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

- Surface graph $\mathcal{G}_S = (\mathcal{V}_S, \mathcal{E}_S)$ (Section 3.1.1)
- Structure graph $\mathcal{G}_B = (\mathcal{V}_B, \mathcal{E}_B)$ (Section 3.1.2)
- Protein-ligand binding affinity (Section 5.1)
- Enzyme-catalyzed reaction classification (Section 5.2)
- Molecular superpixel (Section 4)
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 $C_\alpha$ 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
- $\text{MPN}_\theta$ – MPN encoding process with parameter $\theta$
- $\text{MLP}_\theta(x,y)$ – Multilayer perceptron with parameter $\theta$ and input is the concatenation of $x$ and $y$
- $\text{MLP}_\theta(x)$ – When input is only $x$
- $\text{id}(u)$ – Residue identifier of node $u$
- $\text{N}(u)$ – 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 $u_s$

\[
\{h_{us}\} = \text{MPN}_{\theta_S}(G_S, \{f_{us}\}, \{f_{usvs}\}_{v_S \in N(u_S)}).
\]
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

\[
S = \{ h_{uS} | \text{id}(u_S) = \text{id}(u_B) \} 
\]

\[
x_{u_B} = \text{MLP}_\theta(f_{uB}, \sum_S h_{uS}/|S|). 
\]

- With edge features $f_{u_B v_B}$, run $K$ iterations

\[
\{ h_{u_B} \} = \text{MPN}_{\theta_B}(G_B, \{ x_{u_B} \}, \{ f_{u_B v_B} \}_{v_B \in N(u_B)}). 
\]
Structure Message Passing Network

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

\[
c_{GP} = \sum_{u_B \in G_B} h_{u_B}.
\]
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

$$s_a = \text{MLP}_\phi(c_{GP}, c_{GL}).$$
Enzyme-catalyzed Reaction Classification

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

$$p_k = \text{MLP}_\phi(c_G).$$
Molecular Superpixels

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

\[
\max_{\mathcal{M}} - \sum_i \mu_i \sum_j p_{ij}(\mathcal{M}) \log(p_{ij}(\mathcal{M})) - \sum_i p_{Z,\mathcal{M}}(i) \log(p_{Z,\mathcal{M}}(i)) - n_{\mathcal{M}}
\]

- (i.) entropy rate
- (ii.) balancing function

s.t. \( \mathcal{M} \subseteq \mathcal{E}_S \) 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.

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

<table>
<thead>
<tr>
<th>Model</th>
<th># Params</th>
<th>Sequence Identity (60 %)</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>RMSE</td>
<td>Pearson</td>
<td>Spearman</td>
<td>RMSE</td>
<td>Pearson</td>
<td>Spearman</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sequence-based Methods</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Öztürk et al. (2018)</td>
<td>1.93 M</td>
<td>1.908 ± 0.145</td>
<td>0.384 ± 0.014</td>
<td>0.387 ± 0.016</td>
<td>1.793 ± 0.048</td>
<td>0.669 ± 0.010</td>
<td>0.670 ± 0.008</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bepler and Berger (2019)</td>
<td>48.8 M</td>
<td>1.864 ± 0.009</td>
<td>0.269 ± 0.002</td>
<td>0.285 ± 0.019</td>
<td>1.891 ± 0.004</td>
<td>0.249 ± 0.006</td>
<td>0.275 ± 0.008</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rao et al. (2019)</td>
<td>93.0 M</td>
<td>1.680 ± 0.055</td>
<td>0.487 ± 0.029</td>
<td>0.462 ± 0.051</td>
<td>1.633 ± 0.016</td>
<td>0.568 ± 0.033</td>
<td>0.571 ± 0.021</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Elnaggar et al. (2020)</td>
<td>2.4M</td>
<td>1.592 ± 0.009</td>
<td>0.398 ± 0.027</td>
<td>0.409 ± 0.029</td>
<td>1.641 ± 0.016</td>
<td>0.595 ± 0.014</td>
<td>0.588 ± 0.009</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Surface-based Methods</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gainza et al. (2020)</td>
<td>0.62 M</td>
<td>1.583 ± 0.132</td>
<td>0.416 ± 0.111</td>
<td>0.412 ± 0.126</td>
<td>1.426 ± 0.017</td>
<td>0.709 ± 0.008</td>
<td>0.701 ± 0.011</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Structure-based Methods</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hermosilla et al. (2021)</td>
<td>5.80 M</td>
<td>1.592 ± 0.012</td>
<td>0.365 ± 0.024</td>
<td>0.373 ± 0.019</td>
<td>1.473 ± 0.024</td>
<td>0.667 ± 0.011</td>
<td>0.675 ± 0.019</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HOLOPROT (θ)</td>
<td>1.44 M</td>
<td>1.523 ± 0.028</td>
<td>0.489 ± 0.019</td>
<td>0.491 ± 0.020</td>
<td>1.365 ± 0.038</td>
<td>0.749 ± 0.014</td>
<td>0.742 ± 0.011</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HOLOPROT (ϕ)</td>
<td>1.28 M</td>
<td>1.516 ± 0.014</td>
<td>0.491 ± 0.016</td>
<td>0.493 ± 0.014</td>
<td>1.416 ± 0.022</td>
<td>0.724 ± 0.011</td>
<td>0.715 ± 0.006</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- 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.

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

- 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.

<table>
<thead>
<tr>
<th>Model</th>
<th>Ligand Binding Affinity</th>
<th>Enzyme Class</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RMSE</td>
<td>Pearson</td>
</tr>
<tr>
<td>Structure</td>
<td>1.476 ± 0.027</td>
<td>0.51 ± 0.029</td>
</tr>
<tr>
<td>Surface</td>
<td>1.482 ± 0.015</td>
<td>0.512 ± 0.022</td>
</tr>
<tr>
<td>HOLOPROT (●)</td>
<td>1.464 ± 0.006</td>
<td>0.509 ± 0.002</td>
</tr>
<tr>
<td>HOLOPROT (●●)</td>
<td>1.491 ± 0.004</td>
<td>0.491 ± 0.014</td>
</tr>
<tr>
<td>HOLOPROT (●●●)</td>
<td>1.491 ± 0.027</td>
<td>0.503 ± 0.005</td>
</tr>
</tbody>
</table>

- 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