M3Epi: an Hierarchical and Relation-aware Graph Neural Network for Antibody-aware Epitope Prediction

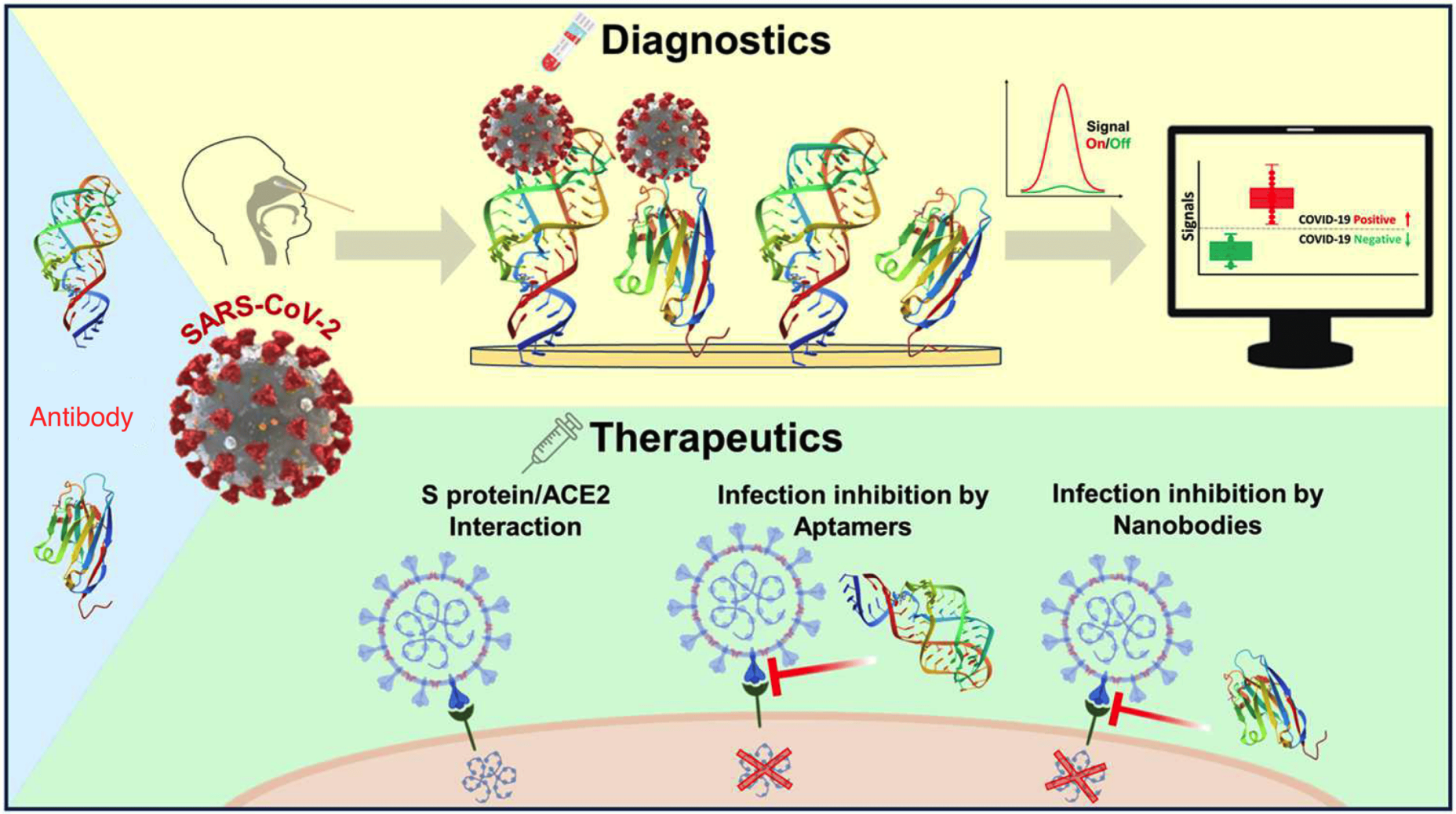
Mansoor Ahmed

UCLA CGSI-2025

Agenda

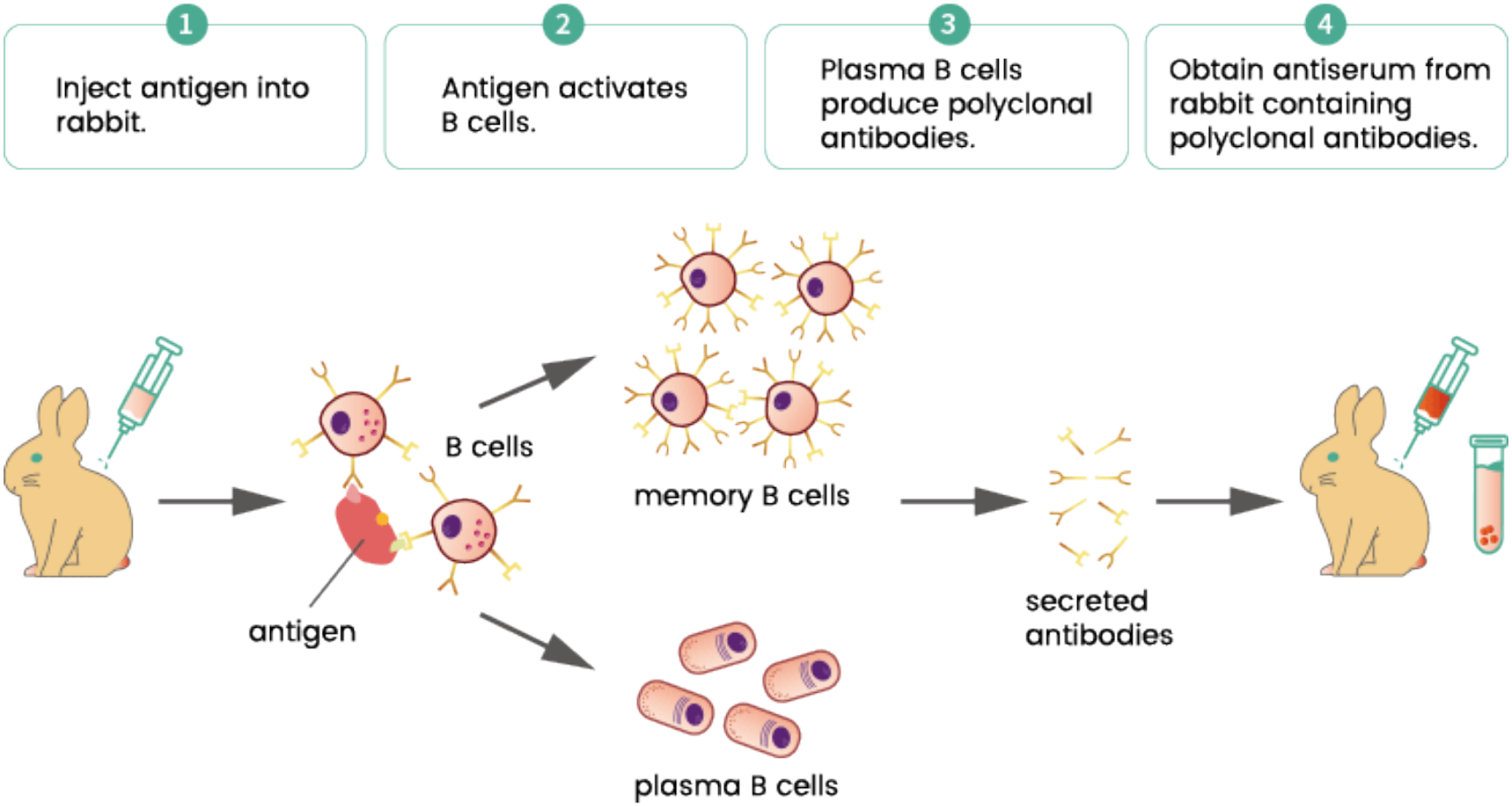
1. **Motivation for the "antibody design" problem**
2. Formulating the "epitope prediction" problem
3. Model architecture and dataset
4. Results and ablation studies

Why are we interested in Antibodies?



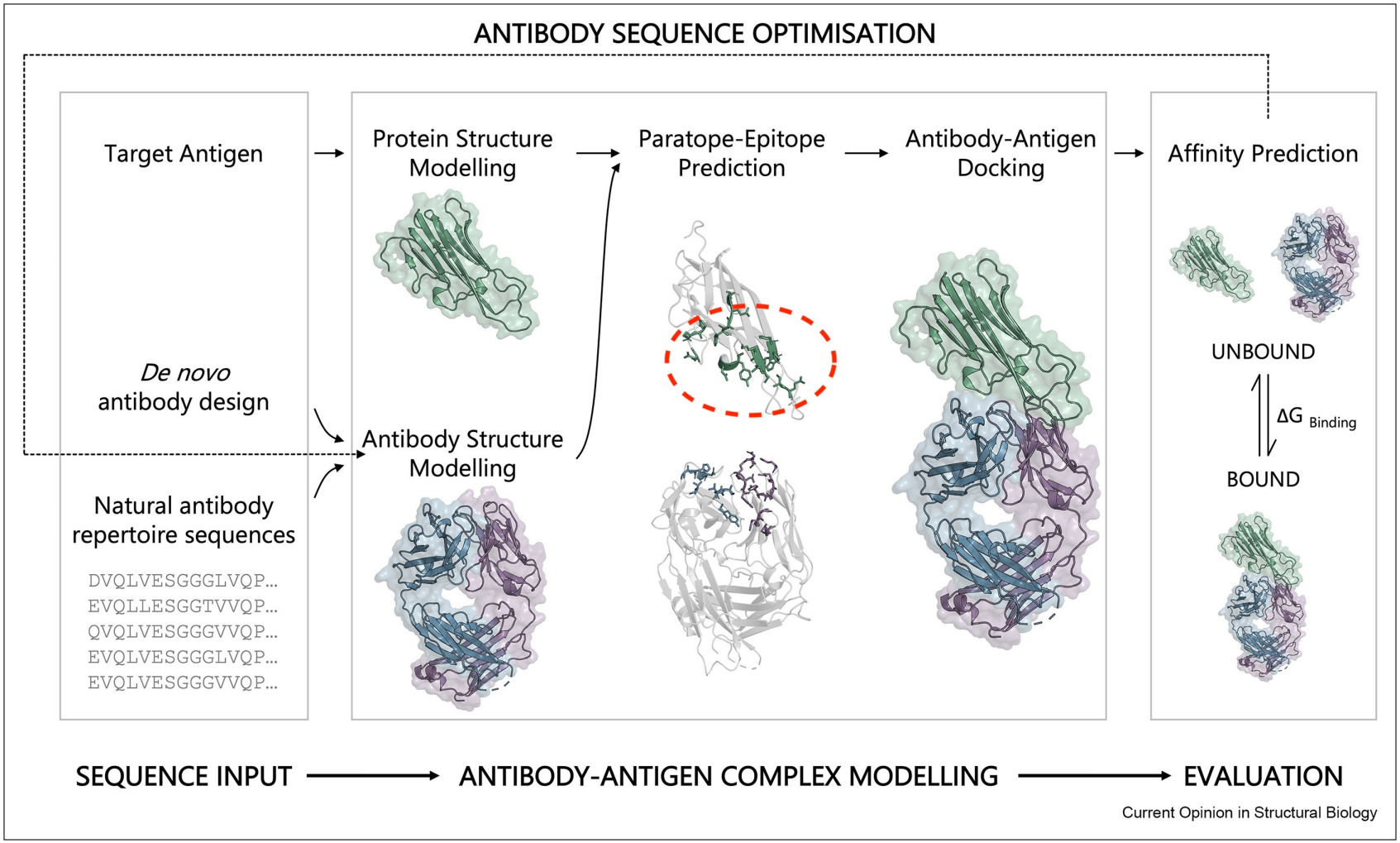
(Source: Park et al. [2024])

What do we currently have? *In-vivo* methods



Source: Lumen Learning

We can do better – why not "design" antibodies?



Source: Hummer et al. [2022]

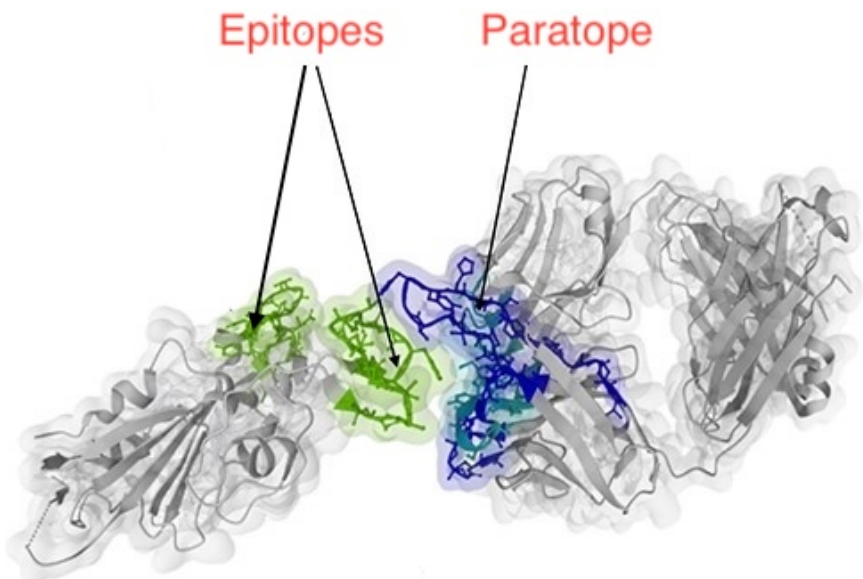
Some Definitions

Antigen

- Toxin, bacteria, or virus
- Induces an immune response producing antibodies
- Epitope: regions on antigens recognized by antibodies

Antibody

- Large and Y-shaped protein produced by B cells
- Identifies and neutralizes antigens
- Paratope: antibody binding site



Agenda

1. Motivation for the "antibody design" problem
2. **Formulating the "epitope prediction" problem**
3. Model architecture and dataset
4. Results and ablation studies

Current Challenges for Epitope Prediction

- Non-conformational/sequential – sequence-based approaches fail to capture the spatial relationship
- Ineffective protein representation and limited datasets
- Multiple epitopes on a single antigen

Table: Current baseline methods.

Method	F1 Score	MCC Score	Precision	Recall	AUROC	AUPRC
EpiGraph	0.247	0.240	0.145	0.852	**0.819**	**0.279**
EpiScan	0.197	0.043	-0.115	**0.912**	0.593	0.229
WALLE	0.145	0.210	0.235	0.258	0.635	-
MIPE	0.169	0.176	**0.317**	0.248	0.774	0.213

Current Challenges for Epitope Prediction

- Non-conformational/sequential – sequence-based approaches fail to capture the spatial relationship
- Ineffective protein representation and limited datasets
- Multiple epitopes on a single antigen

Table: Current baseline methods.

Method	F1 Score	MCC Score	Precision	Recall	AUROC	AUPRC
EpiGraph	0.247	0.240	0.145	0.852	**0.819**	**0.279**
EpiScan	0.197	0.043	-0.115	**0.912**	0.593	0.229
WALLE	0.145	0.210	0.235	0.258	0.635	-
MIPE	0.169	0.176	**0.317**	0.248	0.774	0.213

Our idea: develop GNNs that can:

1. reason on multiple levels of protein (atom, residue, & edge)
2. capture multi-relational edge relationships within the protein
3. incorporate protein geometry by constraining rotational and translational equivariance in E(3)-space

Problem Formulation

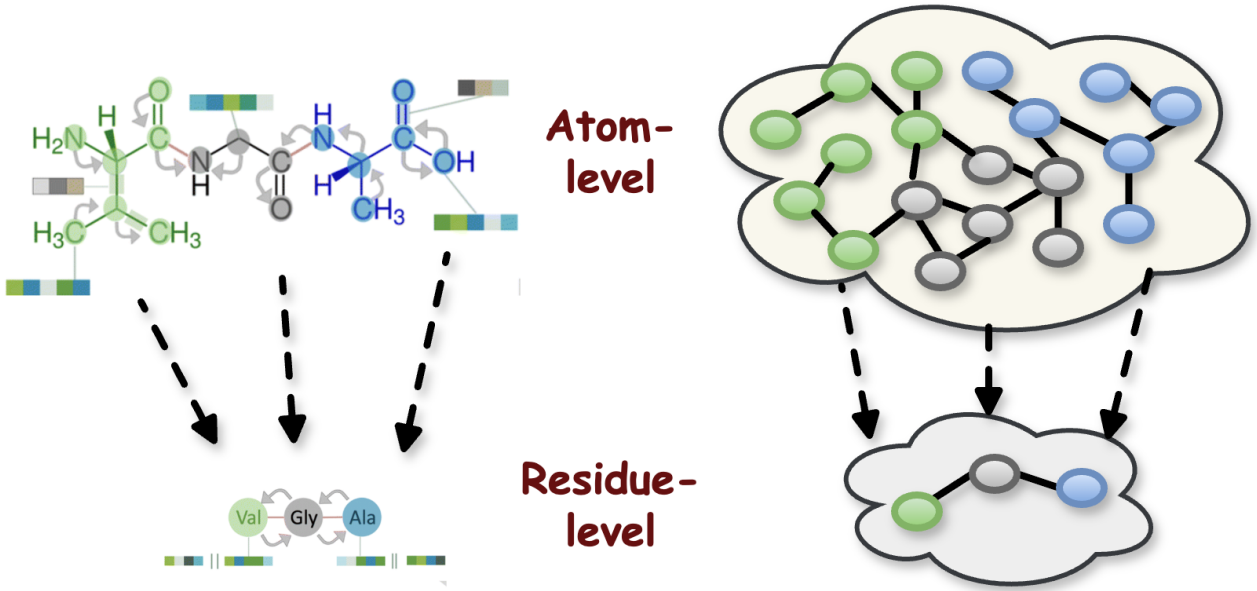
- **Input**: Two disjoint sets of atom, residue, and edge graphs for antibody and antigen:
  - Atom graph: $\mathcal{G}_a = (\mathcal{V}_a, \mathcal{E}_a)$, where adjacency matrix (edge list) $\mathcal{E}_a$ is based on residue proximity (distance < 4.5Å) and atom nodes are encoded into a vector
  - Residue graph: $\mathcal{G}_r = (\mathcal{V}_r, \mathcal{E}_r, \mathcal{R}_r)$, where $\mathcal{E}_r$ is based on sequential and spatial proximity and $\mathcal{R}_r$ represents the set of edge-relations
  - Edge (line) graph, $\mathcal{G}_e = (\mathcal{V}_e, \mathcal{E}_e)$, is the complement of the residue graph $\mathcal{G}_r^{'}$
- **Output**: Nodes and edges between interacting antibody and antigen residues.

Problem Formulation

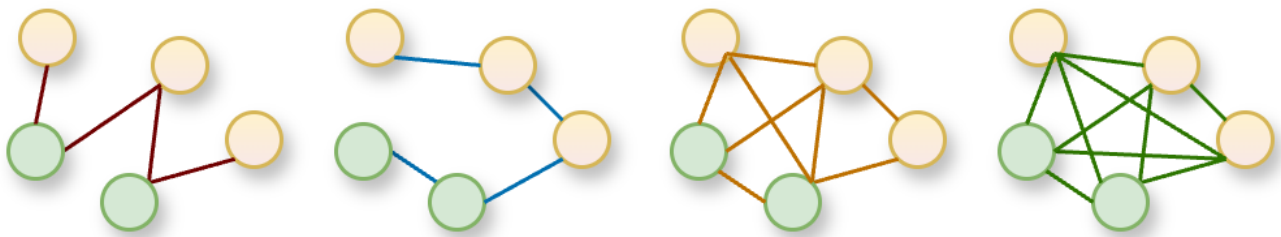
- **Input**: Two disjoint sets of atom, residue, and edge graphs for antibody and antigen:
  - Atom graph: $\mathcal{G}_a = (\mathcal{V}_a, \mathcal{E}_a)$, where adjacency matrix (edge list) $\mathcal{E}_a$ is based on residue proximity (distance < 4.5Å) and atom nodes are encoded into a vector
  - Residue graph: $\mathcal{G}_r = (\mathcal{V}_r, \mathcal{E}_r, \mathcal{R}_r)$, where $\mathcal{E}_r$ is based on sequential and spatial proximity and $\mathcal{R}_r$ represents the set of edge-relations
  - Edge (line) graph, $\mathcal{G}_e = (\mathcal{V}_e, \mathcal{E}_e)$, is the complement of the residue graph $\mathcal{G}_r^{'}$
- **Output**: Nodes and edges between interacting antibody and antigen residues.

Given two disjoint antibody and antigen graphs, predict the binding nodes in the antigen – essentially a binary classification problem!!

Illustration of the Input Data



sequential edges: r_1, r_2 spatial edges: r_3, r_4



Equivariance and Invariance in E(3) Space

- Introduce geometric (spatial) inductive bias using equivariant graph neural networks (EGNN) Jiao et al. [2023]
- Let $\mathbf{X} \in \mathbb{R}^{3\times m}$ be a collection of m atom coordinates and $\mathbf{h} \in \mathbb{R}^{d}$ be non-geometric features (charges, residue indices, etc.).

A mapping $f : (\mathbf{X}, \mathbf{h}) \mapsto (\mathbf{X}', \mathbf{h}')$ is $\mathrm{E}(3)$-*equivariant* if

$$f\big(g \cdot \mathbf{X}, \mathbf{h}\big) = g \cdot f\big(\mathbf{X}, \mathbf{h}\big), \quad \forall\, g \in \mathrm{E}(3) \quad \text{(rotation / reflection)}$$

Equivariance and Invariance in E(3) Space

- Introduce geometric (spatial) inductive bias using equivariant graph neural networks (EGNN) Jiao et al. [2023]
- Let $\mathbf{X} \in \mathbb{R}^{3\times m}$ be a collection of m atom coordinates and $\mathbf{h} \in \mathbb{R}^{d}$ be non-geometric features (charges, residue indices, etc.).

A mapping $f : (\mathbf{X}, \mathbf{h}) \mapsto (\mathbf{X}', \mathbf{h}')$ is $\mathrm{E}(3)$-*equivariant* if

$$f\big(g \cdot \mathbf{X}, \mathbf{h}\big) = g \cdot f\big(\mathbf{X}, \mathbf{h}\big), \quad \forall\, g \in \mathrm{E}(3) \quad \text{(rotation / reflection)}$$

At depth ℓ, an EGNN layer performs updates as follows:

$$\mathbf{m}_{ij}^{(\ell)} = \phi_m\Big(\mathbf{h}_i^{(\ell-1)}, \mathbf{h}_j^{(\ell-1)}, e_{ij}, \|\mathbf{x}_i^{(\ell-1)} - \mathbf{x}_j^{(\ell-1)}\|^2\Big),$$

$$\mathbf{x}_i^{(\ell)} = \mathbf{x}_i^{(\ell-1)} + \frac{1}{N_i} \sum_{j:(i,j)\in\mathcal{E}_a} (\mathbf{x}_i^{(\ell-1)} - \mathbf{x}_j^{(\ell-1)})\phi_x\big(\mathbf{m}_{ij}^{(\ell)}\big),$$

$$\mathbf{h}_i^{(\ell)} = \phi_h\Big(\mathbf{h}_i^{(\ell-1)}, \sum_{j:(i,j)\in\mathcal{E}_a} \mathbf{m}_{ij}^{(\ell)}\Big),$$

Formal Definition of the Tasks

- **Task 1: Epitope Node Classification** (Ahmed et al. [2025])
  Epitopes are antigen residues in contact with the antibody (distance < 4.5Å)

  Binary classification of antigen nodes:
  - Label $= 1$: Epitope residue
  - Label $= 0$: Non-epitope residue

$$f(v; G_B, G_A) = \begin{cases} 1 & \text{if } v \text{ is an epitope,} \\ 0 & \text{otherwise.} \end{cases}$$

Formal Definition of the Tasks

- **Task 1: Epitope Node Classification** (Ahmed et al. [2025])
  Epitopes are antigen residues in contact with the antibody (distance < 4.5Å)

  Binary classification of antigen nodes:
  - Label $= 1$: Epitope residue
  - Label $= 0$: Non-epitope residue

$$f(v; G_B, G_A) = \begin{cases} 1 & \text{if } v \text{ is an epitope,} \\ 0 & \text{otherwise.} \end{cases}$$

- **Task 2: Bipartite Link Prediction**
  Predict interactions between antibody and antigen residues.

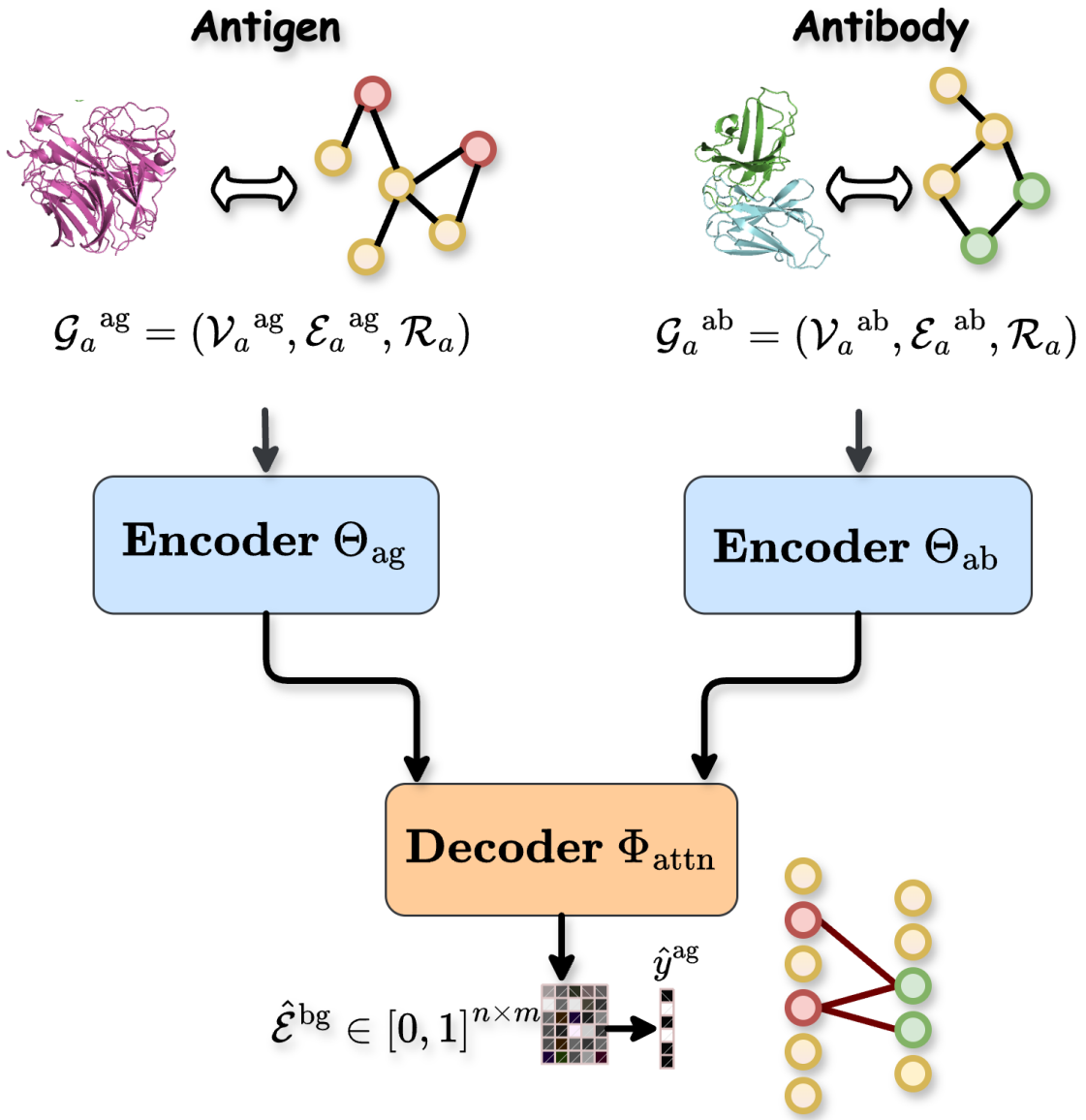
  Binary classification of edges in the antigen-antibody bipartite graph:
  - Label $= 1$: Residues are in contact (distance < 4.5Å)
  - Label $= 0$: Residues are not in contact

$$g(v_a, v_b; K_{m,n}) = \begin{cases} 1 & \text{if } v_a \text{ and } v_b \text{ are in contact,} \\ 0 & \text{otherwise.} \end{cases}$$

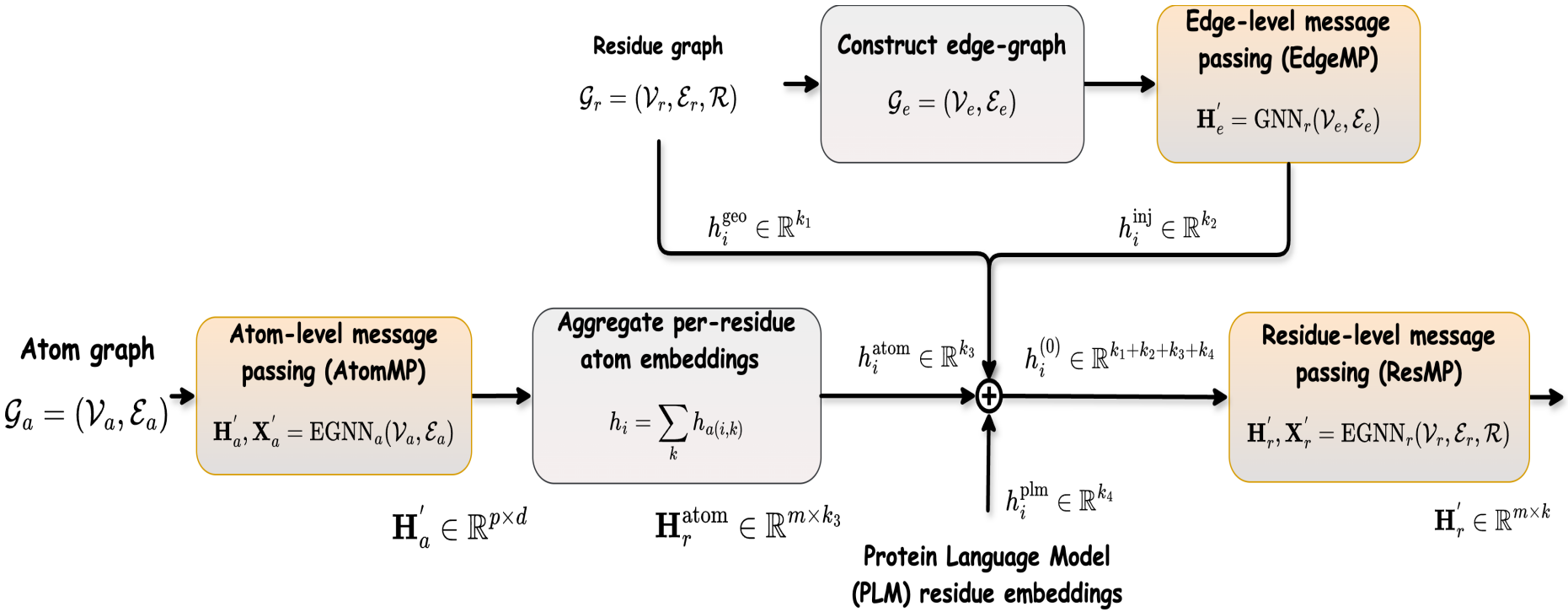
Agenda

1. Motivation for the "antibody design" problem
2. Formulating the "epitope prediction" problem
3. **Model architecture and dataset**
4. Results and ablation studies

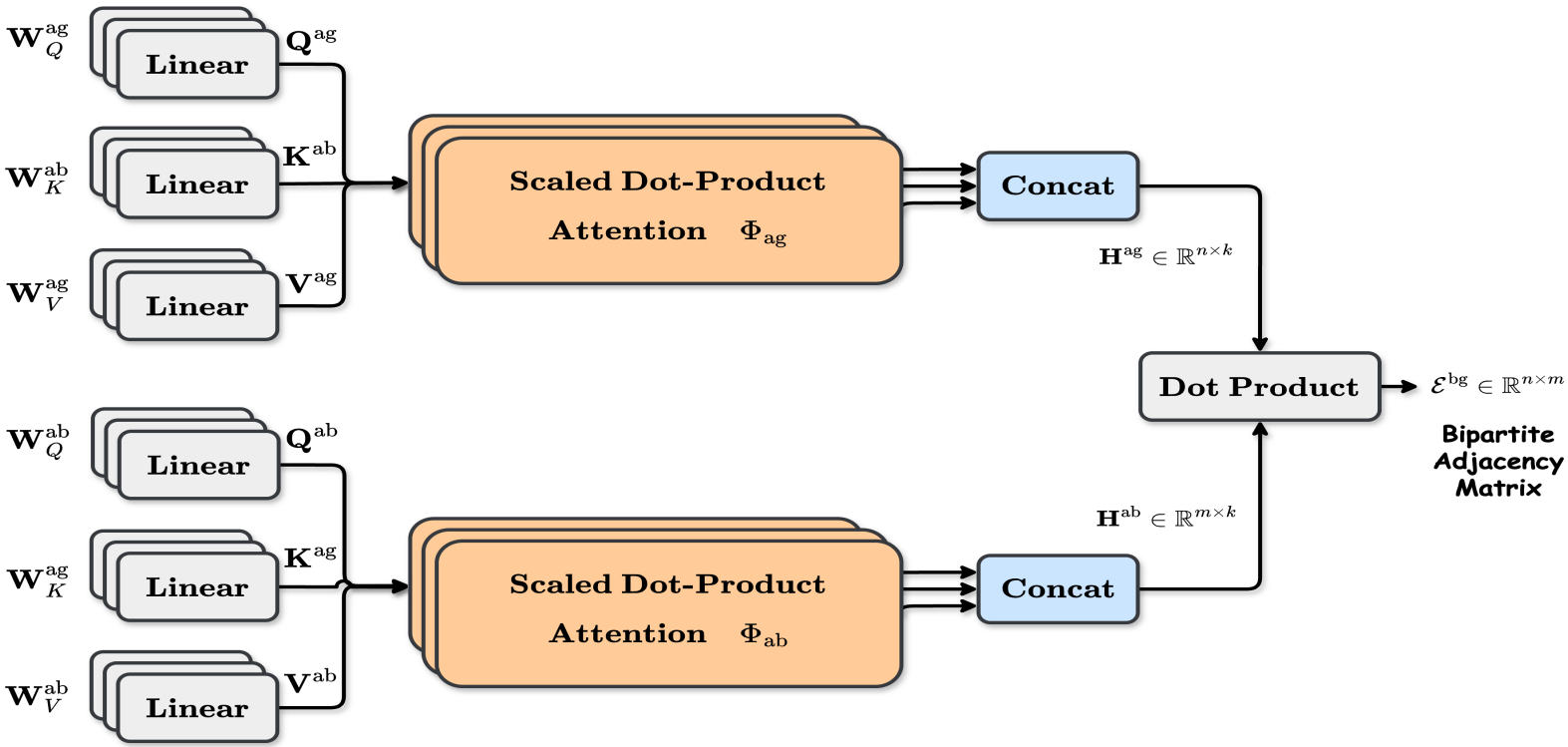
Model Architecture: Overall Framework



Model Architecture: Hierarchical Encoder



Model Architecture: Cross Attention Decoder



Loss Functions

Primary loss functions

1. **Binary Cross-Entropy Loss**:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{|\mathcal{E}^{\text{bg}}|} \sum_{(i,j)\in\mathcal{E}^{\text{bg}}} y_{ij} \log \hat{y}_{ij} + (1 - y_{ij}) \log(1 - \hat{y}_{ij}),$$

2. **Gradient-Weighted NCE Loss (Ji et al. [2024])**:

$$\mathcal{L}_{\text{GW-NCE}} = -\sum_{i\in\mathcal{V}_r^{\text{Ag}}} \log \frac{\sum_{j\in\mathcal{P}(i)} e^{s_{ij}/\tau}}{\sum_{j\in\mathcal{P}(i)\cup\mathcal{N}(i)} \alpha_{ij} e^{s_{ij}/\tau}},$$

Dual auxiliary "force" losses

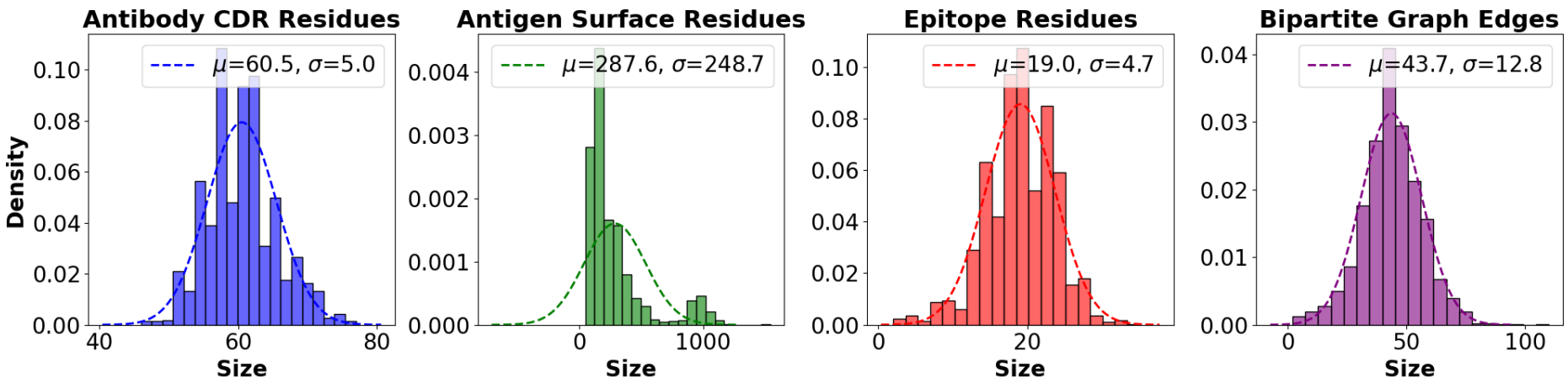
1. **Residue–level inter-graph loss** :
$\hat{d}_{i,j}^{\text{inter}} = \|h(c_i) - h(a_j)\|, \quad L_{\text{inter}} = \sum_{(i,j)\in\mathcal{I}} \text{Huber}\big(\hat{d}_{i,j}^{\text{inter}},\, d_{i,j}^{\text{true}}\big)$

2. **Atom–level intra-residue loss** capturing local side–chain geometry :

$$\hat{d}_{i,k,k'}^{\text{intra}} = \|h(v_{i,k}^{\text{Ag}}) - h(v_{i,k'}^{\text{Ag}})\|, \qquad L_{\text{intra}} = \sum_i \sum_{k<k'} \text{Huber}\big(\hat{d}_{i,k,k'}^{\text{intra}},\, d_{i,k,k'}^{\text{true}}\big),$$

AsEP Benchmark Dataset

- Currated and benchmarked by Liu et al. [2024] using Antibody Database (AbDb) and Protein Data Bank (PDB) at NeurIPS-2024
- **Size**: 1,723 antibody-antigen complexes with labeled binding sites



Agenda

1. Motivation for the "antibody design" problem
2. Formulating the "epitope prediction" problem
3. Model architecture and dataset
4. **Results**

Results: Baseline Comparison and Ablation Studies

Method	F1 Score	MCC Score	Precision	Recall	AUROC	AUPRC
EpiGraph	0.247	0.240	0.145	0.852	**0.819**	**0.279**
EpiScan	0.197	0.043	-0.115	**0.912**	0.593	0.229
WALLE	0.145	0.210	0.235	0.258	0.635	-
MIPE	0.169	0.176	**0.317**	0.248	0.774	0.213
M3Epi	**0.314 ± 0.013**	**0.292 ± 0.006**	0.209 ± 0.012	0.635 ± 0.010	0.736 ± 0.003	0.156 ± 0.008

- Encoder:
  - single-level vs. hierarchical (multi-level)
  - uni-relational (homogeneous) vs. multi-relational (heterogeneous) graphs
  - Vanilla GNN modules (GAT, GCN, GIN) vs. equivariant GNN blocks (EGCN, ET, SchNet)
- Interaction decoder: ATTENTION versus lightweight DOT product
- Loss functions: contrastive objective (vanilla InfoNCE versus the gradient-weighted GW-NCE) with auxiliary losses
- Self-supervised pre-training with Langevin energy modeling, multi-view contrastive learning, and graph self-prediction tasks (on-going)

References

Mansoor Ahmed, Sarwan Ali, Avais Jan, Imdad Ullah Khan, and Murray Patterson. Improved graph-based antibody-aware epitope prediction with protein language model-based embeddings. *bioRxiv*, 2025. doi: 10.1101/2025.02.12.637989. URL `https://www.biorxiv.org/content/early/2025/02/17/2025.02.12.637989`.

Alissa Hummer, Brennan Abanades, and Charlotte Deane. Advances in computational structure-based antibody design. *Current Opinion in Structural Biology*, 74:102379, 2022.

Cheng Ji, Zixuan Huang, Qingyun Sun, Hao Peng, Xingcheng Fu, Qian Li, and Jianxin Li. Regcl: rethinking message passing in graph contrastive learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 8544–8552, 2024.

Rui Jiao, Jiaqi Han, Wenbing Huang, Yu Rong, and Yang Liu. Energy-motivated equivariant pretraining for 3d molecular graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 8096–8104, 2023.

Chunan Liu, Lilian Denzler, Yihong Chen, Andrew Martin, and Brooks Paige. AsEP: Benchmarking deep learning methods for antibody-specific epitope prediction. *arXiv preprint arXiv:2407.18184*, 2024.

Ki Sung Park, Tae-In Park, Jae Eon Lee, Seo-Yeong Hwang, Anna Choi, and Seung Pil Pack. Aptamers and nanobodies as new bioprobes for sars-cov-2 diagnostic and therapeutic system applications. *Biosensors*, 14(3):146, 2024.

Thanks to:

1. **Murray Patterson (PI)**
2. **Hemanth Venkateswara**
3. **Huriong Chai**

Algorithmic Biology Lab at GSU:
`https://alibilab.github.io/`

Questions & Suggestions!!