Multi-Scale Representation Learning on Proteins

Vignesh Ram Somnath, Charlotte Bunne, Andreas Krause

Presented by

Mohimenul Karim

Overview

- Multi-scale graph construction of a protein-HOLOPROT

- Connects surface to structure and sequence
    - Surface capture the coarser details
    - Sequence (primary component) and structure (secondary and tertiary component) capture finer details

- Tests the representation on two different tasks-
    - Ligand-binding affinity (regression)
    - Protein function prediction (classification)

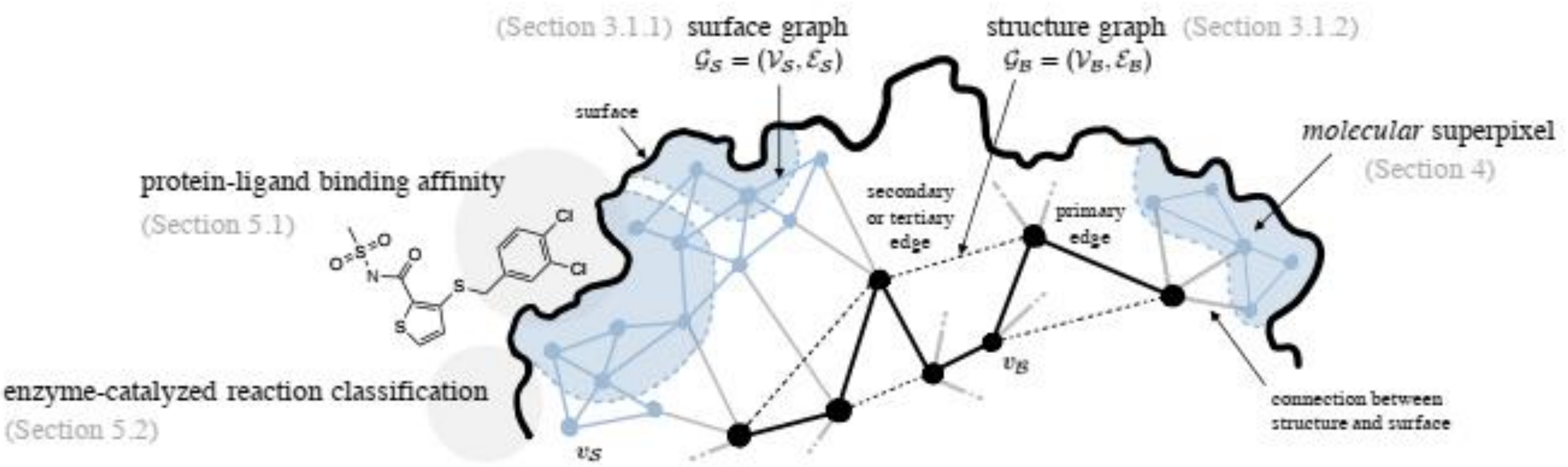
Previous work & Challenges

- Understanding the role and function of proteins is important for studying human diseases
- Representation incorporating the complex nature of the protein is necessary
- Previous study focused on either sequence, structure or surface
- Similar sequence can have completely different structure
- Structures with similar catalyzing property can behave differently towards drugs

Intuition and Design

- Interaction between protein and ligand is controlled by molecular surface contacts
- Hence, important to incorporate surface in the representation
- HOLOPROT consists of a surface and structure layer
- Layers are represented as graphs
- Layers are connected with explicit edges
- Learns representation by integrating the encoding from the layer below
- Propagating information helps to learn higher-level geometric and chemical properties

Multi-scale Protein Representation


(Section 3.1.1) surface graph
$\mathcal{G}_S = (\mathcal{V}_S, \mathcal{E}_S)$
structure graph (Section 3.1.2)
$\mathcal{G}_B = (\mathcal{V}_B, \mathcal{E}_B)$
surface
protein-ligand binding affinity
(Section 5.1)
Cl
Cl
O
O
S
O
N
S
S
secondary
or tertiary
edge
primary
edge
molecular superpixel
(Section 4)
enzyme-catalyzed reaction classification
(Section 5.2)
v_S
v_B
connection between
structure and surface

Multi-scale Protein Representation

- Represent protein P as graph G_p
- Two layers that capture different scales
  - Surface layer
  - Structure layer

Multi-scale Protein Representation

Surface layer

- Represented as a graph G_s
- Surface node u_s has a feature vector f_{us} (charge, hydrophobicity etc.)
- Each node has a residue identifier
- Surface nodes u_s and v_s have an edge if they are part of a triangulation

Structure layer

- Represented as a graph G_B
- Each node u_B corresponds to a residue r
- Two nodes $u_{B,}$ v_B have an edge based on a certain distance between the C_α atoms of the nodes

Multi-scale Protein Representation (Multi-scale Graph)

- The multi-scale graph is obtained by connecting the surface node and the backbone nodes
- The above mentioned nodes have an edge if they have the same residue identifier r
- The graph is encoded by the multi-scale message passing network

Multi-scale Encoder

- Uses one message passing neural network (MPN) for each layer in the multi-scale graph
- MPN_θ – MPN encoding process with parameter θ
- $\mathrm{MLP}_\theta(x,y)$ –Multilayer perceptron with parameter θ and input is the concatenation of x and y
- $\mathrm{MLP}_\theta(x)$ – When input is only x
- id(u) – Residue identifier of node u
- N(u) – Neighbors of node u

Surface Message Passing Network

- Encode the surface layer G_s
- Inputs to the MPN
  - Node features f_{us}
  - Edge features f_{usvs}
- MPN propagates messages between nodes for K iterations
- Output – A representation h_{us} for each surface node u_s

$$\{\mathbf{h}_{u_S}\} = \mathrm{MPN}_{\theta_S}(\mathcal{G}_S, \{\mathbf{f}_{u_S}\}, \{\mathbf{f}_{u_S v_S}\}_{v_S \in \mathcal{N}(u_S)}).$$

Structure Message Passing Network

- Preparation of input to MPN
  - For each node u_B: Concatenate f_{uB} and mean of surface node vector with the same residue identifier
  - Use MLP

$$\mathcal{S} = \{\mathbf{h}_{u_\mathcal{S}} | \mathrm{id}(u_\mathcal{S}) = \mathrm{id}(u_\mathcal{B})\}$$

$$\mathbf{x}_{u_\mathcal{B}} = \mathrm{MLP}_\theta(\mathbf{f}_{u_\mathcal{B}}, \Sigma_\mathcal{S}\, \mathbf{h}_{u_\mathcal{S}}/|\mathcal{S}|).$$

- With edge features f_{uBvB}, run K iterations

$$\{\mathbf{h}_{u_\mathcal{B}}\} = \mathrm{MPN}_{\theta_\mathcal{B}}(\mathcal{G}_\mathcal{B}, \{\mathbf{x}_{u_\mathcal{B}}\}, \{\mathbf{f}_{u_\mathcal{B} v_\mathcal{B}}\}_{v_\mathcal{B} \in \mathcal{N}(u_\mathcal{B})}).$$

Structure Message Passing Network

- Graph representation c_{GP}
  - Aggregation of structure node representation

$$\mathbf{c}_{\mathcal{G}_{\mathcal{P}}} = \sum_{u_{\mathcal{B}} \in \mathcal{G}_{\mathcal{B}}} \mathbf{h}_{u_{\mathcal{B}}}.$$

Task Specific Training

- Multi-scale encoding method is evaluated for two different tasks
  - Protein-ligand binding affinity regression
  - Enzyme-catalyzed reaction classification

Protein-ligand Binding Affinity Prediction

- Depends on the interaction of a protein encoded using HOLOPROT and a ligand (small molecules in most cases)
- Use MPN to encode the ligand represented as graph G_L and aggregate the node features
- Obtain the graph representation c_{GL}
- Concatenate the graph representations of protein and ligand
- Use MLP to obtain the prediction

$$s_a = \text{MLP}_\phi(c_{\mathcal{G}_\mathcal{P}}, c_{\mathcal{G}_\mathcal{L}}).$$

Enzyme-catalyzed Reaction Classification

- Use MLP
- Input
  - The graph representation c_{GP} of the protein obtained via HOLOPROT

$$p_k = \mathrm{MLP}_{\phi}(c_{\mathcal{G}}).$$

Molecular Superpixels

- Segments on the protein surface capturing higher-level fingerprint features
- Improve computational and memory efficiency
- Achieved via optimizing the objective function

$$\max_{\mathcal{M}} -\underbrace{\sum_i \mu_i \sum_j p_{ij}(\mathcal{M}) \log\left(p_{ij}(\mathcal{M})\right)}_{\text{(i.) entropy rate}} - \underbrace{\sum_i p_{Z_{\mathcal{M}}}(i) \log\left(p_{Z_{\mathcal{M}}}(i)\right) - n_{\mathcal{M}}}_{\text{(ii.) balancing function}}$$

$$\text{s.t. } \mathcal{M} \subseteq \mathcal{E}_S \text{ and } n_{\mathcal{M}} \geq k,$$

Also check the second last paragraph of Section 4

Evaluation (Protein-ligand Binding Affinity Prediction)

- Dataset
  - PDBBIND database
  - 4709 biomolecular complexes
- Baselines
  - Sequence based
  - Structure based
  - Geometric deep learning on protein molecular surfaces

Evaluation (Protein-ligand Binding Affinity Prediction)

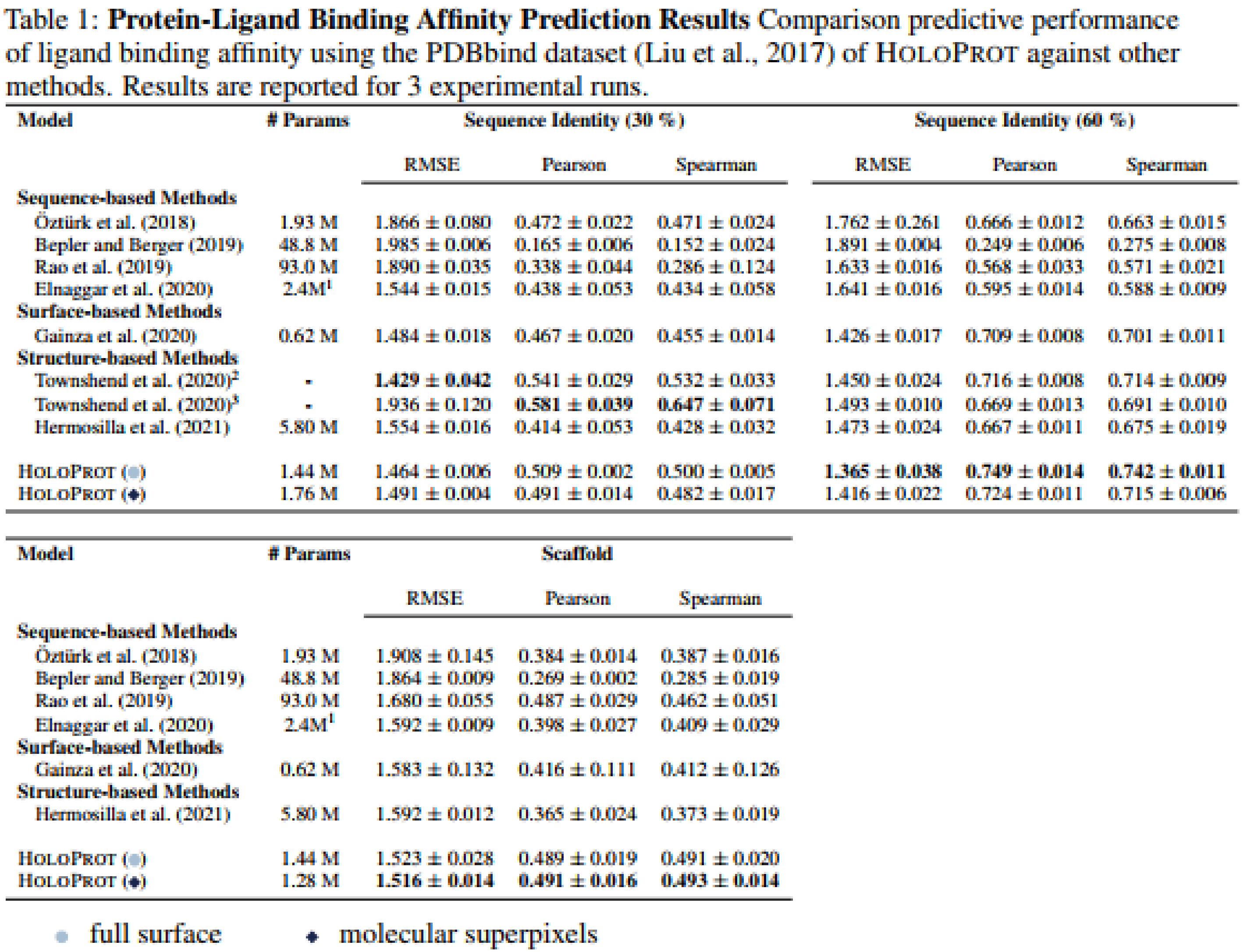


Table 1: **Protein-Ligand Binding Affinity Prediction Results** Comparison predictive performance of ligand binding affinity using the PDBbind dataset (Liu et al., 2017) of HoloProt against other methods. Results are reported for 3 experimental runs.

Model	# Params	Sequence Identity (30 %)			Sequence Identity (60 %)		
		RMSE	Pearson	Spearman	RMSE	Pearson	Spearman
Sequence-based Methods							
Öztürk et al. (2018)	1.93 M	1.866 ± 0.080	0.472 ± 0.022	0.471 ± 0.024	1.762 ± 0.261	0.666 ± 0.012	0.663 ± 0.015
Bepler and Berger (2019)	48.8 M	1.985 ± 0.006	0.165 ± 0.006	0.152 ± 0.024	1.891 ± 0.004	0.249 ± 0.006	0.275 ± 0.008
Rao et al. (2019)	93.0 M	1.890 ± 0.035	0.338 ± 0.044	0.286 ± 0.124	1.633 ± 0.016	0.568 ± 0.033	0.571 ± 0.021
Elnaggar et al. (2020)	2.4M[1]	1.544 ± 0.015	0.438 ± 0.053	0.434 ± 0.058	1.641 ± 0.016	0.595 ± 0.014	0.588 ± 0.009
Surface-based Methods							
Gainza et al. (2020)	0.62 M	1.484 ± 0.018	0.467 ± 0.020	0.455 ± 0.014	1.426 ± 0.017	0.709 ± 0.008	0.701 ± 0.011
Structure-based Methods							
Townshend et al. (2020)[2]	-	**1.429 ± 0.042**	0.541 ± 0.029	0.532 ± 0.033	1.450 ± 0.024	0.716 ± 0.008	0.714 ± 0.009
Townshend et al. (2020)[3]	-	1.936 ± 0.120	**0.581 ± 0.039**	**0.647 ± 0.071**	1.493 ± 0.010	0.669 ± 0.013	0.691 ± 0.010
Hermosilla et al. (2021)	5.80 M	1.554 ± 0.016	0.414 ± 0.053	0.428 ± 0.032	1.473 ± 0.024	0.667 ± 0.011	0.675 ± 0.019
HoloProt (●)	1.44 M	1.464 ± 0.006	0.509 ± 0.002	0.500 ± 0.005	**1.365 ± 0.038**	**0.749 ± 0.014**	**0.742 ± 0.011**
HoloProt (◆)	1.76 M	1.491 ± 0.004	0.491 ± 0.014	0.482 ± 0.017	1.416 ± 0.022	0.724 ± 0.011	0.715 ± 0.006

Model	# Params	Scaffold		
		RMSE	Pearson	Spearman
Sequence-based Methods				
Öztürk et al. (2018)	1.93 M	1.908 ± 0.145	0.384 ± 0.014	0.387 ± 0.016
Bepler and Berger (2019)	48.8 M	1.864 ± 0.009	0.269 ± 0.002	0.285 ± 0.019
Rao et al. (2019)	93.0 M	1.680 ± 0.055	0.487 ± 0.029	0.462 ± 0.051
Elnaggar et al. (2020)	2.4M[1]	1.592 ± 0.009	0.398 ± 0.027	0.409 ± 0.029
Surface-based Methods				
Gainza et al. (2020)	0.62 M	1.583 ± 0.132	0.416 ± 0.111	0.412 ± 0.126
Structure-based Methods				
Hermosilla et al. (2021)	5.80 M	1.592 ± 0.012	0.365 ± 0.024	0.373 ± 0.019
HoloProt (●)	1.44 M	1.523 ± 0.028	0.489 ± 0.019	0.491 ± 0.020
HoloProt (◆)	1.28 M	**1.516 ± 0.014**	**0.491 ± 0.016**	**0.493 ± 0.014**

● full surface ◆ molecular superpixels

Evaluation (Enzyme-catalyzed Reaction Classification)

- Dataset
  - 37428 proteins from 384 EC numbers
- Baselines
  - Sequence based
  - Partially pretrained on millions of sequences
  - Geometric deep learning based

Evaluation (Enzyme-catalyzed Reaction Classification)

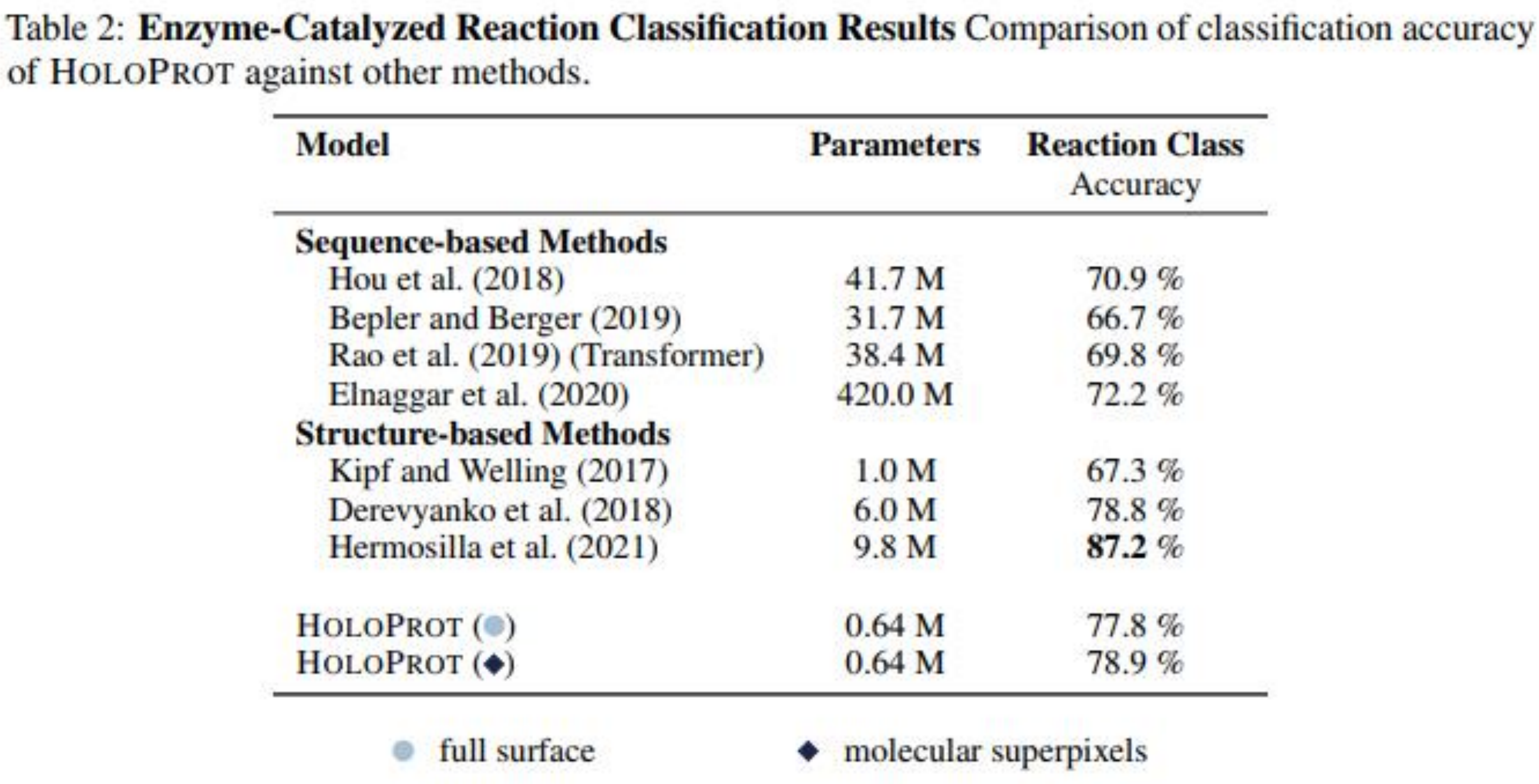


Table 2: **Enzyme-Catalyzed Reaction Classification Results** Comparison of classification accuracy of HoloProt against other methods.

Model	**Parameters**	**Reaction Class** Accuracy
Sequence-based Methods		
Hou et al. (2018)	41.7 M	70.9 %
Bepler and Berger (2019)	31.7 M	66.7 %
Rao et al. (2019) (Transformer)	38.4 M	69.8 %
Elnaggar et al. (2020)	420.0 M	72.2 %
Structure-based Methods		
Kipf and Welling (2017)	1.0 M	67.3 %
Derevyanko et al. (2018)	6.0 M	78.8 %
Hermosilla et al. (2021)	9.8 M	**87.2** %
HoloProt (●)	0.64 M	77.8 %
HoloProt (◆)	0.64 M	78.9 %

● full surface ◆ molecular superpixels

Ablation Studies

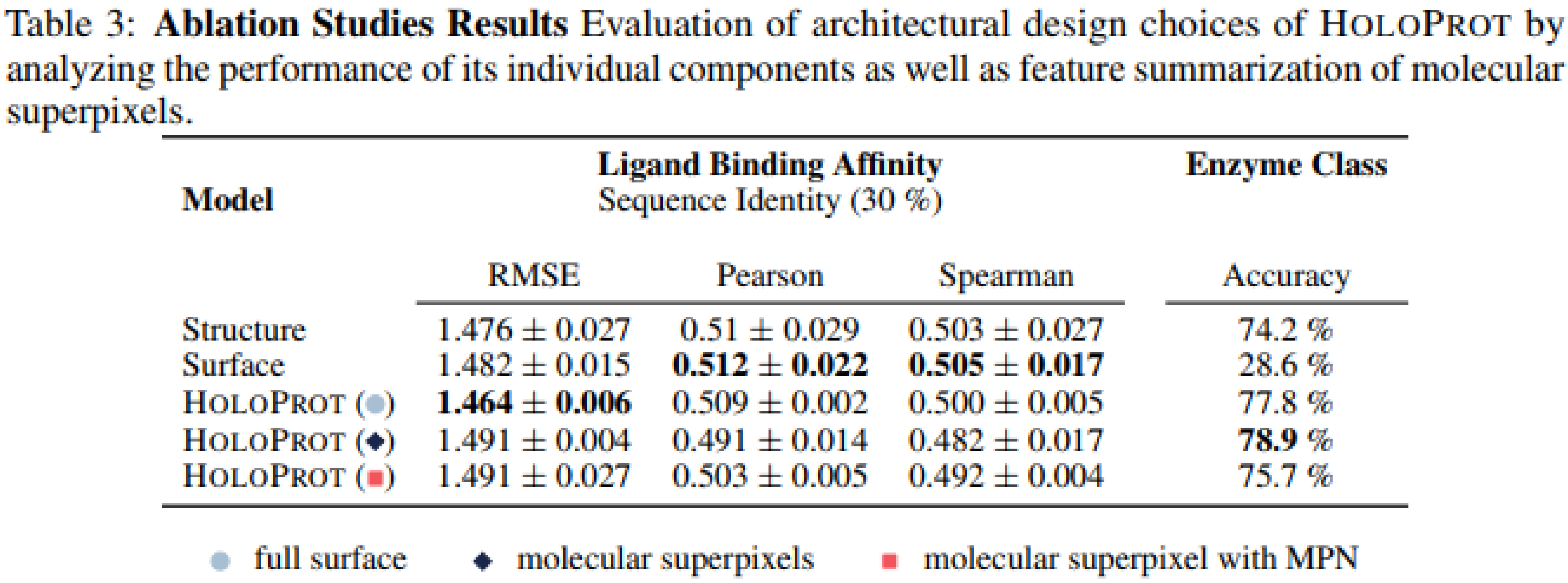


Table 3: **Ablation Studies Results** Evaluation of architectural design choices of HoloProt by analyzing the performance of its individual components as well as feature summarization of molecular superpixels.

Model	**Ligand Binding Affinity** Sequence Identity (30 %)			**Enzyme Class**
	RMSE	Pearson	Spearman	Accuracy
Structure	1.476 ± 0.027	0.51 ± 0.029	0.503 ± 0.027	74.2 %
Surface	1.482 ± 0.015	**0.512 ± 0.022**	**0.505 ± 0.017**	28.6 %
HoloProt (●)	**1.464 ± 0.006**	0.509 ± 0.002	0.500 ± 0.005	77.8 %
HoloProt (◆)	1.491 ± 0.004	0.491 ± 0.014	0.482 ± 0.017	**78.9** %
HoloProt (■)	1.491 ± 0.027	0.503 ± 0.005	0.492 ± 0.004	75.7 %

● full surface ◆ molecular superpixels ■ molecular superpixel with MPN

Limitations of The Work

- Relies on existing protein structures, although there are a lot of protein sequence data
- Requires precomputed surface meshes resulting in an additional preprocessing step

Thank you