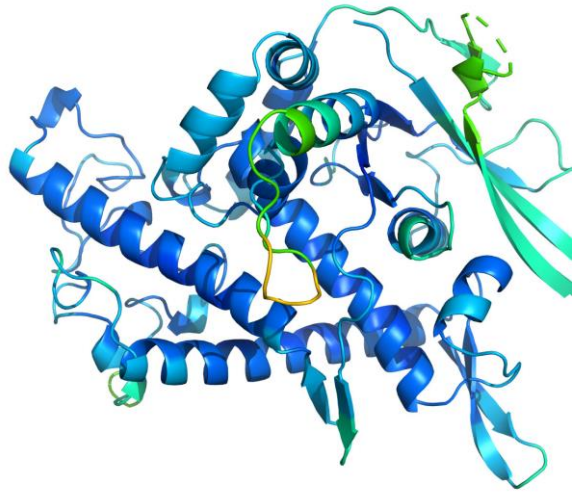
# Geometric Graph Learning for Predicting Protein Mutation Effect

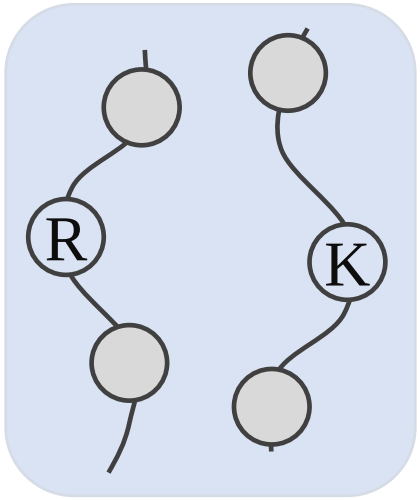
Kangfei Zhao[*#], Yu Rong[#], Biaobin Jiang[#], Jianheng Tang[$],
Hengtong Zhang[#], Jeffrey Xu Yu[◊], Peilin Zhao[#]
[*]Beijing Insitute of Technology
[#]Tencent AI Lab
[$]Hong Kong University of Science and Technology
[◊] The Chine University of Hong Kong

CIKM 2023

# Protein Mutation effect Prediction

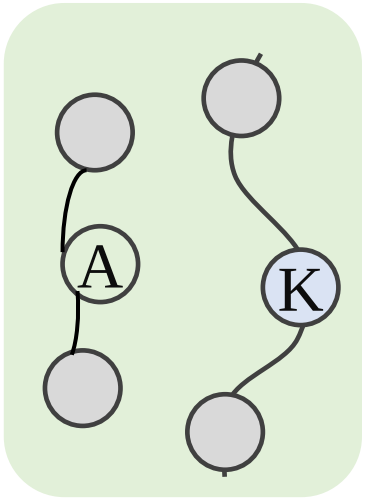


A Wild-type Protein



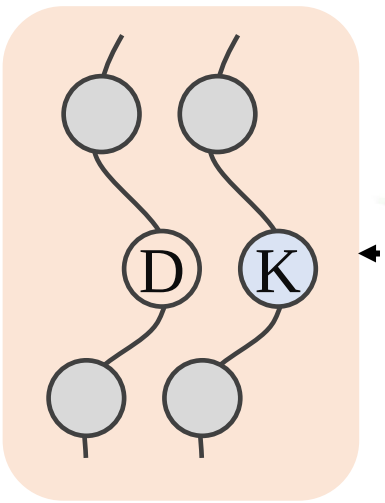

Affinity=0.01

Mutant Proteins



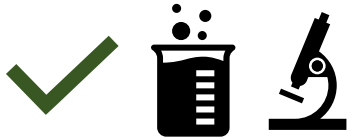

Affinity 0.13

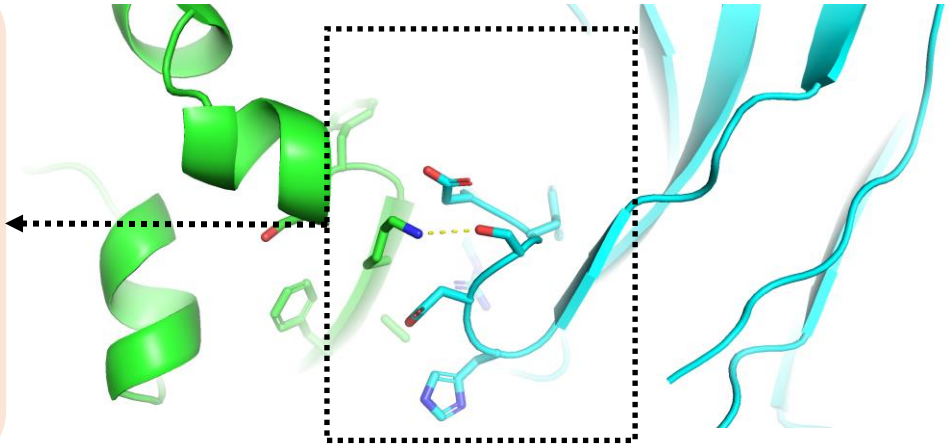
ΔAffinity=0.12




Affinity=2.10

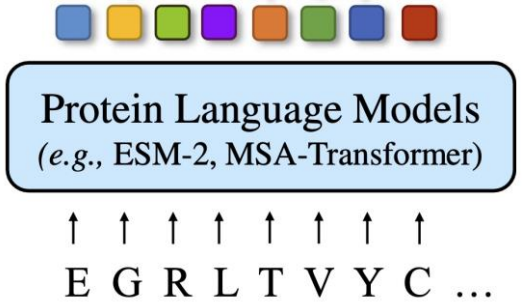
**ΔAffinity=2.09**



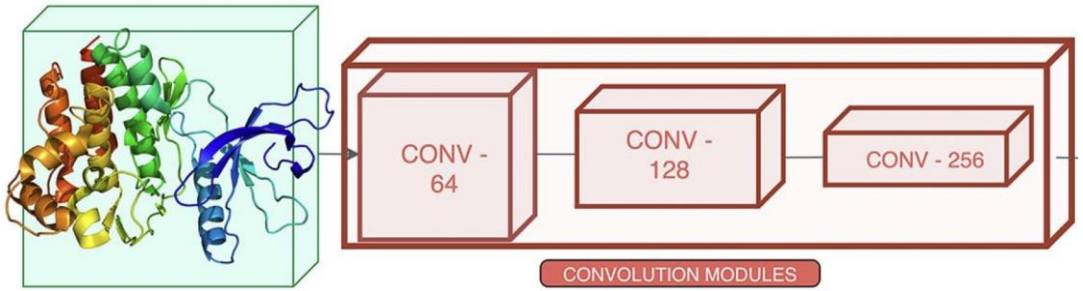


# Background: Protein Rrepresenation

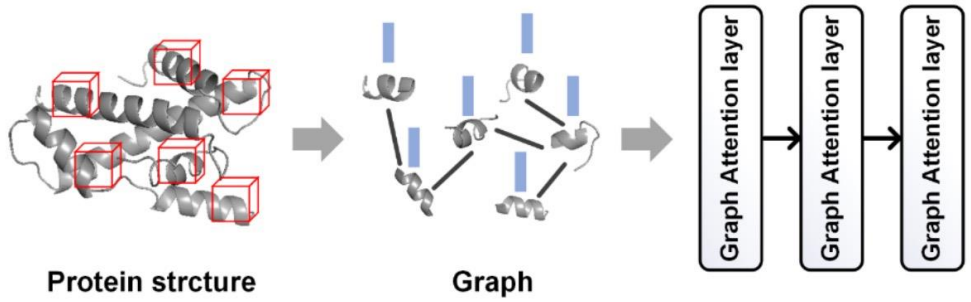
- Protein features: elementary biophysical and sequence-derived features
- Sequence modeling for amino acid residue sequence
  - RNN, Pretrain LLM




- 3D structure modeling
  - 3D convolution

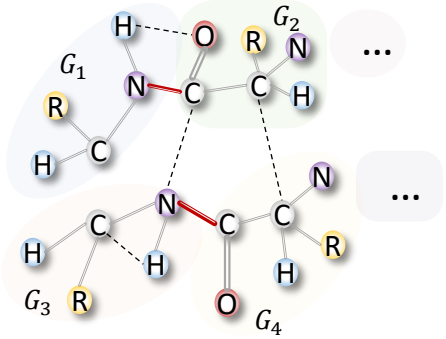


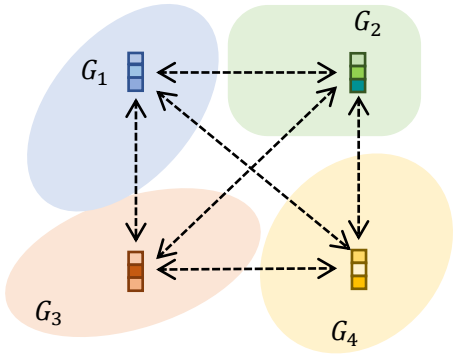

- 3D graph learning



# Graph Learning for Protein Representation

- Single-level: Atom-level vs. Amino acid residue-level

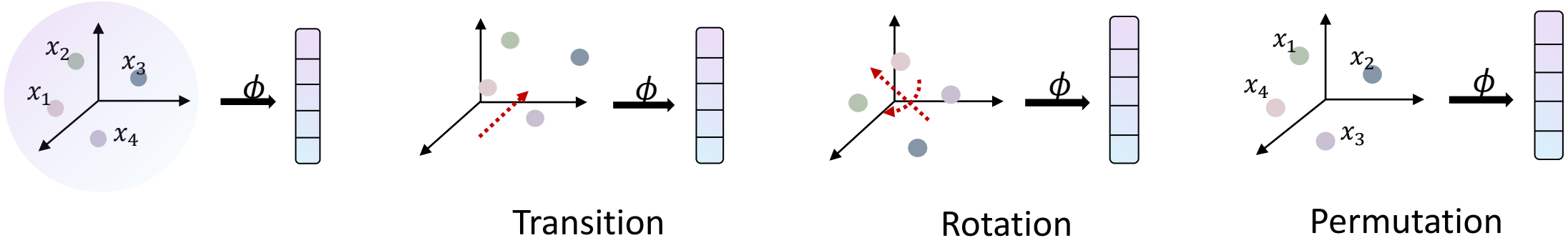


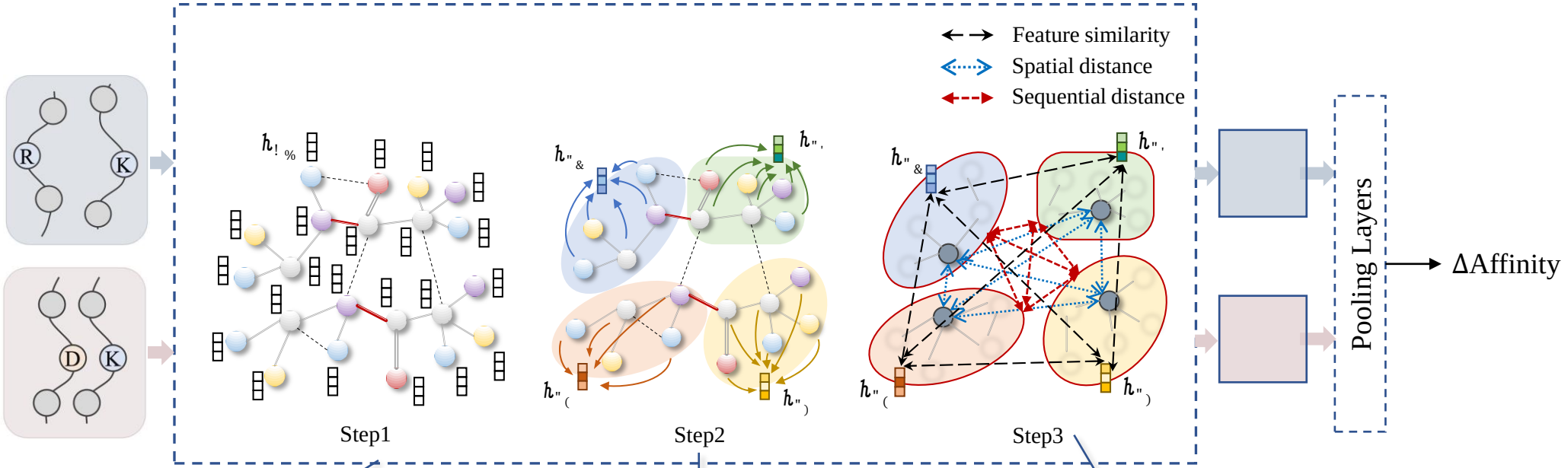

😢 Cannot model the structural hierarchy of proteins

- Inducive biases: invariant to 3D transformations

😢 Cannot preserve the inductive biases



# Overview: Hierarchical Graph Invariant Network




Equivalent Graph Neural Network

$$m_{ij} = \phi_e(h_{a_i}^{(k)}, h_{a_j}^{(k)}, \|c_i - c_j\|^2), \forall a_i, a_j \in G_1 \cup \cdots \cup G_L$$

$$m_i = \sum_{a_j \in G_1 \cup \cdots \cup G_L \setminus \{a_i\}} m_{ij}, \forall a_i \in G_1 \cup \cdots \cup G_L$$

$$h_{a_i}^{(k+1)} = h_{a_i}^{(k)} + m_i, \forall a_i \in G_1 \cup \cdots \cup G_L$$

Sum Pooling

$$h_{G_i}^{(0)} = \sum_{a_j \in G_i} h_{a_j}^{(K)}, \forall G_i \in \mathcal{G}$$

Geometric Attention Network

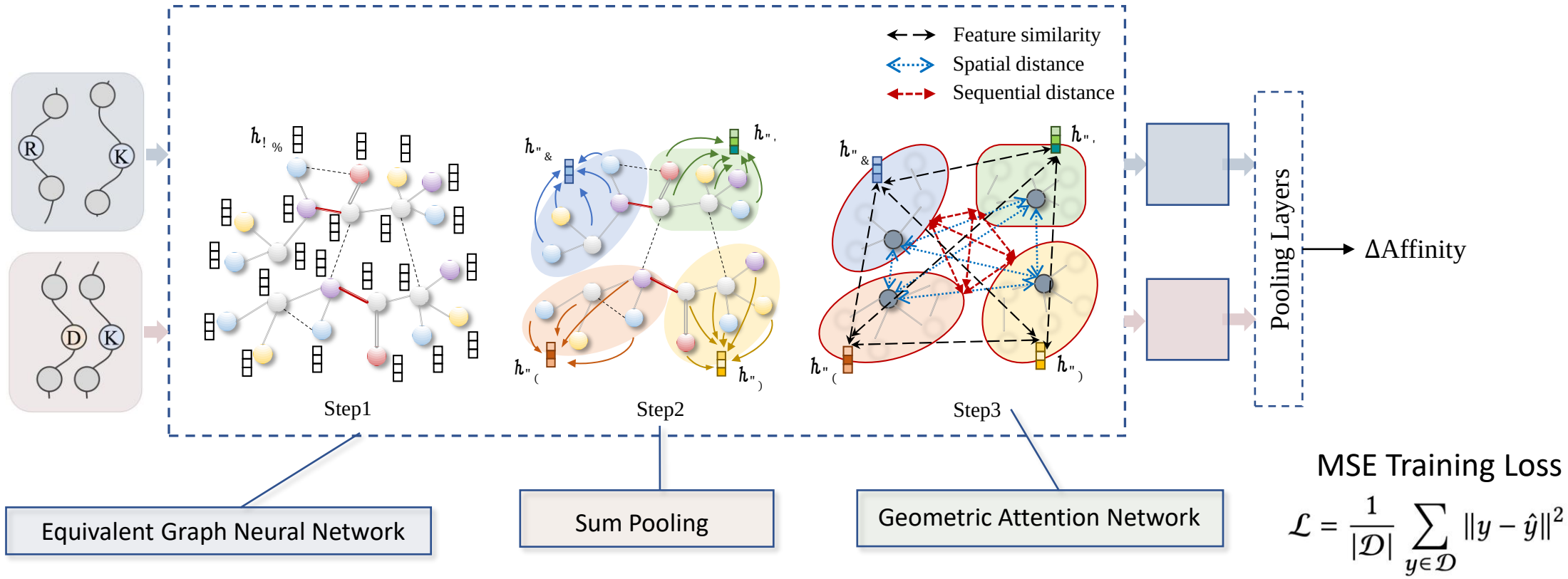
$$u_i = \sum_{G_j \in \mathcal{G} \setminus \{G_i\}} \alpha_{ij} \text{Linear}(h_{G_j}^{(k)}), \forall G_i \in \mathcal{G}$$

$$v_i = \sum_{G_j \in \mathcal{G} \setminus \{G_i\}} \alpha_{ij} \text{Linear}(z_{ij}), \forall G_i \in \mathcal{G}$$

$$w_i = \sum_{G_j \in \mathcal{G} \setminus \{G_i\}} \alpha_{ij} O_i^T (c_i^{C_\beta} - c_j^{C_\beta}), \forall G_i \in \mathcal{G}$$

$$h_{G_i}^{(k+1)} = h_{G_i}^{(k)} + \text{Linear}(u_i \| v_i \| w_i), \forall G_i \in \mathcal{G}$$

# Overview: Hierarchical Graph Invariant Network




- Invariant to transition, rotation of atom coordinates and permutation of atom indices

$$f(\Pi[\boldsymbol{h}_a^{(0)}], \Pi[Q\boldsymbol{c} + g]) = f(\boldsymbol{h}_a^{(0)}, \boldsymbol{c})$$

# Experimental Studies: Setup

- 9 Baseline Approaches
  - 2 ML models: GBDT, SVR
  - 3 Sequence models: PIPR, BertPIPR, ECNet
  - 4 Graph based models: GeoPPI, HGAT, EGNN, GAN
  - Ours: 3-layer EGNN and 3-layer GAN
- 3 Protein Datasets
  - Envision: functional fitness changes
  - SKEMPI: binding free energy changes
  - SARA-COV-2: human antibody affinity against COVID-19 virus changes
- Protein Structures
  - Protein Databank and EvoEF2
- Evaluation Metrics
  - MAE, MSE, STD, Spearman coefficient (R)

| Variants | Envision | SKEMPI2 | SARS-COV-2 |
|---|---|---|---|
| # proteins | 6,899 | 6,323 | 349 |
| # wild-type | 6 | 348 | 35 |
| # mutation points | 1 | 1 ~ 27 | 1 ~ 7 |
| # chains | 1 | 2 ~ 8 | 3 |
| Species | human, rat, mouse, etc. | human, rat, mouse, etc. | human |
| Range of target | [-0.38, 1.57] | [-9.51, 12.30] | [-2.61, 2.77] |

Profile of the Datasets

# Experimental Studies: Effectiveness

| Category | Method | Envision | | | | SKEMPI2 | | | | SARS-COV-2 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | MSE ↓ | MAE ↓ | STD ↓ | R ↑ | MSE ↓ | MAE ↓ | STD ↓ | R ↑ | MSE ↓ | MAE ↓ | STD ↓ | R ↑ |
| Classical ML | *SVR* | 0.1053 | 0.2666 | 0.1850 | 0.4959 | 3.7153 | 1.3134 | 1.4108 | 0.4044 | 2.5270 | 1.4029 | 0.7476 | 0.0171 |
| | *GBDT* | 0.0536 | 0.1761 | 0.1504 | 0.7887 | 1.3610 | 0.8073 | 0.8422 | 0.7339 | 2.2586 | 1.3318 | 0.6963 | 0.1500 |
| Sequence based | *PIPR* | 0.0817 | 0.2210 | 0.1814 | 0.6480 | 2.5766 | 1.0735 | 1.1934 | 0.5786 | 10.5463 | 2.9059 | 1.4498 | 0.0059 |
| | *BertPIPR* | 0.0703 | 0.2020 | 0.1814 | 0.7070 | 2.2282 | 0.9937 | 1.1139 | 0.6020 | 9.6517 | 2.7249 | 1.4922 | -0.1183 |
| | *ECNet* | 0.0741 | 0.2004 | 0.1839 | 0.7268 | 1.6146 | 0.8656 | 0.9291 | 0.6793 | 2.5311 | 1.3859 | 0.8084 | -0.0193 |
| Graph based | *GeoPPI* | 0.1197 | 0.2749 | 0.2100 | 0.5473 | 2.4170 | 1.1389 | 1.0582 | 0.5704 | 2.7779 | 1.4011 | 0.9026 | 0.2833 |
| | *GAN* | 0.0748 | 0.2041 | 0.1821 | 0.7318 | 1.8722 | 0.9736 | 0.9614 | 0.6876 | 2.3947 | 1.2806 | 0.8688 | 0.1052 |
| | *HGAT* | 0.1005 | 0.2398 | 0.2072 | 0.6269 | 1.4467 | 0.8278 | 0.8726 | 0.7037 | 2.1370 | 1.2259 | 0.7963 | 0.3548 |
| | *EGNN* | 0.0924 | 0.2346 | 0.1932 | 0.6506 | 6.4225 | 1.7459 | 1.8369 | 0.2194 | 5.3932 | 1.9169 | 1.3110 | -0.2347 |
| | *HGIN* | 0.0694 | 0.1715 | 0.2000 | 0.7931 | 1.1646 | 0.7172 | 0.8064 | 0.7646 | 1.3841 | 0.9271 | 0.7242 | 0.5832 |

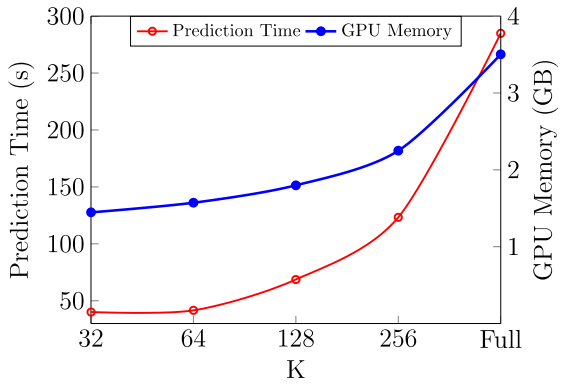
The MSE, MAE, STD of Absolute Error and Spearman coefficient (R) on 3 Protein Datasets

# Experimental Studies: Ablation Studies

| Variants | SKEMPI2 | | | |
|---|---|---|---|---|
| | MSE ↓ | MAE ↓ | STD ↓ | R ↑ |
| Node (Message) | 1.3433 | 0.7861 | 0.8517 | 0.7360 |
| Node + Seq. (Message) | 1.1930 | 0.7301 | 0.8123 | 0.7540 |
| Node + Spat. (Message) | 1.3317 | 0.7777 | 0.8526 | 0.7314 |
| Node (Att. Bias) | 1.6182 | 0.8738 | 0.9245 | 0.7017 |
| Node + Seq. (Att. Bias) | 1.3442 | 0.7784 | 0.8592 | 0.7265 |
| Node + Spat. (Att. Bias) | 1.2410 | 0.7478 | 0.8258 | 0.7525 |
| 2-layer EGNN | 1.1793 | 0.7266 | 0.8070 | 0.7551 |
| 4-layer EGNN | 1.1495 | 0.7119 | 0.8016 | 0.7676 |
| 2-layer GAN | 1.3355 | 0.7761 | 0.8562 | 0.7412 |
| 4-layer GAN | 1.1516 | 0.6887 | 0.8230 | 0.7648 |

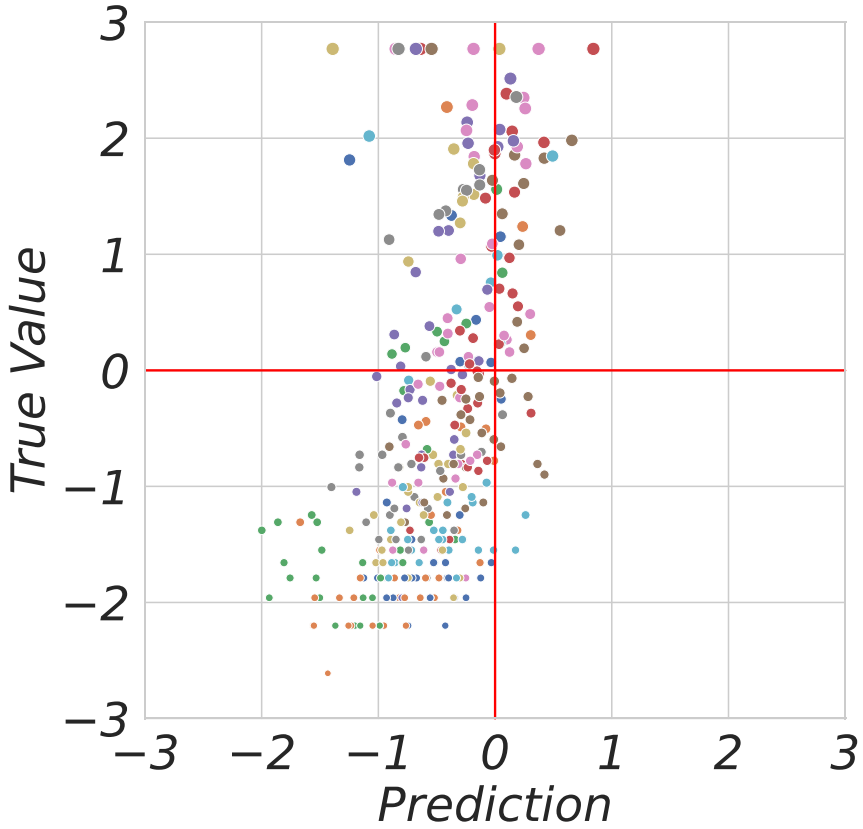
Ablation Studies on SKEMPI2

| K | MSE ↓ | MAE ↓ | STD ↓ | R ↑ |
|---|---|---|---|---|
| 32 | 1.1713 | 0.7156 | 0.8120 | 0.7582 |
| 64 | 1.1216 | 0.6995 | 0.7952 | 0.7698 |
| 128 | 1.1178 | 0.7063 | 0.7867 | 0.7665 |
| 256 | 1.1519 | 0.7095 | 0.8053 | 0.7640 |
| Full | 1.1646 | 0.7172 | 0.8064 | 0.7646 |

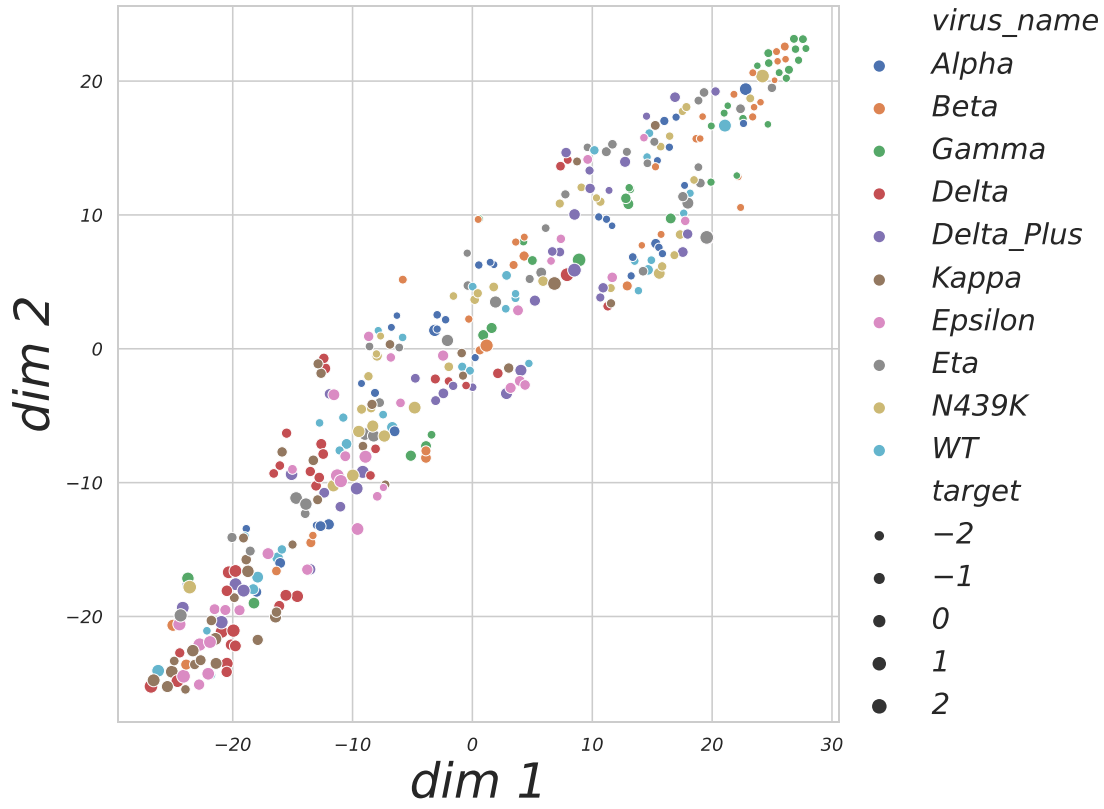



Varying the K nearest neighbors of in atom-level

# Experimental Studies: Case Study



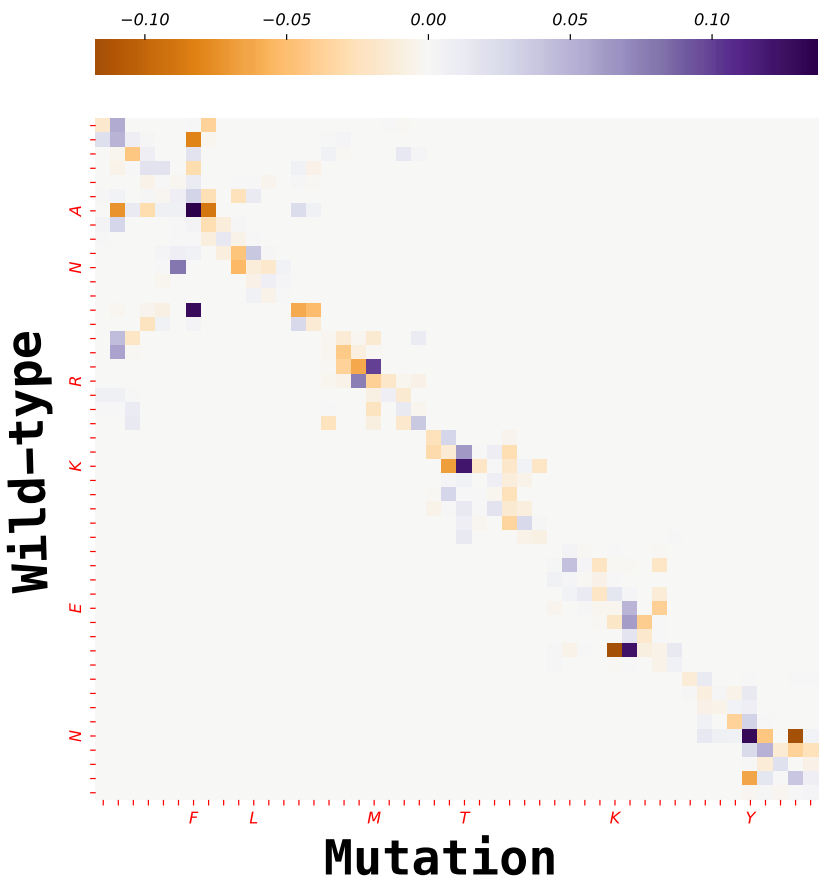

Prediction vs. True Value



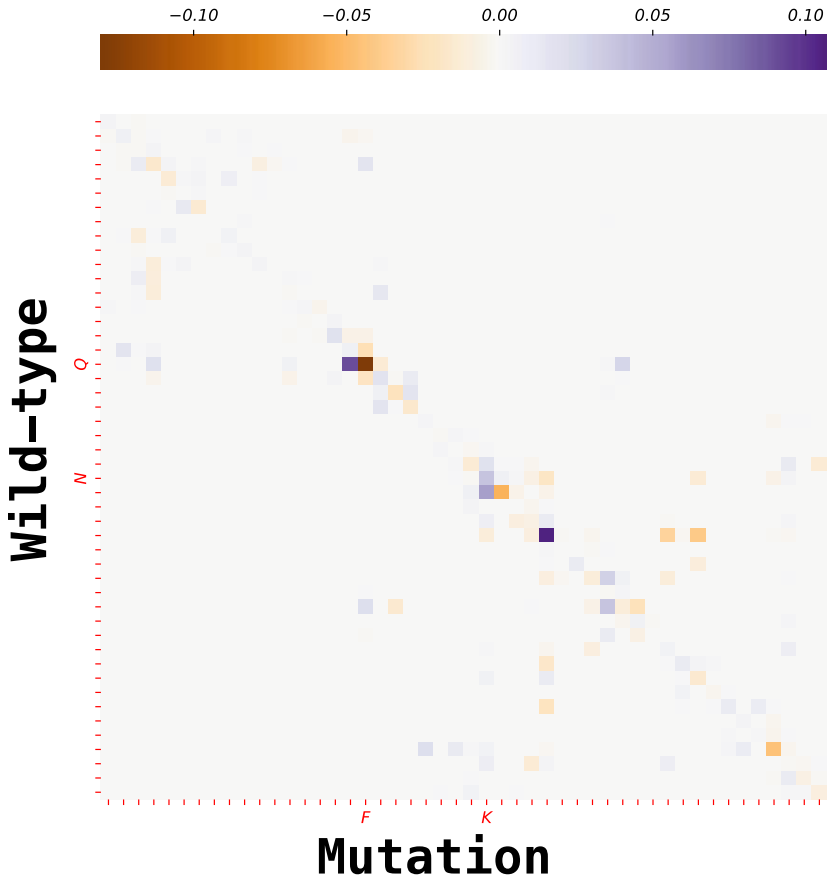

Protein embedding by T-SNE

# Experimental Studies: Case Study

Attention heatmap




Good-performed Case




Bad-performed Case

# Thank you!

Q & A

zkf1105@gmail.com