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

Võ Tuấn Kiệt

Ngày 22 tháng 5 năm 2024

# Research Context

The first stage of drug discovery is design novel drug-like compounds with high binding affinities to target proteins. This process consist of two sub-task :

- Searching for candidates
- Measuring drug-target affinities (DTA)

However, the method that involves atomistic molecular dynamics simulations are computationally expensive and time-consuming, making them infeasible for large-scale sets of protein-ligand complexes.
**Question** : Are there any methods that accelerate and automate these tasks using computational methods and machine-learning techniques ?

# DGMs

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.
**Disadvantages**: slow, requires specific property networks to be trained for each target protein.
**Solutions** : incorporate prior knowledge of protein structures, but several existing studies rely solely on the information of specific binding sites of target proteins and are limited when the binding sites are not determined.

# Proteins Representation Structure

Proteins are macromolecules that can be represented in term of :

- Sequences of amino acids (i.e. primary structure)
- 2D graphs at residue level constructed by neares neighbors from folding information(i.e. tertiary structure)
- 3D point clouds at atom level

# Proteins Representation Learning

- Sequence-based methods : protein sequence is regarded as a long sequence of tokens (i.e.k-mers) that are fed to a transformer-based language model.

  **Advantage**: capture the relationships among distant residues in a long protein sequence

  **Disadvantage**: not able to exploit the geometric relations among residues
- GNNs and CNNs-based : operate on relational and geometric structures of proteins.

  **Advantage**: can learn spatial information about protein structures

  **Disadvantage**: limit in capturing long-range interactions in large protein structures.
- In recent years, pre-trained large language models for scientific discovery have achieved remarkable results in protein science.

Question : Are there any methods that apply language models and also facilitate the disadvantages of methods discussed above ?

# Contributions

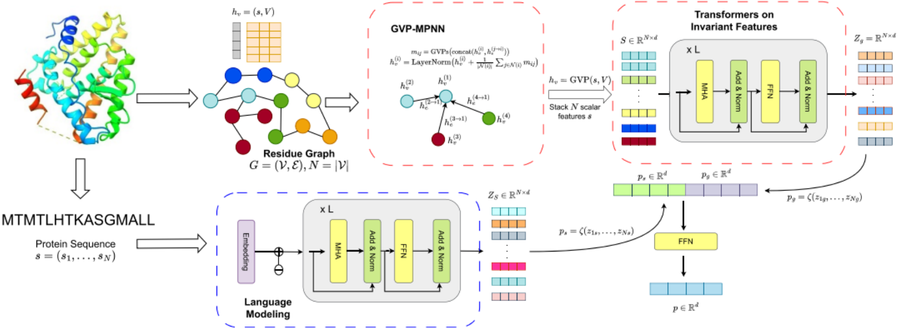
- **TargetVAE** :
  - generate chemically valid, drug-like molecules with high binding scores to an arbitrarily given protein structure.
  - directly condition on the entire structure of any protein target
  - generate ligard candidates with high binding affinity without prior knowledge of any binding site
- **Protein Multimodal Network(PMN)**
  - unifies different modalities of protein.
  - produce a protein embedding that can serve as the prior for generative model (i.e TargetVAE).
  - accurately estimate protein-lingand biding affinity.
  - replace the computationally expense in the evaluation of ligand generation.

# PNN

Proteins can be seen as long-range graphs.

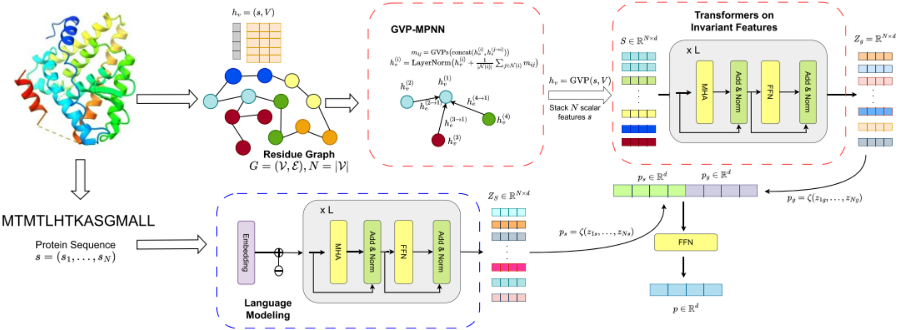
Encode both sequences and 3D graphs of residues then combine them to create a unified representation for large protein.

Transformer-based model can efficiently capture a protein's local and global information.



$h_v = (s, V)$
Residue Graph
$G = (\mathcal{V}, \mathcal{E}), N = |\mathcal{V}|$
GVP-MPNN
$m_{ij} = \mathrm{GVPs}(\mathrm{concat}(h_v^{(i)}, h_e^{(j \to i)}))$
$h_v^{(i)} = \mathrm{LayerNorm}(h_v^{(i)} + \frac{1}{|\mathcal{N}(i)|} \sum_{j \in \mathcal{N}(i)} m_{ij})$
$h_v^{(1)}$
$h_v^{(2)}$
$h_v^{(3)}$
$h_v^{(4)}$
$h_e^{(2 \to 1)}$
$h_e^{(3 \to 1)}$
$h_e^{(4 \to 1)}$
$h_v = \mathrm{GVP}(s, V)$
Stack $N$ scalar features $s$
Transformers on Invariant Features
$S \in \mathbb{R}^{N \times d}$
x L
MHA
Add & Norm
FFN
Add & Norm
$Z_g = \mathbb{R}^{N \times d}$
$p_s \in \mathbb{R}^d$
$p_g \in \mathbb{R}^d$
$p_g = \zeta(z_{1_g}, \ldots, z_{N_g})$
MTMTLHTKASGMALL
Protein Sequence
$s = (s_1, \ldots, s_N)$
Embedding
x L
MHA
Add & Norm
FFN
Add & Norm
Language Modeling
$Z_S \in \mathbb{R}^{N \times d}$
$p_s = \zeta(z_{1_s}, \ldots, z_{N_s})$
FFN
$p \in \mathbb{R}^d$

# Model Structure


$h_v = (s, V)$
Residue Graph
$G = (\mathcal{V}, \mathcal{E}), N = |\mathcal{V}|$
GVP-MPNN
$m_{ij} = \mathrm{GVPs}(\mathrm{concat}(h_v^{(i)}, h_e^{(j\to i)}))$
$h_v^{(i)} = \mathrm{LayerNorm}(h_v^{(i)} + \frac{1}{|\mathcal{N}(i)|} \sum_{j\in\mathcal{N}(i)} m_{ij})$
$h_v^{(2)}$
$h_v^{(1)}$
$h_e^{(2\to1)}$
$h_e^{(4\to1)}$
$h_e^{(3\to1)}$
$h_v^{(4)}$
$h_v^{(3)}$
$h_v = \mathrm{GVP}(s, V)$
Stack $N$ scalar features $s$
Transformers on Invariant Features
$S \in \mathbb{R}^{N\times d}$
x L
MHA
Add & Norm
FFN
Add & Norm
$Z_g = \mathbb{R}^{N\times d}$
MTMTLHTKASGMALL
Protein Sequence
$s = (s_1, \ldots, s_N)$
Embedding
x L
MHA
Add & Norm
FFN
Add & Norm
Language Modeling
$Z_S \in \mathbb{R}^{N\times d}$
$p_s = \zeta(z_{1s}, \ldots, z_{Ns})$
$p_s \in \mathbb{R}^d$
$p_g \in \mathbb{R}^d$
$p_g = \zeta(z_{1g}, \ldots, z_{Ng})$
FFN
$p \in \mathbb{R}^d$

# Local Message-Passing

Replace dense layers by GVPs in MPNN to operate invariant features

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

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

# Global Interaction

1. Utilize GVP module to update $h_v = (s, V)$ as $(s^{'}, V^{'}) = GVP((s, V))$
2. Pass $S \in \mathbb{R}^{N \times d}$(row i indicates d-dimentional scalar feature $s_i$ of node $i$) to a L-layer Transformers encoder :

$$Q_l = Z_{l-1} W_l^Q, K_l = Z_{l-1} W_l^K, V_l = Z_{l-1} W_l^V, \quad (1)$$

$$H_l = MultiheadAttention(Q_l, K_l, V_l), \quad (2)$$

$$Z_l = LayerNorm(Z_{l-1} + FFN(H_l)) \quad (3)$$

$Z_0 \triangleq S, \{W_l^Q, W_l^K, W_l^V\}_{l=1}^L \in \mathbb{R}^{d \times d_k}, Z_g \triangleq Z_l$ denote the final node embeddings.

3. aggregate node embeddings by row-wise Aggregator $p_g = \zeta(Z_g) \in \mathbb{R}^d$.

# Language Modeling On Protein Sequence

A protein can be represented as a sequence $s = (s_0, s_1, ..., s_n)$, $s_i \in \mathbb{R}^{20}$ is a one-hot vector indicates one in a total of 20 types of residues.

Ultilize Transformer-based language model to compute the text representaion of this protein sequence with initial embeddings $Z_0 = [z_1, z_2, ..., z_n] \in \mathbb{R}^{n \times d}$, $z_i \in \mathbb{R}^d = Embedd(s_i) + p_i$, $p_i$ is the positional encoding.

Use the pre-trained Transformer protein language models proposed by Lin to extract protein-level embeddings for each protein

$p_s = \zeta(Z_s) \in \mathbb{R}^d$ is the global representaion for entire protein sequence

# Multi-modal Fusion

Concatenate geometric and sequential features and process them by a feed-forward network : $p = FFN(p_s \oplus p_g)$

# Binding Affinity Prediction

**Goal**: Predict the binding affinities between ligands and their target proteins.

Represent small molecules as 2D graphs G=(V,E) and use GAT,MPNN to learn their representaion.At the layer t, embedding vector of node $i \in V$ :

$$x_i^t = \alpha_{i,i} W x_i^{t-1} + \sum_{j \in (i)} \alpha_{i,j} W x_j^{t-1},$$

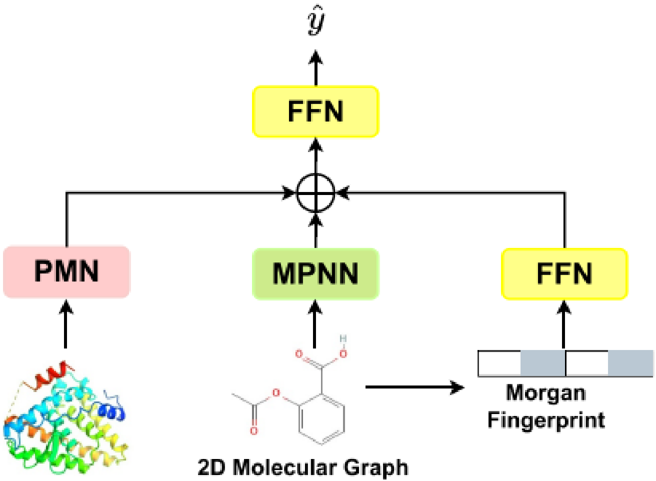
$$a_{i,j} = \frac{exp(\sigma(a^T[Wx_i \oplus Wx_j]))}{\sum_{k \in N(i) \bigcup j} exp(\sigma(a^T[Wx_i \oplus Wx_k]))}$$

Combine features from all sources, including sequence, geometric information of proteins, and grapp-based feature ligands, then pass them to FFN to make predictions :

$$h = p \oplus l$$

$$\hat{y} = \phi(h)$$

# Binding Affinity Prediction



# Target-Aware Ligand Generation

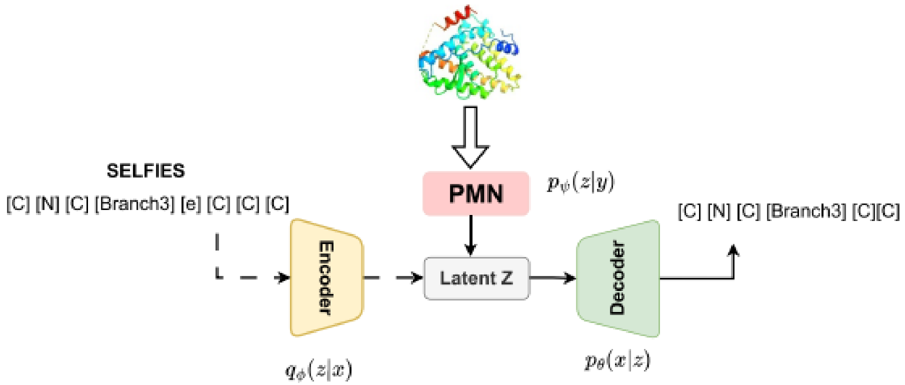
Generate ligands $\hat{l} = (\hat{l}_1, \hat{l}_2, ..., \hat{l}_n)$ by computing n independent probability vectors $y = (y_1, y_2, ...y_n), y_i \in \mathbb{R}^{|S|}$. $\hat{l}_i = S_j, j = argmax_{0 \leqslant j < |S|}(y_i)$ $\phi, \theta, \omega$ denote the encoder,decoder,an prior network in a conditional VAE framework.

- $\phi$ : RNN that endoes a molecule into a latent $z \in \mathbb{R}^d$
- $\theta$ : RNN that autoregressively generate each token of SELFIES string
- $\psi$ : Protein multi-modal network that encode both sequence and geometric information of target protein.

Use two MLP $\mu_\omega, \sigma_\psi$ to compute the mean and variance of a Gaussian distribution over the latent vector $z_l \sim \mathbb{N}(\mu_\psi(p), \sigma_\psi(p)$, which is the inferred latent embeddings of the ligand l.
Both conditional and unconditional VAE are optimized based on equation :

$$log(p_{\theta,\psi})(x|p) \geq O_{for} \triangleq \mathbb{E}_{q_\phi}[log(p_{\theta,\psi})(x|z)] - KL[q_\phi(z|p)||p_\psi(z|p)]$$

# Target-Aware Ligand Generation



# Experiments

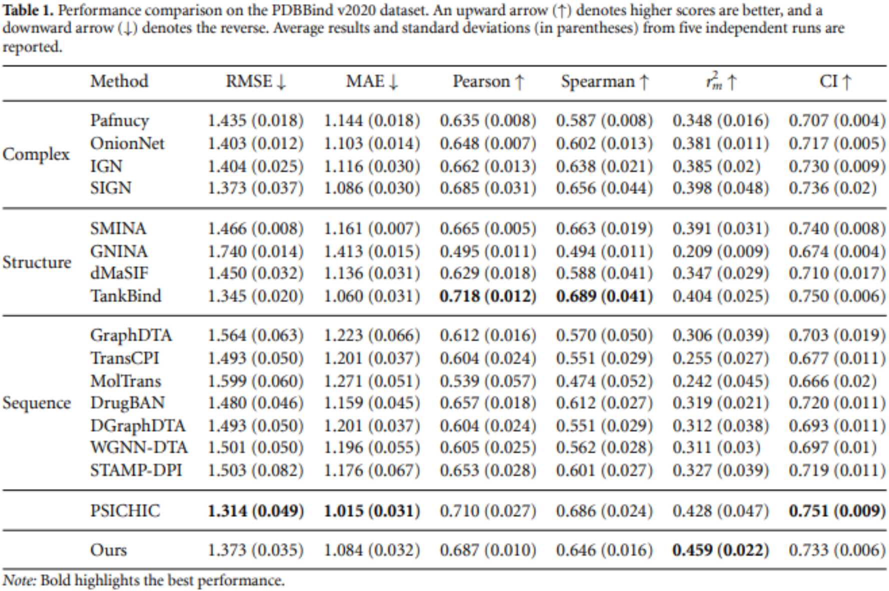


**Table 1.** Performance comparison on the PDBBind v2020 dataset. An upward arrow (↑) denotes higher scores are better, and a downward arrow (↓) denotes the reverse. Average results and standard deviations (in parentheses) from five independent runs are reported.

| | 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.

# Experiments

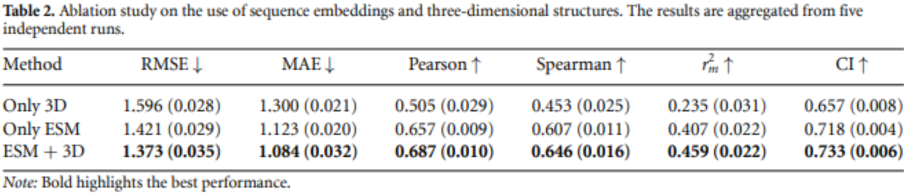


**Table 2.** Ablation study on the use of sequence embeddings and three-dimensional structures. The results are aggregated from five independent runs.

| 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.

# Experiments

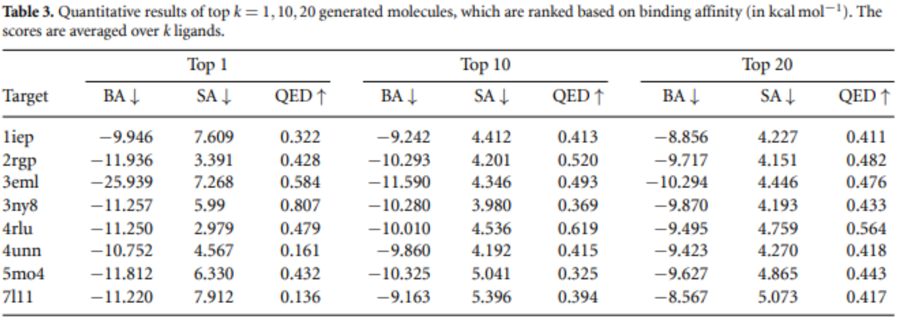


**Table 3.** Quantitative results of top $k = 1, 10, 20$ generated molecules, which are ranked based on binding affinity (in kcal mol$^{-1}$). The scores are averaged over $k$ ligands.

| | 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 |