Geometric Deep Learning for Protein Science and Drug Discovery

Dr. Truong Son Hy

Assistant Professor
Department of Mathematics and Computer Science, Indiana State University

Introduction

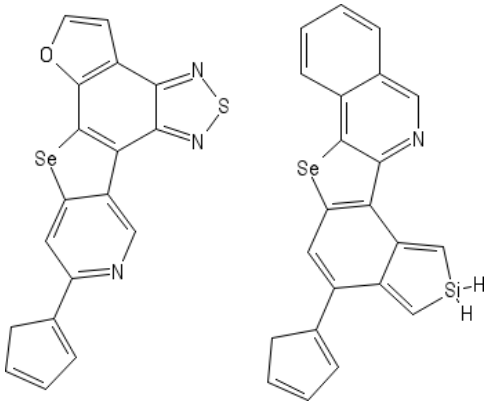
HySonLab – AI for Science

Advanced machine learning and deep learning for scientific problems.

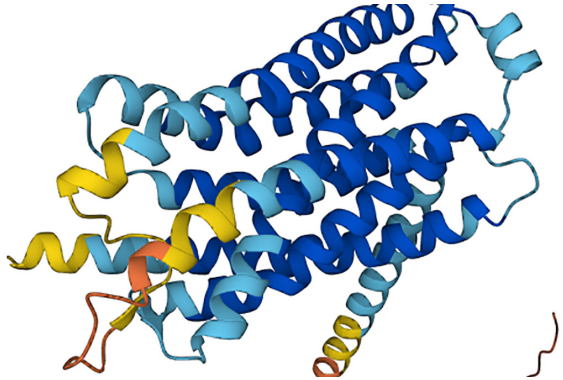
For our talk today:

1. Introduction to Geometric Deep Learning
   - Graph neural networks & Equivariant neural networks
   - Limitations of GNNs and our solutions
2. Multiresolution Graph Transformers
   - Hierarchical and long-range interactions modeling
   - Protein, peptide and polymer properties prediction
3. Protein Multimodal Representation Learning, LLM & Generative AI
   - Unify different modalities of protein representations into a single model
   - Generate ligands with high binding affinity
   - Unsupervised symmetry-preserving multimodal pretraining
4. Directed Evolution for Protein Optimization

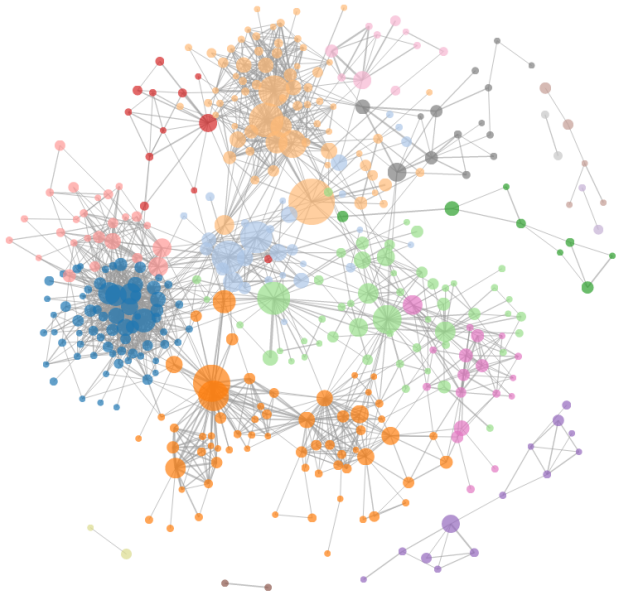
Motivation for graph learning & geometric graphs



Molecules



Macro-Molecules



Citation network

Knowledge Graph

Graph Neural Networks (GNNs)

DFT = Density Functional Theory

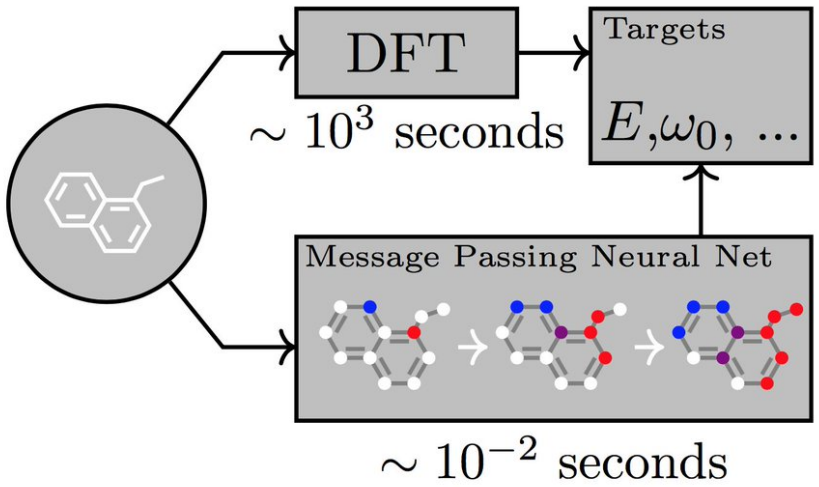


Figure 1. A Message Passing Neural Network predicts quantum properties of an organic molecule by modeling a computationally expensive DFT calculation.

Figure taken from (Gilmer et al., 2017).

Our works:

- Predicting molecular properties with covariant compositional networks, Journal of Chemical Physics, Volume 148, Issue 24
- Cormorant: Covariant Molecular Neural Networks, NeurIPS 2019

Message Passing Neural Networks (MPNN)

Given an input graph / network $G = (V, E)$:

1. Initially, each vertex v of the graph is associated with a feature representation ℓ_v (label) or f_v^0. This feature representation can also be called as a *message*.

Message Passing Neural Networks (MPNN)

Given an input graph / network $G = (V, E)$:

1. Initially, each vertex v of the graph is associated with a feature representation ℓ_v (label) or f_v^0. This feature representation can also be called as a *message*.
2. Iteratively, at iteration t, each vertex collects / aggregates all messages of the previous iteration $\{f_{v_1}^{t-1}, .., f_{v_k}^{t-1}\}$ from other vertices in its neighborhood $\mathcal{N}(v) = \{v_1, .., v_k\}$, and then produces a new message f_v^t via some *hashing function* $\psi(.)$.

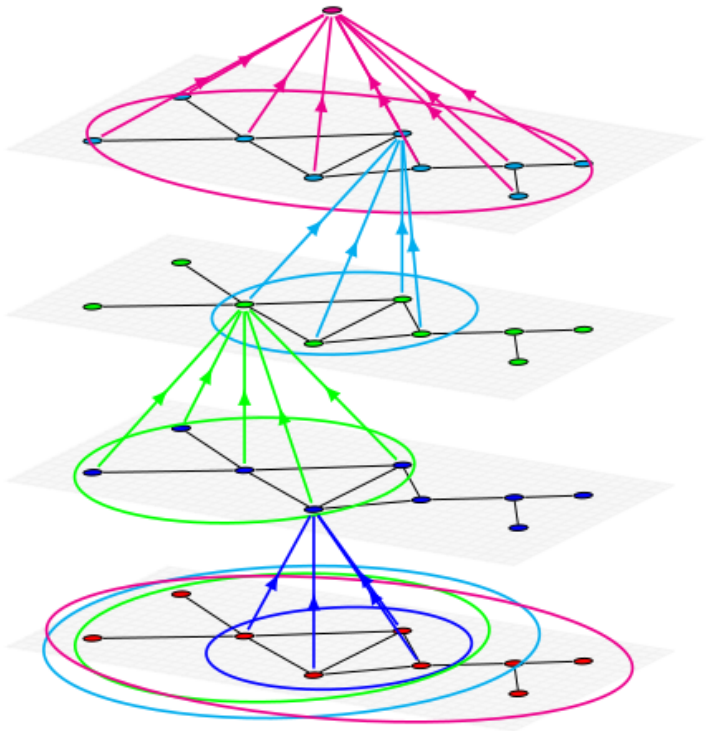
Message Passing Neural Networks (MPNN)

Given an input graph / network $G = (V, E)$:

1. Initially, each vertex v of the graph is associated with a feature representation ℓ_v (label) or f_v^0. This feature representation can also be called as a *message*.
2. Iteratively, at iteration t, each vertex collects / aggregates all messages of the previous iteration $\{f_{v_1}^{t-1}, .., f_{v_k}^{t-1}\}$ from other vertices in its neighborhood $\mathcal{N}(v) = \{v_1, .., v_k\}$, and then produces a new message f_v^t via some *hashing function* $\psi(.)$.
3. The graph representation $\phi(G)$ is obtained by aggregating all messages in the last iteration of every vertex. $\phi(G)$ is then used for downstream application.

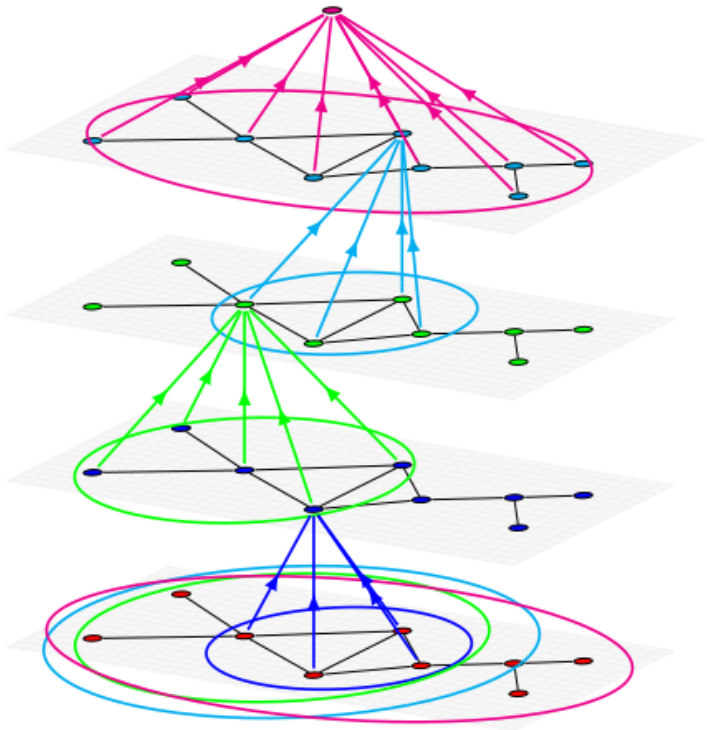
Message Passing Neural Networks (MPNN)

1: **for** $v \in V$ **do**
2: $\quad f_v^0 \leftarrow \ell_v$
3: **end for**
4: **for** $t = 1 \rightarrow T$ **do**
5: $\quad$ **for** $v \in V$ **do**
6: $\quad\quad f_v^t \leftarrow \psi\left(\{f_i^{t-1}\}_{i \in \mathcal{N}(v)}\right)$
7: $\quad$ **end for**
8: **end for**
9: $\phi(G) \leftarrow \psi\left(\{f_v^T\}_{v \in V}\right)$



Message Passing Neural Networks (MPNN)

1: **for** $v \in V$ **do**
2: $\quad f_v^0 \leftarrow \ell_v$
3: **end for**
4: **for** $t = 1 \rightarrow T$ **do**
5: $\quad$ **for** $v \in V$ **do**
6: $\quad\quad f_v^t \leftarrow \psi\left(\{f_i^{t-1}\}_{i \in \mathcal{N}(v)}\right)$
7: $\quad$ **end for**
8: **end for**
9: $\phi(G) \leftarrow \psi\left(\{f_v^T\}_{v \in V}\right)$



Note

This procedure is used in Weisfeiler–Lehman **graph isomorphism** test (NP-complete problem).

Message Passing Neural Networks (MPNN)

With learnable parameters:

for $v \in V$ **do**
  $f_v^0 \leftarrow \ell_v$
end for
for $t = 1 \rightarrow T$ **do**
  for $v \in V$ **do**
    $f_v^t \leftarrow \psi\left(\{f_i^{t-1}\}_{i \in \mathcal{N}(v)}; W^t\right)$
  end for
end for
$\phi(G) \leftarrow \psi\left(\{f_v^T\}_{v \in V}; W^{T+1}\right)$

Message Passing Neural Networks (MPNN)

With learnable parameters:

1: **for** $v \in V$ **do**
2: $\quad f_v^0 \leftarrow \ell_v$
3: **end for**
4: **for** $t = 1 \rightarrow T$ **do**
5: $\quad$ **for** $v \in V$ **do**
6: $\quad\quad f_v^t \leftarrow \psi\left(\{f_i^{t-1}\}_{i \in \mathcal{N}(v)}; W^t\right)$
7: $\quad$ **end for**
8: **end for**
9: $\phi(G) \leftarrow \psi\left(\{f_v^T\}_{v \in V}; W^{T+1}\right)$

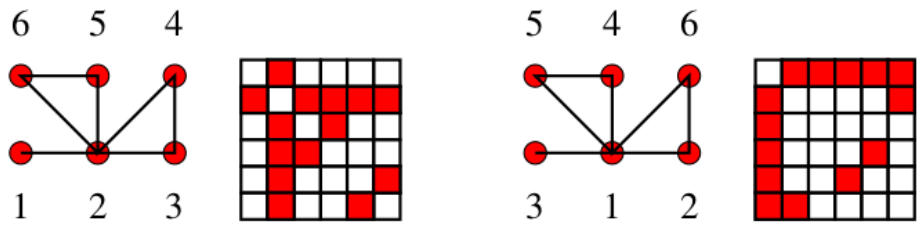
Given a graph properties $y(G) \in \mathbb{R}^d$ to regress, we have the optimization:

$$\min_{\{W^t\}_{t=1}^{T+1}} ||y(G) - \phi(G)||_2^2$$

The gradient with respect to $\{W^t\}_{t=1}^{T+1}$ can be computed via Back-propagation.

Invariance

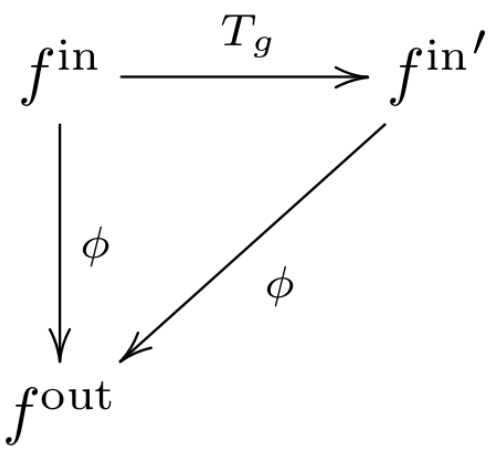
We renumber the vertices by a permutation $\sigma : \{1, 2, .., 6\} \mapsto \{1, 2, .., 6\}$. The adjacency matrices of G (left) and G' (right) are different, but topologically they represent the same graph:

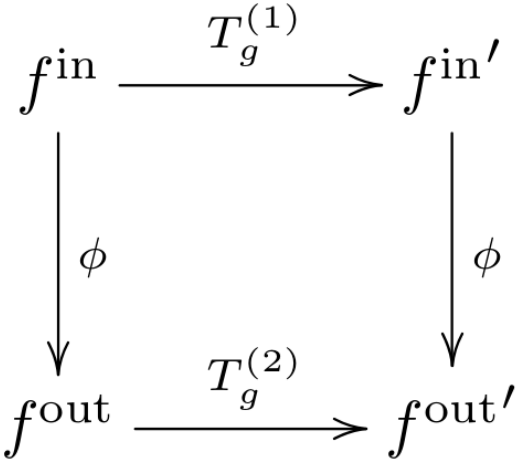


Therefore, ϕ must be **invariant** wrt permutation, i.e. $\phi(G) = \phi(G')$.

Invariance vs. Equivariance

T_g is an action of a group G on the space of inputs and outputs.





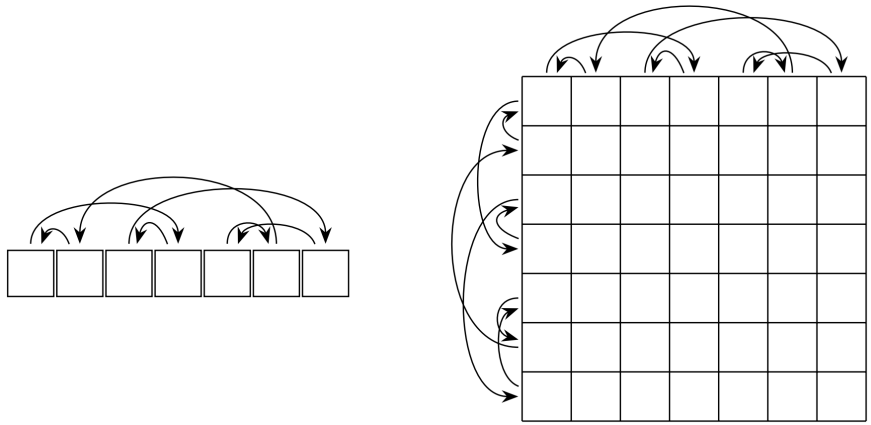
Invariance: $\phi(T_g(f)) = \phi(f)$

Equivariance: $\phi(T_g^{(1)}(f)) = T_g^{(2)}(\phi(f))$

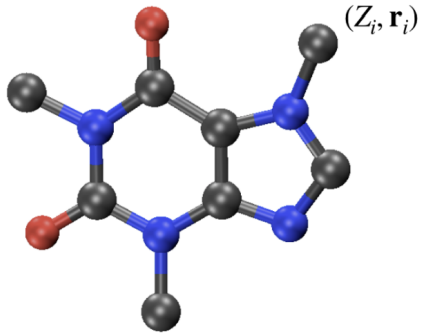
Our works:

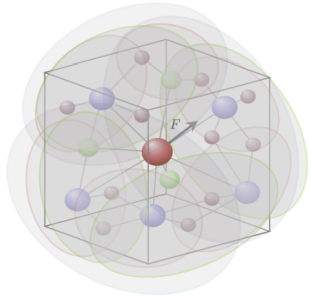
- Predicting molecular properties with covariant compositional networks, Journal of Chemical Physics, Volume 148, Issue 24. $G = \mathbb{S}_n$
- Covariant compositional networks for learning graphs, ICLR 2018. $G = \mathbb{S}_n$
- Cormorant: Covariant Molecular Neural Networks, NeurIPS 2019. $G = \mathrm{SO}(3)$

Symmetry preservation for geometric graphs



A permutation group $\mathbb{S}_n$'s action on node order and adjacency matrix.





$F(\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_m)$

Molecular data specified by a set of charge-position pairs (Z_i, r_i) for each atom. This problem is invariant to rotations and the atomic representation must be SO(3)-equivariant.

Limitations of GNNs and Message Passing

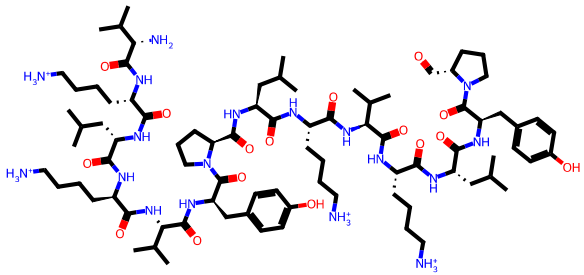
GNNs are powerful but have several limitations theoretically and practically:

- **Long-range** modeling (i.e. graphs with large diameters) [1, 2]
- Modeling highly **symmetric** structures [3, 4]
- Over-smoothing & Over-squashing (we are working on it on the theoretical front!)

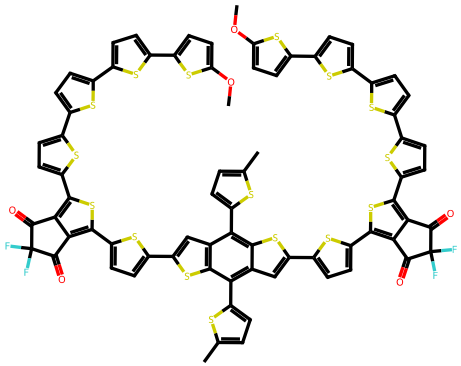
Our works:

1. On the Connection Between MPNN and Graph Transformer, ICML 2023
2. Multiresolution graph transformers and wavelet positional encoding for learning long-range and hierarchical structures, Journal of Chemical Physics, Volume 159, Issue 3
3. Predicting molecular properties with covariant compositional networks, Journal of Chemical Physics, Volume 148, Issue 24
4. Covariant compositional networks for learning graphs, ICLR 2018

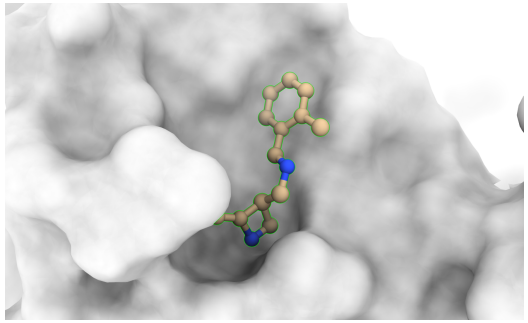
Long-range graphs



(a) Peptide



(b) Polymer

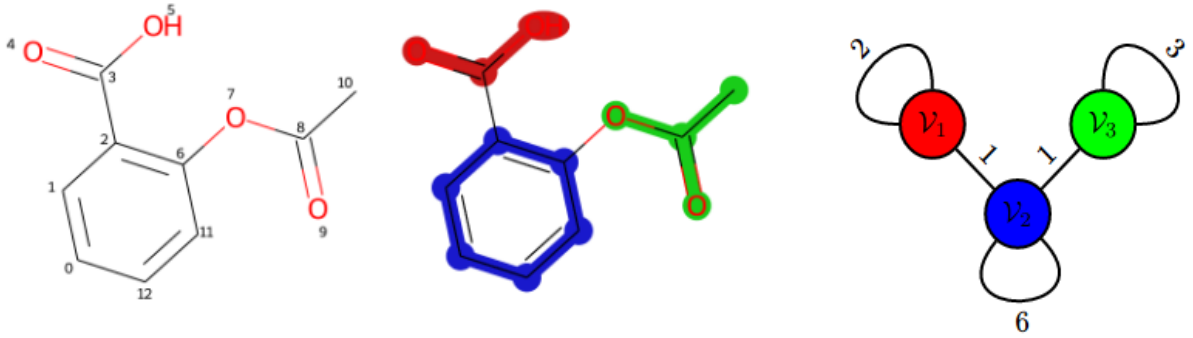


(c) Protein-Ligand

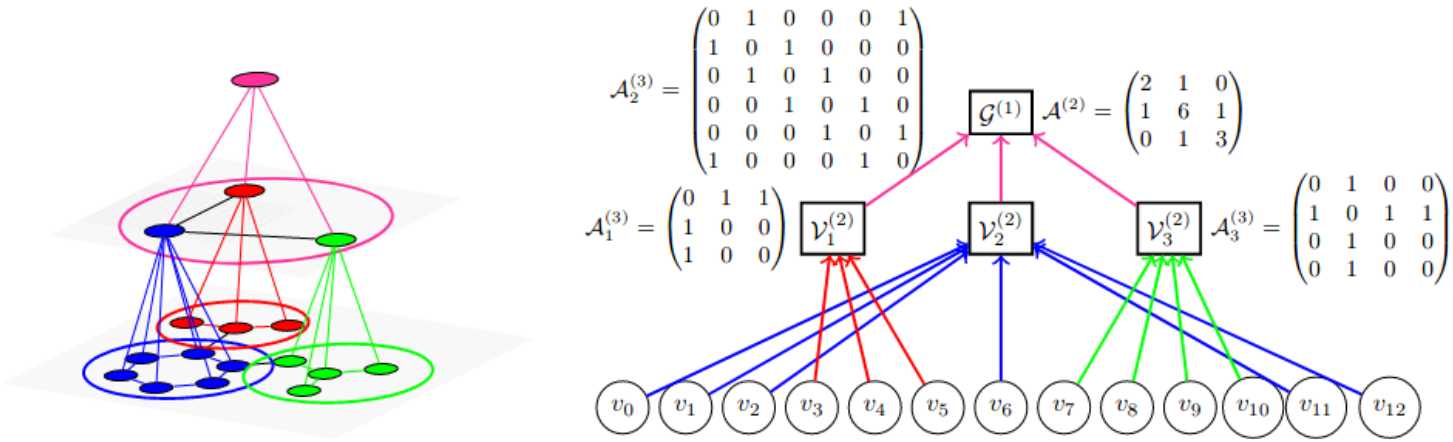
Figure: Macromolecules that are actually long-range graphs.

- Macromolecules have **hierarchical structures** and comprise multiple **long-range** dependencies among distant atoms.
- We want to predict functions of peptides, properties of polymers calculated from Density Functional Theory (DFT), and protein-ligand binding affinity.
- Conventional GNNs **fail** with long-range graphs (I will discuss our solutions and theoretical results!).

Multiresolution Graph Networks (MGN)



Aspirin $C_9H_8O_4$, its 3-cluster partition and the coarsen graph.

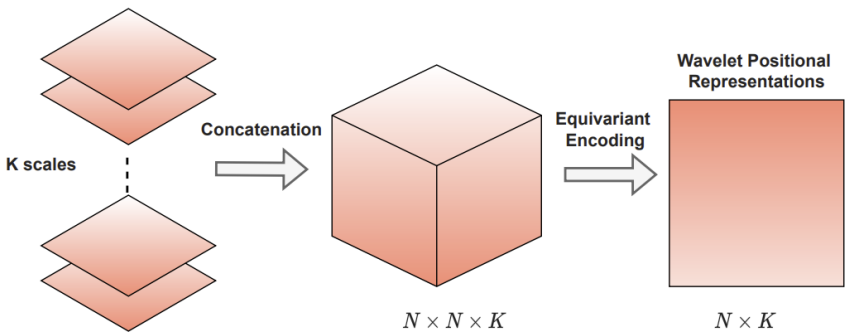


3-level Multiresolution Graph Networks learning to cluster on Aspirin.

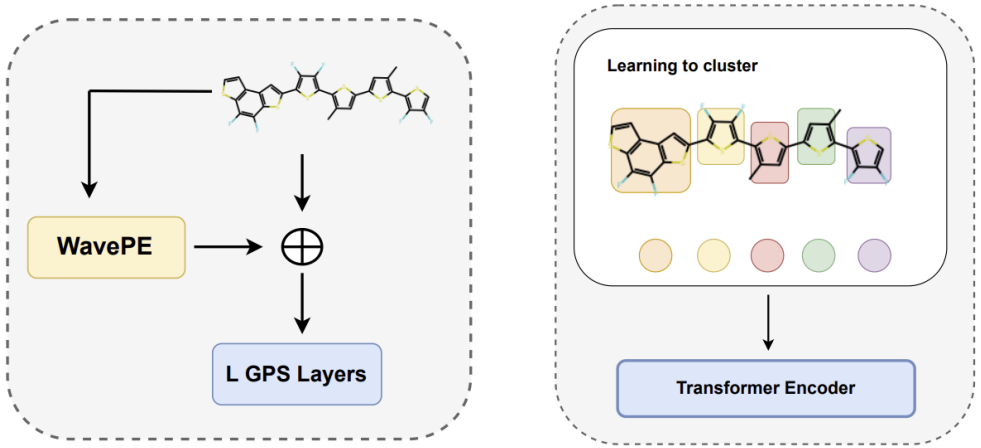
Our work: Multiresolution equivariant graph variational autoencoder, Machine Learning: Science and Technoglogy, Volume 4, Number 1

New proposal for long-range graphs

Wavelet positional encoding



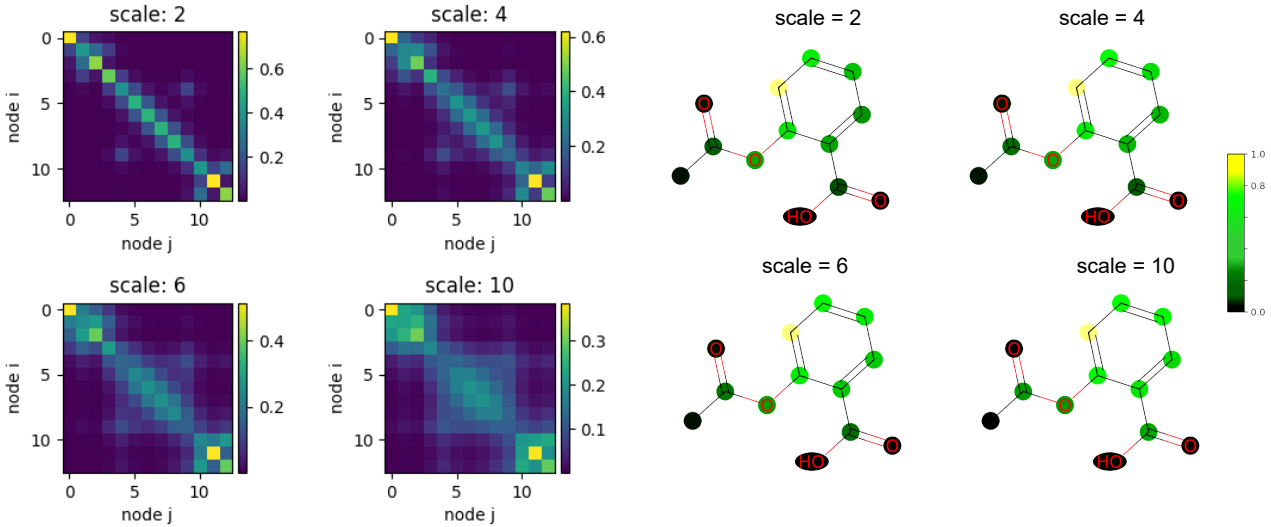
Multiresolution Graph Transformer



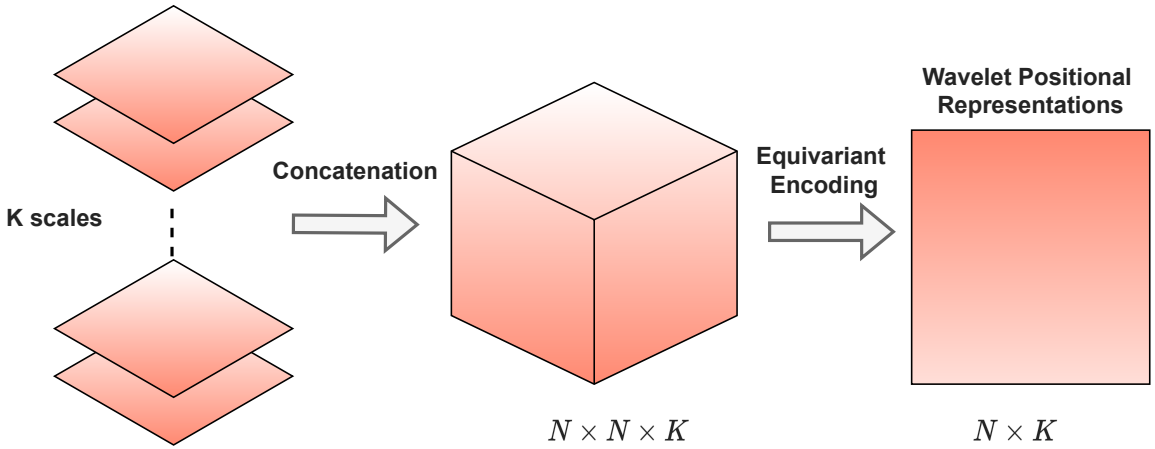
- Transformers are effective for computing the interactions between distant atoms via **self-attention** mechanisms.
- To adapt Transformer-like architectures to graphs, we need **positional encoding** (PE) schemes that embody the local structures.

Our work: Multiresolution Graph Transformers and Wavelet Positional Encoding for Learning Long-Range and Hierarchical Structures, Journal of Chemical Physics, Volume 159, Issue 3

Wavelet Positional Encoding

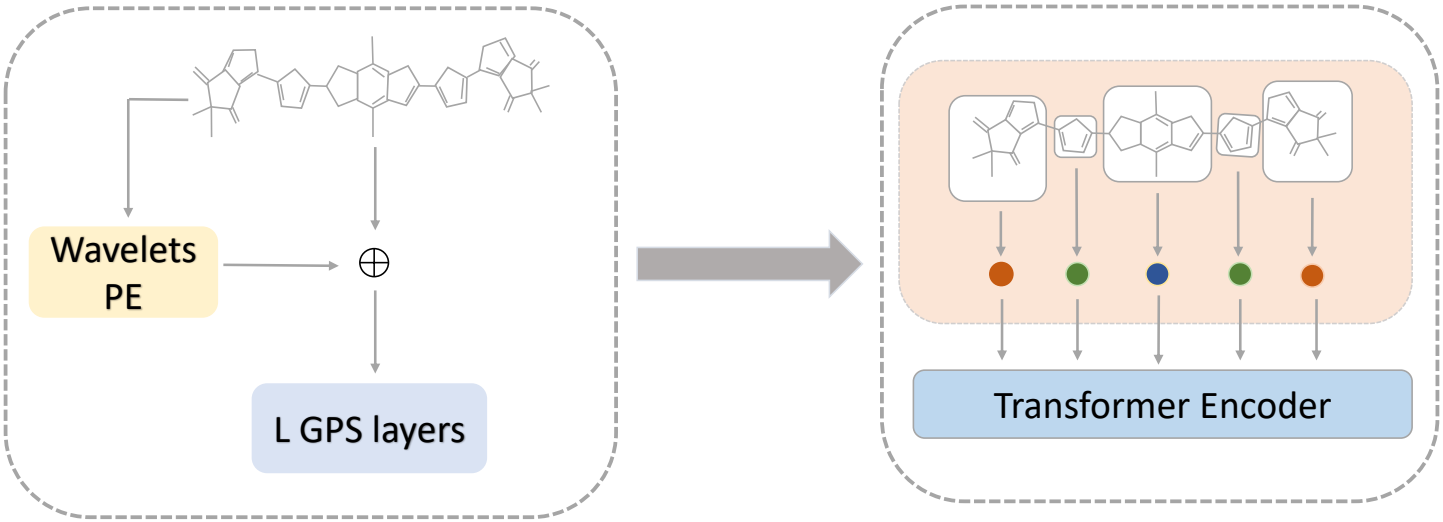


Graph wavelets with scaling parameters on Aspirin molecular graph



Encode the wavelet tensor into node-level PE via $\mathbb{S}_n$-equivariant neural networks

Multiresolution Graph Transformer (MGT)



- We propose a **learning-to-cluster** algorithm that coarsens graphs iteratively to build a multiresolution (i.e. multiple of resolutions) representation of the input graph.
- We employ the **graph transformer** model learning **on each resolution** and we integrate our **wavelet positional encoding**.

Peptides Property Prediction

Model	Params	Peptides-struct	Peptides-func
		MAE ↓	AP ↑
GCN	508k	0.3496 ± 0.0013	0.5930 ± 0.0023
GINE	476k	0.3547 ± 0.0045	0.5498 ± 0.0079
GatedGCN	509k	0.3420 ± 0.0013	0.5864 ± 0.0077
GatedGCN + RWPE	506k	0.3357 ± 0.0006	0.6069 ± 0.0035
Transformer + LapPE	488k	0.2529 ± 0.0016	0.6326 ± 0.0126
GPS	—	0.2500 ± 0.0005	0.6535 ± 0.0041
SAN + LapPE	493k	0.2683 ± 0.0043	0.6384 ± 0.0121
SAN + RWPE	500k	0.2545 ± 0.0012	0.6562 ± 0.0075
MGT + WavePE (ours)	499k	**0.2453 ± 0.0025**	**0.6817 ± 0.0064**

Peptides-func: a multi-label graph **classification** dataset with 10 classes based on the peptide function: Antibacterial, Antiviral, cell-cell communication, etc.

Peptides-struct: a multi-label graph **regression** dataset based on the 3D structure of the peptide: inertia mass, inertia valence, length, sphericity, and plane best fit.

Polymer Property Prediction

We **achieve the chemical accuracy** in estimating the molecular properties of polymers that are calculated from Density Functional Theory, while outperforming all other competitive baselines.

Model	Params	Property		
		GAP	HOMO	LUMO
DFT error		1.2	2.0	2.6
Chemical accuracy		0.043	0.043	0.043
GCN	527k	0.1094 ± 0.0020	0.0648 ± 0.0005	0.0864 ± 0.0014
GCN + Virtual Node	557k	0.0589 ± 0.0004	0.0458 ± 0.0007	0.0482 ± 0.0010
GINE	527k	0.1018 ± 0.0026	0.0749 ± 0.0042	0.0764 ± 0.0028
GINE + Virtual Node	557k	0.0870 ± 0.0040	0.0565 ± 0.0050	0.0524 ± 0.0010
GPS	600k	0.0467 ± 0.0010	0.0322 ± 0.0020	0.0385 ± 0.0006
Transformer + LapPE	700k	0.2949 ± 0.0481	0.1200 ± 0.0206	0.1547 ± 0.0127
MGT + LapPE (ours)	499k	**0.0378 ± 0.0004**	**0.0270 ± 0.0010**	0.0300 ± 0.0006
MGT + RWPE (ours)	499k	0.0384 ± 0.0015	0.0274 ± 0.0005	**0.0290 ± 0.0007**
MGT + WavePE (ours)	499k	0.0387 ± 0.0011	0.0283 ± 0.0004	0.0290 ± 0.0010

HOMO: Highest Occupied Molecular Orbital, energy of the highest occupied electronic state (eV)

LUMO: Lowest Unoccupied Molecular Orbital, energy of the lowest unoccupied electronic state (eV)

GAP: difference / gap between HOMO and LUMO (eV)

Protein-Ligand Binding Affinity Prediction: ATOM3D

Understanding the multiscale structure of protein complexes is important in estimating their fitness and functionality.

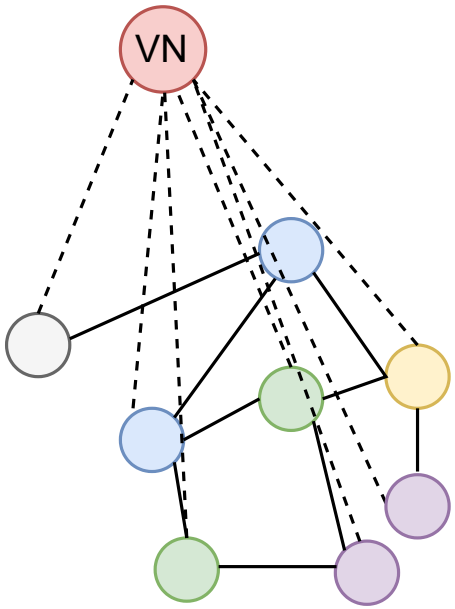
Method	3D-CNN	GNN	ENN	GVP-GNN	MGT + WavePE
RMSE ↓	**1.416 ± 0.021**	1.570 ± 0.025	1.568 ± 0.012	1.594 ± 0.073	1.436 ± 0.066

We show the effectiveness of our model in capturing the long-range and hierarchical structures of proteins. Our multiresolution graph transformer is competitive on ATOM3D benchmark in predicting **protein-ligand binding affinity** (i.e. estimating $pK = -\log(K)$, where K is the binding affinity in Molar units), even without knowing the 3D structure.

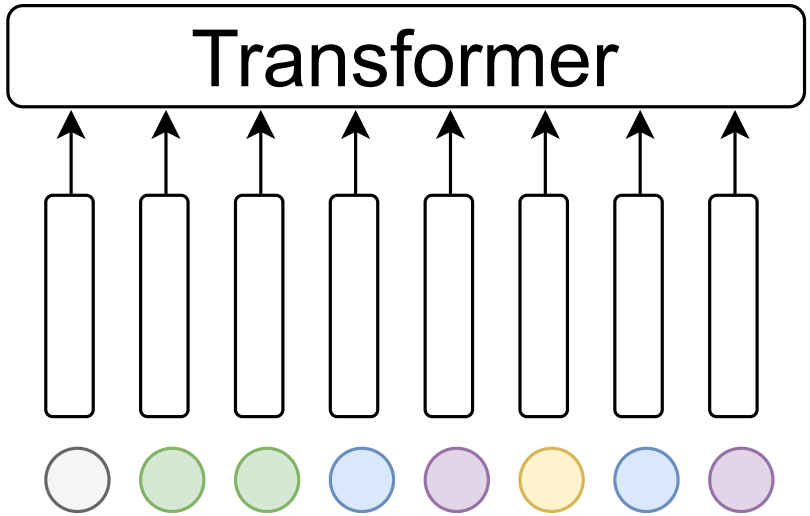
Software & Datasets

https://github.com/HySonLab/Multires-Graph-Transformer

Is there an alternative to Graph Transformer?



(a)



(b)

(a) MPNN + VN = we augment the graph with a virtual node (VN) connecting to all other nodes. VN acts as a "bridge" that reduces the maximum shortest path to 2.

(b) Graph Transformer = we treat each node embedding as a token and apply a Transformer on the sequence of node embeddings/tokens.

Our theoretical results

From our paper **On the Connection Between MPNN and Graph Transformer** at ICML 2023:

Theorem 1

MPNN + VN can simulate (not just approximate) equivariant DeepSets: $\mathbb{R}^{n\times d} \to \mathbb{R}^{n\times d}$. This implies that MPNN + VN of $\mathcal{O}(1)$ depth and $\mathcal{O}(n^d)$ width is permutation equivariant universal, and can approximate self-attention layer and transformers arbitrarily well.

Our theoretical results

From our paper **On the Connection Between MPNN and Graph Transformer** at ICML 2023:

Theorem 1

MPNN + VN can simulate (not just approximate) equivariant DeepSets: $\mathbb{R}^{n\times d} \to \mathbb{R}^{n\times d}$. This implies that MPNN + VN of $\mathcal{O}(1)$ depth and $\mathcal{O}(n^d)$ width is permutation equivariant universal, and can approximate self-attention layer and transformers arbitrarily well.

Theorem 2

Given any graph G of size n with node features $\boldsymbol{X} \in \mathcal{X}$, and a self-attention layer $\boldsymbol{L}$ on G (fix $\boldsymbol{W}_K, \boldsymbol{W}_Q, \boldsymbol{W}_V$), there exists a $\mathcal{O}(n)$ layer of heterogeneous MPNN + VN with the specific aggregate/update/message function that can approximate $\boldsymbol{L}$ on $\mathcal{X}$ arbitrarily well.

MPNN + VN on Long-Range Graph Benchmark (LRGB)

Model	# Params.	Peptides-functional		Peptides-structural	
		Test AP before VN	Test AP after VN ↑	Test MAE before VN	Test MAE after VN ↓
GCN	508k	0.5930±0.0023	0.6623±0.0038	0.3496±0.0013	**0.2488±0.0021**
GINE	476k	0.5498±0.0079	0.6346±0.0071	0.3547±0.0045	0.2584±0.0011
GatedGCN	509k	0.5864±0.0077	0.6635±0.0024	0.3420±0.0013	0.2523±0.0016
GatedGCN+RWSE	506k	0.6069±0.0035	**0.6685±0.0062**	0.3357±0.0006	0.2529±0.0009
Transformer+LapPE	488k	0.6326±0.0126	-	0.2529±0.0016	-
SAN+LapPE	493k	0.6384±0.0121	-	0.2683±0.0043	-
SAN+RWSE	500k	0.6439±0.0075	-	0.2545±0.0012	-

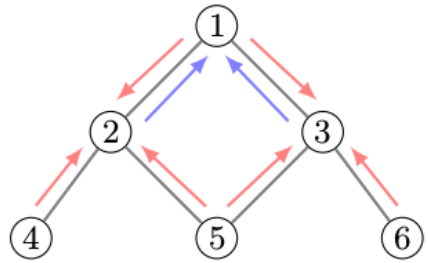
AP = Average Precision MAE = Mean Average Error

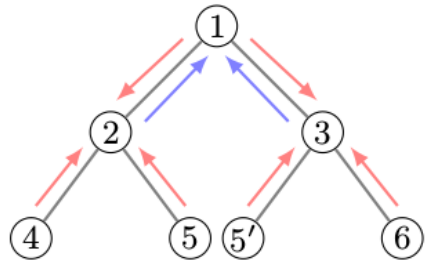
- Peptides-functional and Peptides-structural are two datasets of LRGB.
- Previously GT shows a large margin over MPNN.
- **Simply adding VN is enough to make simple MPNN outperform Graph Transformers!**

Limitation of GNNs on highly symmetric structures

The summing operator **limits** the representative power of MPNNs such that each node loses their identity after being aggregated. For example:

These two graphs are **not** isomorphic, but message passing scheme **fails** to distinguish whether 5 and 5' are the same vertex or not.



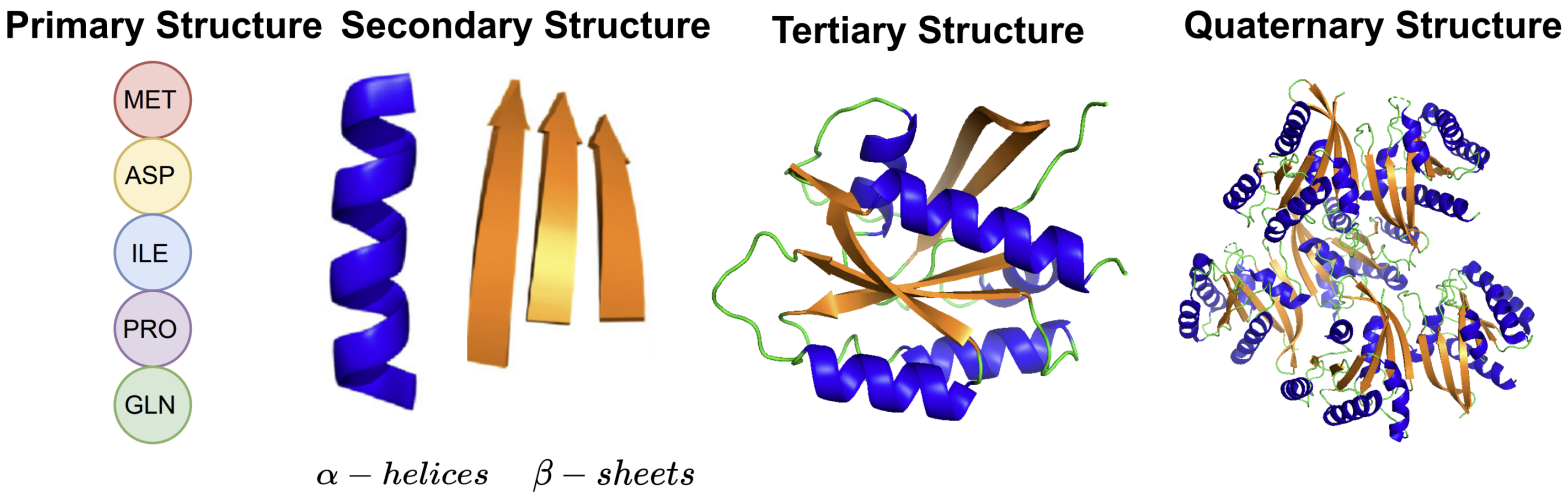


Weisfeiler-Lehman isomorphism test fails for highly symmetric structures such as regular graphs.

Solution: Higher-order equivariant models

- Covariant compositional networks for learning graphs, ICLR 2018
- Predicting molecular properties with covariant compositional networks, Journal of Chemical Physics, Volume 148, Issue 24

Protein Multimodal Representation Learning



Each of these structural levels corresponds to a specific modality of representation:

- Primary & Secondary:

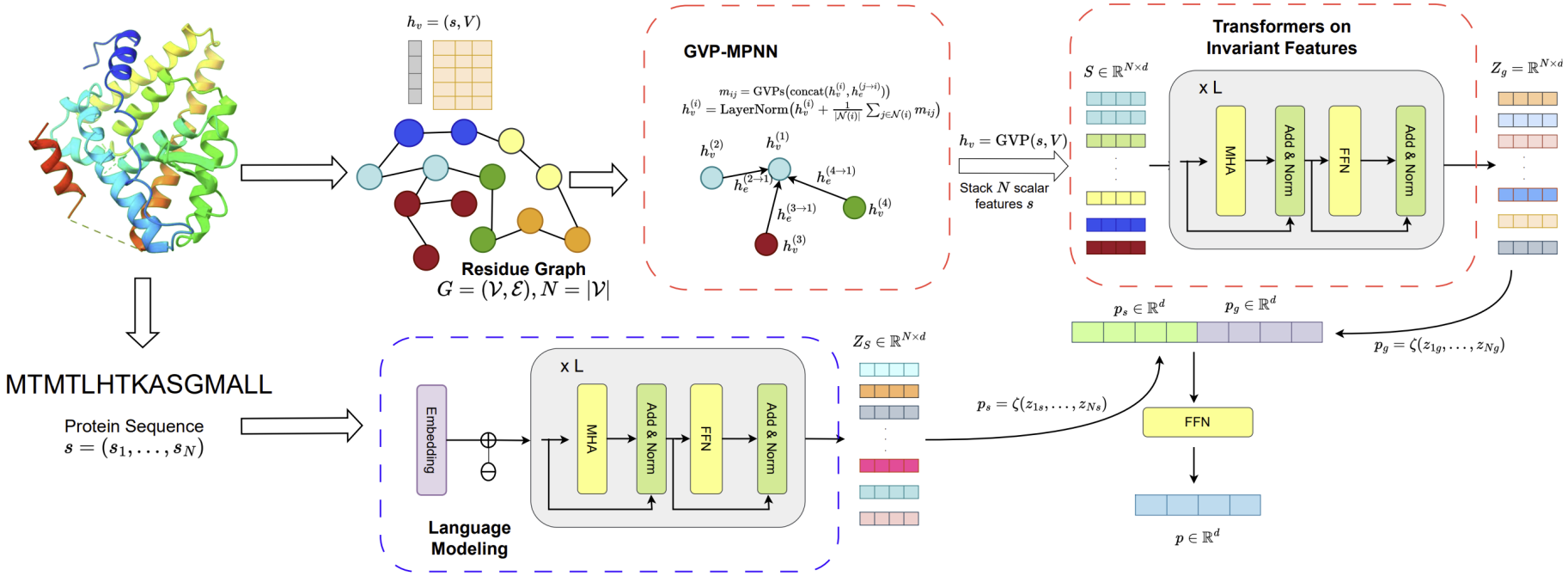
  Sequence $\rightarrow$ **Large Language Models**

- Tertiary & Quaternary:

  2D/3D graph, 3D point cloud $\rightarrow$ **Geometric Deep Learning**

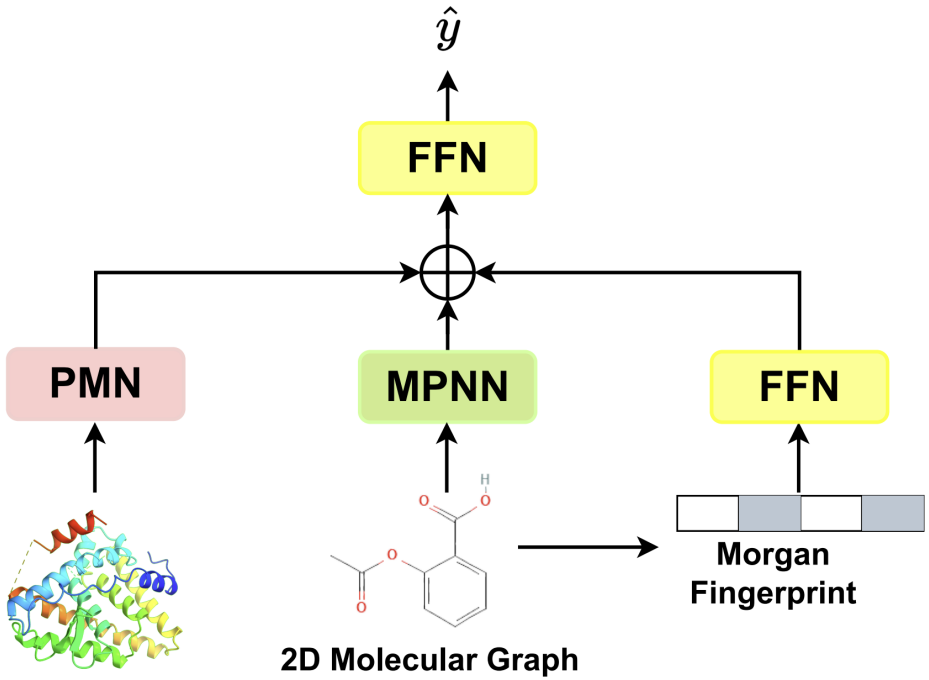
Protein Multimodal Representation Learning (Supervised)

Our **Protein Multimodal Network** (PMN) learns to produce a unified representation of the protein structures including primary structure (i.e. amino-acid sequence) and tertiary structure (i.e. 3D structure) by **Large Language Models** and $E(3)$-**Equivariant Graph Neural Networks**.



Our work: Target-aware Variational Auto-encoders for Ligand Generation with Multimodal Protein Representation Learning, Machine Learning for Structural Biology Workshop, NeurIPS 2023

Protein-Ligand Binding Affinity Prediction: Architecture



A framework for predicting binding affinities between proteins and ligands.

PMN = Protein Multimodal Networks
MPNN = Message Passing Neural Networks
FFN = Feed-Forward Networks

Protein-Ligand Binding Affinity Prediction: DAVIS & KIBA

Approach	DAVIS			KIBA		
	MSE ↓	CI ↑	r_m^2 ↑	MSE ↓	CI ↑	r_m^2 ↑
KronRLS [28]	0.379	0.871	0.407	0.411	0.782	0.342
SimBoost [29]	0.282	0.872	0.644	0.222	0.836	0.629
SimCNN-DTA [67]	0.319	0.852	0.595	0.274	0.821	0.573
DeepDTA [30]	0.261	0.878	0.63	0.194	0.863	0.673
WideDTA [64]	0.886	0.262	—	0.875	0.179	—
AttentionDTA [68]	0.216	0.893	0.677	0.155	0.882	0.755
MATT-DTI [69]	0.227	0.891	0.683	**0.150**	0.882	0.756
GraphDTA [33]	0.258	0.884	0.656	0.162	0.879	0.736
FusionDTA [70]	0.220	0.903	0.666	0.167	**0.891**	0.699
BiCompDTA [65]	0.237	0.904	0.696	0.167	**0.891**	0.757
PMN (ours)	**0.202**	**0.906**	**0.739**	0.153	0.874	**0.767**
std	± 0.007	± 0.003	± 0.011	± 0.002	± 0.003	± 0.003

Experimental results on DAVIS and KIBA dataset. Our results are averaged over five runs. Combining both sequential and topological information of proteins (i.e. multimodal) is necessary!

MSE = Mean Square Error
CI = Concordance Index
r_m^2 = Correlation Score

Protein-Ligand Binding Affinity Prediction: PDB v2020

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)**

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

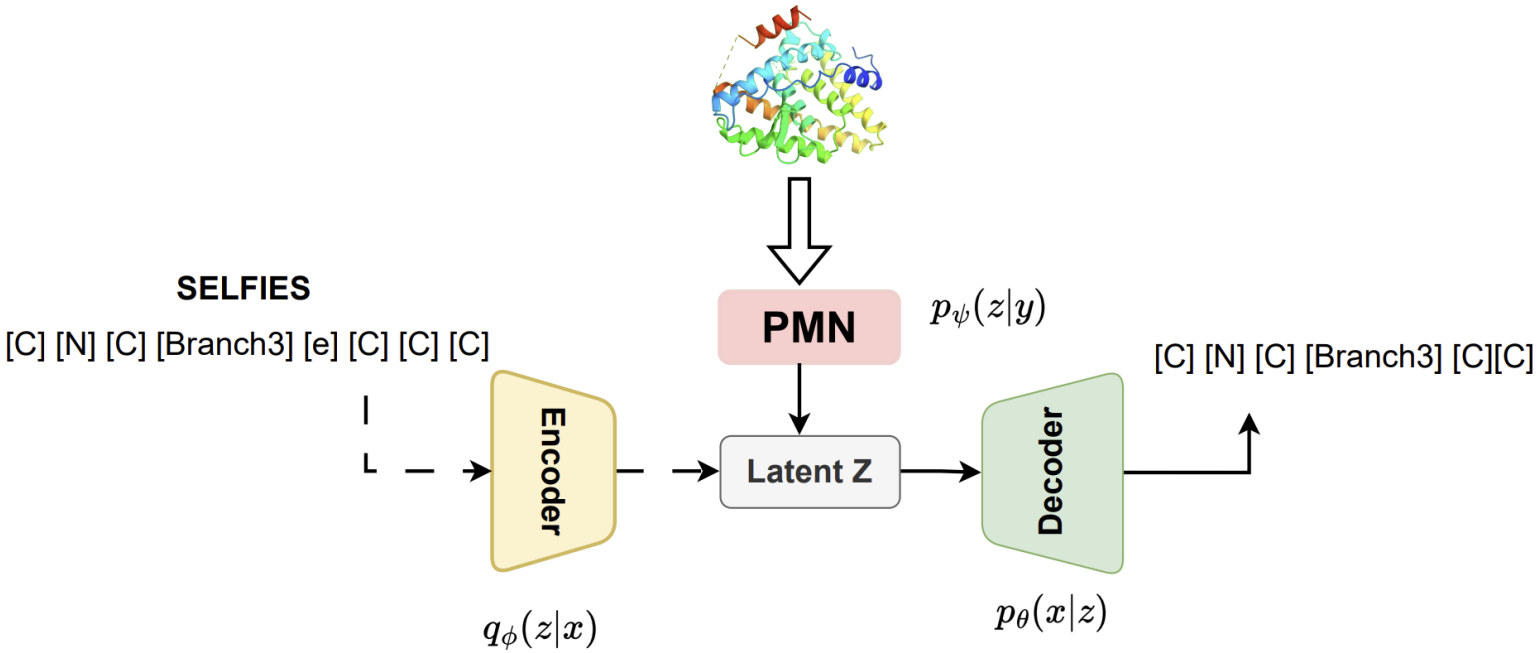
	Method	RMSE ↓	MAE ↓	Pearson ↑	Spearman ↑	r_m^2 ↑	CI ↑
Complex	Pafnucy	1.435 (0.018)	1.144 (0.018)	0.635 (0.008)	0.587 (0.008)	0.348 (0.016)	0.707 (0.004)
	OnionNet	1.403 (0.012)	1.103 (0.014)	0.648 (0.007)	0.602 (0.013)	0.381 (0.011)	0.717 (0.005)
	IGN	1.404 (0.025)	1.116 (0.030)	0.662 (0.013)	0.638 (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)

Performance comparison on the PDBBind v2020 dataset.
(We are still working on it!)

ESM = Evolutionary Scale Modeling (i.e. Protein Language Model)
RMSE = Root Mean Square Error
MAE = Mean Average Error
Pearson, Spearman, r_m^2, CI = Correlation Scores

Generative AI: Protein-binding Ligand Generation

TargetVAE - a Conditional Variational Autoencoder - 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.

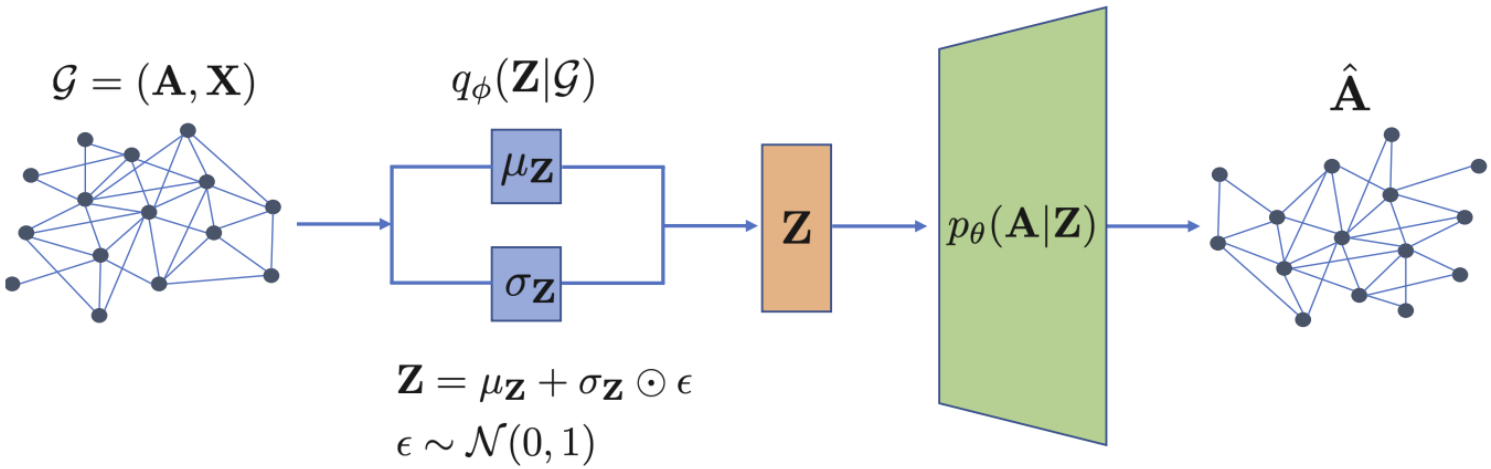


Our work: Target-aware Variational Auto-encoders for Ligand Generation with Multimodal Protein Representation Learning, recently accepted at **Machine Learning: Science and Technology** journal and presented at **NeurIPS 2023**.

Generative AI: Multiresolution Equivariant Graph VAE

Instead of **string-based** VAE (generating SELFIES / SMILES), we also proposed **graph-based** VAE (generating molecular graph) to:

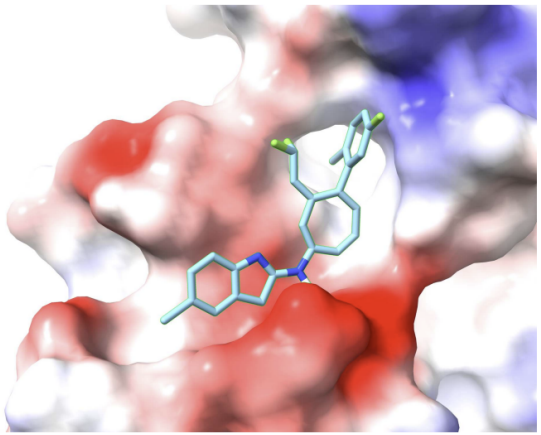
- Generate molecules in **multiple levels of resolutions** (i.e. multiresolution),
- While **preserving the permutation symmetry**.



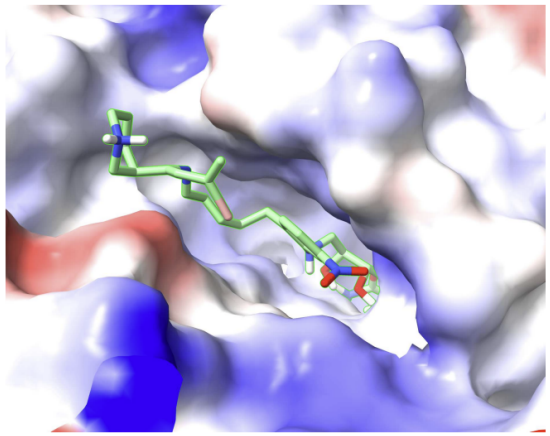
Our works:

- Multiresolution Equivariant Graph Variational Autoencoder, Machine Learning: Science and Technology, Volume 4, Number 1.
- The general theory of permutation equivarant neural networks and higher order graph variational encoders, Preprint, 2020.

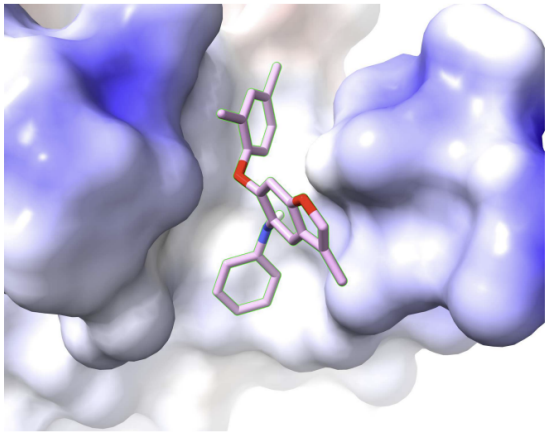
Generative AI: Protein-binding Ligand Generation



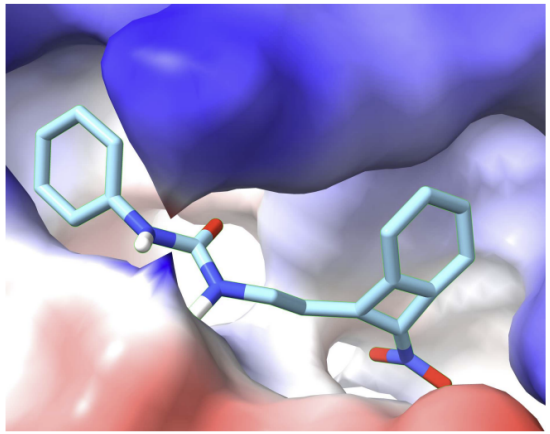
(a) 1iep, 3.59, -9.48 kcal/mol



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



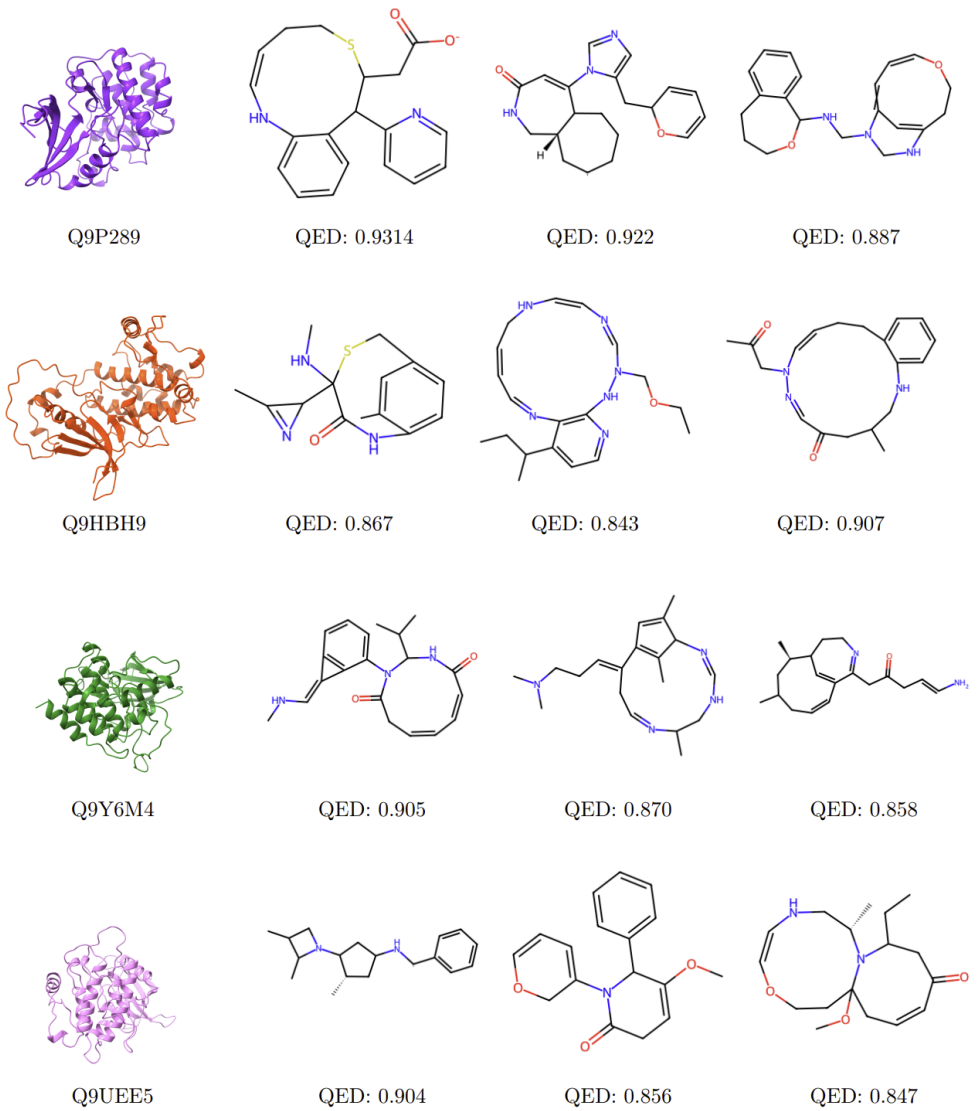
(c) 3eml,2.20,-8.29 kcal/mol



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

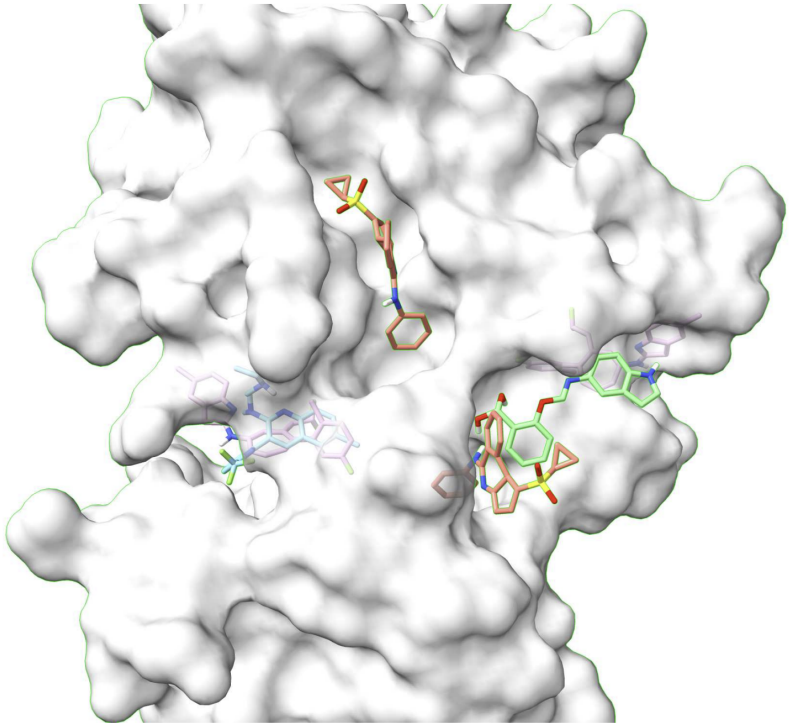
Each figure is associated with the name of each target protein, synthetic accessibility, and binding affinity in kcal/mol of the generated ligand.

Generative AI: Protein-binding Ligand Generation



Some generated ligands with high QED (i.e. Drug-likeness score).

Generative AI: Protein-binding Ligand Generation



Multiple generated ligands with different poses bind to a given target protein.
Our TargetVAE can generate ligands for a protein without the prior knowledge of the binding pocket.

Our source code & datasets:

https://github.com/HySonLab/Ligand_Generation

Drug discovery pipeline

1. **In-silico:** HySonLab is working on generative AI, molecular dynamic simulations.
2. **In-vitro & In-vivo:** HySonLab is collaborating with Department of Biology at Indiana State University (Prof. Tak & Prof. Cho). We are testing generated drug candidates (by AI) on **fruit flies**. We are interested in mutant proteins of KRAS, NRAS and HRAS.

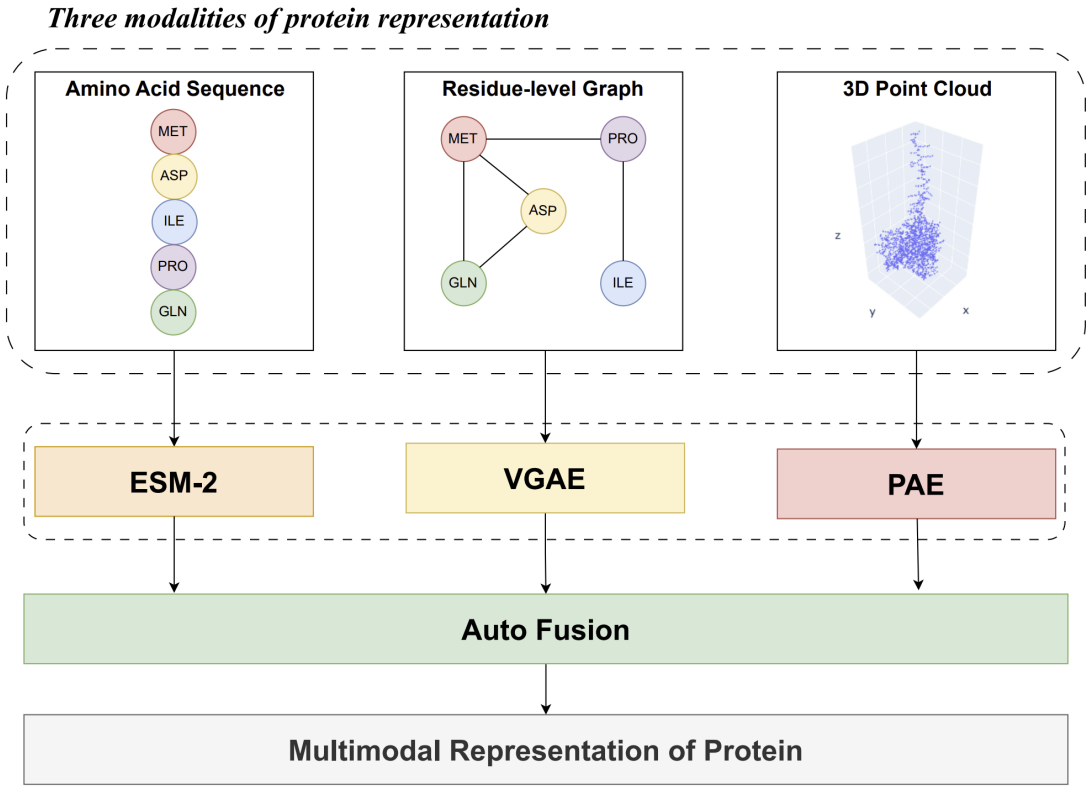
Drug discovery pipeline

1. **In-silico:** HySonLab is working on generative AI, molecular dynamic simulations.
2. **In-vitro & In-vivo:** HySonLab is collaborating with Department of Biology at Indiana State University (Prof. Tak & Prof. Cho). We are testing generated drug candidates (by AI) on **fruit flies**. We are interested in mutant proteins of KRAS, NRAS and HRAS.
3. **Clinical trial:** HySonLab is collaborating with Australian National University (Prof. Nhung Nghiem) for potential clinical trials once approved (in 2-3 years).

We are working hard on **Step 1** and **Step 2**. We aim to deliver novel AI methods with ready-to-use software packages for pharmaceutical industry, while finding new potential drugs.

Unsupervised Protein Representation Learning

We combine three **pretraining** models to exploit the vast amount of unannotated / unlabeled protein data: **ESM-2:** Evolutionary Scale Modeling (without MSA), **VGAE:** Graph Variational Autoencoder, and **PAE:** PointNet Autoencoder.



Our work: Multimodal Pretraining for Unsupervised Protein Representation Learning, Preprint, 2024.

Unsupervised Protein Representation Learning

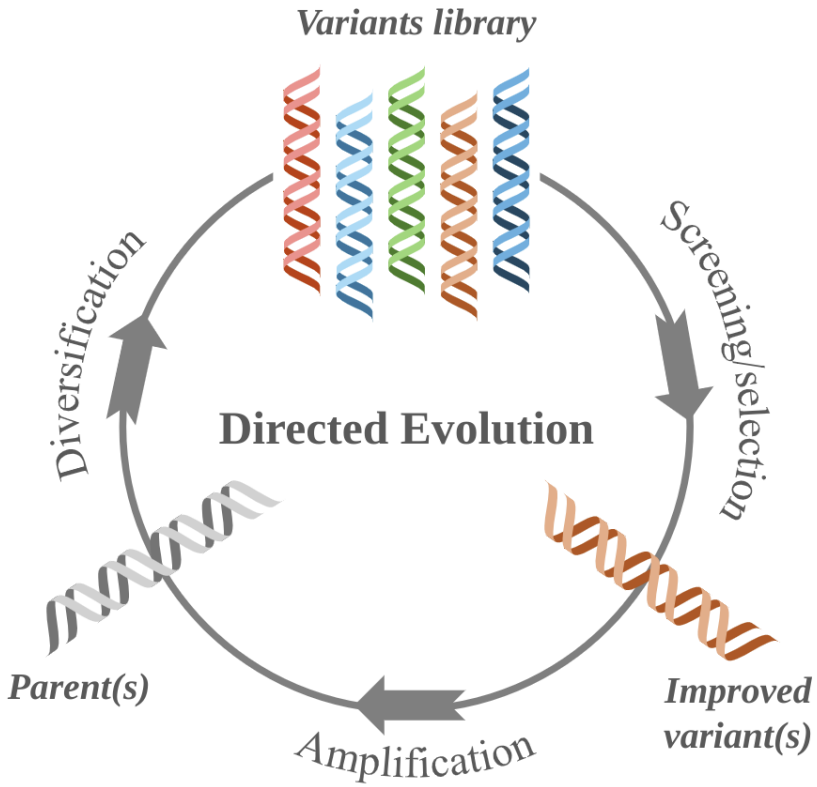
1. We extract the protein embedding of proteins from PDB v2020 from our unsupervised pretraining model.
2. Then, we apply a simple Gaussian Process for predicting protein-ligand binding affinity. **Better than several supervised methods!**

Approach	RMSE ↓	MAE ↓	Pearson ↑	Spearman ↑	r_m^2 ↑	CI ↑
Pafnucy [61]	1.435	1.144	0.635	0.587	0.348	0.707
OnionNet [62]	1.403	1.103	0.648	0.602	0.381	0.717
IGN [63]	1.404	1.116	0.662	0.638	0.385	0.73
SIGN [64]	1.373	1.086	0.685	0.656	0.398	0.736
SMINA [65]	1.466	1.161	0.665	0.663	0.391	0.74
GNINA [66]	1.740	1.413	0.495	0.494	0.209	0.674
dMaSIF [67]	1.450	1.136	0.629	0.588	0.347	0.71
TankBind [68]	1.345	1.060	**0.718**	**0.689**	0.404	0.750
GraphDTA [69]	1.564	1.223	0.612	0.570	0.306	0.703
TransCPI [70]	1.493	1.201	0.604	0.551	0.255	0.677
MolTrans [71]	1.599	1.271	0.539	0.474	0.242	0.666
DrugBAN [72]	1.480	1.159	0.657	0.612	0.319	0.72
DGraphDTA [73]	1.493	1.201	0.604	0.551	0.312	0.693
WGNN-DTA [74]	1.501	1.196	0.605	0.562	0.311	0.697
STAMP-DPI [75]	1.503	1.176	0.653	0.601	0.327	0.719
PSICHIC [50]	**1.314**	**1.015**	0.710	0.686	**0.428**	**0.751**
ESM (ours)	1.380	1.120	0.659	0.619	0.247	0.719
VGAE (ours)	1.485	1.209	0.594	0.547	0.246	0.690
PAE (ours)	1.466	1.158	0.604	0.557	0.248	0.696
Multimodal (ours)	1.372	1.104	0.663	0.634	0.246	0.726

Our source code & pretrained model:

https://github.com/HySonLab/Protein_Pretrain

Directed Evolution for Protein Optimization



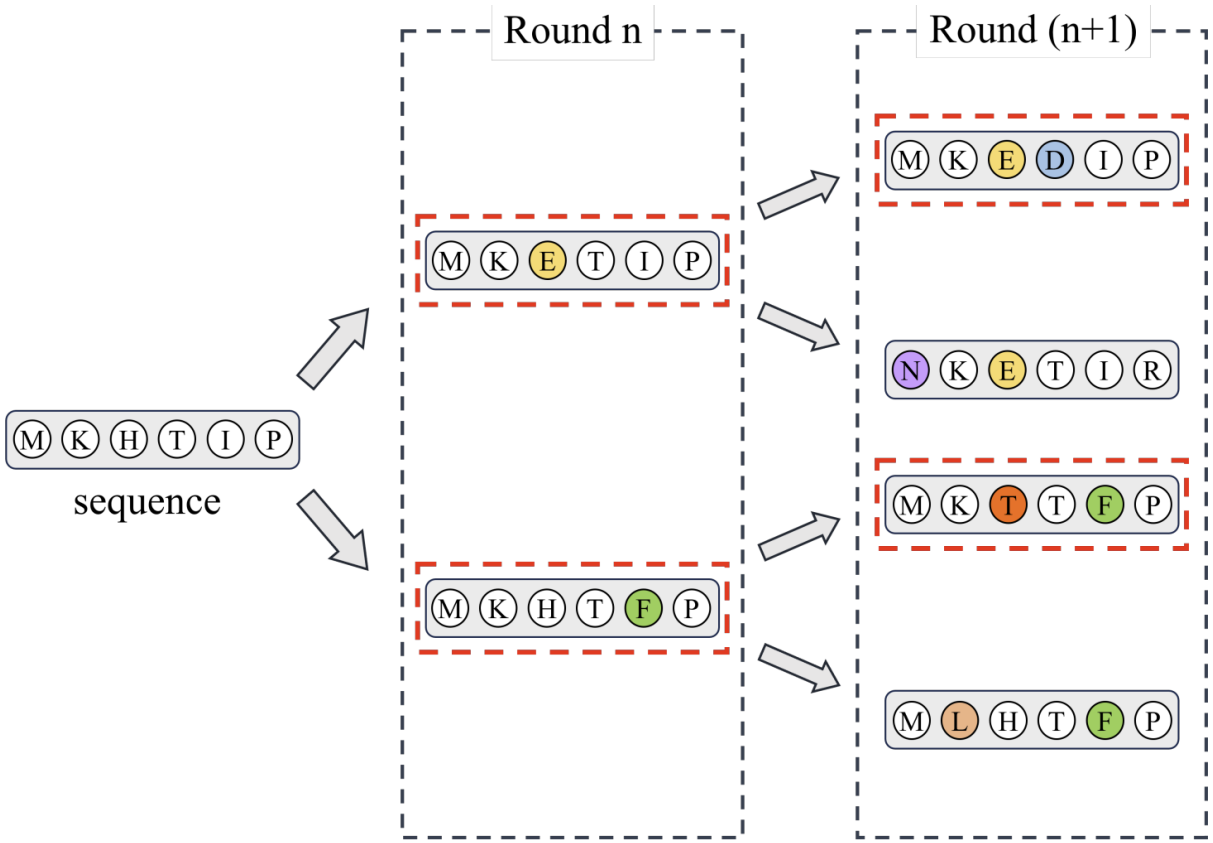
Directed Evolution (DE) begins with parent sequences, introduces mutations to diversify the population, and then selects top-scoring proteins based on a desired fitness. These selections are then amplified to enhance their properties.

Our work: Protein Design by Directed Evolution Guided by Large Language Models, Preprint, 2024.

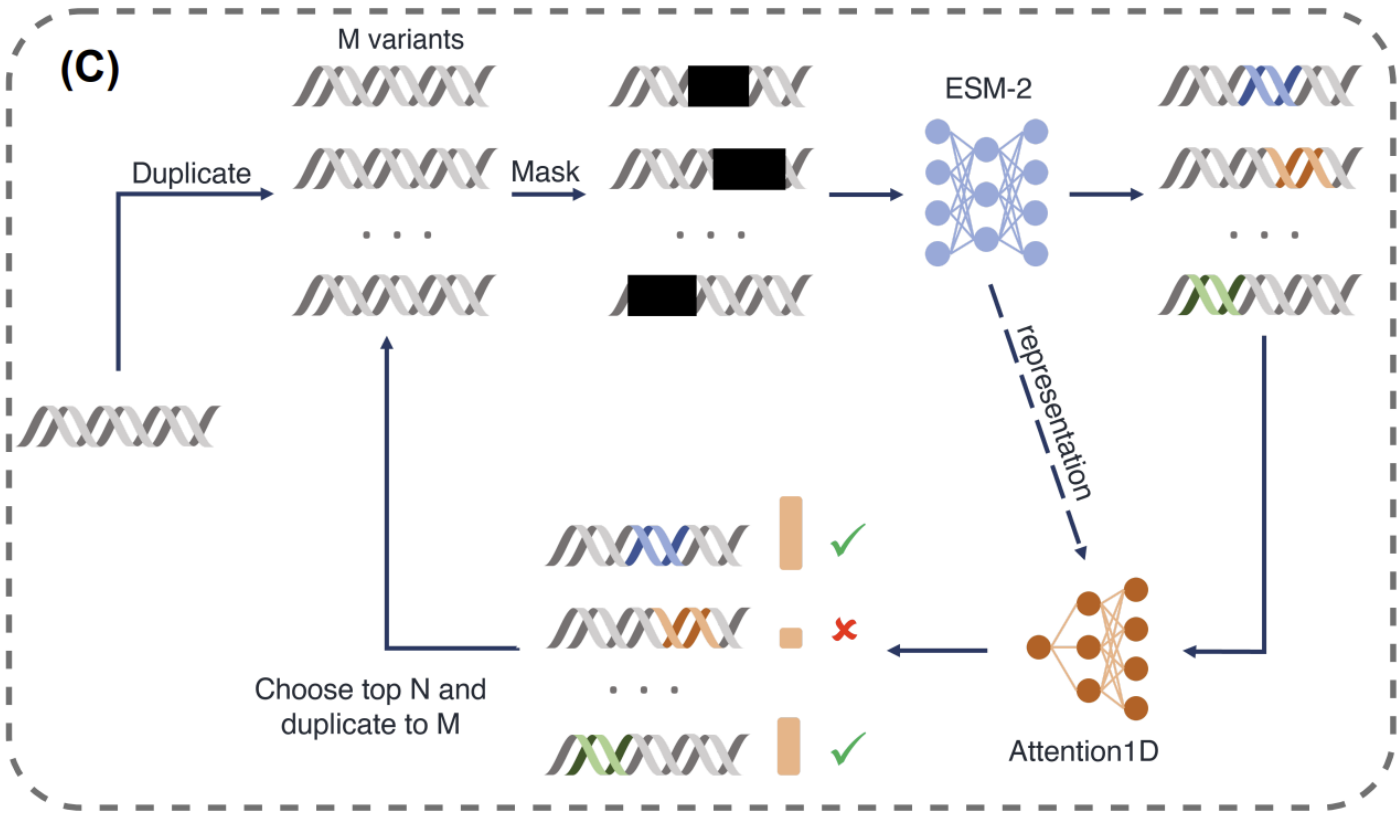
Machine Learning guided Directed Evolution (MLDE)

Main idea:

- Instead of random mutations, our idea is to use **Large Language Model** (LLM) to select and predict possible / likely mutations.
- We train a Machine Learning model to **predict the protein fitness** in order to replace the time-consuing process of wet-lab evaluation.



Machine Learning guided Directed Evolution (MLDE)



Our proposed MLDE framework: The workflow begins by identifying a protein with activity for a target function. Once the starting point is identified, diversity is introduced by mutagenesis, and the resulting variants are screened (in-silico) for function.

Machine Learning guided Directed Evolution (MLDE)

Five protein engineering benchmarks: Green Fluorescent Proteins (avGFP), Adeno-associated Viruses (AAV), Aliphatic Amide Hydrolase (AMIE), TEM-1 β-Lactamase (TEM), and Ubiquitination Factor Ube4b (E4B).

Models	avGFP	AAV	TEM	E4B	UBE2I	Average
CMA-ES†	4.492	-3.417	0.375	-0.768	2.461	0.629
FBGAN†	1.251	-4.227	0.006	0.369	0.208	-0.479
DbAS†	3.548	4.327	0.003	-1.286	3.088	1.936
CbAS†	3.550	4.336	0.106	-1.000	3.263	2.051
PEX†	3.764	3.265	0.121	5.019	-0.474	2.339
GFlowNet-AL†	5.062	1.205	1.552	3.155	3.576	2.910
IsEM-Pro†	6.185	4.813	1.850	5.737	**4.536**	4.624
MLDE	**11.796 ± 0.676**	**6.585 ± 0.367**	**6.731 ± 0.120**	**10.311 ± 0.519**	2.978 ± 0.099	**7.680 ± 0.356**
MLDE (AFS)	11.670 ± 0.671	6.171 ± 0.374	6.675 ± 0.126	10.176 ± 0.485	2.874 ± 0.098	7.513 ± 0.351
MLDE-random	10.800 ± 0.501	6.231 ± 0.570	6.635 ± 0.150	10.161 ± 0.718	2.744 ± 0.013	7.314 ± 0.390

We outperform all competing methods (4/5 benchmarks) including evolutionary search algorithm, probabilistic modeling, and recent generative models from DL.

Source code & Datasets:

https://github.com/HySonLab/Directed_Evolution

Summary

Summary of our talk today:

1. Introduction to Geometric Deep Learning
2. Multiresolution Graph Transformers for Long-Range Interactions
3. Protein Multimodal Network & Generative AI
4. Unsupervised Protein Foundation Model
5. Machine Learning guided Directed Evolution for Protein Optimization

HySonLab is also working on several other interesting directions:

- Drug repurposing
- Biomedical NLP & Knowledge Graph
- ML for Operations Research
- Materials discovery by Generative AI

Please visit my group's github for many useful packages:

`https://github.com/HySonLab/`

Thank you for your attention!

Acknowledgments

Special thanks to students and collaborators of HySonLab who contributed to our recent results which I am presenting today!

For prospective students/interns/residents:

- Email me at sonpascal93@gmail.com
- Include your CV, academic transcript, github, etc.
- Your ideas, proposals, suggestions, etc.

For potential collaborators: I am always happy to hear from you and your ideas by any mean of contact.