla, and Stan Z. Li. Learning to predict mutational effects of protein-protein interactions by microenvironment-aware hierarchical prompt learning. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=g89jAdrnAF.

Minghao Xu, Xinyu Yuan, Santiago Miret, and Jian Tang. Protst: Multi-modality learning of protein sequences and biomedical texts. In *International Conference on Machine Learning*, pp. 38749–38767. PMLR, 2023.

Fang Yang, Kunjie Fan, Dandan Song, and Huakang Lin. Graph-based prediction of protein-protein interactions with attributed signed graph embedding. *BMC bioinformatics*, 21:1–16, 2020.

Ziduo Yang, Weihe Zhong, Lu Zhao, and Calvin Yu-Chian Chen. Mgraphdta: deep multiscale graph neural network for explainable drug–target binding affinity prediction. *Chemical science*, 13(3): 816–833, 2022.

Weining Yuan, Guanxing Chen, and Calvin Yu-Chian Chen. Fusiondta: attention-based feature polymerizer and knowledge distillation for drug-target binding affinity prediction. *Briefings in Bioinformatics*, 23(1):bbab506, 2022.

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. In *The Eleventh International Conference on Learning Representations*, 2023.

Yangtian Zhang, Huiyu Cai, Chence Shi, and Jian Tang. E3bind: An end-to-end equivariant network for protein-ligand docking. In *The Eleventh International Conference on Learning Representations*, 2023a.

Zuobai Zhang, Chuanrui Wang, Minghao Xu, Vijil Chenthamarakshan, Aurelie Lozano, Payel Das, and Jian Tang. A systematic study of joint representation learning on protein sequences and structures. *arXiv preprint arXiv:2303.06275*, 2023b.

Zuobai Zhang, Minghao Xu, Arian 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*, 2023c.

Zuobai Zhang, Minghao Xu, Aurelie Lozano, Vijil Chenthamarakshan, Payel Das, and Jian Tang. Pre-training protein encoder via siamese sequence-structure diffusion trajectory prediction. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023d.

Jiale Zhao, Wanru Zhuang, Jia Song, Yaqi Li, and Shuqi Lu. Pre-training protein bi-level representation through span mask strategy on 3d protein chains. In *International Conference on Machine Learning*, pp. 21791–21805. PMLR, 2024.

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 LLM USAGE STATEMENT

LLMs were used solely for language editing and refinement of the manuscript text. The LLM did not contribute to research ideation, experimental design, data analysis, or content generation beyond linguistic improvements. All scientific concepts, results, and interpretations remain entirely the work of the authors.

The authors reviewed all LLM-suggested changes to ensure accuracy and alignment with the intended meaning, and no text was accepted without human verification.

B LIMITATIONS.

First, the current model is limited to protein-protein and protein-ligand interactions, excluding other biomolecules such as DNA and RNA. Given AlphaFold 3's demonstrated ability to model a broader range of biomolecular interaction types, extending our approach to encompass all biomolecular interactions is promising for future research. Second, the computational resource constraints have limited our ability to further scale up the model size. We believe that larger models, combined with advanced pre-training objectives, could lead to stronger foundation models. Additionally, while our study provides initial insights into scaling laws for structure foundation models, further exploration with various model sizes is necessary to better understand these trends. Finally, our framework requires fine-tuning for downstream tasks, which poses challenges in scenarios with extremely limited data. Future work should explore zero-shot inference capabilities to broaden the applicability of structure foundation models.

C BROADER IMPACTS

Our work has significant broader impacts for computational biology and drug discovery. Positively, ADiT provides a powerful, general-purpose framework for modeling biomolecular interactions, potentially accelerating therapeutic discovery and advancing biological understanding. However, ethical concerns, such as the potential misuse in drug development, also warrant careful consideration. Responsible deployment and transparency are essential to maximize benefits while mitigating risks.

D ALGORITHMS

The pseudo codes for "Hierarchical Representation Learning" and "Diffusion Transformer" are shown in Alg 1 and Alg 2.

Algorithm 1 Hierarchical Representation Learning

Input: token condition $\boldsymbol{c}^{\text{token}}$, pair representation $\boldsymbol{z}^{\text{token}}$ and atom condition $\boldsymbol{c}^{\text{atom}}$, single representation $\boldsymbol{s}^{\text{atom}}$ and pair representation $\boldsymbol{z}^{\text{atom}}$
Output: updated atom single representation $\boldsymbol{s}^{\text{atom}}$
 # Step 1: Update atom representations
$\boldsymbol{s}^{\text{atom}} \leftarrow \text{DiffusionTransformer}(\boldsymbol{s}^{\text{atom}}, \boldsymbol{c}^{\text{atom}}, \boldsymbol{z}^{\text{atom}})$

 # Step 2: Update token representations
$\boldsymbol{s}^{\text{token}} \leftarrow \text{Atom2Token}(\boldsymbol{s}^{\text{atom}})$
$\boldsymbol{s}^{\text{token}} \leftarrow \text{Linear}(\boldsymbol{c}^{\text{token}}) + \boldsymbol{s}^{\text{token}}$
$\boldsymbol{s}^{\text{token}} \leftarrow \text{DiffusionTransformer}(\boldsymbol{s}^{\text{token}}, \boldsymbol{c}^{\text{token}}, \boldsymbol{z}^{\text{token}})$

 # Step 3: Update atom representations again
$\boldsymbol{s}^{\text{atom}} \leftarrow \text{Token2Atom}(\boldsymbol{s}^{\text{token}}) + \boldsymbol{s}^{\text{atom}}$
$\boldsymbol{s}^{\text{atom}} \leftarrow \text{DiffusionTransformer}(\boldsymbol{s}^{\text{atom}}, \boldsymbol{c}^{\text{atom}}, \boldsymbol{z}^{\text{atom}})$
Return $\boldsymbol{s}^{\text{atom}}$

Algorithm 2 Diffusion Transformer

Input: single representation $\boldsymbol{s}$, pair representation $\boldsymbol{z}$, condition $\boldsymbol{c}$, number of DiT blocks N_{block}
Output: updated representation $\boldsymbol{s}$
for $i = 1$ **to** N_{block} **do**
  $\boldsymbol{s} \leftarrow \text{AdaLN}(\boldsymbol{s}, \boldsymbol{c})$
  $\boldsymbol{s} \leftarrow \text{MultiheadAttn}(\boldsymbol{s}, \boldsymbol{z}) + \boldsymbol{s}$
  $\boldsymbol{s} \leftarrow \text{Transition}(\boldsymbol{s}, \boldsymbol{c}) + \boldsymbol{s}$
end for
Return S

E EXPERIMENTAL DETAILS

All experiments in this paper are performed NVIDIA Tesla A100 GPUs with 40G memory. The dataset statistics are presented in Table 6.

E.1 PRE-TRAINING CONFIGURATION

We implement and pre-train three variants of ADiT: ADiT-S (12M parameters), ADiT-M (35M parameters), and ADiT-L (253M parameters), using the PDB dataset. Detailed training hyperparameters are listed in Table 5. Notably, ADiT-L adopts the same number of atom- and token-level transformer layers as AlphaFold 3. To accommodate GPU memory constraints, we apply random sequential or spatial cropping to the protein inputs. Additionally, we employ a data augmentation strategy that centers the protein at the origin (based on its center of mass) and applies random rotations. These augmentations help the model learn more robust and generalizable representations.

Table 5: Model Hyperparameters.

Model	ADiT-S	ADiT-M	ADiT-L
Model Configuration			
Atom Repr Dim	128	128	128
Atom Pair Dim	16	16	16
Token Repr Dim	384	384	768
Token Pair Dim	32	64	128
Atom Enc Layers	3	3	3
Atom Enc Heads	4	4	4
Token Attn Layers	3	12	24
Token Attn Heads	8	8	16
Atom Dec Layers	3	3	3
Atom Dec Heads	4	4	4
Pre-Trainig Setup			
Optimizer	Adam	Adam	Adam
Learning Rate	1e-4	1e-4	1e-4
Batch Size per GPU	6	3	1
Grad. Acc. Steps	1	2	2
Number of GPUs	4	4	32
Effective Batch Size	24	24	64
Training Steps	180k	180k	180k
Dataset Setup			
Truncation Length	250	250	200
Random Rotation	yes	yes	yes

Table 6: Dataset statistics.

Dataset	Split	Train	Validation	Test	Data Type
PDB	Random	142,509 clusters	7,500 clusters	–	Unlabeled Structure
LBA	Identity 30%	3,507	466	490	Protein-Ligand Affinity
	Identity 60%	3,678	460	460	
Davis	Random	24,044	3,005	3,005	Drug-Target Affinity
SKEMPIv2	Fold 0	4,765	–	1,929	Protein-Protein Affinity
	Fold 1	4,282	–	2,412	
	Fold 2	4,341	–	2,353	
HER2	–	–	–	419	Antibody-Antigen Affinity

E.2 Protein-Ligand Binding Affinity Prediction

Baselines We employ a diverse set of baselines, including B&B (Bepler & Berger, 2019), DeepDTA (Öztürk et al., 2018), DeepAffinity (Karimi et al., 2019), MaSIF (Gainza et al., 2020), Atom3D-GNN (Townshend et al., 2021), IEConv (Hermosilla et al., 2021), Holoprot (Somnath et al., 2021), ProtNet (Wang et al., 2023), Uni-Mol (Zhou et al., 2023), GeoSSL (Liu et al., 2023), ProFSA (Gao et al., 2023), and GET (Kong et al., 2024). Most of the reported results for the baseline methods are taken from (Gao et al., 2023), while we independently evaluate GET (Kong et al., 2024) on the LBA-60 dataset using the same configuration as in the LBA-30 setting reported by the original authors.

To fine-tune Protenix, we modified its Diffusion Module by replacing the noisy coordinates with ground-truth coordinates. We then extracted the output from the last Transformer layer of the AtomAttentionDecoder. This output was passed through a LayerNorm, followed by mean pooling to obtain per-token representations, and then sum pooling to derive a complex-level representation, following the same aggregation strategy used in ADiT. As for the hyperparameters, we kept the architectural settings of Protenix unchanged, since its architecture is fixed. However, we carefully

tuned the training hyperparameters to avoid underestimating the baseline performance. Specifically, we used a learning rate of 1e-4, an effective batch size of 16, dropout set to 0.0, and no weight decay.

Training For both the 30% and 60% sequence identity threshold splits, we use a batch size of 16, a fine-tuning learning rate of 3e-5, a weight decay of 1e-5, and apply random structural rotation as data augmentation. For 60% sequence identity threshold split, we additionally apply a dropout rate of 0.5. The total epoches on LBA-30 and LBA-60 are 3 and 99. The training process can be efficiently handled using one or two A100 GPU.

E.3 Drug-Target Binding Affinity Prediction

Baselines The baselines are: DeepDTA (Öztürk et al., 2018), MT-DTI (Shin et al., 2019), GraphDTA (Nguyen et al., 2021), rzMLP-DTA (Qiu et al., 2021), EnsembleDLM (Kao et al., 2021), MGraphDTA (Yang et al., 2022), FusionDTA (Yuan et al., 2022) and NHGNN-DTA (He et al., 2023). The reported results are taken from (He et al., 2023).

Training We generate unbound protein structures using AlphaFold 2 and molecular structures using RDKit. Since these structures do not initially reside in the same spatial framework, we remove the edges between proteins and ligands to ensure proper alignment. For all five random splits, we use an effective batch size of 8, a learning rate of 1e-4 for ADiT-S and 3e-5 for ADiT-M and ADiT-L, a dropout rate of 0.2 and train for 120 epochs without rotation augmentation. To handle long proteins, we randomly truncate them to a max length of 150. The training process can be efficiently handled using a 4 A100 GPU (for ADiT-L).

E.4 Protein-Protein Binding Affinity Prediction

Baselines We employ a diverse set of baselines for comparison, including ESM-IF (Hsu et al., 2022), self-attention-based End-to-End network (End-to-End) (Jumper et al., 2021), DDGPred (Shan et al., 2022), MIF-Network (Yang et al., 2020), RDE-Network (Luo et al., 2023), DiffAffinity (Liu et al., 2024), Prompt-DDG (Wu et al., 2024), and Surface-VQMAE (Wu & Li, 2024). The reported results are sourced from (Wu et al., 2024; Wu & Li, 2024).

Training We truncate sequences based on the distance between the current residue and mutation-related residues, limiting the remaining sequence length to 50 residues. Training is conducted with an effective batch size of 8, a learning rate of 3e-5, a dropout rate of 0.0, and a total of 60 epochs. The training process can be efficiently handled using a single A100 GPU.

E.5 Antibody-Antigen Binding Affinity Prediction

Baselines We compare ADiT against several strong baselines for antibody affinity prediction, including Bind-ddG (Shan et al., 2022), and GearBind (Cai et al., 2024). We also include the physics-based tool FoldX (Delgado et al., 2019) as an additional baseline. The performance of most baselines is reported from (Cai et al., 2024).

Case study Affinity maturation aims to enhance the binding affinity of an antibody to its target antigen. As the search space for antibodies is too vast for wet-lab enumeration, *in silico* methods are increasingly being used to increase success rates. We can formulate the *in silico* affinity maturation task as a virtual screening problem, where we use a model to rank mutants of the starting antibody by their binding affinity to the target antigen, and validate the top-ranked mutants in wet lab. The two antibodies Anti-5T4 UdAb and CR3022 were chosen in (Cai et al., 2024) because (1) they cover two common antibody formats, nanobody and antibody; (2) they cover two antigens: UdAb targets an oncofetal antigen 5T4, while CR3022 targets SARS-CoV-1 and multiple variants of SARS-CoV-2. The 780 mutations for Anti-5T4 UdAb are obtained from saturation single-point mutagenesis of residues 1, 3, 25, 27-33, 39-45, 52-57, 59, 91-93, 95, 99-103, 105, 110, 112, 115-117 (*i.e.* mutating one of these sites to every possible amino acid type; each site yields 19 mutants), which are residues close to the antibody interface to 5T4. The 1420 mutations for CR3022 are obtained from saturation single-point mutagenesis of residues H25-35, H50-66, H99-108, L24-40, L56-62, L95-103.

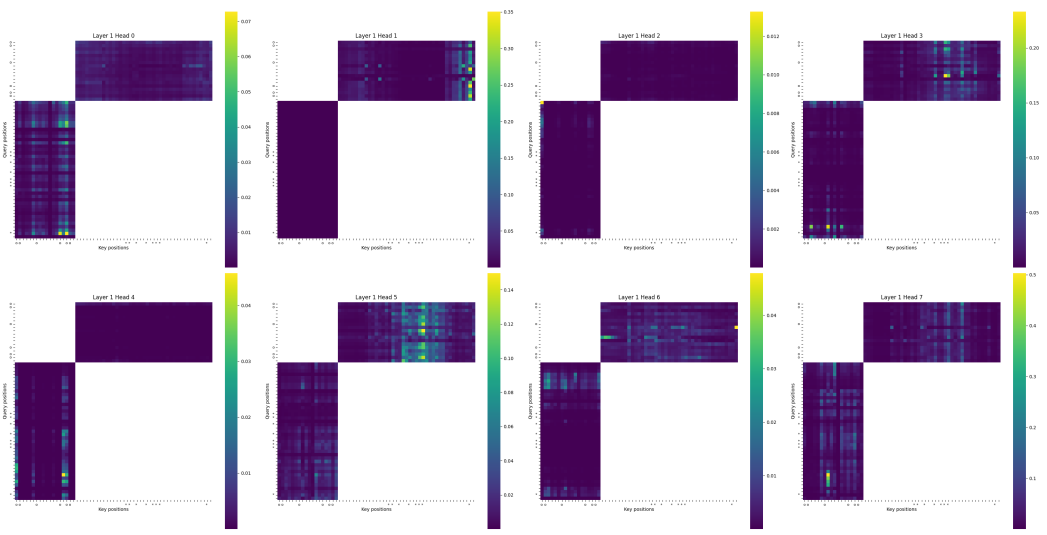


Figure 5: Attention score visualization for the `5f63` pocket-ligand complex, comprising 18 residue tokens and 41 ligand atom tokens. The plot highlights residue-atom interaction patterns. Here, o on the x- or y-axis denotes the closest one-third of residue tokens to the ligand, and $*$ on the x- or y-axis denotes the closest one-fifth of ligand atom tokens to the protein.

F Enzyme Commission (EC) Number Annotation Task

To show ADiT's generalization and applicability to non-binding tasks, we run additional experiments on the Enzyme Commission (EC) number annotation task (Gligorijević et al., 2021), which involves predicting whether a protein can catalyze a chemical reaction. This task is evaluated using the Fmax metric. The results, presented in Table 7, show that our method consistently outperforms the baseline, even under challenging data splits, demonstrating ADiT's applicability beyond binding-related functions.

Table 7: Enzyme Commission (EC) Number Annotation Task.

Method	30%	40%	50%	70%
DeepFRI	0.470	0.505	0.545	0.600
GearNet-Edge (Multiview Contrast)	0.744	0.769	**0.808**	**0.848**
ADiT-S	**0.751**	**0.773**	**0.806**	**0.843**

G Interpretability

Interpretability is a crucial next step toward understanding the mechanisms underlying our model's effectiveness. In this work, due to project scope, we primarily focus on establishing performance and scalability. A comprehensive and systematic interpretability investigation is left for future research. As a preliminary exploration, we conduct a simple case study examining the attention patterns learned by our trained model. We randomly select the `5f63` pocket-ligand complex from the LBA60 test set, which consists of 18 residues in the binding pocket and 41 atoms in the ligand. In this analysis, we extract attention scores from the token-level transformer, since the atom encoder and decoder employ only local block attention.

Several observations arise from this initial investigation:

First, we observe that the average attention weight for intra-chain pairs is **0.0253**, compared to only **0.0056** for inter-chain pairs. This suggests that, although the model captures cross-chain interactions,

it places greater emphasis on intra-chain relationships, which may play a more critical role in binding context representation.

Second, the model's attention mechanism appears to effectively "focus" on key interfacing residues and atoms. We visualize the cross-chain attention scores from the second layer (layer index=1, zero-based) in Figure5. To identify potentially important residues and atoms, we highlight the closest one-third of residue tokens to the ligand and the closest one-fifth of ligand atom tokens to the protein. As shown in the results for Head 0 and Head 4, these highlighted residues exhibit higher attention weights (columns appear brighter). Similarly, the highlighted ligand atoms receive higher attention in Head 1, Head 3, and Head 5.