M^{2}N: A Progressive Macro-to-Micro 3D Modeling Scheme for Unveiling Drug-Target Affinity

Tianxu Lv[1], Jie Zhu[1], Jinyi Liu[1], Shiyun Nie[1], Hongnian Tian[1], Yang Xiao[2], Yuan Liu[1*], Lihua Li[3*], Xiang Pan[1,4*]

[1] School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi 214122, China
[2] Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen 518071, China
[3] Institute of Biomedical Engineering and Instrumentation, Hangzhou Dianzi University, Hangzhou 310018, China
[4] The PRC Ministry of Education Engineering Research Center of Intelligent Technology for Healthcare, Wuxi, Jiangsu 214122, China
lyuan1800@jiangnan.edu.cn, lilh@hdu.edu.cn, xiangpan@jiangnan.edu.cn

Abstract

Accurate drug-target affinity (DTA) prediction holds significant potential in the field of artificial intelligence (AI)-based drug discovery. However, existing methods primarily operate at a *single scale*, specifically at the *macro* (residue) scale for target proteins and the *micro* (atom) scale for drugs, which limits their ability to provide information at *micro* (atom) scale for targets and *macro* (functional group, FG) scale for drugs. This limitation hinders a comprehensive understanding of the binding patterns and properties of drug-target pairs. In this paper, we propose a progressive **M**acro-**to**-**M**icro 3D Modeling **N**etwork (M²N) that enables *macro* (residue/FG) to *micro* (atom) scale unified modeling, termed *cross-scale*, to predict DTA. Specifically, M²N operates drugs by learning their chemical properties and structural characteristics from a 3D FG graph to a 3D atom graph. Correspondingly, M²N encodes proteins from a 3D residue graph to a 3D atom graph to exploit their sequence, evolutionary, and geometric representations. Such *cross-scale* 3D modeling scheme allows for *coarse-to-fine* embedding optimization, followed by an adaptive fusion module to dynamically integrate the refined features by end-to-end learning. Extensive experiments on two datasets indicate that M²N not only outperforms state-of-the-art methods under various conditions, but also provides a new paradigm for target and drug unified modeling.

Introduction

Predicting drug-target affinity (DTA) is vital in drug discovery, as it is essential to identify potential drug candidates and understand their interactions with target proteins (Li et al. 2022). Traditional biochemical assays are effective but time-consuming and costly. Docking software offers a computational alternative but requires pockets and structures (Pinzi and Rastelli 2019). Thus, it is necessary to develop novel *in silico* methods for efficient drug lead screening.

Recently, various deep learning-based approaches have been proposed for DTA prediction (Zhang et al. 2023), including 1D sequence-based methods (Öztürk, Özgür, and Ozkirimli 2018; Öztürk, Ozkirimli, and Özgür 2019), 2D structure-based methods (Nguyen et al. 2021; Jiang et al. 2020; Ma et al. 2023; Bi et al. 2023), and hybrid methods (Yuan, Chen, and Chen 2022). Sequence-based methods focus on the sequential nature of protein amino acid sequences and drug SMILES strings, leveraging techniques from natural language processing (NLP) for DTA prediction. In comparison, 2D structure-based methods extend the analysis to the topological and geometrical aspects of drug molecules and proteins. Drug molecular graphs are typically constructed from SMILES strings with atoms as nodes and bonds as edges, while protein graphs are built based on contact maps. Thus, graph neural networks (GNNs) are often employed to learn structural information from these graph representations. Hybrid methods aim to integrate both sequence and structure features, combining the strengths of the two approaches to enhance DTA prediction performance.

However, the aforementioned methods are primarily engaged in *single-scale* modeling of proteins and molecules. For proteins, they often operate at the *macro* (residue) *scale*, which cannot provide high-resolution information at the *micro* (atom) *scale*. Recent studies have demonstrated the potential of multi-scale protein modeling (Zheng et al. 2024), which offers a comprehensive understanding of protein structures and functions. For drug molecules, current models typically focus on *micro* (atom) *scale* modeling, neglecting the importance of *macro* (functinal groups, FGs) *scale*. FGs are specific groups of atoms within a molecule that determine the characteristic chemical reactions of the molecule (Liang, Neumann, and Ritter 2013). The presence and arrangement of FGs in a drug can influence its interactions with targets, affecting the drug's efficacy and pharmacokinetics (Lipinski 2004). Several works have shown the effectiveness of FGs in molecule design (Wu et al. 2024a). Thus, a core question for DTA prediction is how to optimally exploit *macro-to-micro scale* representations of targets-drugs.

Based on the above considerations, we propose a novel progressive **M**acro-**to**-**M**icro 3D modeling **N**etwork (M^{2}N), designed to refine the representations of protein/drug from residue/FG-scale to atom-scale for accurate DTA prediction. First, M^{2}N constructs a 3D drug FG-based graph and utilizes a DFG-GTformer to assimilate structural information

[*] Corresponding authors.

at the FG level. This information is subsequently incorporated into a 3D drug atom graph to capture grained atom-level embeddings using DAtom-GTformer. Second, M^2N builds a 3D protein residue-level graph and applies a TR-GTformer to integrate its sequence, evolutionary, and spatial information, which is then incorporated into a 3D protein atom graph to capture detailed embeddings at the atom level using TAtom-GTformer. Afterwards, M^2N designs an adaptive fusion module (AFM) to dynamically integrate the refined features by end-to-end learning. With such designs, M^2N can harmonize the macro-scale and micro-scale representations, ensuring a comprehensive understanding of the molecular interactions.

Specific contributions of this work are as follows:

- This is the first study to comprehensively exploit progressive 3D cross-scale modeling of drug-targets (*i.e.*, from residue/FG-scale to atom-scale) to predict DTA.
- A novel and effective M^2N model is proposed to integrate multi-level and mutli-scale representations in a coarse-to-fine manner for DTA prediction.
- Extensive experiments on two DTA datasets demonstrate the viability, effectiveness, and superiority of our M^2N.

Related Work

Protein Modeling for DTA Prediction

Protein modeling methods can be classified into three categories: sequence-based, structure-based, and hybrid approaches. Sequence-based methods often treat protein sequences as sentences, where each word corresponds to an individual amino acid residue (Öztürk, Özgür, and Ozkirimli 2018; Wang et al. 2021). In this way, techniques from the field of NLP, such as CNNs (Öztürk, Özgür, and Ozkirimli 2018), RNNs (Mukherjee, Ghosh, and Basuchowdhuri 2022), and Transformers (Monteiro, Oliveira, and Arrais 2022), have proven effective in capturing the contextual semantics of protein sequences for DTA prediction. However, given the complex nature of proteins, structure-based approaches have emerged as powerful tools for extracting high-quality protein representations (Jiang et al. 2020; Ma et al. 2023). These methods typically represent the protein structure as a graph, using GNNs to exploit its spatial information (Bi et al. 2023). Moreover, a handful of hybrid methods combine sequence and structure features to improve DTA performance (Yuan, Chen, and Chen 2022). However, these methods only focus on single-scale protein modeling (*i.e.*, residue-scale), failing to exploit inherent relationships of cross-scale modeling (from residue-scale to atom-scale), which provides a deeper insights into the connections between protein functions and structures.

Drug Modeling for DTA Prediction

Drug modeling methods primarily rely on the utilization of SMILES strings. Some methods treat SMILES as linear 1D sequences, harnessing the contextual information of atoms and bonds within a molecule (Öztürk, Özgür, and Ozkirimli 2018). Alternatively, SMILES can be converted into 2D molecular graphs with atoms as nodes and bonds as edges. This graph-based representation highlights the topological structure of the molecule, and GNNs are often applied to process these 2D structural representations (Nguyen et al. 2021). In addition, several methods consider 3D molecule structures for feature extraction (Luo, Liu, and Peng 2023). However, these methods overlook the modeling for FGs, which has an important impact on the drug's interactions with targets. Thus, it is crucial to integrate the role of FGs into molecular modeling for DTA prediction.

Proposed Method

Problem Formulation Given a collection of targets $T = \{t_0, t_1, ..., t_n\}$ and drugs $D = \{d_0, d_1, ..., d_m\}$, the binding affinity between drugs and targets can be described by a matrix $Z = \{z_{ij}, i = 0, 1, ..., n, j = 0, 1, ..., m\}$, where z_{ij} represents the binding affinity between target t_i and drug d_j. The DTA prediction aims to learn a function $f : (t_i, d_j) \rightarrow \hat{z}_{ij}$ from the training set Z_{train}, so that for any drug-target pair $(t_i, d_j) \in Z_{test}$, $\hat{z}_{ij}$ is the estimated binding affinity.

Overview The framework of M^2N is displayed in Figure 1a, which consists of three main functional parts: macro-to-micro drug modeling module (M^2D), macro-to-micro target modeling module (M^2T), and adaptive fusion module (AFM). M^2D utilizes a cascaded graph transformer to learn and refine drug embeddings from the FG scale to the atom scale, while target representations are captured and optimized in M^2T from the target scale to the atom scale. AFM performs adaptive fusion to exploit the potential complementarity and relationships among the refined embeddings.

Macro-to-Micro Drug Modeling

Construction and Learning of Drug Functional Group Graphs in Macro-scale. We present a methodology for constructing drug FG graphs derived from molecular structures, wherein nodes represent FGs and edges denote the spatial proximity between these groups. First, we define a catalogue of FGs alongside their corresponding SMARTS (SMiles ARbitrary Target Specification) (Arús-Pous et al. 2020) patterns. FGs dictate the reactivity of a molecule, which plays a pivotal role in drug-target binding and interactions. Given a drug molecule d, the collection of FGs is:

$$d = \bigcup_{i=1}^{p} \mathrm{fg}_i \tag{1}$$

where p is the number of FGs in the drug molecule, and fg_i represents the i-th FG. Subsequently, to construct the drug FG graph, we preprocess the input molecule by neutralizing its charged atoms and extract its FGs using the defined SMARTS patterns. For each FG, we calcluate its centroid, defined as the average position of its constituent atoms. These centroids form the foundation for calculating the spatial proximity between FGs. We then compute the pairwise Euclidean distances between the centroids of all FGs as $d_{ij} = \|c_i - c_j\|_2$ where c_i and c_j are the centroids of fg_i and fg_j, respectively, and d_{ij} represents the Euclidean distance between them. To establish the edges of the drug FG graph, we select the top k (k_1) shortest distances and

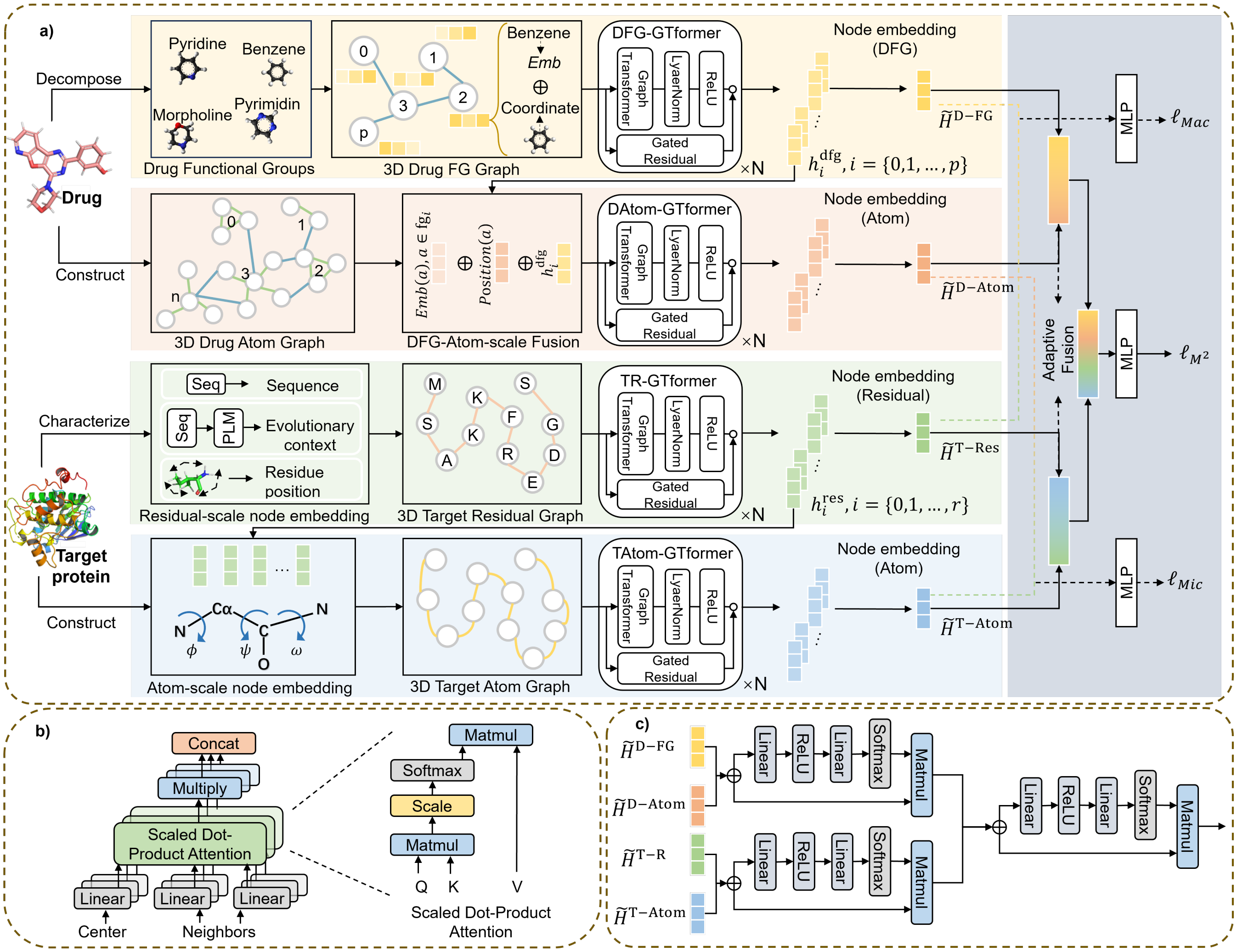

Figure 1: **a)** The overall framework of M^{2}N, composed of a M^{2}D, a M^{2}T, and an AFM. The M^{2}D exploits and refines drug representations from a FG-scale graph to an atom-scale graph, while the M^{2}T learns protein features from a residue-scale graph to an atom-scale graph. Subsequently, the refined cross-scale embeddings are fused in the AFM to exploit the potential complementarity and relationships. **b)** The details of the graph transformer block. **c)** The feature fusion strategy of AFM.

generate an edge between the associated FGs. As thus, the FG graph is defined as $G_{\mathrm{DFG}} = (V_{\mathrm{DFG}}, E_{\mathrm{DFG}})$, where each node $v_i \in V_{\mathrm{DFG}}$ corresponds to an embedding of fg_i, and each edge e_{ij} connects FGs based on their spatial proximity.

To learn high-order representations from the constructed FG graphs, we propose a DFG-GTformer leveraging the power of graph transformer (GT) (Shi et al. 2021), as shown in Figure 1b. Given node features $H^l = \{h_1^l, h_2^l, ..., h_p^l\}$, the multi-head attention for each edge from j to i is calculated:

$$q_i^l = W_q^l h_i^l + b_q^l, \quad k_j^l = W_k^l h_j^l + b_k^l, \quad e_{ij} = W_e e_{ij} + b_e \tag{2}$$

$$\alpha_{ij}^l = \frac{\langle q_i^l, k_j^l + e_{ij}\rangle}{\sum_{u\in\Omega(i)}\langle q_i^l, k_u^l + e_{iu}\rangle}, \quad \langle q, k\rangle = \exp\left(\frac{q^T k}{\sqrt{d}}\right) \tag{3}$$

where W_q^l, W_k^l, and W_e are learnable parameters, b_q^l, b_k^l, and b_e are bias terms, q_i^l and k_i^l are the query and key vectors for i-th and j-th nodes, e_{ij} represents the edge features, $\Omega(i)$ represents a set of neighbors of node i, and α_{ij}^l is the attention weight. Afterwards, we perform the message aggregation from the neighbor j to the source i:

$$\hat{h}_i^{l+1} = \sum_{j\in\Omega(i)} \alpha_{ij}^l (v_j^l + e_{ij}), \qquad v_j^l = W_v^l h_j^l + b_v^l \tag{4}$$

where W_v^l and b_v^l are the learnable parameter and bias term, respectively. Then a gated residual connection (Chen et al. 2020) is added to prevent the GNN from oversmoothing:

$$r_i^l = W_r^l h_i^l + b_r^l, \quad \beta_i^l = \mathrm{sigmoid}(W_g^l[\hat{h}_i^{l+1}; r_i^l; \hat{h}_i^{l+1} - r_i^l]) \tag{5}$$

$$h_i^{l+1} = \mathrm{ReLU}(\mathrm{LayerNorm}((1 - \beta_i^l)\hat{h}_i^{l+1} + \beta_i^l r_i^l)) \tag{6}$$

After propagating messages in the G_{DFG} through multiple GT layers, the updated node features $H^{\mathrm{DFG}} = \{h_1^{\mathrm{dfg}}, h_2^{\mathrm{dfg}}, ..., h_p^{\mathrm{dfg}}\}$ that capture the relations between FGs and their spatial proximity are acquired. Then we conduct feature aggregation to obtain the FG-scale drug embedding:

$$\widetilde{H}^{\mathrm{D-FG}} = \frac{1}{p}\sum_{i=1}^{p} h_i^{\mathrm{dfg}} \tag{7}$$

where $\widetilde{H}^{\mathrm{D-FG}}$ encapsulates the spatial arrangement, reactivity, and interactions of the drug's FGs.

Knowledge Transfer from Macro-scale to Micro-scale within Drug Atom Graphs. In order to investigate the fine-grained relationships between the atoms within the molecule, we construct the drug atom graph $G_{\mathrm{DA}} = (V_{\mathrm{DA}}, E_{\mathrm{DA}})$, where each node represents an atom and the edges represent the bonds between atoms. Specifically, each node is enriched with a feature set that encapsulates diverse information:

- Atomic properties. These include the atom symbol, the number of adjacent atoms, the number of adjacent hydrogens, the implicit value of the atom, and whether the atom is in an aromatic structure.
- Atomic spatial positions. The 3D coordinates of the atom in the molecular structure provide information about the spatial arrangement of the atoms.
- Corresponding FG features. The associated FG features H^{DFG} acquired from the FG-graph are integrated into the atom features, providing a more comprehensive representation of the atom's chemical environment.

Formally, given an atom a, the DFG-Atom-scale hybrid fusion process can be defined as:

$$h_a = \mathrm{concat}(Emb(a); Position(a); h_i^{\mathrm{dfg}}), a \in \mathrm{fg}_i \tag{8}$$

where $Emb(a)$, $Position(a)$, and h_i^{dfg} represent the atomic properties, spatial positions, and FG features, respectively. To capture atom-scale interactions within the molecule, we employ the DAtom-Transformer to process this graph, acquiring the atom-scale drug embedding $\widetilde{H}^{\mathrm{D-Atom}}$.

Macro-to-Micro Target Modeling

Inspired by recent advances that encode proteins from residue-scale to atom-scale (Zheng et al. 2024; Abramson et al. 2024), we propose a cross-scale target modeling approach to capture the complex hierarchical structure of proteins. This approach aims to integrate information from different scales, capturing both the primary sequence information and the three-dimensional spatial arrangement of residues and atoms within the protein.

Incorporating Sequence, Evolutionary, and Spatial Features into Macro-scale Target Residue Graphs. To exploit the intricate residue-scale interactions within the protein, a residue graph $G_{\mathrm{T-Res}}$ is constructed where each node represents an amino acid and the edges denote the spatial proximity between residues. To comprehensively characterize residues, we develop three types of features:

- Sequence features. One-hot encoding to represent the type of amino acid of the residue.
- Evolutionary features. We utilize ESM (Lin et al. 2023), a protein language model (PLM) trained on 250 million sequences, to create embeddings for each residue. The ESM embeddings are known to capture structural, functional, and evolutionary characteristics of the protein.
- Spatial features. These features are represented as the central coordinates of the backbone of each residue

Then we establish edges in the residue graph $G_{\mathrm{T-Res}}$ based on the spatial proximity between residues. In particular, for each residue, we identify its k (k_2) nearest neighbors based on the central coordinates of the residues' backbones, which captures the local spatial context of each residue. Once the protein residue graph is constructed, we employ a TRes-GTformer to model the complex interactions and dependencies among residues. The output of the TRes-GTformer is an updated set of residue embeddings that encapsulate the collective influence of sequence, evolutionary, and spatial features, as well as the inter-residue spatial relationships: $h_i^{\mathrm{res}} = \mathrm{TRes} - \mathrm{GTformer}(G_{\mathrm{T-Res}}), i = 1, 2, ..., r$, where r is the number of residues. To obtain the residue-scale protein representation, the feature aggregation is performed:

$$\widetilde{H}^{\mathrm{T-Res}} = \frac{1}{r}\sum_{i=1}^{r} h_i^{\mathrm{res}} \tag{9}$$

Representation Refinement from Macro to Micro scale within Target Atom Graphs. Building upon the residue graph, we advance to the construction of the protein atom graph $G_{\mathrm{T-Atom}}$ to further exploit the detailed atomic-level interactions. Each node's feature consist of two distinct components, *i.e.*, learned residue-scale representations and atom-scale geometric features. Specifically, we calculate three dihedral angles (ϕ, ψ, ω) following (Wu et al. 2024b) for each relevant atom based on the backbone coordinates, which are pivotal for understanding the local conformation of the protein's 3D structure. The geometric features are encoded using a vector of cosine and sine values:

$$h_{\mathrm{geometric}} = (\mathrm{sin}\phi, \mathrm{sin}\psi, \mathrm{sin}\omega, \mathrm{cos}\phi, \mathrm{cos}\psi, \mathrm{cos}\omega) \tag{10}$$

We further build edge features that characterize the spatial relationships based on (Ingraham et al. 2019). Specifically, we calculate an orientation matrix $\mathbf{O}_i$ that establishes a local coordinate system:

$$\mathbf{O}_i = [\mathbf{b}_i, \mathbf{n}_i, \mathbf{b}_i \times \mathbf{n}_i] \tag{11}$$

$$\mathbf{u}_i = \frac{\mathbf{c}_i - \mathbf{c}_{i-1}}{\|\mathbf{c}_i - \mathbf{c}_{i-1}\|}, \mathbf{b}_i = \frac{\mathbf{u}_i - \mathbf{u}_{i+1}}{\|\mathbf{u}_i - \mathbf{u}_{i+1}\|}, \mathbf{n}_i = \frac{\mathbf{u}_i \times \mathbf{u}_{i+1}}{\|\mathbf{u}_i \times \mathbf{u}_{i+1}\|} \tag{12}$$

where $\mathbf{c}_i \in \mathbb{R}^3$ is the central coordinates of atoms of i-th residue, $\mathbf{b}_i$ is the negative bisector of the angle, and $\mathbf{n}_i$ is a unit vector perpendicular to the plane. Then we calculate spatial edge features $\mathbf{e}_{ij}$ that reflect the distance, direction, and orientation information as follows:

$$\mathbf{e}_{ij} = \Big(\mathbf{rbf}(\|\mathbf{c}_j - \mathbf{c}_i\|), \mathbf{O}_i^T \tfrac{\mathbf{c}_j - \mathbf{c}_i}{\|\mathbf{c}_j - \mathbf{c}_i\|}, \mathbf{q}(\mathbf{O}_i^T \mathbf{O}_j)\Big) \tag{13}$$

where $\mathbf{rbf}$ represents a radial basis function encoding the Euclidean distance, the second denotes the direction of $\mathbf{c}_j$ relative to the local frame of $\mathbf{c}_i$, and the third is a quaternion encoding of the orientation difference between the two local frames. Such design allows for a detailed characterization of the local atomic environment within the protein structure, providing a rich set of features for protein-related works, such as DTA prediction.

To understand the intricate details of protein conformation and function, we employ a TAtom-GTformer to extract knowledge from the atom-scale interactions and structural features within protein structures. Through a series of GT layers, the model effectively captures the complex dependencies and interactions among atoms. After the feature aggregation, which integrates information from local neighborhoods into a unified representation for each atom, we acquire the atom-scale protein embedding $\widetilde{H}^{\mathrm{T-Atom}}$.

Adaptive Fusion Module

For each drug-target pair (t, d), we have derived their cross-scale features $\widetilde{H}^{\mathrm{D-FG}}$, $\widetilde{H}^{\mathrm{D-Atom}}$, $\widetilde{H}^{\mathrm{T-Res}}$, and $\widetilde{H}^{\mathrm{T-Atom}}$. Then we feed them into the AFM to harmoniously align and combine the diverse chemical and spatial information encapsulated within each feature set, thereby providing a comprehensive molecular signature, as shown in Figure 1c. Specifically, we first align and refine the concatenated cross-scale drug features $\widetilde{H}^D$ and protein features $\widetilde{H}^T$ using a shared refinement scheme:

$$H^x = \alpha_x \cdot \widetilde{H}^x, \alpha_x = \varphi(\mathrm{ReLU}(W^x \widetilde{H}^x + b^x)), x \in D, T \tag{14}$$

where φ represents the Softmax operation, $\cdot$ denotes the element-wise multiplication and α_x is the weight matrix operated on the cross-scale features $\widetilde{H}^x$. Then we concatenate the refined drug and protein representations into $\widetilde{H}$ and apply another layer of adaptive weighting:

$$H = \alpha_3 \cdot \widetilde{H}, \alpha_3 = \varphi(\mathrm{ReLU}(W^{DT} \widetilde{H} + b^{DT})) \tag{15}$$

where H is a sophisticated integration of the drug and protein features, capturing the complex interplay between molecular structure and chemical properties at multi-scale.

Then we feed the integrated features into a multi-layer perceptron (MLP) to generate the final predictions for DTA. Additionally, we implement auxiliary prediction schemes for macro-scale (FG/Residue-scale) and micro-scale (Atom-scale) features, each processed through separate MLP branches to predict DTA at their respective scales. Based on the multi-scale predictions, the training objective of our M^{2}N is to minimize the loss function:

$$\ell = \ell_{M^2} + \lambda_1 \ell_{Mac} + \lambda_2 \ell_{Mic} \tag{16}$$

where each of the loss components (ℓ_{M^2}, ℓ_{Mac}, ℓ_{Mic}) is computed using the mean squared error (MSE) loss, λ_1 and λ_2 are hyper-parameters that balance the contributions.

Experiments

Datasets We conduct a comprehensive set of experiments on two benchmark datasets, namely DAVIS(Davis et al. 2011) and KIBA(Tang et al. 2014). Since the original datasets only include the SMILES strings of drugs and target sequences, we gather the corresponding 3D structural data in this study. The processed DAVIS dataset comprises 64 unique drugs and 226 unique targets, with 14,464 kinase dissociation constant K_d values serving as drug-target affinities. In line with (Nguyen et al. 2021), the values of K_d are converted to the logarithmic scale as $pK_d = -log_{10}(K_d/10^9)$, falling within the range of 5.0 to 10.8. The refined KIBA dataset includes 1,986 unique drugs and 160 unique proteins, along with 89,958 KIBA scores as affinities, which span a range from 0.0 to 17.2 and are derived based on the integration of kinase inhibitor bioactivities.

Baselines To validate the efficacy of our M^{2}N, we evaluate it against five state-of-the-art sequence-based techniques: **GraphDTA** (Nguyen et al. 2021), **MolTrans** (Huang et al. 2021), **DeepGLSTM** (Mukherjee, Ghosh, and Basuchowdhuri 2022), **FusionDTA** (Yuan, Chen, and Chen 2022), and **MFRDTA** (Hua et al. 2023). Additionally, we perform experiments with three structure-based techniques: **DGraphDTA** (Jiang et al. 2020), **MSFDTA** (Ma et al. 2023), and **HiSIFDTA** (Bi et al. 2023).

Implementation details The developed model is implemented utilizing Pytorch (Paszke et al. 2019) along with Pytorch Geometric (Fey and Lenssen 2019). The experiments in this study are executed on a platform comprising two NVIDIA GeForce RTX 4090 GPUs to accelerate the training process. All experiments use a five-fold cross-validation approach, and the mean scores of the results are presented as the final result. In addition, several metrics including mean squared error (MSE), Pearson's correlation coefficient (PCC) and concordance index (CI) are used to evaluate the prediction performance of the models. All hyperparameters of M^{2}N are listed in Table 1.

To comprehensively evaluate the generalization and robustness of the models, we consider three distinct and challenging scenarios as outlined below:

- S1: The target protein within the drug-target pair has not been encountered during the training phase.
- S2: The drug molecule within the drug-target pair has not been encountered during the training phase.
- S3: The target protein and the drug molecule are both novel, representing the most rigorous evaluation.

Hyperparameter	Value(s)
Epoch	200
Batch size	32
Optimizer	AdamW
Learning rate	0.0005
Weight decay	0.0001
Balancing coefficient λ_1	0.1
Balancing coefficient λ_2	0.2
Output dimension of graph embedding	256
Number of GT layers in each GTformer	3
Number of neighbours per node k_1 in G_{DFG}	5
Number of neighbours per node k_2 in $G_{\mathrm{T-Res}}$	25

Table 1: Hyperparameter settings of our M^{2}N model.

Dataset	Metric	GraphDTA	MolTrans	DeepGLSTM	FusionDTA	MFRDTA	DGraphDTA	MSFDTA	HiSIFDTA	M^2N
DAVIS	MSE (↓)	0.647	0.424	0.459	0.462	0.448	0.432	0.368	0.359	**0.316**
	PCC (↑)	0.427	0.468	0.521	0.548	0.558	0.561	0.604	0.608	**0.639**
	CI (↑)	0.752	0.761	0.771	0.788	0.795	0.798	0.801	0.803	**0.816**
KIBA	MSE (↓)	0.984	0.396	0.528	0.418	0.430	0.382	0.358	0.364	**0.331**
	PCC (↑)	0.269	0.606	0.579	0.569	0.558	0.617	0.638	0.632	**0.656**
	CI (↑)	0.600	0.720	0.687	0.669	0.795	0.729	0.735	0.731	**0.742**

Table 2: Comparison of M^2N and baselines for DTA prediction on two datasets under S1 setting.

Dataset	Metric	GraphDTA	MolTrans	DeepGLSTM	FusionDTA	MFRDTA	DGraphDTA	MSFDTA	HiSIFDTA	M^2N
DAVIS	MSE (↓)	0.812	0.925	0.808	0.784	0.799	0.825	0.791	0.792	**0.778**
	PCC (↑)	0.366	0.242	0.371	0.408	0.373	0.358	0.401	0.399	**0.419**
	CI (↑)	0.671	0.648	0.684	0.711	0.694	0.656	0.705	0.701	**0.721**
KIBA	MSE (↓)	0.892	0.528	0.507	0.464	0.716	0.458	0.432	**0.415**	0.421
	PCC (↑)	0.458	0.585	0.592	0.618	0.412	0.635	0.642	0.649	**0.668**
	CI (↑)	0.644	0.731	0.744	0.753	0.726	0.761	0.766	0.768	**0.773**

Table 3: Comparison of M^2N and baselines for DTA prediction on two datasets under S2 setting.

Comparison with Baselines

Tables 2-4 report the performances of the proposed method and recent state-of-the-art approaches on the DAVIS and KIBA datasets under three challenging scenarios. The best score in each row is in **bold** and the second best is underlined. Experimental results demonstrate that the proposed HeHiDTAG comprehensively outperforms other models on the DAVIS dataset when faced with unseen targets or drugs, which verifies the effectiveness and generalization of our framework. In three challenging scenarios, our method achieves PCC of 3.1% (S1), 1.1% (S2), and 8.7% (S3) gain over the best baseline models. This signifies that our M^2N is capable of capturing subtle yet critical patterns in drug-target interactions that other models may overlook. Besides, it can be observed that our M^2N also achieves promising performance on the KIBA dataset. In comparison with the results of GraphDTA, the mean CI under three experimental settings in increased by approximately 14.2% (S1), 12.9% (S2), and 5.9% (S3), respectively. These observations further verify the robustness of our model. Among all baselines, the performance of protein structure-based methods (*e.g.*, MSFDTA, HiSIFDTA) is better than that of protein sequence-based methods (*e.g.*, GraphDTA, MolTrans) in general. This indicates that simply extracting representations from target sequences is not sufficient to capture the inherent properties and the spatial arrangement and three-dimensional conformation of proteins play a crucial role in their affinity to drugs. In comparison, the proposed method not only leverages both sequence and structure features of proteins and drugs, but also exploits the multiscale representations in a coarse-to-fine manner and adaptively fuse them for end-to-end learning. This approach allows for a more comprehensive understanding of the drug-target interaction landscape, leading to improved predictive accuracy.

Ablation Study

Effectiveness of designed components To investigate the impact of each designed component in the proposed approach, we conduct an extensive ablation analysis by evaluating different M^2N variants as follows:

- **M^2N without the drug FG-scale representations (w/o DFG)**: This variant excludes the functional group-level information from the drug representation.
- **M^2N without the drug atom-scale representations (w/o DAtom)**: This variant removes the atom-scale information from the drug representation.
- **M^2N without the target residue-scale representations (w/o TRes)**: This configuration removes the residue-level features from the protein representation.
- **M^2N without the target atom-scale representations (w/o TAtom)**: The atom-scale details of the protein are omitted in this variant.
- **M^2N without the adaptive fusion module (w/o AFM)**: This variant replaces the AFM strategy with a direct concentration operation.

Table 5 reports the performance of M^2N and its five variants on DAVIS and KIBA datasets under S1 scenarios. It can be observed that all variants of M^2N produce the decreased performances, verifying that all components can contribute to DTA prediction. Besides, we have the following observations: (1) M^2N (w/o DAtom) outperforms the M^2N (w/o DFG), which demonstrates that integrating atom-scale details provide a more precise description of the molecular structure. M^2N (w/o TRes) has more decreased performance than M^2N (w/o TAtom), highlighting the importance of atomic-scale granularity in protein representation. (2) The comparison of M^2N and M^2N (w/o AFM) reveals the importance of the adaptive fusion module in harmonizing features from different scales. (3) M^2N outperforms M^2N (w/o DFG) and M^2N (w/o DAtom), which demonstrates the effectiveness of cross-scale modeling of drugs, including FG-scale and atom-scale information. (4) M^2N performs better than M^2N (w/o TRes) and M^2N (w/o TAtom), indicating the value of cross-scale modeling of proteins, including residue-scale and atom-scale information. The above observations could guide further optimization of the M^2N model, potentially leading to an even more refined approach to cross-scale feature integration.

Dataset	Metric	GraphDTA	MolTrans	DeepGLSTM	FusionDTA	MFRDTA	DGraphDTA	MSFDTA	HiSIFDTA	M^{2}N
DAVIS	MSE (↓)	0.877	0.946	0.891	0.734	0.751	0.742	0.708	0.692	**0.653**
	PCC (↑)	0.228	0.162	0.181	0.243	0.235	0.239	0.245	0.258	**0.345**
	CI (↑)	0.631	0.583	0.596	0.654	0.641	0.651	0.656	0.661	**0.708**
KIBA	MSE (↓)	1.126	0.598	0.582	0.582	0.701	0.565	0.551	0.558	**0.544**
	PCC (↑)	0.359	0.384	0.395	0.426	0.316	0.445	0.467	0.459	**0.492**
	CI (↑)	0.611	0.619	0.614	0.625	0.640	0.637	0.645	0.641	**0.670**

Table 4: Comparison of M^{2}N and baselines for DTA prediction on two datasets under S3 setting.

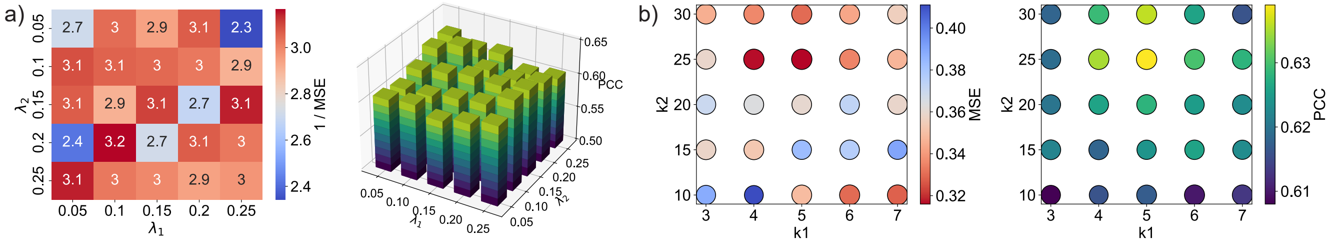

Figure 2: Analysis of hyper-parameter sensitivity concerning the number of neighbors (a) and the balance coefficients (b).

Methods	DAVIS			KIBA		
	MSE (↓)	PCC (↑)	CI (↑)	MSE (↓)	PCC (↑)	CI (↑)
M^{2}N	**0.316**	**0.639**	**0.816**	**0.331**	**0.656**	**0.742**
w/o DFG	0.342	0.615	0.808	0.354	0.626	0.733
w/o TRes	0.336	0.618	0.811	0.362	0.623	0.732
w/o DAtom	0.348	0.611	0.808	0.381	0.618	0.729
w/o TAtom	0.354	0.608	0.802	0.388	0.615	0.727
w/o AFM	0.325	0.621	0.813	0.349	0.637	0.737
w/o PDM	0.321	0.623	0.814	0.342	0.641	0.739
w/o PTM	0.327	0.620	0.814	0.346	0.639	0.738

Table 5: Ablation study of designed components in M^{2}N.

Effectiveness of progressive modeling We further evaluate the progressive modeling strategy of drugs and targets:

- **M^{2}N without Progressive Drug Modeling (w/o PDM)**: This variant constructs the drug atom graph without integrating the DFG features, treating the drug molecules solely based on their atom-scale information.
- **M^{2}N without Progressive Target Modeling (w/o PTM)**: Similarly, this variant constructs the target atom graph without the residue-scale feature integration, focusing only on the atomic representation of the protein.

As reported in Table 5, both M^{2}N (w/o PDM) and M^{2}N (w/o PTM) exhibit reduced performance compared to the full M^{2}N model in S1 setting. This highlights the importance of the hierarchical feature integration process, which aligns with the multi-scale nature of biological systems. By mimicking this natural hierarchy, M^{2}N is able to capture a more comprehensive representation of drug-target interactions.

Hyper-parameter Sensitivity Analysis

Effect of coefficients λ_1 and λ_2 We adjust the coefficients λ_1 and λ_2 in Eq.(16) to investigate the contribution of various auxiliary tasks. Specifically, the values of λ_1 and λ_2 are searched within the range of $\{0.05, 0.10, 0.15, 0.20, 0.25\}$, and M^{2}N attains optimal performance when $\lambda_1 = 0.10$ and $\lambda_2 = 0.20$. The results are summarized in Figure 2a. The key findings are: (1) Higher values of λ_1 and λ_2 result in poorer model performance, as a larger attention on the auxiliary task causes the model to overly focus on it during training. (2) When M^{2}N achieves peak performance, λ_2 exceeds λ_1, indicating that detailed atom-scale information might be more crucial for multi-task learning in the DAVIS dataset.

Effect of the number of neighbours k_1 and k_2 We vary the hyper-parameters k_1 and k_2 to assess the impact of the number of neighbours in G_{DFG} and $G_{\text{T-Res}}$ for DTA prediction. Considering the number of nodes in two graphs, the k_1 is searched in the range of $\{3, 4, 5, 6, 7\}$ and k_2 is explored in the range of $\{10, 15, 20, 25, 20\}$. As shown in Figure 2b, the proposed M^{2}N achieves the best performance when $k_1 = 5$ and $k_2 = 25$ and the following observations are obtained: (1) Higher values of k_1 and k_2 lead to the degradation of model performance due to the oversmoothing phenomenon, where the model representations become excessively generalized, losing the fine-grained details necessary for accurate prediction. (2) Smaller values k_1 and k_2 also have a negative impact on predictive performance. This may be due to an insufficient aggregation of neighbor information, which hinders the model's ability to capture the full context of the molecular structures.

Conclusion

In this work, we propose a progressive macro-to-micro 3D modeling network (M^{2}N) for DTA prediction. By transitioning from functional group to atom-level representations for drugs and from residue to atom-level for proteins, M^{2}N captures cross-scale information crucial for DTA prediction. This cross-scale 3D modeling approach facilitates a coarse-to-fine embedding refinement, followed by an adaptive fusion component to combine the extracted features. Comprehensive experiments demonstrate that M^{2}N not only surpasses the performance of recent leading methods under various scenarios, but introduces a novel perspective for unified modeling of proteins and molecules in drug discovery.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China under grants W2411054, U21A20521 and 62271178, the Postgraduate Research & Practice Innovation Program of Jiangsu Province KYCX23_2524, National Foreign Expert Project of China under Grant G2023144009L, Zhejiang Provincial Natural Science Foundation of China (LR23F010002), Wuxi Health Commission Precision Medicine Project (J202106), Jiangsu Provincial Six Talent Peaks Project (YY-124), and the construction project of Shanghai Key Laboratory of Molecular Imaging (18DZ2260400).

References

Abramson, J.; Adler, J.; Dunger, J.; Evans, R.; Green, T.; Pritzel, A.; Ronneberger, O.; Willmore, L.; Ballard, A. J.; Bambrick, J.; et al. 2024. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 1–3.

Arús-Pous, J.; Patronov, A.; Bjerrum, E. J.; Tyrchan, C.; Reymond, J.-L.; Chen, H.; and Engkvist, O. 2020. SMILES-based deep generative scaffold decorator for de-novo drug design. *Journal of cheminformatics*, 12: 1–18.

Bi, X.; Zhang, S.; Ma, W.; Jiang, H.; and Wei, Z. 2023. HiSIF-DTA: A Hierarchical Semantic Information Fusion Framework for Drug-Target Affinity Prediction. *IEEE Journal of Biomedical and Health Informatics*.

Chen, M.; Wei, Z.; Huang, Z.; Ding, B.; and Li, Y. 2020. Simple and deep graph convolutional networks. In *International conference on machine learning*, 1725–1735. PMLR.

Davis, M. I.; Hunt, J. P.; Herrgard, S.; Ciceri, P.; Wodicka, L. M.; Pallares, G.; Hocker, M.; Treiber, D. K.; and Zarrinkar, P. P. 2011. Comprehensive analysis of kinase inhibitor selectivity. *Nature biotechnology*, 29(11): 1046–1051.

Fey, M.; and Lenssen, J. E. 2019. Fast graph representation learning with PyTorch Geometric. In *International Conference on Learning Representations (RLGM Workshop)*. ICLR.

Hua, Y.; Song, X.; Feng, Z.; and Wu, X. 2023. MFR-DTA: a multi-functional and robust model for predicting drug–target binding affinity and region. *Bioinformatics*, 39(2): btad056.

Huang, K.; Xiao, C.; Glass, L. M.; and Sun, J. 2021. MolTrans: molecular interaction transformer for drug–target interaction prediction. *Bioinformatics*, 37(6): 830–836.

Ingraham, J.; Garg, V.; Barzilay, R.; and Jaakkola, T. 2019. Generative models for graph-based protein design. *Advances in neural information processing systems*, 32.

Jiang, M.; Li, Z.; Zhang, S.; Wang, S.; Wang, X.; Yuan, Q.; and Wei, Z. 2020. Drug–target affinity prediction using graph neural network and contact maps. *RSC advances*, 10(35): 20701–20712.

Li, Z.; Jiang, M.; Wang, S.; and Zhang, S. 2022. Deep learning methods for molecular representation and property prediction. *Drug Discovery Today*, 27(12): 103373.

Liang, T.; Neumann, C. N.; and Ritter, T. 2013. Introduction of fluorine and fluorine-containing functional groups. *Angewandte Chemie International Edition*, 52(32): 8214–8264.

Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; Smetanin, N.; Verkuil, R.; Kabeli, O.; Shmueli, Y.; et al. 2023. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637): 1123–1130.

Lipinski, C. A. 2004. Lead-and drug-like compounds: the rule-of-five revolution. *Drug discovery today: Technologies*, 1(4): 337–341.

Luo, Y.; Liu, Y.; and Peng, J. 2023. Calibrated geometric deep learning improves kinase–drug binding predictions. *Nature Machine Intelligence*, 1–12.

Ma, W.; Zhang, S.; Li, Z.; Jiang, M.; Wang, S.; Guo, N.; Li, Y.; Bi, X.; Jiang, H.; and Wei, Z. 2023. Predicting Drug-Target Affinity by Learning Protein Knowledge From Biological Networks. *IEEE Journal of Biomedical and Health Informatics*, 27(4): 2128–2137.

Monteiro, N. R.; Oliveira, J. L.; and Arrais, J. P. 2022. DTITR: End-to-end drug–target binding affinity prediction with transformers. *Computers in Biology and Medicine*, 147: 105772.

Mukherjee, S.; Ghosh, M.; and Basuchowdhuri, P. 2022. DeepGLSTM: deep graph convolutional network and LSTM based approach for predicting drug-target binding affinity. In *Proceedings of the 2022 SIAM International Conference on Data Mining (SDM)*, 729–737. SIAM.

Nguyen, T.; Le, H.; Quinn, T. P.; Nguyen, T.; Le, T. D.; and Venkatesh, S. 2021. GraphDTA: Predicting drug–target binding affinity with graph neural networks. *Bioinformatics*, 37(8): 1140–1147.

Öztürk, H.; Özgür, A.; and Ozkirimli, E. 2018. DeepDTA: deep drug–target binding affinity prediction. *Bioinformatics*, 34(17): i821–i829.

Öztürk, H.; Ozkirimli, E.; and Özgür, A. 2019. WideDTA: prediction of drug-target binding affinity. *arXiv preprint arXiv:1902.04166*.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Pinzi, L.; and Rastelli, G. 2019. Molecular docking: shifting paradigms in drug discovery. *International journal of molecular sciences*, 20(18): 4331.

Shi, Y.; Huang, Z.; Feng, S.; Zhong, H.; Wang, W.; and Sun, Y. 2021. Masked Label Prediction: Unified Message Passing Model for Semi-Supervised Classification. In Zhou, Z.-H., ed., *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, 1548–1554. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Tang, J.; Szwajda, A.; Shakyawar, S.; Xu, T.; Hintsanen, P.; Wennerberg, K.; and Aittokallio, T. 2014. Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis. *Journal of Chemical Information and Modeling*, 54(3): 735–743.

Wang, K.; Zhou, R.; Li, Y.; and Li, M. 2021. DeepDTAF: a deep learning method to predict protein–ligand binding affinity. *Briefings in Bioinformatics*, 22(5): bbab072.

Wu, J.-N.; Wang, T.; Chen, Y.; Tang, L.-J.; Wu, H.-L.; and Yu, R.-Q. 2024a. t-SMILES: a fragment-based molecular representation framework for de novo ligand design. *Nature Communications*, 15(1): 4993.

Wu, K. E.; Yang, K. K.; van den Berg, R.; Alamdari, S.; Zou, J. Y.; Lu, A. X.; and Amini, A. P. 2024b. Protein structure generation via folding diffusion. *Nature communications*, 15(1): 1059.

Yuan, W.; Chen, G.; and Chen, C. Y.-C. 2022. FusionDTA: attention-based feature polymerizer and knowledge distillation for drug-target binding affinity prediction. *Briefings in Bioinformatics*, 23(1): bbab506.

Zhang, Y.; Hu, Y.; Han, N.; Yang, A.; Liu, X.; and Cai, H. 2023. A survey of drug-target interaction and affinity prediction methods via graph neural networks. *Computers in Biology and Medicine*, 163: 107136.

Zheng, K.; Long, S.; Lu, T.; Yang, J.; Dai, X.; Zhang, M.; Nie, Z.; Ma, W.-Y.; and Zhou, H. 2024. ESM All-Atom: Multi-Scale Protein Language Model for Unified Molecular Modeling. In *International conference on machine learning*.