# Multimodal protein representation learning and target-aware variational auto-encoders for protein-binding ligand generation

IOP Publishing Mach. Learn.: Sci. Technol. 5 (2024) 025021 https://doi.org/10.1088/2632-2153/ad3ee4

MACHINE LEARNING
Science and Technology

OPEN ACCESS

RECEIVED
21 January 2024

REVISED
26 February 2024

ACCEPTED FOR PUBLICATION
15 April 2024

PUBLISHED
25 April 2024



PAPER

# Multimodal protein representation learning and target-aware variational auto-encoders for protein-binding ligand generation

**Nhat Khang Ngo**[1,3] **and Truong Son Hy**[1,2,3,*]

[1] AI Center, FPT Software, Hanoi, Vietnam
[2] Department of Mathematics and Computer Science, Indiana State University, Terre Haute 47809, IN, United States of America
[3] These authors contributed equally to this work.
[*] Author to whom any correspondence should be addressed.

**E-mail:** TruongSon.Hy@indstate.edu and khangnn4@fpt.com

**Keywords:** protein representation learning, protein-ligand binding, ligand generation, geometric deep learning, large language models, multimodal architecture, variational autoencoder

## Abstract

Without knowledge of specific pockets, generating ligands based on the global structure of a protein target plays a crucial role in drug discovery as it helps reduce the search space for potential drug-like candidates in the pipeline. However, contemporary methods require optimizing tailored networks for each protein, which is arduous and costly. To address this issue, we introduce **TargetVAE**, a target-aware variational auto-encoder that generates ligands with desirable properties including high binding affinity and high synthesizability to arbitrary target proteins, guided by a multimodal deep neural network built based on geometric and sequence models, named **Protein Multimodal Network** (PMN), as the prior for the generative model. PMN unifies different representations of proteins (e.g. primary structure—sequence of amino acids, 3D tertiary structure, and residue-level graph) into a single representation. Our multimodal architecture learns from the entire protein structure and is able to capture their sequential, topological, and geometrical information by utilizing language modeling, graph neural networks, and geometric deep learning. We showcase the superiority of our approach by conducting extensive experiments and evaluations, including predicting protein-ligand binding affinity in the PBDBind v2020 dataset as well as the assessment of generative model quality, ligand generation for unseen targets, and docking score computation. Empirical results demonstrate the promising and competitive performance of our proposed approach. Our software package is publicly available at https://github.com/HySonLab/Ligand_Generation.

## 1. Introduction

Drug discovery is a complex and expensive process that involves multiple stages and often takes years of development, with costs running into billions of dollars [1]. The first stage is to design novel drug-like compounds with high binding affinities to target proteins. This process consists of two sub-tasks: searching for candidates and measuring drug-target affinities (DTA). Searching for potential candidates in a huge database of roughly $10^{33}$ chemically valid molecules is a daunting task as current methods often rely on virtual screenings, professional software, and expert evaluation [2, 3]. Besides, DTA are critical measurements for identifying potential candidates and avoiding those that are inefficient for clinical trials. The most reliable technique for predicting DTA involves atomistic molecular dynamics simulations. However, these methods are computationally expensive and time-consuming, making them infeasible for large-scale sets of protein-ligand complexes. Our ultimate objective is to accelerate and automate these two sub-tasks in the first stage of the drug development process, using computational methods and machine-learning techniques.

Deep generative models (DGMs) have been proposed as a promising approach to reducing the workload of wet lab experiments in drug discovery by effectively designing probable drug-like candidates [2–11].

DGMs have demonstrated remarkable results in unconditionally generating or optimizing molecular properties, such as drug-likeness (QED) and synthetic accessibility (SA). However, they are prohibitively slow when enhancing binding affinity or other computationally expensive molecular properties. This is because they need to be trained in reinforcement learning frameworks where the generated molecular graph is modified based on a reward function determined by calling a property network that estimates the binding affinities. While effective and powerful, these approaches require specific property networks to be trained for each target protein, which is challenging due to the vast number of known and unknown proteins [3, 12]. Moreover, binding scores (i.e. labels for supervised learning) are not widely available and time-consuming to approximate with software such as Autodock Vina [13]. Additionally, several existing studies [14–16] incorporate prior knowledge of protein structures to guide conditional generative models to generate bioactive molecules. However, these methods rely solely on the information of specific binding sites of target proteins and are limited when the binding sites are not determined. It is worth noting that determining binding sites, also known as pockets, on a target protein (receptor) is a challenging problem, which consists of many constraints [17].

Proteins are macromolecules that can be represented in terms of sequences of amino acids (i.e. primary structure), 2D graphs at residue level constructed by nearest neighbors from folding information (i.e. tertiary structure), or 3D point clouds at atom level. Recent advanced methods for protein representation learning leverage language models, graph neural networks (GNNs), and convolutional neural networks (CNNs). In sequence-based methods [18–22], a protein sequence is regarded as a long sequence of tokens (i.e. *k*-mers [18]) that are fed to a transformer-based language model. In contrast, GNNs and CNNs-based approaches [23–27] operate on relational and geometric structures of proteins, respectively. While sequence-based approaches can capture the relationships among distant residues in a long protein sequence, they are not able to exploit the geometric relations among them. On the other hand, although GNNs and CNNs can learn spatial information about protein structures, they are limited in their ability to capture long-range interactions in large protein structures due to their reliance on localities. Recent years have also seen the rise of pre-trained large language models for scientific discovery, which have achieved remarkable results in protein science [28, 29]. This motivates future work to apply language models to many downstream tasks in drug discovery, where receptor representations are essential for the search for novel ligands.

### 1.1. Contributions

This work aims at developing a data-driven approach that can help accelerate the drug discovery process. We develop a conditional DGM to generate drug-like ligands that can bind to a target protein when its binding sites are unknown. This approach facilitates the process of determining binding sites in drug discovery, which is computationally expensive. Furthermore, we build a multimodal protein network to leverage multiple data types of proteins. In summary, our contributions are two-fold as follows:

- We build a conditional VAE model, named **TargetVAE**, that can generate chemically valid, drug-like molecules with high binding scores to an arbitrarily given protein structure. Apart from other methods, ours can directly condition on the entire structure of any protein target and design multiple candidates that can bind to it, without requiring the training of a specific property network for each target. It is important to note that our generative model is conditional on the whole protein structure rather than a specific pocket or binding site (i.e. a region of the protein surface where the ligand binds to) as in works of Luo *et al* [16], Guan *et al* [30], Peng *et al* [31], Liu *et al* [32]. Our approach is more flexible because a protein complex can contain multiple binding sites and therefore potentially have different ligands. TargetVAE allows us to efficiently generate ligand candidates with high binding affinity without prior knowledge of any binding site.
- We design a novel architecture, named **Protein Multimodal Network** (PMN), that unifies different modalities of proteins, i.e. sequence of amino-acids and 3D tertiary structure, to improve the performance of predicting binding affinities. The proposed model shows competitive results on the PDBBind v2020 benchmark in comparison with current state-of-the-art methods. Our novel multimodal architecture enables us to efficiently produce a protein embedding that can serve as the prior for our generative model (i.e. TargetVAE), and accurately estimate protein-ligand binding affinity, and potentially replace the computationally expensive molecular dynamical simulation in the evaluation of ligand generation.

## 2. Related work

### 2.1. Protein-ligand binding prediction

Machine learning methods, especially GNNs, have emerged as effective techniques for binding affinity prediction [33]. Using a Kronecker regularized least squares approach (KronRLS), Nascimento *et al* [34] cast the problem as a link prediction task on bipartite networks and compute distinct kernels that indicate the similarities among drugs and targets to make predictions. Apart from binary prediction, He *et al* [35] introduce a framework named SimBoost that can predict continuous binding affinity scores between drugs and targets. In the deep learning era, Öztürk *et al* [36] propose DeepDTA, a deep-learning-based method that uses CNNs to operate on sequence representations of drugs and protein targets. Moreover, Zhao *et al* [37] use generative adversarial networks to learn better feature representations for compounds and proteins. On the other hand, Nguyen *et al* [38]; Voitsitskyi *et al* [39] utilize GNNs to learn the representations of the molecular graphs and protein structures, which are superior to the previous methods operating on sequences and handcrafted features. However, the common limitation of these above approaches is that none of them covers a wide range of representations of proteins. Indeed, our proposed method is the first multimodal model combining sequence, graph, and spatial information of proteins altogether.

#### *2.1.1. Molecule generation*

Previous studies on molecule generation are mostly categorized into SMILES-based and graph-based approaches [40]. Gómez–Bombarelli *et al* [41–43] use recurrent neural networks to build generative models operating on SMILES strings. However, these SMILES-based approaches often generate chemically invalid molecules. Kusner *et al* [44] and Dai *et al* [45] circumvent this issue by augmenting the decoders with grammar and semantic constraints to only generate valid molecules, yet this added information does not fully capture chemical validity in the latent space. Apart from SMILES-based methods, several works [5–10, 46, 47] propose graph generative models to design novel drug-like molecules. For example, Jin *et al* [5]; Jin *et al* [6] can generate molecules with 100% validity, but their methods are relatively slow as chemical rules are verified during the generative process. Although graph-based methods perform well in unconditional molecule generation, they have difficulty generating molecules with optimized properties. For a target-aware generation, reinforcement learning methods are used to systematically modify the generated molecular graphs [4–7]. Distinguishing from other prior works based on SMILES representation, we utilize the recently proposed SELFIES representation [48] to achieve a high chemical validity in generated ligands. Given the success of conditional VAE in image generation [49], our work is the first attempt to introduce a learnable prior based on the whole protein structures.

## 3. PMN

Proteins are complex structures that consist of long chains of residues/amino acids. Each amino acid is a molecule with 3D structures, and a combination of hundreds to thousands of residues determines the unique 3D structure of a specific protein and its functions. It is worth noting that while two residues are distant along the protein sequence, they could be close to each other in three-dimensional space. This is our key observation to design a novel framework that can unify different representations of proteins in an end-to-end learning manner.

In the field of graph learning, the conventional GNNs based on the message passing scheme [50] that propagates and aggregates information of each node to and from its local neighborhoods are incapable of capturing the long-range interactions in a large-diameter or long-range[4] graph [51, 52] while suffering from over-smoothing [53] and over-squashing [54] problems. Furthermore, to obtain a global understanding of the whole input graph, the message-passing scheme needs the number of layers/iterations proportional to the diameter length for distant nodes to 'communicate' with each other. That is computationally infeasible for macromolecules with thousands of atoms or residues like proteins. Meanwhile, the graph Transformers that consider all pairwise node interactions via the self-attention mechanism can successfully capture the long-range dependencies [52, 55, 56]. Since proteins can be seen as long-range graphs, we utilize sequential and graph Transformers to encode both sequences and 3D graphs of residues and combine them to create a unified representation for a large protein, making our model operate on multi-modalities of proteins. Our

[4] Diameter of a graph is defined as the maximum length of the shortest paths among all pairs of nodes.

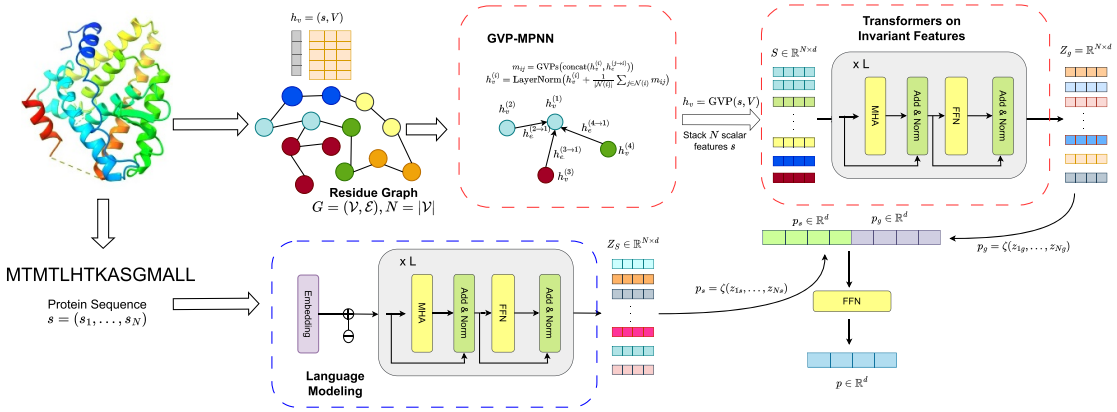



**Figure 1.** Overview of our PMN. The architecture consists of two components. Top: the three-dimensional structure of the input protein is processed by a local encoder built upon a message-passing neural network (MPNN) enhanced with geometric vector perceptron module (GVP) [26]. Subsequently, a Transformer processes the extracted invariant features to capture the global interactions between distant residues. Bottom: a Transformer-based language model is used to process the input protein sequence.

Transformer-based model can efficiently capture a protein's local and global information with a reasonably small number of layers.

### 3.1. Geometric learning on protein

According to figure 1, there are three major components in the 3D modeling part, including a local encoder, a geometric vector perceptron (GVP) module [26], and a global Transformers encoder.

#### *3.1.1. Local message-passing*

We use a message-passing network (MPNN) in which the GVPs [26] replace dense layers to operate on invariant features:

$$m_{ij} = \text{GVPs}\left(h_v^{(i)} \oplus h_e^{(j\to i)}\right), \tag{1}$$

$$h_v^{(i)} = \text{LayerNorm}\left(h_v^{(i)} + \frac{1}{|\mathcal{N}(i)|}\sum_{j\in\mathcal{N}(i)} m_{ij}\right), \tag{2}$$

where $\oplus$ is the concatenation; $m_{ij}$ computed by a module of three GVP layers denotes the message propagated from node $j$ to $i$. Also, $h_v^{(i)}$ and $h_e^{(j\to i)}$ indicate the embeddings of node $i$ and edge $(j \to i)$ and are tuples of scalar and vector features as described appendix A.1. The local encoder outputs a tuple of scalar and vector features for each residue node, which are rotationally invariant and equivariant, respectively.

#### *3.1.2. Global interaction*

We utilize a GVP module to update the tuple $h_v = (s, V)$ as $(s', V') = \text{GVP}\big((s, V)\big)$, and we take the invariant scalar feature $s' \in \mathbb{R}^d$ as the node embedding for successor modules. The resulting tensor $S \in \mathbb{R}^{N\times d}$, in which row $i$ indicates a $d$-dimensional scalar feature $s_i$ of node $i$, is passed to a $L$-layer Transformers encoder:

$$Q_\ell = Z_{\ell-1}W_\ell^Q, \;\; K_\ell = Z_{\ell-1}W_\ell^K, \;\; V_\ell = Z_{\ell-1}W_\ell^V, \tag{3}$$

$$H_\ell = \text{MultiheadAttention}\left(Q_\ell, K_\ell, V_\ell\right), \tag{4}$$

$$Z_\ell = \text{LayerNorm}\left(Z_{\ell-1} + \text{FFN}\left(H_\ell\right)\right). \tag{5}$$

Here, we initialize $Z_0 \triangleq S$; and we have $\{W_\ell^Q, W_\ell^K, W_\ell^V\}_{\ell=1}^L \in \mathbb{R}^{d\times d_k}$ as learnable weight matrices/parameters corresponding to the query $Q_\ell$, key $K_\ell$ and value $V_\ell$ matrices, respectively, of the Transformer at each layer $\ell$;

and $Z_g \triangleq Z_L$ denotes the final node embeddings produced by the network. Notably, this global encoder allows residue nodes to attend to other nodes on a large protein graph, especially those that are distant from them (i.e. long-range modeling). Finally, we aggregate node embeddings by a row-wise *Aggregator* $\zeta$ (e.g. mean, max, sum, etc) to produce an embedding for the protein structure $p_g = \zeta(Z_g) \in \mathbb{R}^d$.

*3.1.3. Language modeling on protein sequence*

A protein can be represented as a sequence $s = (s_0, s_1, \ldots, s_n)$ in which $s_i \in \mathbb{R}^{20}$ is a one-hot vector indicating one in a total of 20 types of residues. We utilize Transformer-based language models, where the layers are the same as in equations (3)–(5), to compute the text representation of this protein sequence with the initial embeddings $Z_0 = [z_1, z_2, \ldots, z_n] \in \mathbb{R}^{n \times d}$ with $z_i \in \mathbb{R}^d$ is calculated as $z_i = \text{Embed}(s_i) + p_i$. Here, $p_i$ is the positional encoding feature added at each token $i$. In this work, for the language modeling component, we use the pre-trained Transformer protein language models proposed by Lin *et al* [29] to extract protein-level embeddings for each protein. Then, we define $p_s = \zeta(Z_s) \in \mathbb{R}^d$ as the global representation for the entire protein sequence.

*3.1.4. Multi-modal fusion*

Finally, to produce the final representations of proteins, we concatenate their geometric and sequential features from the Transformers-based models and process them by a feed-forward network:

$$p = \text{FFN}\left(p_s \oplus p_g\right). \tag{6}$$

Here, $\oplus$ denotes the concatenation between two global vector features $p_s$ and $p_g$.

## 4. Binding affinity prediction and target-aware ligand generation

### 4.1. Problem setup

Given a dataset $D$ of protein-ligand pairs, our objective is to predict the binding affinity and generate novel drug-like ligands that have the potential to bind to a conditioning protein structure. We cast the former as a prediction task based on geometric and relational reasoning on protein and ligand structures, whereas the latter is regarded as a protein-structure conditioned ligand generation. Let $(l, p, s) \in D$ be a pair of protein-ligand where $l$ and $p$ denote the representations of ligands and proteins, respectively, and $s$ indicates the binding score between them. Additionally, figures 2 and 3 depict the overview of our approach in both tasks.

### 4.2. Binding affinity prediction

Figure 2 illustrates our designed architecture for predicting the binding affinities between ligands and their target proteins. Regarding the ligand part, we represent small molecules as 2D graphs $G = (\mathcal{V}, \mathcal{E})$ where nodes $v \in \mathcal{V}$ are the atoms and edges $e \in \mathcal{E}$ indicate their covalence bonds. Then, we use graph attention networks (GAT), a MPNN, to learn their representations. In particular, at the layer $t$, the embedding vector $\mathbf{x}_i^t$ of node $i \in \mathcal{V}$ is updated as:

$$\mathbf{x}_i^t = \alpha_{i,i} W \mathbf{x}_i^{t-1} + \sum_{j \in \mathcal{N}(i)} \alpha_{i,j} W \mathbf{x}_j^{t-1}, \tag{7}$$

where the attention coefficients $\alpha_{i,j}$ are computed as

$$\alpha_{i,j} = \frac{\exp\left(\sigma\left(\mathbf{a}^\top \left[W\mathbf{x}_i \oplus W\mathbf{x}_j\right]\right)\right)}{\sum_{k \in \mathcal{N}(i) \cup \{i\}} \exp\left(\sigma\left(\mathbf{a}^\top \left[W\mathbf{x}_i \oplus W\mathbf{x}_k\right]\right)\right)}. \tag{8}$$

Here, $\sigma$ denotes non-linear activations, $\mathcal{N}(i)$ is the set of neighbors of node $i$, and $\oplus$ denotes concatenation, and $\mathbf{a}$ and $W$ are the weight vector and matrix, respectively. At the end of the network, we aggregate all the node embeddings on the molecular graphs to produce its global embedding as $\ell = \zeta\{x_i | x_i \in \mathcal{V}\}$. In addition to graph-based features, Morgan fingerprints can also be used as molecule features. Finally, we combine features from all sources, including sequence, geometric information of proteins, and graph-based features of ligands, and pass them to a feed-forward network (FFN) to make predictions:

$$h = p \oplus \ell, \tag{9}$$

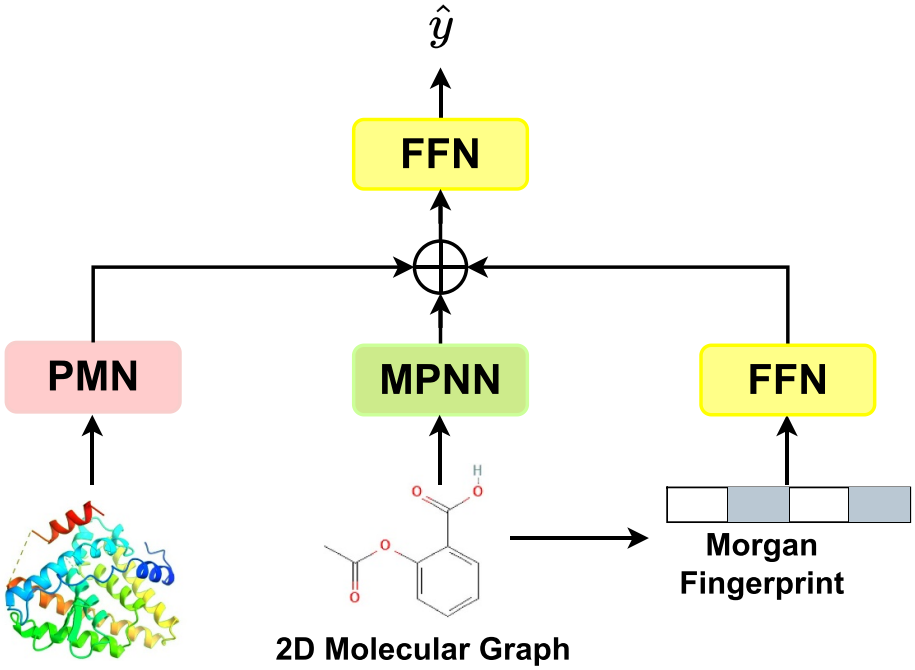



**Figure 2.** A framework for predicting binding affinities between proteins and ligands. The protein multi-modal network (PMN) computes a unified representation from the sequence and structure of the input protein. Besides, a message-passing neural network (MPNN) and feed-forward network (FFN) are used to process the 2D graph and Morgan Fingerprint of the input ligand, respectively. Finally, all the representations are combined and passed to a FFN that predicts the binding affinity score.

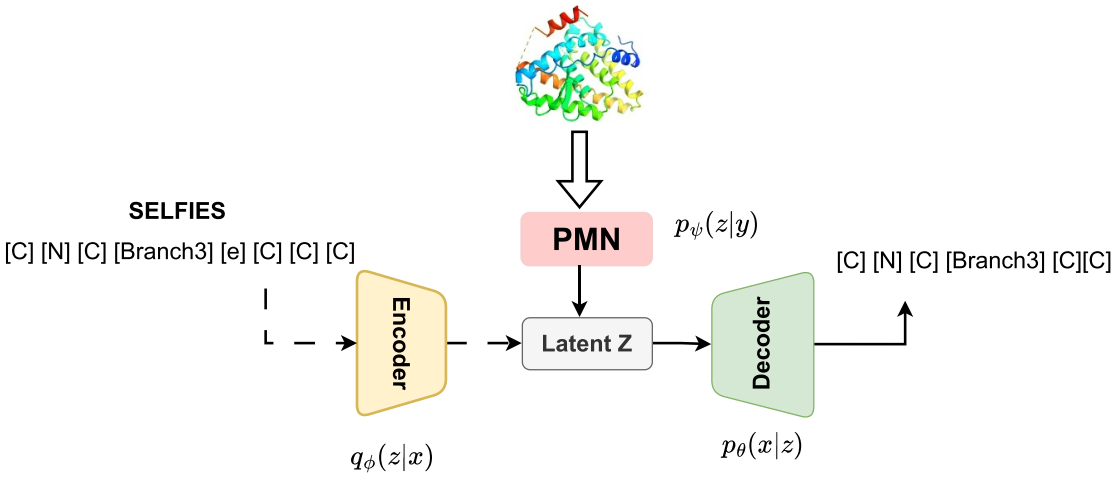



**Figure 3.** TargetVAE with an encoder, decoder, and a prior network. The PMN prior network computes the conditions from protein structures for constructing the latent space of the VAE framework, which learns to generate SELFIES representations of molecules.

$$\hat{y} = \phi(h)\,. \tag{10}$$

Here, $h$ is the combination of all sources of inputs, and $\phi$ is a feed-forward network that predicts the binding affinity score $\hat{y}$.

### 4.3. Target-aware ligand generation

Although there exist many machine-learning approaches that generate drug-like molecules, it is challenging for graph-based or smiles-based methods to generate chemically valid ligands with high probability. Meanwhile, SELFIES (SELF-referenced Embedded Strings) [48] is a string-based representation of molecules 100% robust to molecular validity. A ligand $l$ can be defined as a string of $l = (l_1, l_2, \ldots, l_n)$ in which $l_i$ is a SELFIES token, which belongs to a predefined symbol set $S$ derived from the training dataset. We generate

ligands $\hat{l} = (\hat{l}_1, \hat{l}_2, \ldots, \hat{l}_n)$ by computing $n$ independent probability vectors $y = (y_1, y_2, \ldots, y_n)$, $y_i \in \mathbb{R}^{|S|}$. Each new token $\hat{l}_i$ is defined as $\hat{l}_i = S_j$ where $j = \underset{0 \leqslant j < |S|}{\text{argmax}}(y_i)$.

Let $\phi, \theta$, and $\psi$ denote the encoder, decoder, and prior network in a conditional VAE framework, respectively. In particular, $\phi$ is a recurrent neural network (RNN) that encodes a molecule into a latent vector $z \in \mathbb{R}^d$, and $\theta$ is also an RNN that autoregressively generates each token of SELFIES string. The prior network $\psi$ is our proposed protein muli-modal network in which both sequence and geometric information of a target protein are encoded, and we use two multi-layer perceptions denoted as $\mu_\psi$ and $\sigma_\psi$ to compute the mean and variance of a Gaussian distribution over the latent vector $z_l \sim \mathcal{N}(\mu_\psi(p), \sigma_\psi(p))$, which is the inferred latent embedding of the ligand $l$. All the networks are jointly optimized based on equation (11). After training, given a target protein $p$, a ligand $\hat{l}$ is generated by sampling a latent vector $z \sim \mathcal{N}(\mu_\psi(p), \sigma_\psi(p))$, which is fed to the decoder $\theta$ to decode into a SELFIES representation.

#### *4.3.1. Conditional inference with pretrained unconditional VAE*

In addition to validity, the diversity of generated sets of ligands is also an important criterion in drug discovery. While classical conditional VAE trained on protein-ligand pairs can generate novel and valid molecules, the diversity and uniqueness of these samples are relatively low due to the limited amount of available data. We address this issue by adapting the work proposed by Harvey *et al* [49] from the computer vision domain to diversify the latent variables. In this framework, the decoder $\theta$ of the generative model is independent with the condition $p$ as $p_{\theta,\psi}(x, z|p) = p_\theta(x|z)p_\psi(z|p)$, allowing $\theta$ to re-use weights of the decoder $\theta^*$ of an unconditional VAE as both have the identical architecture. We train the model to optimize the objective as:

$$\log p_{\theta,\psi}(x|p) \geqslant O_{\text{for}} \triangleq \mathbb{E}_{q_\phi}\left[\log p_{\theta,\psi}(x|z)\right] - \text{KL}\left[q_\phi(z|p)\,||\,p_\psi(z|p)\right]. \tag{11}$$

Different from equation (13), both $q_\phi$ and $p_\theta$ in equation (11) are not conditioned by the auxiliary covariate $y$. This allows conditional VAEs to use weights of $\phi^*$ and $\theta^*$ of a pre-trained VAE, which is trained on a diverse set of unconditional molecules.

## 5. Results

### 5.1. Binding affinity prediction on PDBBind 2020

#### *5.1.1. Experimental setup*

The objective of the task is to predict the binding affinities that reflect how small molecules (ligands) bind to their target proteins. We evaluate the performance of our approach on the PDBBind v2020 [57] with time-split training and testing sets. In particular, the dataset contains 19 119 pre-processed complexes, and there are 1152 of them were discovered in 2019 or later. We follow the split procedure described by Stärk *et al* [17] that randomly selects 125 unique proteins and collects all new complexes containing them to create the final test set, resulting in a total of 363 samples. For the complexes discovered before 2019, those that have ligands in the test set are removed, and this gives 17 347 complexes in total. This set is respectively split into 16 379 and 968 complexes for training and validation subsets such that they share no ligands.

#### *5.1.2. Implementation details*

In this task, we use three layers of GVP with a hidden dimension of 128 to extract the local geometric information of proteins, while also considering their interactions with the ligands, which are encoded by a three-layer GAT network with the same dimension. Regarding the language modeling component, we extract the residue-level embeddings of the pre-trained ESM-2 proposed by [29] with a hidden dimension of 1280. The features are fused and passed to a four-layer MLP with hidden dimensions of 1024 512, and 256 for making the predictions. We use batch normalization and a dropout rate of 0.05 across layers. The models are trained in 50 epochs with a batch size of 32 and a learning rate of 0.0001.

#### *5.1.3. Main results*

According to table 1, our proposed approach outperforms other sequence-based methods by a large margin, demonstrating the necessity of modeling three-dimensional structures of proteins in machine learning. Compared with structure-based and complex-based methods, the model shows comparable performances across all metrics with lower standard deviations. Furthermore, the model performs on par with other more sophisticated state-of-the-art methods, namely PSICHIC [58] and TankBind. It is worth noting that

**Table 1.** Performance comparison on the PDBBind v2020 dataset. An upward arrow (↑) denotes higher scores are better, and a downward arrow (↓) denotes the reverse. Average results and standard deviations (in parentheses) from five independent runs are reported.

| | Method | RMSE ↓ | MAE ↓ | Pearson ↑ | Spearman ↑ | $r_m^2$ ↑ | CI ↑ |
|---|---|---|---|---|---|---|---|
| Complex | Pafnucy | 1.435 (0.018) | 1.144 (0.018) | 0.635 (0.008) | 0.587 (0.008) | 0.348 (0.016) | 0.707 (0.004) |
| | OnionNet | 1.403 (0.012) | 1.103 (0.014) | 0.648 (0.007) | 0.602 (0.013) | 0.381 (0.011) | 0.717 (0.005) |
| | IGN | 1.404 (0.025) | 1.116 (0.030) | 0.662 (0.013) | 0.638 (0.021) | 0.385 (0.02) | 0.730 (0.009) |
| | SIGN | 1.373 (0.037) | 1.086 (0.030) | 0.685 (0.031) | 0.656 (0.044) | 0.398 (0.048) | 0.736 (0.02) |
| Structure | SMINA | 1.466 (0.008) | 1.161 (0.007) | 0.665 (0.005) | 0.663 (0.019) | 0.391 (0.031) | 0.740 (0.008) |
| | GNINA | 1.740 (0.014) | 1.413 (0.015) | 0.495 (0.011) | 0.494 (0.011) | 0.209 (0.009) | 0.674 (0.004) |
| | dMaSIF | 1.450 (0.032) | 1.136 (0.031) | 0.629 (0.018) | 0.588 (0.041) | 0.347 (0.029) | 0.710 (0.017) |
| | TankBind | 1.345 (0.020) | 1.060 (0.031) | **0.718 (0.012)** | **0.689 (0.041)** | 0.404 (0.025) | 0.750 (0.006) |
| Sequence | GraphDTA | 1.564 (0.063) | 1.223 (0.066) | 0.612 (0.016) | 0.570 (0.050) | 0.306 (0.039) | 0.703 (0.019) |
| | TransCPI | 1.493 (0.050) | 1.201 (0.037) | 0.604 (0.024) | 0.551 (0.029) | 0.255 (0.027) | 0.677 (0.011) |
| | MolTrans | 1.599 (0.060) | 1.271 (0.051) | 0.539 (0.057) | 0.474 (0.052) | 0.242 (0.045) | 0.666 (0.02) |
| | DrugBAN | 1.480 (0.046) | 1.159 (0.045) | 0.657 (0.018) | 0.612 (0.027) | 0.319 (0.021) | 0.720 (0.011) |
| | DGraphDTA | 1.493 (0.050) | 1.201 (0.037) | 0.604 (0.024) | 0.551 (0.029) | 0.312 (0.038) | 0.693 (0.011) |
| | WGNN-DTA | 1.501 (0.050) | 1.196 (0.055) | 0.605 (0.025) | 0.562 (0.028) | 0.311 (0.03) | 0.697 (0.01) |
| | STAMP-DPI | 1.503 (0.082) | 1.176 (0.067) | 0.653 (0.028) | 0.601 (0.027) | 0.327 (0.039) | 0.719 (0.011) |
| | PSICHIC | **1.314 (0.049)** | **1.015 (0.031)** | 0.710 (0.027) | 0.686 (0.024) | 0.428 (0.047) | **0.751 (0.009)** |
| | Ours | 1.373 (0.035) | 1.084 (0.032) | 0.687 (0.010) | 0.646 (0.016) | **0.459 (0.022)** | 0.733 (0.006) |

*Note:* Bold highlights the best performance.

**Table 2.** Ablation study on the use of sequence embeddings and three-dimensional structures. The results are aggregated from five independent runs.

| Method | RMSE ↓ | MAE ↓ | Pearson ↑ | Spearman ↑ | $r_m^2$ ↑ | CI ↑ |
|---|---|---|---|---|---|---|
| Only 3D | 1.596 (0.028) | 1.300 (0.021) | 0.505 (0.029) | 0.453 (0.025) | 0.235 (0.031) | 0.657 (0.008) |
| Only ESM | 1.421 (0.029) | 1.123 (0.020) | 0.657 (0.009) | 0.607 (0.011) | 0.407 (0.022) | 0.718 (0.004) |
| ESM + 3D | **1.373 (0.035)** | **1.084 (0.032)** | **0.687 (0.010)** | **0.646 (0.016)** | **0.459 (0.022)** | **0.733 (0.006)** |

*Note:* Bold highlights the best performance.

PSICHIC also leverages the residue-level embeddings extracted from the pre-trained ESM; however, Koh *et al* [58] use these embeddings to construct 2D graphs of proteins. This does not preserve the SE(3)-symmetry (rotations and translations), which is an important property in learning three-dimensional structures.

#### *5.1.4. Ablation studies*

To comprehensively evaluate the effectiveness of protein multi-modal learning, we conducted an ablation study to assess the necessity of sequence information in protein modeling. Specifically, we re-trained our model without the sequence embeddings from pre-trained ESM, keeping the 3D modeling component unchanged. The numerical results in table 2 show that exploiting both sequence and geometric information can lead to superior performance in protein-ligand binding affinity prediction. The model that combines both sources of information outperforms its 3D modeling counterpart by a large margin in Pearson, Spearman, $r_m^2$, and CI scores. Additionally, we observe that using only ESM embeddings can also achieve comparable results to other state-of-the-art methods, indicating the importance of protein primary structure, which is the most common protein data modality.

### 5.2. Target-aware ligand generation

#### *5.2.1. Experimental setup and implementation details*

This task aims to generate small drug-like molecules that bind to given target proteins with unknown binding sites. We train a prior network that encodes geometric and sequence information of a target by optimizing the objective in equation (11). We use the train set of PDBBind 2020 is used, as described in section 5.1.2. For docking simulation, we adopt AutoDock Vina [13] to compute binding affinity. The score, named Vina Score, characterizes the free energy changes of binding processes in kcal $mol^{-1}$. We test our approach with nine target proteins, including G-protein coupling receptors and kinases from DUD-E [59] and the SARS-Conv-2 main protease [60]. Notably, these targets are unseen to the model during the training stage. For implementation, we use the architecture and parameters of a pre-trained RNN-based VAE

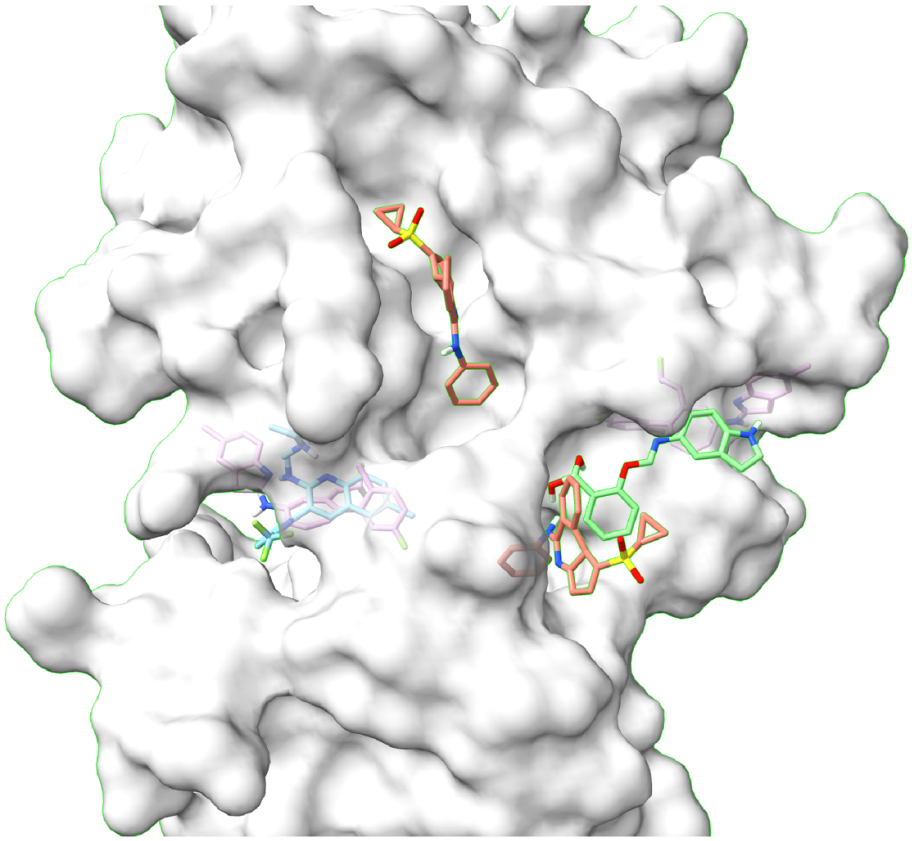


**Figure 4.** Multiple generated ligands with different poses bind to the 1iep target.

proposed by Gao *et al* [61] in which the dimensions of latent vectors are 128. For each test target, we generate 100 molecules and estimate their drug-likeness (QED), synthetic accessibility (SA), and binding affinity (Vina Score)d. The first two metrics are calculated by RDkit, and docked poses are generated by the AutoDock Vina [13] (Vina score). Additionally, we use Obabel [62] to optimize the three-dimensional atom positions of the generated molecules.

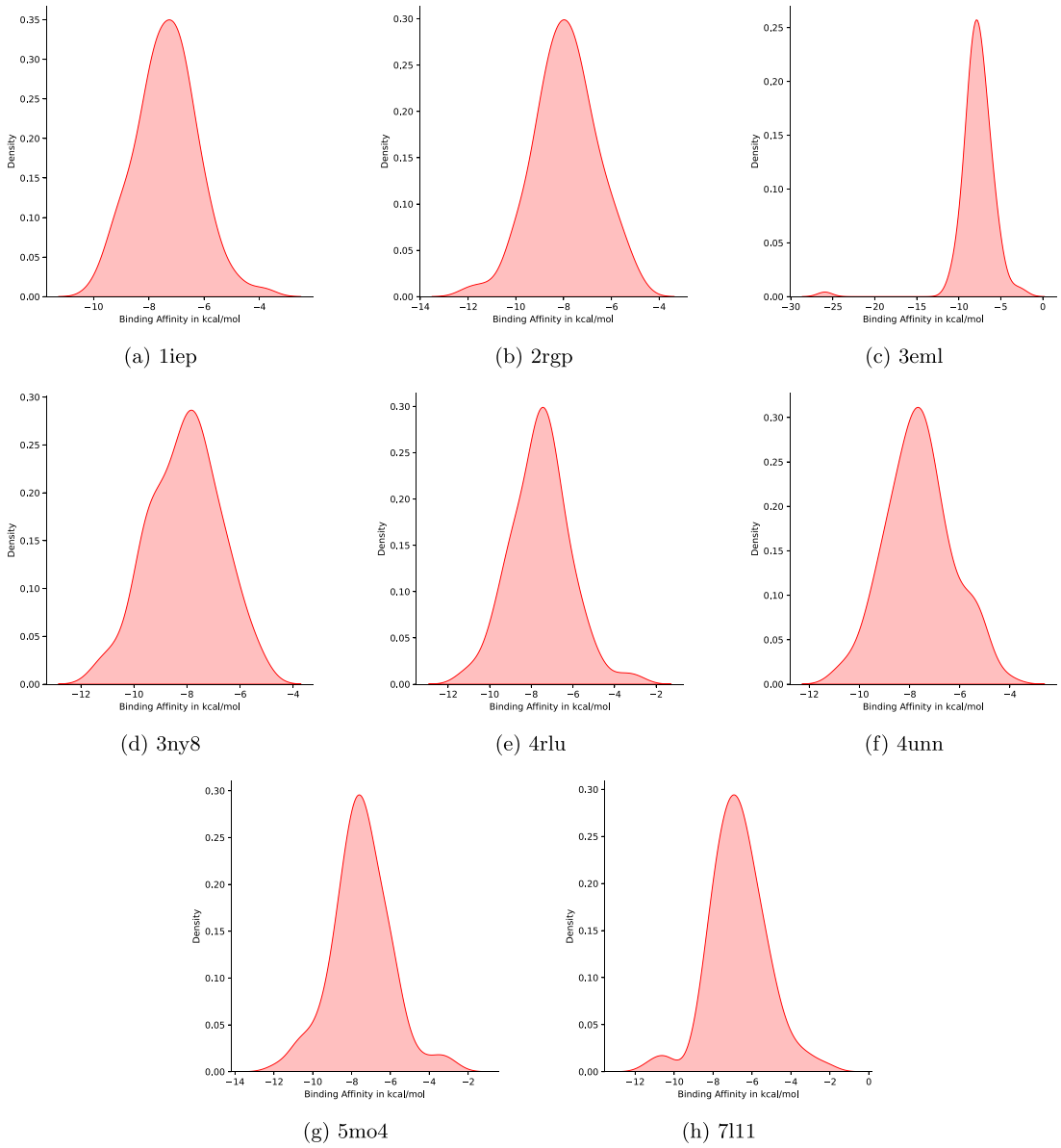
### *5.2.2. Results*

#### *5.2.2.1. Qualitative results*

Apart from pocket-based methods [14–16], our generative model is conditioned on the entire protein structures, meaning that the knowledge of binding sites is unknown to the model. Using protein multi-modalities, as well as modeling the long-range interactions among residues enables us to extract representations that incorporate both geometric and sequence meanings of a target protein. These informative conditions, as shown in figure 4, allow the generative model to generate multiple ligands that can bind to different sites with different poses on a given target protein. Furthermore, we choose some of the generated ligands and visualize how they bind to their targets in figure 6.

#### *5.2.2.2. Quantitative results*

Figure 5 illustrates the distribution of binding affinities of the generated molecules for their corresponding receptors. We used Autodock Vina to compute the docking scores between the ligands and receptors, with lower scores indicating stronger binding. The histograms show that our method can generate ligands with high binding affinities (low docking scores). The average binding affinity for each target protein is less than $-6$ kcal mol$^{-1}$. Moreover, table 3 shows the average scores for binding affinities (BA), drug-likeness (QED), and synthetic accessibility (SA) of the top 1, 10, and 20 generated molecules ranked according to their binding affinities to corresponding targets. We observe that TargetVAE can generate molecules with high binding affinity while maintaining low SA scores ranging from 4.0 to 8.0. However, we also note that our




(a) 1iep (b) 2rgp (c) 3eml

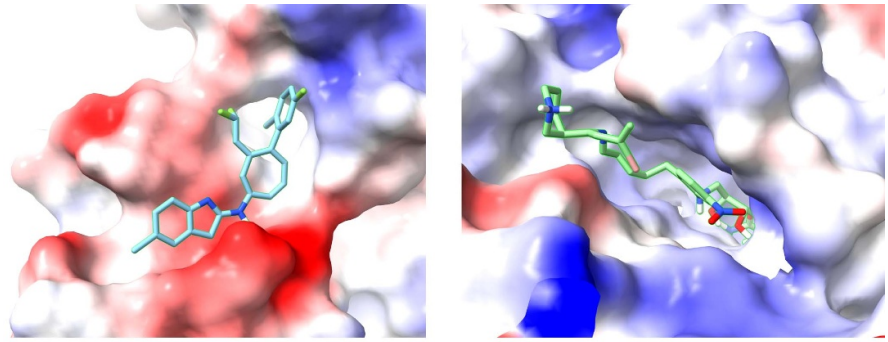
(d) 3ny8 (e) 4rlu (f) 4unn

(g) 5mo4 (h) 7l11

**Figure 5.** The binding affinity (in kcal $mol^{-1}$) distribution of the generated molecules on different target proteins.

**Table 3.** Quantitative results of top $k = 1, 10, 20$ generated molecules, which are ranked based on binding affinity (in kcal $mol^{-1}$). The scores are averaged over $k$ ligands.

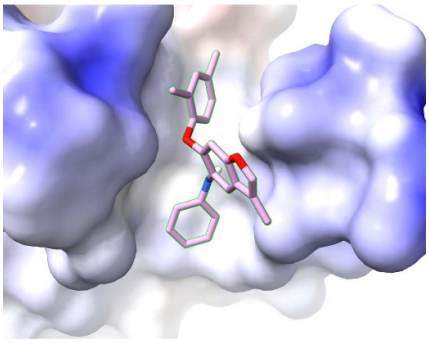
| | Top 1 | | | Top 10 | | | Top 20 | | |
|---|---|---|---|---|---|---|---|---|---|
| Target | BA ↓ | SA ↓ | QED ↑ | BA ↓ | SA ↓ | QED ↑ | BA ↓ | SA ↓ | QED ↑ |
| 1iep | −9.946 | 7.609 | 0.322 | −9.242 | 4.412 | 0.413 | −8.856 | 4.227 | 0.411 |
| 2rgp | −11.936 | 3.391 | 0.428 | −10.293 | 4.201 | 0.520 | −9.717 | 4.151 | 0.482 |
| 3eml | −25.939 | 7.268 | 0.584 | −11.590 | 4.346 | 0.493 | −10.294 | 4.446 | 0.476 |
| 3ny8 | −11.257 | 5.99 | 0.807 | −10.280 | 3.980 | 0.369 | −9.870 | 4.193 | 0.433 |
| 4rlu | −11.250 | 2.979 | 0.479 | −10.010 | 4.536 | 0.619 | −9.495 | 4.759 | 0.564 |
| 4unn | −10.752 | 4.567 | 0.161 | −9.860 | 4.192 | 0.415 | −9.423 | 4.270 | 0.418 |
| 5mo4 | −11.812 | 6.330 | 0.432 | −10.325 | 5.041 | 0.325 | −9.627 | 4.865 | 0.443 |
| 7l11 | −11.220 | 7.912 | 0.136 | −9.163 | 5.396 | 0.394 | −8.567 | 5.073 | 0.417 |



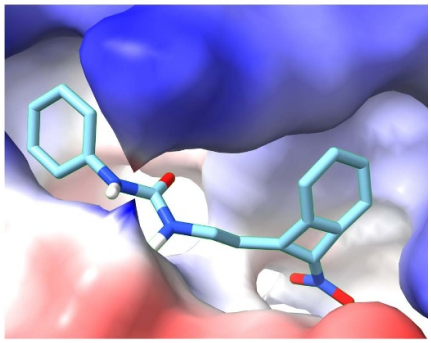
(a) 1iep, 3.59, -9.48 kcal/mol



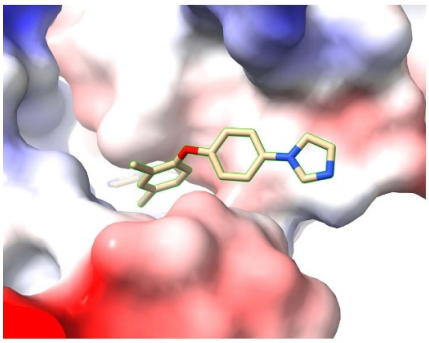
(b) 2rgp, 4.28, -8.17 kcal/mol



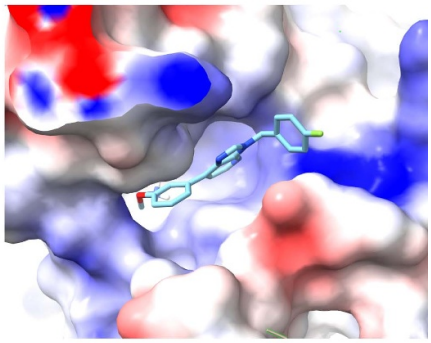
(c) 3eml,2.20,-8.29 kcal/mol



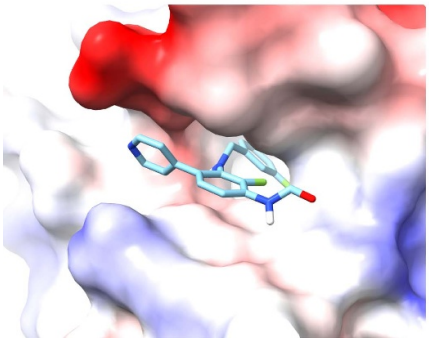
(d) 3ny8, 2.69,-8.78 kcal/mol



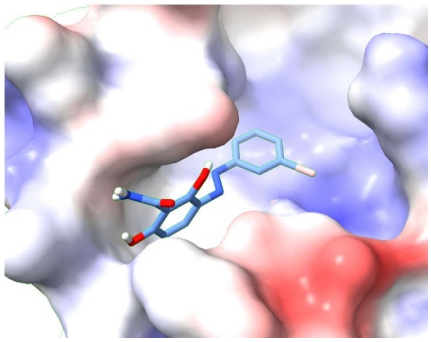
(e) 4rlu, 1.98,-7.99 kcal/mol



(f) 4unn, 2.65, -7.96 kcal/mol



(g) 5mo4, 2.43, -8.67 kcal/mol



(h) 7l11, 2.69, -6.55 kcal/mol

**Figure 6.** Three-dimensional poses of protein-ligand complexes generated by AutoDock Vina and visualized by Chimerax. Each figure is associated with the name of each target protein, synthetic accessibility, and binding affinity in kcal $mol^{-1}$ of the generated ligand.

approach is limited in generating compounds with low drug-likeness (QED). We hypothesize that this is because generating ligands that simultaneously satisfy all objectives is challenging, and our model only focuses on binding affinity. This means that there may be a trade-off between the metrics.

## 6. Conclusion

This paper proposes **Protein Multimodal Network** (PMN), a novel neural architecture that learns to combine multiple levels of information (i.e. multimodal) of proteins including protein sequence (i.e. from the primary structure) as well as residue-level graph and geometry (i.e. from the 3D tertiary structure) into a unified representation. Our experiments in predicting protein-ligand binding affinity on the PDBBind v2020 dataset have shown that this new multimodal representation is highly effective in capturing both local and global structural information of protein, with competitive performance against current state-of-the-art methods.

Furthermore, we build a conditional variational autoencoder named **TargetVAE** that can generate new ligands that can bind to a target protein and have desirable properties such as high binding affinity and high synthesizability, etc. It is important to note that in our case, the binding sites are not known in advance; therefore, we utilize our pre trained PMN to encode the whole protein structure as a prior / condition to guide the generation process. We evaluate our generative model and the quality of generated ligands both quantitatively and qualitatively. The candidate ligands we generate have shown great potential in in-silico simulation.

We believe our two main contributions, PMN and TargetVAE, are highly applicable in practice. In general, PMN can help scientists estimate several protein functionalities and properties, as well as produce a sophisticated protein embedding that can be used for other downstream tasks. PMN in combination with TargetVAE allows us to generate new ligands for new proteins such as mutant proteins. In our future work, we plan to examine these candidates further *in vitro* and *in vivo* experiments.

## Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: https://github.com/HySonLab/Ligand_Generation. The data that we used for this study is publicly available at https://github.com/vtarasv/3d-prot-dta [39] and https://github.com/HannesStark/EquiBind [17]. We release our data processing pipeline and software along with the installation instructions at https://github.com/HySonLab/Ligand_Generation. All experimental results and visualizations are reproducible given our software release.

## Funding

Not applicable.

## Scientific contribution statement

In this work, we propose **Protein Multimodal Network** (PMN), a novel neural architecture that learns to combine multiple levels of information (i.e. multimodal) of proteins including protein sequence (i.e. from primary structure) as well as residue-level graph and geometry (i.e. from 3D tertiary structure) into a unified representation. Given the encoded protein structure by PMN as a prior, we build a conditional variational autoencoder named **TargetVAE** that can generate new ligands that can bind to a target protein and have desirable properties such as high binding-affinity and high synthesizability, etc. We believe our two main contributions, PMN and TargetVAE, are highly applicable in practice: (i) PMN can help scientists to estimate several protein functionalities and properties, as well as produce a sophisticated protein embedding that can be used for other downstream tasks; and (ii) PMN in combination with TargetVAE allows us to generate new ligands for new proteins such as mutant proteins.

## Appendix. Background

### A.1. Rotational invariant features

According to Jing *et al* [26], geometric features of a residue node on the protein structure can be represented as a tuple $(s, V)$ of scalar features $s \in \mathbb{R}^n$ and vector features $V \in \mathbb{R}^{\mu \times 3}$. Respectively, $s$ and $V$ are invariant and equivariant with respect to geometric transformations in Euclidean space. In addition, to make the information propagate effectively from the vector channel to the scalar channel, Jing *et al* [26] propose GVP as a replacement for conventional dense layers in GNNs, enabling them to operate on geometric vectors and structural information of 3D large protein graphs.

The module transforms an input tuple $(s, V)$ of scalar features $s \in \mathbb{R}^n$ and vector features $V \in \mathbb{R}^{\mu \times 3}$ into a new tuple $(s', V') \in \mathbb{R}^m \times \mathbb{R}^{\nu \times 3}$. According to algorithm 1, GVP consists of two separate

linear transformations $W_m$ and $W_h$ that work on the scalar and vector features respectively, followed by nonlinearities $\sigma$ and $\sigma^+$. Before being transformed, the scalar feature $s$ is concatenated with the $L_2-$norm of the vector feature $V$. This enables GVP to extract the rotation-invariant information from the input vector $V$. Moreover, an additional transformation $W_\mu$ is used to control the dimensionality of the output vector $V'$, making it independent of the number of norms extracted. Albeit simple, GVP is an effective module that guarantees desired properties of invariance/equivariance and expressiveness. The scalar and vector outputs of GVP are invariant and equivariant respectively, with respect to an arbitrary composition $R$ of rotations and reflections in 3D Euclidean space. In other words, if $\text{GVP}(s,V)=(s',V')$, then $\text{GVP}(s,R(V))=(s',R(V'))$.

**Algorithm 1.** Geometric vector perceptron.

**Input:** Scalar and vector features $(s,V)\in\mathbb{R}^n\times\mathbb{R}^{\mu\times 3}$
**Output:** Scalar and vector features $(s',V')\in\mathbb{R}^m\times\mathbb{R}^{\nu\times 3}$
$h\leftarrow\max(\mu,\nu)$
**GVP**:
$V_h\leftarrow W_hV \quad \in\mathbb{R}^{h\times 3}$
$V_\mu\leftarrow W_\mu V_h \quad \in\mathbb{R}^{\mu\times 3}$
$s_h\leftarrow\|V_h\|_2(\text{row-wise}) \quad \in\mathbb{R}^h$
$v_\mu\leftarrow\|V_\mu\|_2(\text{row-wise} \quad \in\mathbb{R}^\mu$
$s_{h+n}\leftarrow\text{concat}(s_h,s) \quad \in\mathbb{R}^{h+n}$
$s_m\leftarrow W_ms_{h+n}+b \quad \in\mathbb{R}^m$
$s'\leftarrow\sigma(s_m) \quad \in\mathbb{R}^m$
$V'\leftarrow\sigma^+(v_\mu)\odot V_\mu(\text{row-wise multiplication}) \quad \in\mathbb{R}^{\mu\times 3}$
**return** $(s',V')$

### A.2. Variational auto-encoders

A variational auto-encoder (VAE) is regarded as an auto-encoding variational Bayes model [63] that comprises two components, including a generative model and an inference model (also known as probabilistic encoder). The former uses a probabilistic decoder $p_\theta(x|z)$ and a prior $p_\psi(z)$ to define a joint distribution $p_{\theta,\psi}(x,z)=p_\theta(x|z)p_\psi(z)$ between latent variables $z$ and data $x$; in addition, Kingma and Welling [63] let $p_\psi(z)$ be isotropic Gaussian. An ideal generative model should learn to maximize the log-likelihood $\log p_{\theta,\psi}(x)=\log\int p_{\theta,\psi}(x,z)dz$. However, this is intractable as marginalization over the latent space is usually infeasible with realistic data. VAE alleviates this issue by using an encoder $q_\phi(z|x)$ to approximate the true posterior distribution of the latent space and maximize the evidence lower bound (ELBO) over each training sample $x$:

$$\log p_{\theta,\psi}(x)\geqslant\mathbb{E}_{q_\phi}\left[\log p_{\theta,\psi}(x|z)\right]-\text{KL}\left[q_\phi(z|x)\,||p_\psi(z)\right]. \tag{12}$$

In conditional VAE, the generative component is augmented by auxiliary covariates $y$. Given a condition $y$, the generative model defines a conditional joint distribution of $z$ and $x$ as $p_{\theta,\psi}(x,z|y)=p_\theta(x|y,z)p_\psi(z|y)$. Similarly, the condition inputs are integrated into the encoder as $q_\phi(z|x,y)$. These two extensions establish a prominent conditional VAE model [64–67] that is trained to maximize the conditional ELBO as:

$$\log p_{\theta,\psi}(x|y)\geqslant O_{\text{cond}}\triangleq\mathbb{E}_{q_\phi}\left[\log p_{\theta,\psi}(x|y,z)\right]-\text{KL}\left[q_\phi(z|x,y)\,||p_\psi(z|y)\right]. \tag{13}$$

## ORCID iD

Truong Son Hy https://orcid.org/0000-0002-5092-3757

## References

[1] Hughes J P, Rees S, Kalindjian S B and Philpott K L 2011 Principles of early drug discovery *Br. J. Pharmacol.* **162** 1239–49
[2] Verkhivker G M *et al* 2001 Binding energy landscapes of ligand-protein complexes and molecular docking: principles, methods and validation experiments *Combinatorial Library Design and Evaluation* (Taylor & Francis Group) pp 177–216
[3] Burley S K *et al* 2019 RCSB protein data bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy *Nucleic Acids Res.* **47** 464–74
[4] You J, Liu B, Ying Z, Pande V and Leskovec J 2018 Graph convolutional policy network for goal-directed molecular graph generation *Advances in Neural Information Processing Systems* vol 31, ed S Bengio, H Wallach, H Larochelle, K Grauman, N Cesa-Bianchi and R Garnett
[5] Jin W, Barzilay R and Jaakkola T 2018 Junction tree variational autoencoder for molecular graph generation *Proc. 35th Int. Conf. on Machine Learning* (*Proc. Machine Learning Research* vol 80) ed J Dy and A Krause pp 2323–32
[6] Jin W, Barzilay D R and Jaakkola T 2020 Hierarchical generation of molecular graphs using structural motifs *Proc. 37th Int. Conf. on Machine Learning* (*Proc. Machine Learning Research* vol 119), ed H Daumé III and A Singh pp 4839–48

[7] Luo S, Guan J, Ma J and Peng J 2021 A 3D generative model for structure-based drug design *Advances in Neural Information Processing Systems* ed A Beygelzimer, Y Dauphin, P Liang and J W Vaughan
[8] Simonovsky M and Komodakis N 2018 GraphVAE: towards generation of small graphs using variational autoencoders (arXiv:1802.03480)
[9] De Cao N and Kipf T 2018 MolGAN: an implicit generative model for small molecular graphs *ICML 2018 Workshop on Theoretical Foundations and Applications of Deep Generative Model*
[10] Luo Y, Yan K and Ji S 2021 GraphDF: a discrete flow model for molecular graph generation *Proc. 38th Int. Conf. on Machine Learning* (*Proc. Machine Learning Research* vol 139) ed M Meila and T Zhang pp 7192–203
[11] Gapsys V, Hahn D F, Tresadern G, Mobley D L, Rampp M and Groot B L 2022 Pre-exascale computing of protein-ligand binding free energies with open source software for drug design *J. Chem. Inf. Model.* **62** 1172–7
[12] Berman H M, Westbrook J, Feng Z, Gilliland G, Bhat T N, Weissig H, Shindyalov I N and Bourne P E 2000 The protein data bank *Nucleic Acids Res.* **28** 235–42
[13] Trott O and Olson A J 2010 Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading *J. Comput. Chem.* **31** 455–61
[14] Guan J, Qian W W, Peng X, Su Y, Peng J and Ma J 2023 3D equivariant diffusion for target-aware molecule generation and affinity prediction *The 11th Int. Conf. on Learning Representations*
[15] Schneuing A *et al* 2022 Structure-based drug design with equivariant diffusion models (arXiv:2210.13695)
[16] Luo S, Guan J, Ma J and Peng J 2021 A 3D generative model for structure-based drug design *Advances in Neural Information Processing Systems* vol 34, ed M Ranzato, A Beygelzimer, Y Dauphin, P S Liang and J W Vaughan pp 6229–39
[17] Stärk H, Ganea O, Pattanaik L, Barzilay D R and Jaakkola T 2022 EquiBind: geometric deep learning for drug binding structure prediction *Proc. 39th Int. Conf. on Machine Learning* (*Proc. Machine Learning Research* vol 162) ed K Chaudhuri, S Jegelka, L Song, C Szepesvari, G Niu and S Sabato pp 20503–21
[18] Notin P, Dias M, Frazer J, Hurtado J M, Gomez A N, Marks D and Gal Y 2022 Tranception: protein fitness prediction with autoregressive transformers and inference-time retrieval *Proc. 39th Int. Conf. on Machine Learning* (*Proc. Machine Learning Research* vol 162) ed K Chaudhuri, S Jegelka, L Song, C Szepesvari, G Niu and S Sabato pp 16990–7017
[19] Brandes N, Ofer D, Peleg Y, Rappoport N and Linial M 2022 ProteinBERT: a universal deep-learning model of protein sequence and function *Bioinformatics* **38** 2102–10
[20] Asgari E, McHardy A C and Mofrad M R 2019 Probabilistic variable-length segmentation of protein sequences for discriminative motif discovery (DiMotif) and sequence embedding (ProtVecX) *Sci. Rep.* **9** 1–16
[21] Wu Z, Johnston K E, Arnold F H and Yang K K 2021 Protein sequence design with deep generative models *Curr. Opin. Chem. Biol.* **65** 18–27
[22] Yang K K, Wu Z, Bedbrook C N and Arnold F H 2018 Learned protein embeddings for machine learning *Bioinformatics* **34** 2642–8
[23] Anderson B, Hy T S and Kondor R 2019 Cormorant: covariant molecular neural networks *Advances in Neural Information Processing Systems* vol 32, ed H Wallach, H Larochelle, A Beygelzimer, F Alché-Buc, E Fox and R Garnett
[24] Townshend R J *et al* 2020 ATOM3D: tasks on molecules in three dimensions (arXiv:2012.04035)
[25] Jing B, Eismann S, Soni P N and Dror R O 2021 Equivariant graph neural networks for 3D macromolecular structure (arXiv:2106.03843)
[26] Jing B, Eismann S, Suriana P, Townshend R J L and Dror R 2021 Learning from protein structure with geometric vector perceptrons *Int. Conf. on Learning Representations*
[27] Zhao C, Liu T and Wang Z 2022 PANDA2: protein function prediction using graph neural networks *NAR Genom. Bioinf.* **4** 004
[28] Madani A *et al* 2023 Large language models generate functional protein sequences across diverse families *Nat. Biotechnol.* **41** 1–8
[29] Lin Z *et al* 2022 Language models of protein sequences at the scale of evolution enable accurate structure prediction *bioRxiv Preprint* (posted online 21 July 2022) (available at: www.science.org/doi/10.1126/science.ade2574)
[30] Guan J, Qian W W, Peng X, Su Y, Peng J and Ma J 2023 3D equivariant diffusion for target-aware molecule generation and affinity prediction *Int. Conf. on Learning Representations*
[31] Peng X, Luo S, Guan J, Xie Q, Peng J and Ma J 2022 Pocket2Mol: efficient molecular sampling based on 3D protein pockets *Proc. 39th Int. Conf. on Machine Learning* (*Proc. Machine Learning Research* vol 162) ed K Chaudhuri, S Jegelka, L Song, C Szepesvari, G Niu and S Sabato pp 17644–55
[32] Liu M, Luo Y, Uchino K, Maruhashi K and Ji S 2022 Generating 3D molecules for target protein binding *Proc. 39th Int. Conf. on Machine Learning* (*Proc. Machine Learning Research* vol 162) ed K Chaudhuri, S Jegelka, L Song, C Szepesvari, G Niu and S Sabato pp 13912–24
[33] Scantlebury J *et al* 2023 A small step toward generalizability: training a machine learning scoring function for structure-based virtual screening *J. Chem. Inf. Model.* **63** 2960–74
[34] Nascimento A C, Prudêncio R B and Costa I G 2016 A multiple kernel learning algorithm for drug-target interaction prediction *BMC Bioinform.* **17** 1–16
[35] He T, Heidemeyer M, Ban F, Cherkasov A and Ester M 2017 Simboost: a read-across approach for predicting drug–target binding affinities using gradient boosting machines *J. Cheminform.* **9** 24
[36] Öztürk H, Özgür A and Ozkirimli E 2018 DeepDTA: deep drug-target binding affinity prediction *Bioinformatics* **34** 821–9
[37] Zhao L, Wang J, Pang L, Liu Y and Zhang J 2020 GANsDTA: predicting drug-target binding affinity using GANs *Front. Genet.* **10** 1243
[38] Nguyen T, Le H, Quinn T P, Nguyen T, Le T D and Venkatesh S 2020 GraphDTA: predicting drug-target binding affinity with graph neural networks *Bioinformatics* **37** 1140–7
[39] Voitsitskyi T *et al* 2023 3DProtDTA: a deep learning model for drug-target affinity prediction based on residue-level protein graphs *RSC Adv.* **13** 10261–72
[40] Merz Jr. K M, De Fabritiis G and Wei G-W 2020 Generative models for molecular design *J. Chem. Inf. Model.* **60** 5635–6
[41] Gómez-Bombarelli R, Wei J N, Duvenaud D, Hernández-Lobato J M, Sánchez-Lengeling B, Sheberla D, Aguilera-Iparraguirre J, Hirzel T D, Adams R P and Aspuru-Guzik A 2018 Automatic chemical design using a data-driven continuous representation of molecules *ACS Cent. Sci.* **4** 268–76
[42] Segler M H S, Kogej T, Tyrchan C and Waller M P 2018 Generating focused molecule libraries for drug discovery with recurrent neural networks *ACS Cent. Sci.* **4** 120–31
[43] Gao K, Nguyen D D, Tu M and Wei G-W 2020 Generative network complex for the automated generation of drug-like molecules *J. Chem. Inf. Model.* **60** 5682–98

[44] Kusner M J, Paige B and Hernández-Lobato J M 2017 Grammar variational autoencoder *Proc. 34th Int. Conf. on Machine Learning* (*Proc. Machine Learning Research* vol 70) ed D Precup and Y W Teh pp 1945–54
[45] Dai H, Tian Y, Dai B, Skiena S and Song L 2018 Syntax-directed variational autoencoder for structured data *Int. Conf. on Learning Representations*
[46] Thiede E H, Hy T S and Kondor R 2020 The general theory of permutation equivarant neural networks and higher order graph variational encoders (arXiv:2004.03990)
[47] Hy T S and Kondor R 2023 Multiresolution equivariant graph variational autoencoder *Mach. Learn.: Sci. Technol.* **4** 015031
[48] Krenn M, Háse F, Nigam A, Friederich P and Aspuru-Guzik A 2020 Self-referencing embedded strings (selfies): a 100% robust molecular string representation *Mach. Learn.: Sci. Technol.* **1** 045024
[49] Harvey W, Naderiparizi S and Wood F 2022 Conditional image generation by conditioning variational auto-encoders *Int. Conf. on Learning Representations*
[50] Gilmer J, Schoenholz S S, Riley P F, Vinyals O and Dahl G E 2017 Neural message passing for quantum chemistry *Proc. 34th Int. Conf. on Machine Learning* (*ICML'17*) vol 70 pp 1263–72
[51] Dwivedi V P, Rampášek L, Galkin M, Parviz A, Wolf G, Luu A T and Beaini D 2022 Long range graph benchmark *Advances in Neural Information Processing Systems* vol 35, ed S Koyejo, S Mohamed, A Agarwal, D Belgrave, K Cho and A Oh pp 22326–40
[52] Ngo N K, Hy T S and Kondor R 2023 Multiresolution graph transformers and wavelet positional encoding for learning long-range and hierarchical structures *J. Chem. Phys.* **159** 034109
[53] Chen D, Lin Y, Li W, Li P, Zhou J and Sun X 2020 Measuring and relieving the over-smoothing problem for graph neural networks from the topological view *Proc. AAAI Conf. on Artificial Intelligence* vol 34 pp 3438–45
[54] Topping J, Giovanni F D, Chamberlain B P, Dong X and Bronstein M M 2022 Understanding over-squashing and bottlenecks on graphs via curvature *Int. Conf. on Learning Representations*
[55] Kim J, Nguyen D T, Min S, Cho S, Lee M, Lee H and Hong S 2022 Pure transformers are powerful graph learners *Advances in Neural Information Processing Systems* ed A H Oh, A Agarwal, D Belgrave and K Cho
[56] Cai C, Hy T S, Yu R and Wang Y 2023 On the connection between mpnn and graph transformer *Int. Conf. of Machine Learning*
[57] Liu Z, Su M, Han L, Liu J, Yang Q, Li Y and Wang R 2017 Forging the basis for developing protein–ligand interaction scoring functions *Acc. Chem. Res.* **50** 302–9
[58] Koh H Y, Nguyen A T, Pan S, May L T and Webb G I 2023 PSICHIC: physicochemical graph neural network for learning protein-ligand interaction fingerprints from sequence data *bioRxiv Preprint* https://doi.org/10.1101/2023.09.17.558145 (posted online 21 July 2022)
[59] Mysinger M M, Carchia M, Irwin J J and Shoichet B K 2012 Directory of useful decoys, enhanced (dud-e): better ligands and decoys for better benchmarking *J. Med. Chem.* **55** 6582–94
[60] Zhang C-H *et al* 2021 Potent noncovalent inhibitors of the main protease of SARS-CoV-2 from molecular sculpting of the drug perampanel guided by free energy perturbation calculations *ACS Cent. Sci.* **7** 467–75
[61] Gao W, Fu T, Sun J and Coley C 2022 Sample efficiency matters: a benchmark for practical molecular optimization *Advances in Neural Information Processing Systems* vol 35, ed S Koyejo, S Mohamed, A Agarwal, D Belgrave, K Cho and A Oh pp 21342–57
[62] O'Boyle N M, Banck M, James C A, Morley C, Vandermeersch T and Hutchison G R 2011 Open babel: an open chemical toolbox *J. Cheminf.* **3** 1–14
[63] Kingma D P and Welling M 2013 Auto-encoding variational bayes (arXiv:1312.6114)
[64] Sohn K, Lee H and Yan X 2015 Learning structured output representation using deep conditional generative models *Advances in Neural Information Processing Systems* vol 28, ed C Cortes, N Lawrence, D Lee, M Sugiyama and R Garnett
[65] Zheng C, Cham T-J and Cai J 2019 Pluralistic image completion *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*
[66] Ivanov O, Figurnov M and Vetrov D 2019 Variational autoencoder with arbitrary conditioning *Int. Conf. on Learning Representations*
[67] Wan Z, Zhang J, Chen D and Liao J 2021 High-fidelity pluralistic image completion with transformers (arXiv:2103.14031)