

AsEP: Benchmarking Deep Learning Methods for Antibody-specific Epitope Prediction

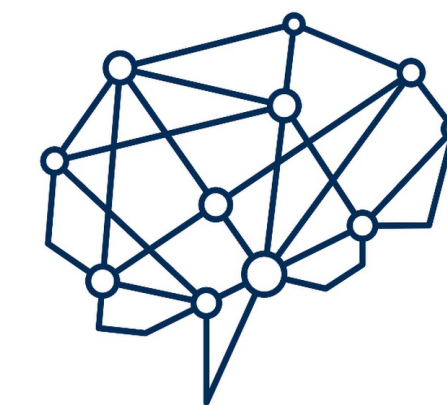
WALLE: a hybrid method leveraging PLMs and GNNs

ChuNan Liu Lilian Denzler Yihong Chen

Andrew C.R. Martin Brooks Paige



powered by 



UCL CENTRE FOR
ARTIFICIAL INTELLIGENCE



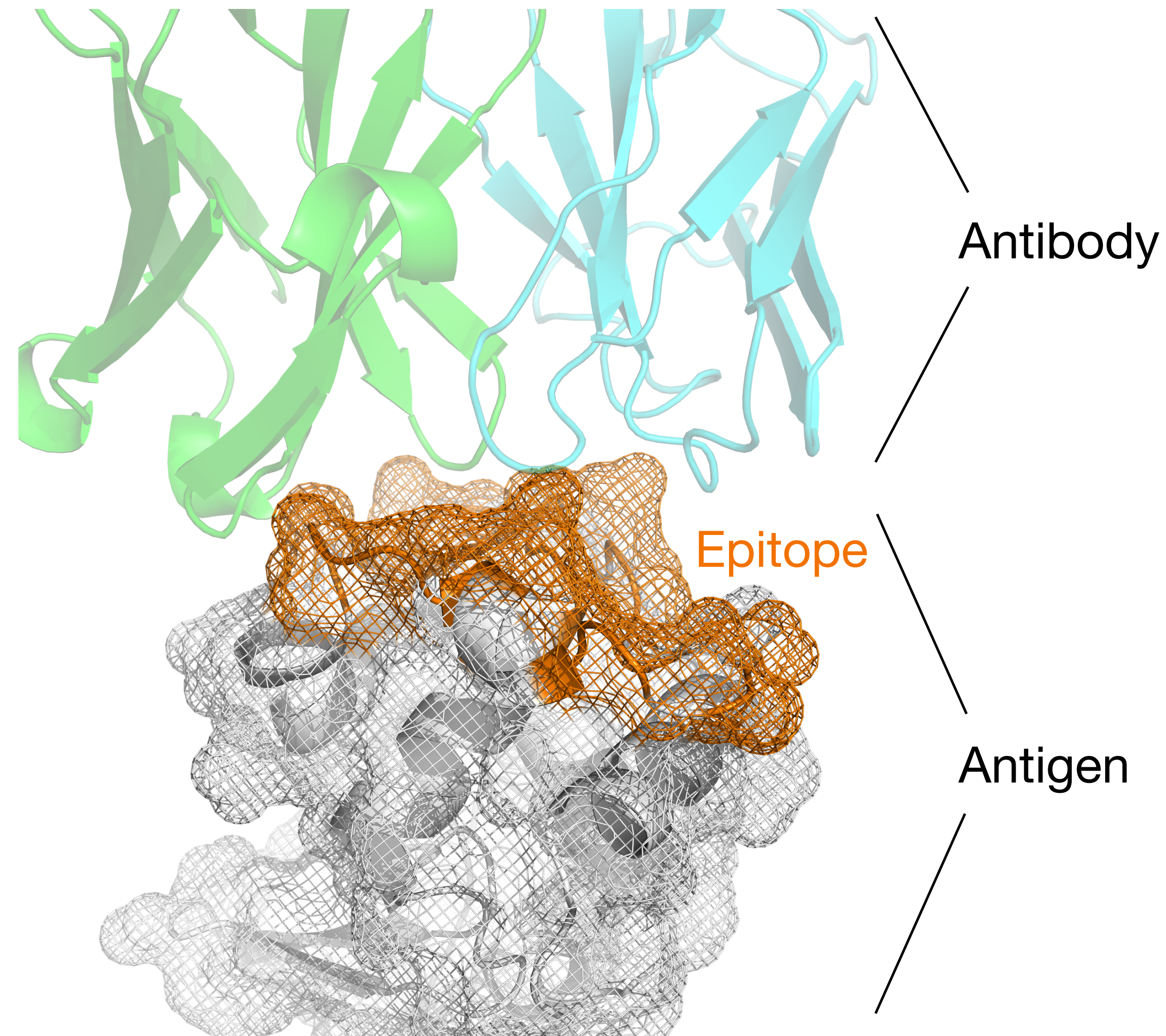
ISMB
Institute of Structural
and Molecular Biology



Epitope

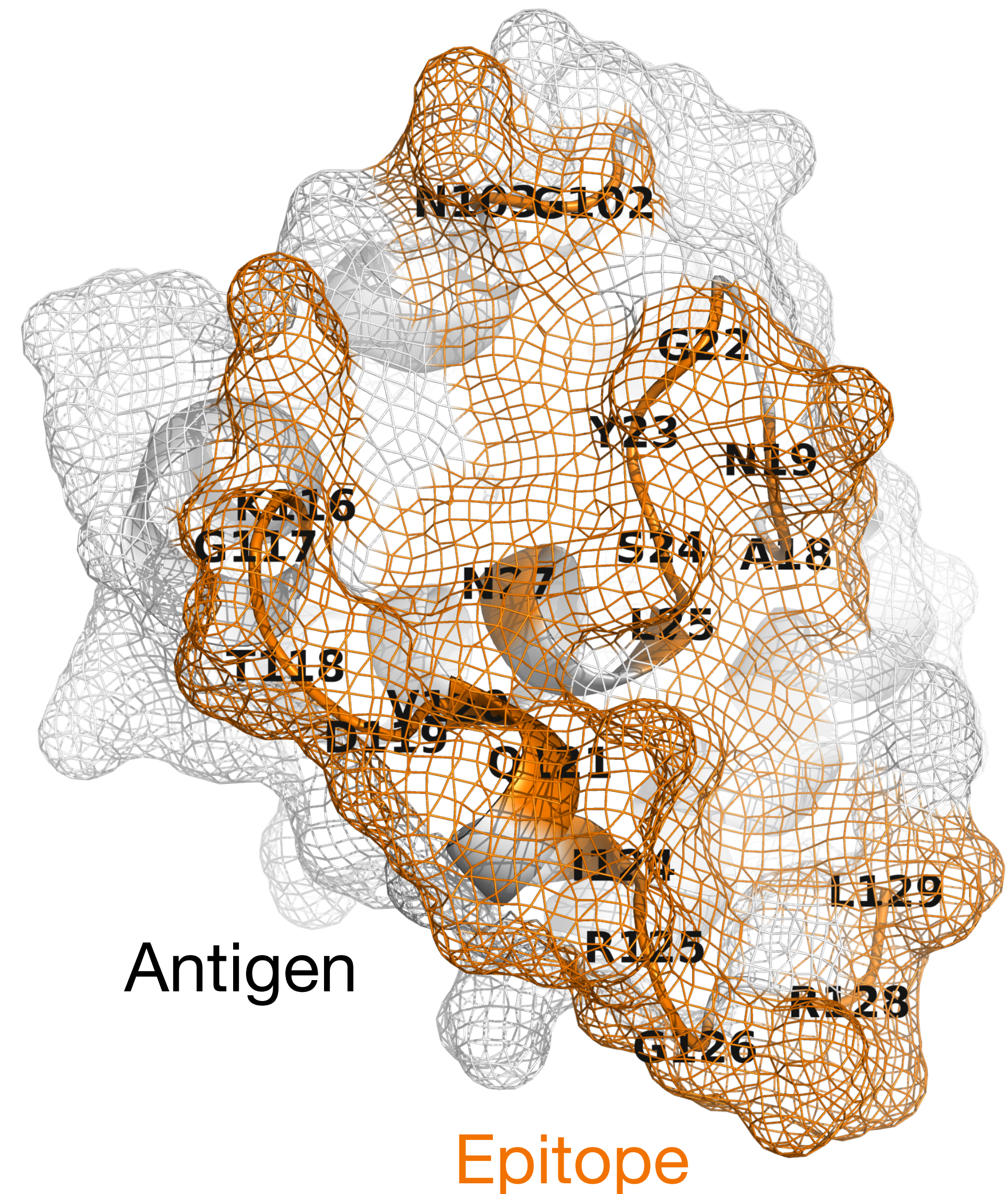
Most B-cell epitopes are discontinuous

- Epitopes are regions of the antigen surface that directly interact with the antibody.



Most B-cell epitopes are discontinuous

- Epitopes are regions of the antigen surface that directly interact with the antibody.
- Two types of epitopes:
 1. **Linear epitopes:** made up of a continuous sequence of amino acids
 2. **Conformational epitopes:** made up of a discontinuous sequence of amino acids, account for **~90%** of cases ([Ferdous et al., 2019](#))
- Risks of not considering epitopes in antibody development
 - Immunogenicity
 - Lack of specificity (off-target)
 - Risk of escape mutations



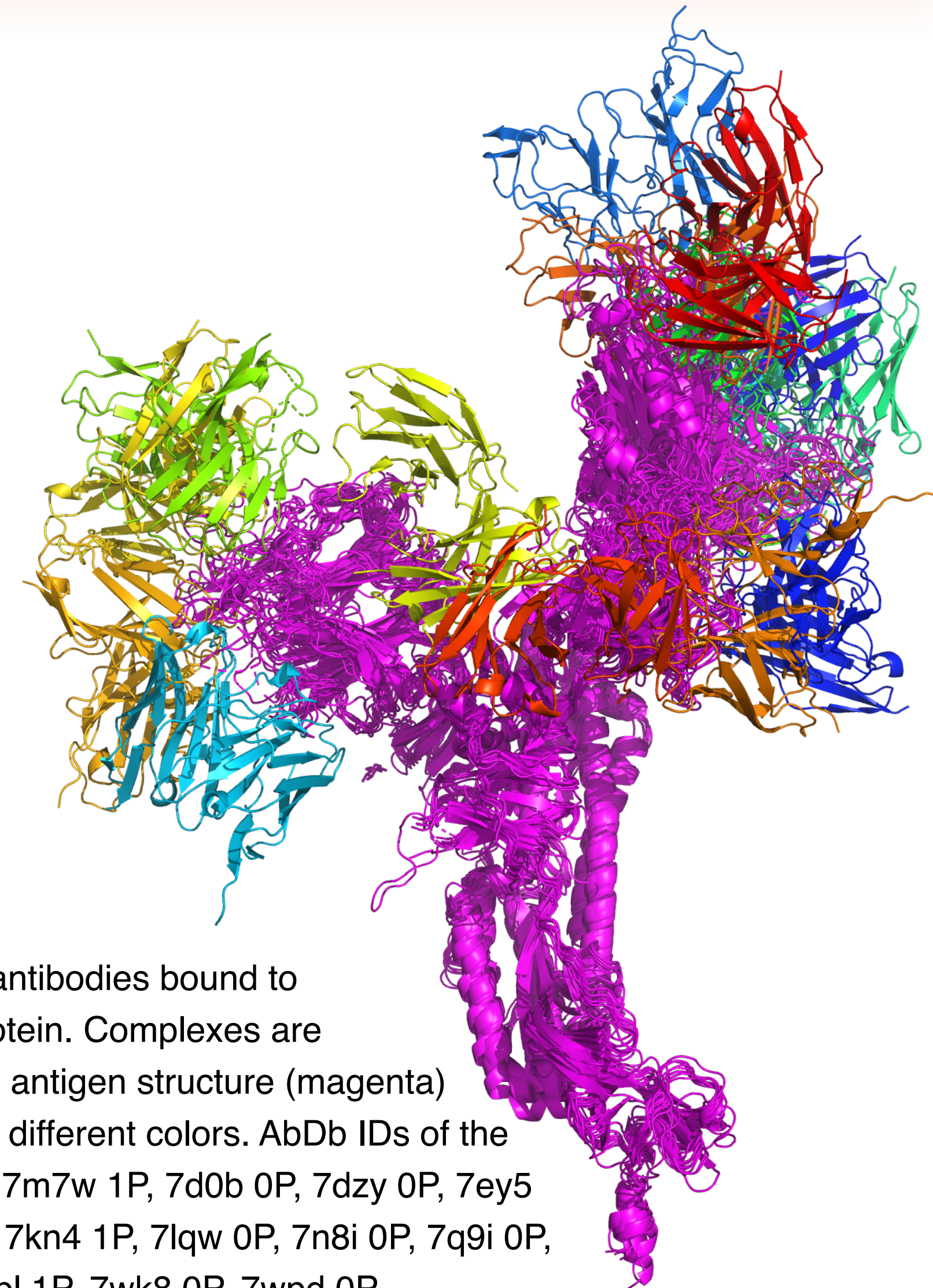
III Motivation

One antigen can have multiple epitopes depending on the antibodies

Table 2: Summary of Features Used in Benchmarking Methods.

	Antibody	Structure	PLM	Graph
WALLE	✓	✓	✓	✓
EpiPred	✓	✓	×	✓
ESMFold	✓	×	✓	×
MaSIF-site	×	✓	×	✓
ESMBind	×	×	✓	×

Antibody: Antibody is taken into consideration when predicting epitope nodes;
Structure: Topological information from protein structures;
PLM: Representation from Protein Language Models;
Graph: Graph representation of protein structures.



(b) Sixteen different antibodies bound to coronavirus spike protein. Complexes are superimposed on the antigen structure (magenta) and antibodies are in different colors. AbDb IDs of the complexes: 7k8s 0P, 7m7w 1P, 7d0b 0P, 7dzy 0P, 7ey5 1P, 7jv4 0P, 7k8v 1P, 7kn4 1P, 7lqw 0P, 7n8i 0P, 7q9i 0P, 7rq6 0P, 7s0e 0P, 7upl 1P, 7wk8 0P, 7wpd 0P.

Existing datasets are limited in size

Table S1: Comparison of Dataset Sizes Across Different Methods

Method	Dataset Size
WALLE (AsEP)	1723 AbAg complexes
Wang et al. 2022 (Wang et al., 2022)	258 AbAg complexes
SAGERank (Sun et al., 2023)	287 AbAg complexes
CSM-AB (Myung et al., 2021)	472 AbAg complexes
Bepipred3.0 (Clifford et al., 2022)	582 AbAg complexes

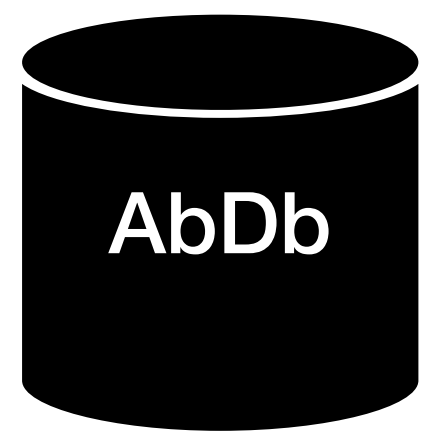
AsEP: Antibody-specific Epitope Prediction

AbAg: antibody-antigen

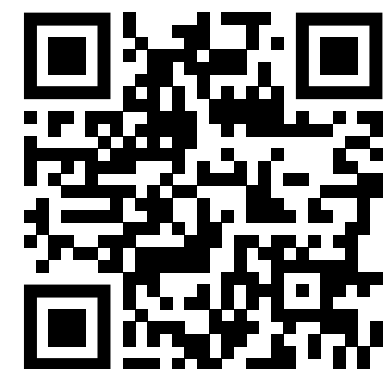
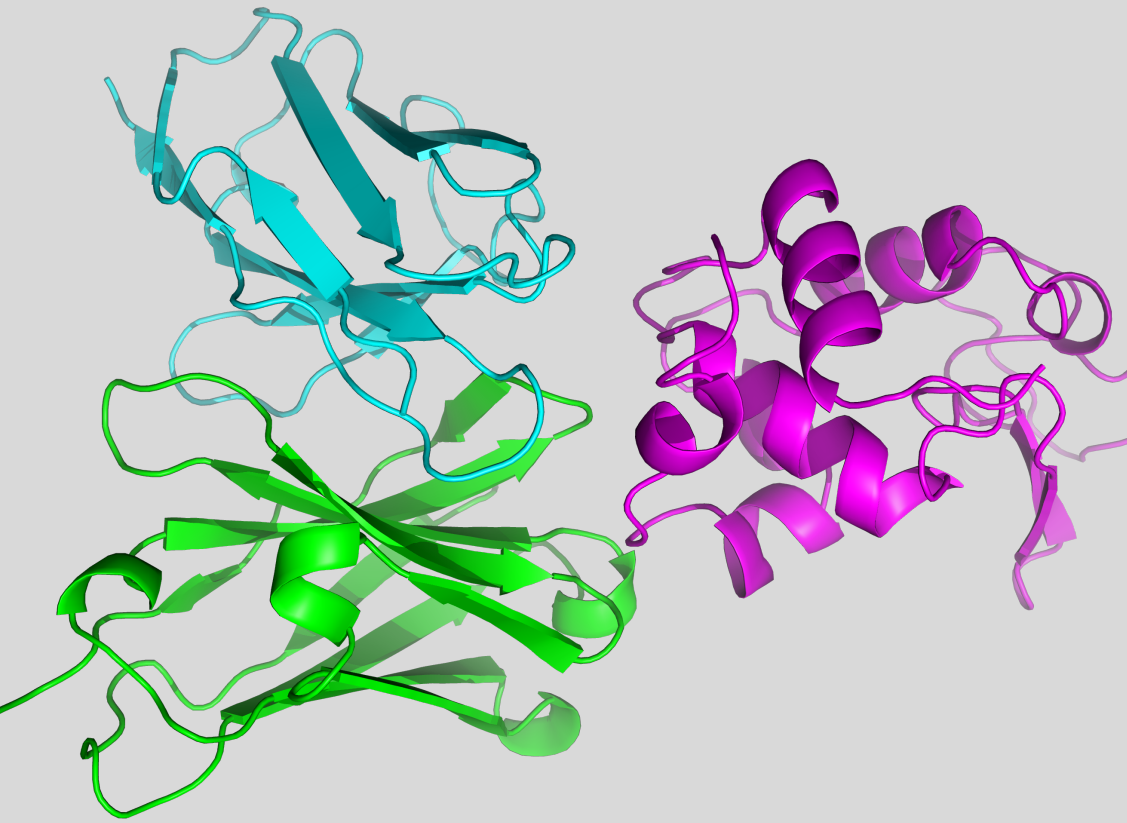
2024 Nov 11: **8,987** structures containing an antibody in the Protein Data Bank **before removing duplicates**

Source: **abYbank / SACS**
(<http://www.abYbank.org/sacs/>)

Dataset construction - 1723 antibody-antigen complexes



Snapshot:
2022September26

11,767 AbAg complexes

- Only antibodies with both VH and VL domains
- Only single-chain protein antigen, at least 50 residues long
- No unresolved CDR residues


VH ———

VL ———

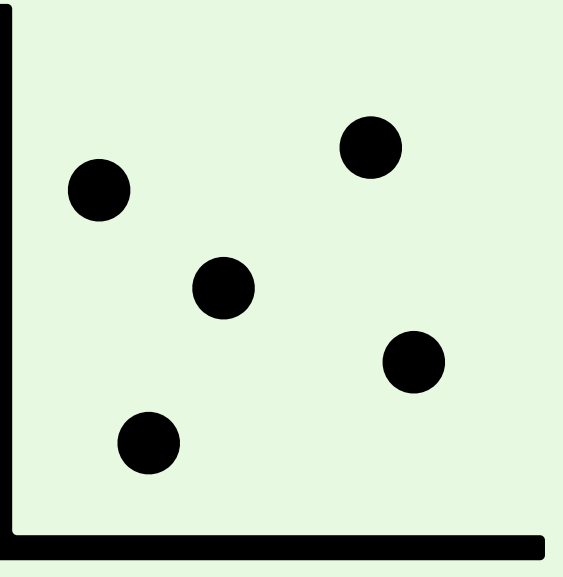
Ag ———

Extract sequences

- `easy-linclust` mode
- `-cov-mode` set to 0
- Use default coverage cutoff at 80%
- `-min-sequence-id` cutoff: Ab 100% and Ag 70%



MMseqs2 clustering and removing duplicates

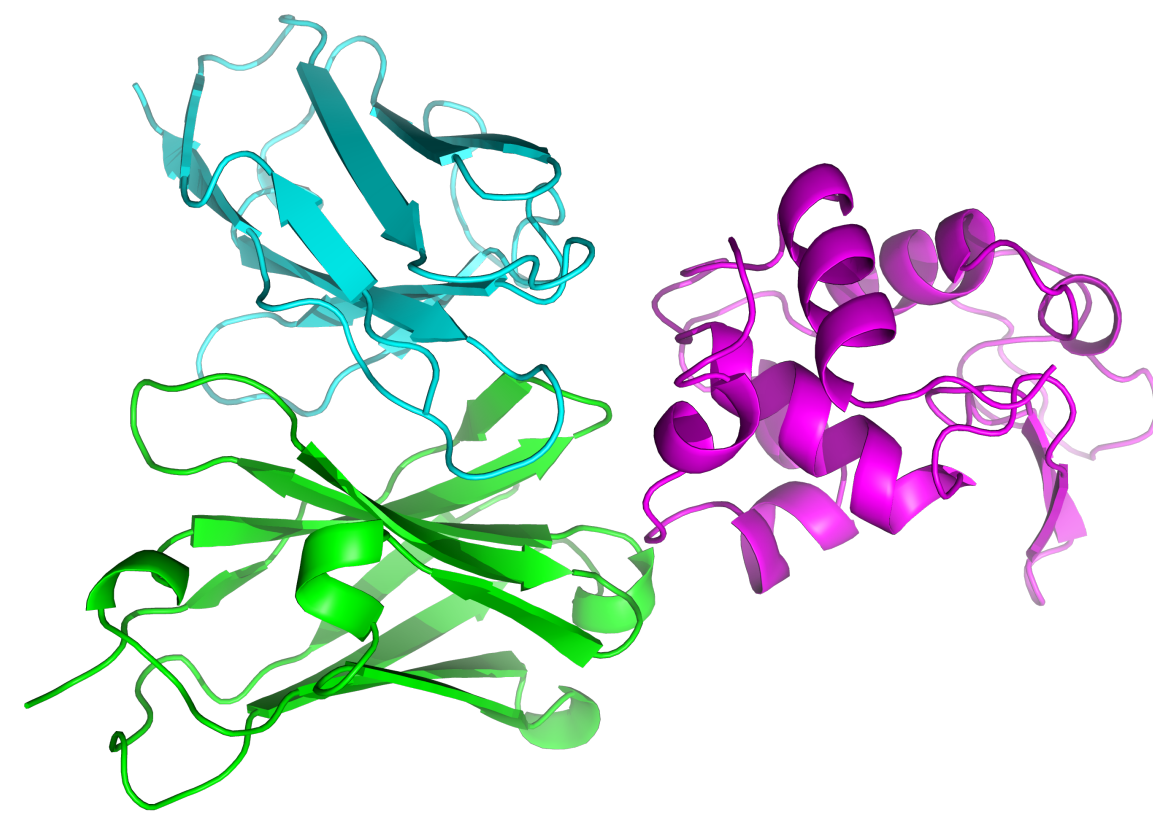


1,723 Representatives

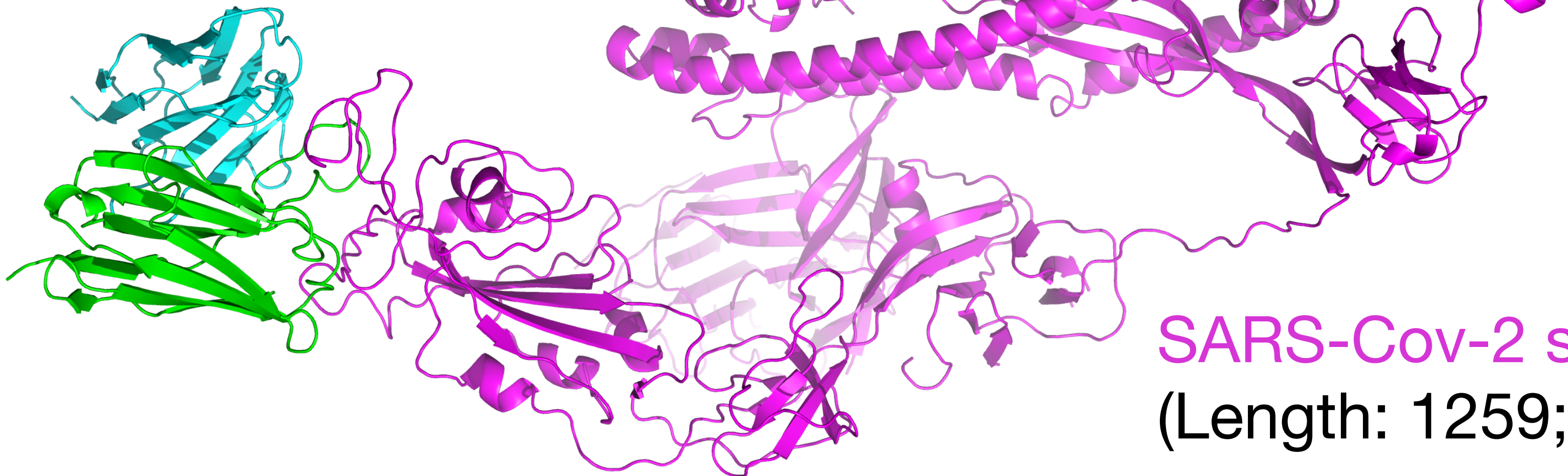
- Remove antibody-antigen complexes with duplicate VH, VL, and Ag cluster labels
- Removed 2 complexes containing unknown and non-canonical CDR residues
- Led to 1,723 complexes

AsEP - Two types of dataset splits

Dataset split 1: epitope/antigen surface ratio



Hen egg white lysozyme
(Length: 107; PDB: 1A2Y)



SARS-Cov-2 spike protein
(Length: 1259; PDB: 7K8S)

Antibody-Antigen Binding Interface Analysis in the Big Data Era

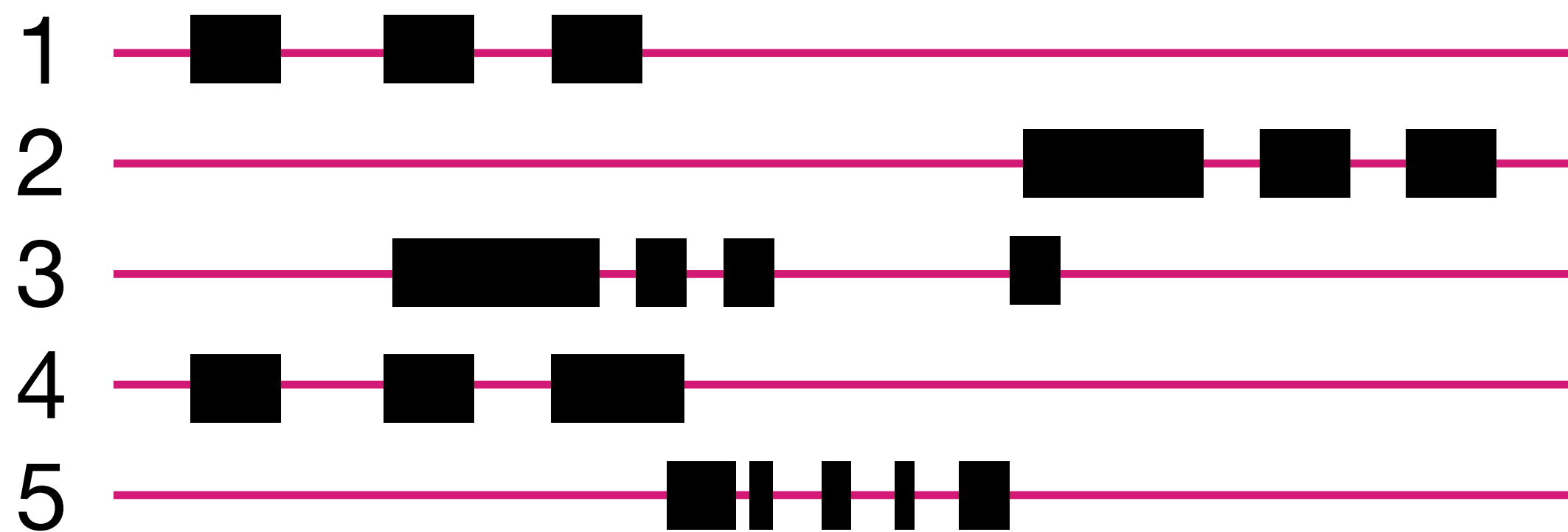
Pedro B. P. S. Reis^{1,2†}, German P. Barletta^{1,3,4†}, Luca Gagliardi¹, Sara Fortuna¹, Miguel A. Soler^{1,5} and Walter Rocchia^{1*}*

¹CONCEPT Lab, Istituto Italiano di Tecnologia, Genova, Italy, ²Bioisi, University of Lisbon, Lisbon, Portugal, ³Universidad Nacional de Quilmes/CONICET, Quilmes, Argentina, ⁴The Abdus Salam International Centre for Theoretical Physics (ICTP), Trieste, Italy, ⁵Dipartimento di Scienze Matematiche, Informatiche e Fisiche, Universita' di Udine, Udine, Italy

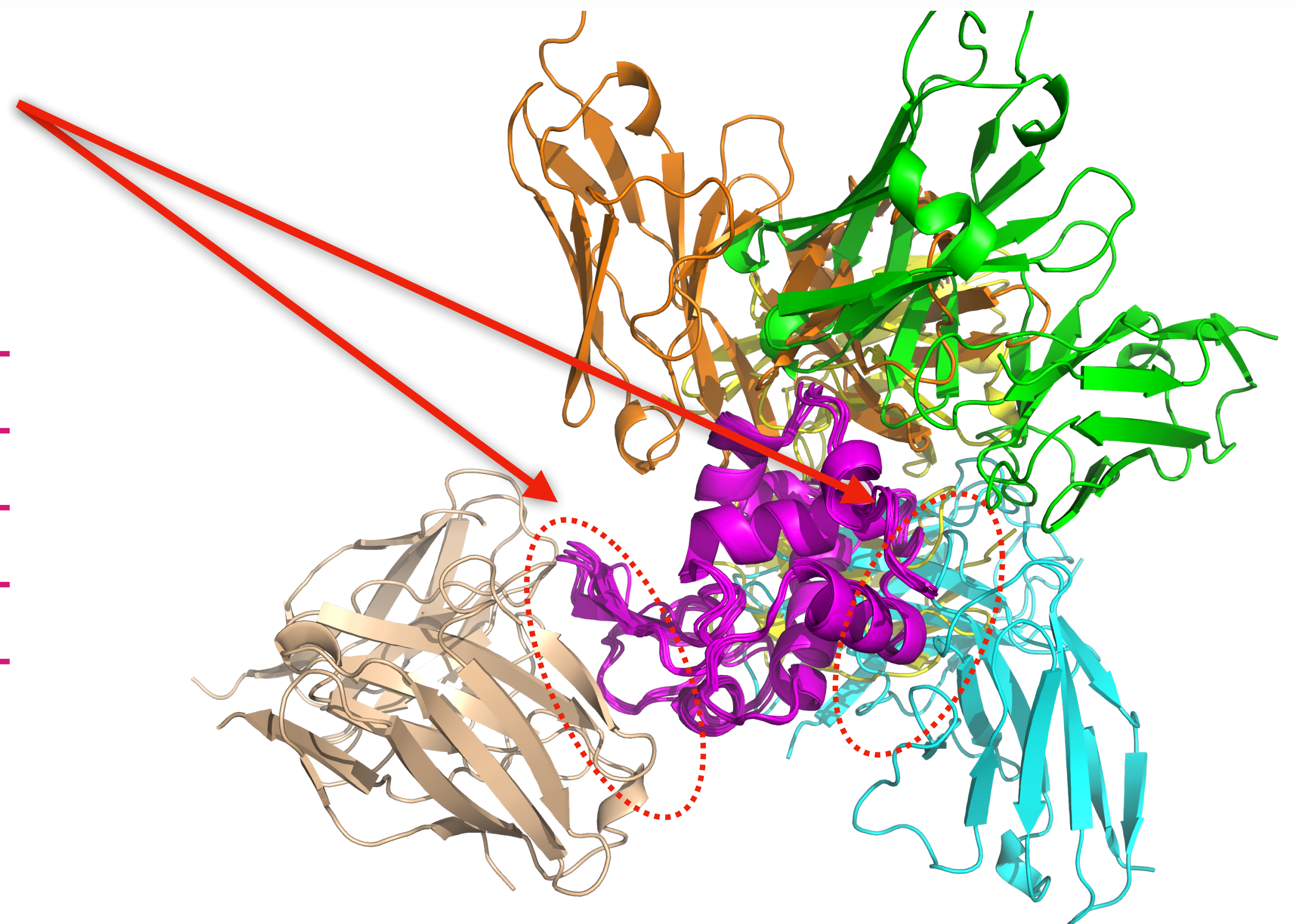
On average, epitopes contain
 14.6 ± 4.9 residues

Dataset split 2: epitope group

- Goal: evaluate generalizability



1. Align antigen sequences as MSA
2. Map epitopes to MSA columns
3. Identity threshold 75%



(a) Five different antibodies bound to hen egg white lysozyme. Complexes are superimposed on the antigen structure (magenta). AbDb IDs of the complexes and their color: 1g7i 0P (green), 2yss 0P (cyan), 1dzb 1P (yellow), 4tsb 0P (orange), 2iff 0P (wheat). Antigens are colored in magenta.

Distance-based interface definition

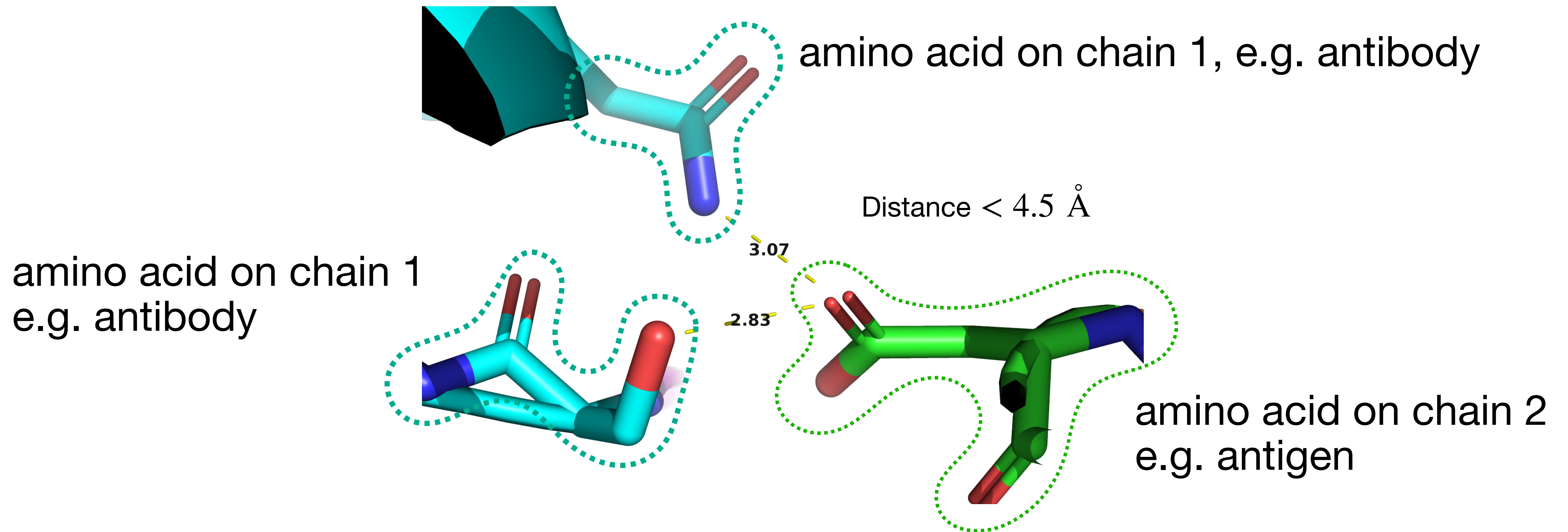


Figure 1: An example illustrating interacting residues. The two dashed lines indicate distances between non-hydrogen atoms from different interacting residues across two protein chains, with each chain's carbon atoms colored cyan and green.

Experiment setup

Represent protein structures as graphs

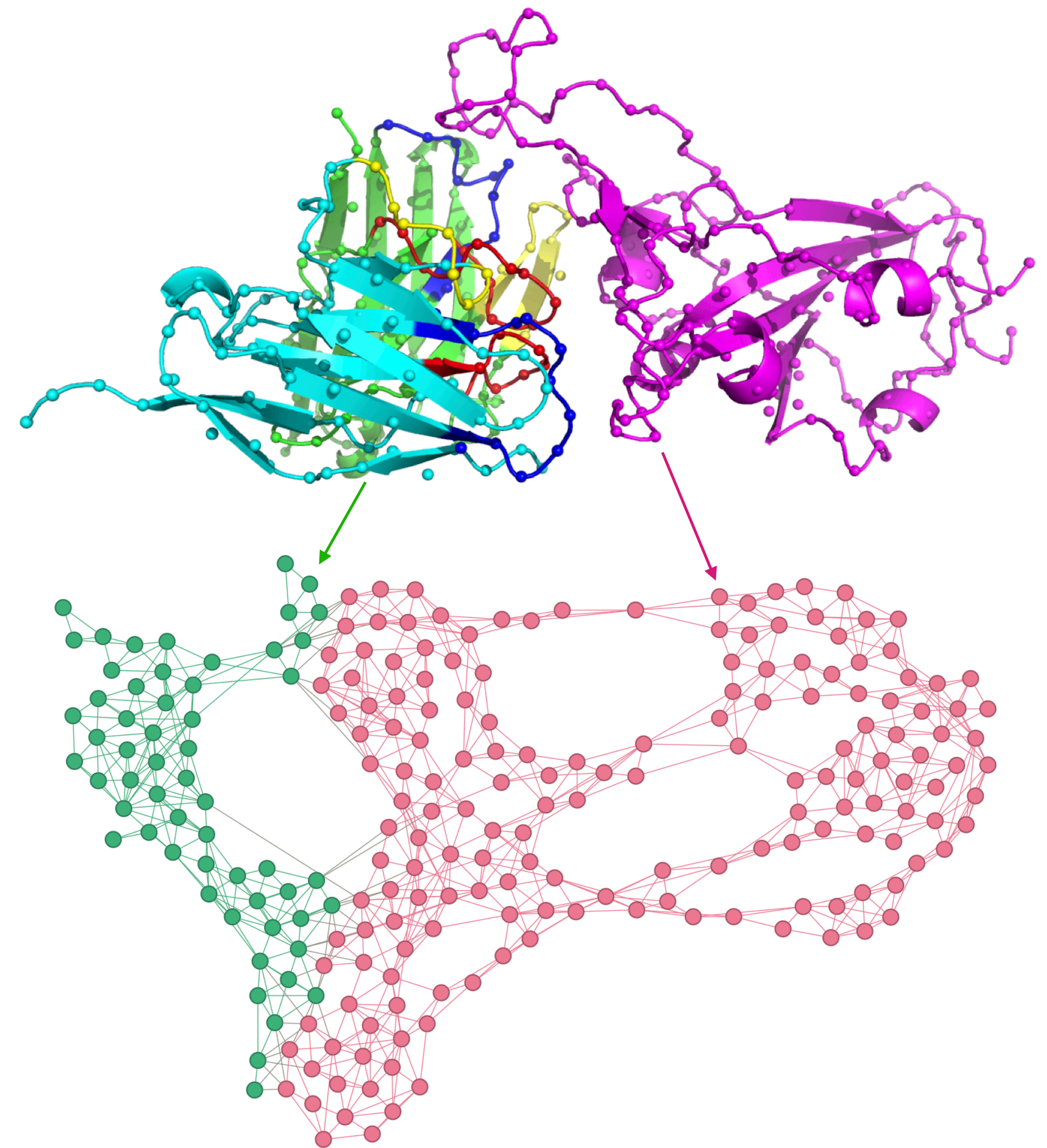
Top: molecular structure of an Ab-Ag complex (PDB code: 7KFW). Spheres denote the α -carbon atoms of each amino acid.

Color scheme: **Antigen**, **Heavy FR**, **Light FR**, **CDR1**, **CDR2**, **CDR3**.

Bottom: the corresponding graph. **Green** vertices are antibody CDR residues. **Pink** vertices are antigen surface residues.

Nodes represent protein residues and are encoded into vector spaces using a customizable embedding function, such as a protein language model.

Edges are defined by residue proximity and are labeled 1 if the Euclidean distance between the non-hydrogen atoms from a pair of residues is less than **4.5Å**.



Experiment setup

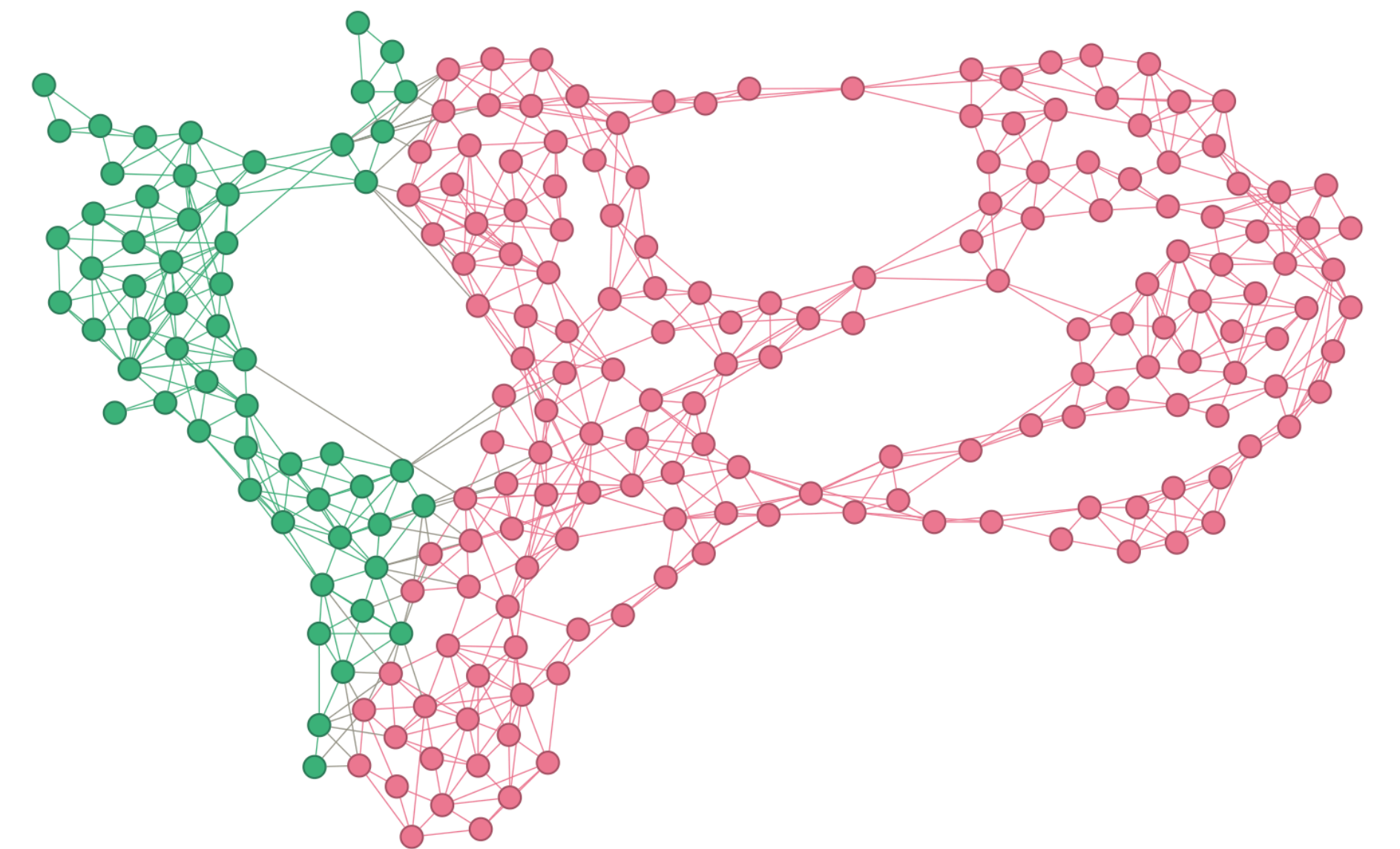
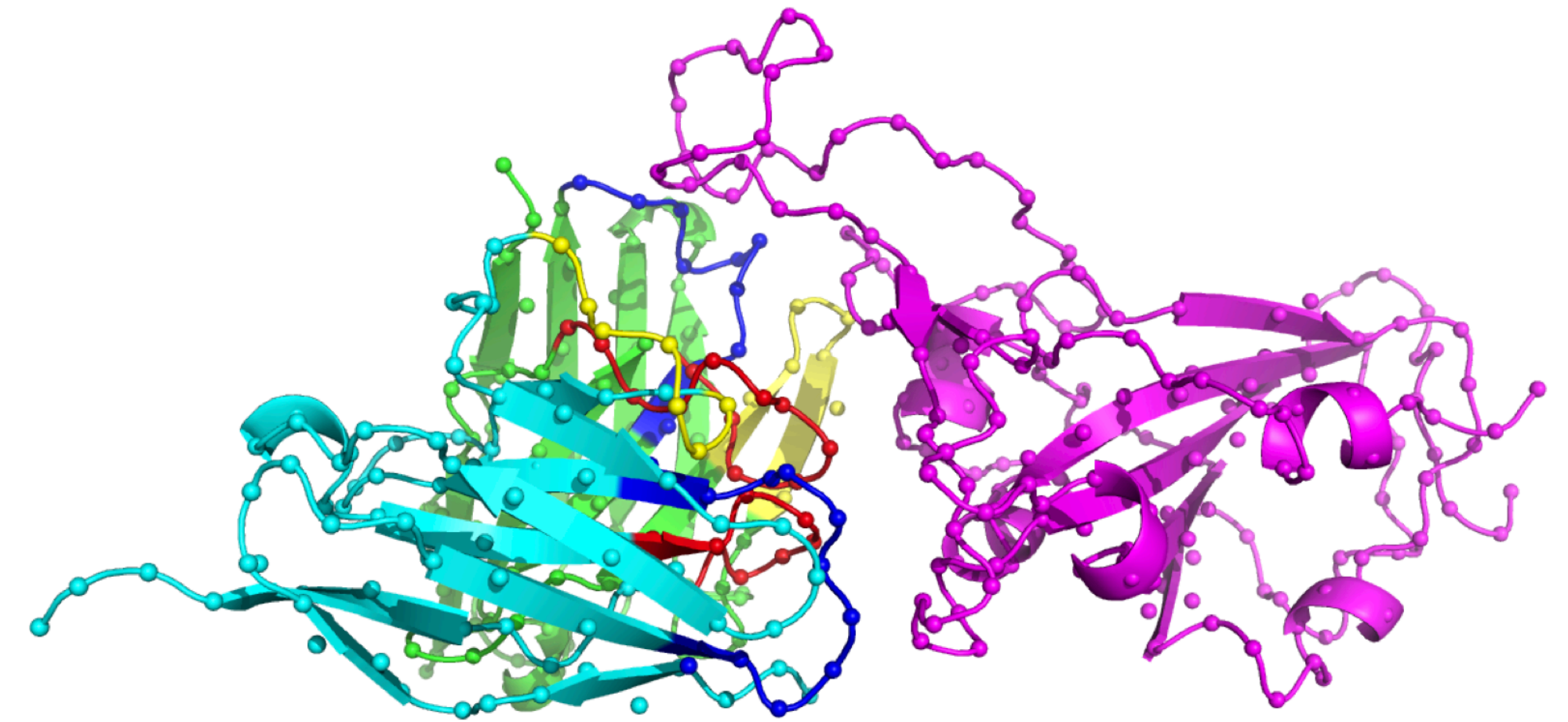
Question formulation - two tasks

Inputs: Disjoint graphs

- **Antibody graph** $G_A = (V_A, E_A)$ combining CDR residues from the heavy and light chains
- **Antigen graph** $G_B = (V_B, E_B)$ surface residues of the antigen

Tasks:

1. **Epitope Prediction:** Classify antigen nodes as epitope or non-epitope.
2. **Bipartite Link Prediction:** Predict interaction links between antibody and antigen nodes indicating direct contact.



AsEp dataset

PyTorch interface (<https://github.com/biochunan/AsEP-dataset>)



README MIT license

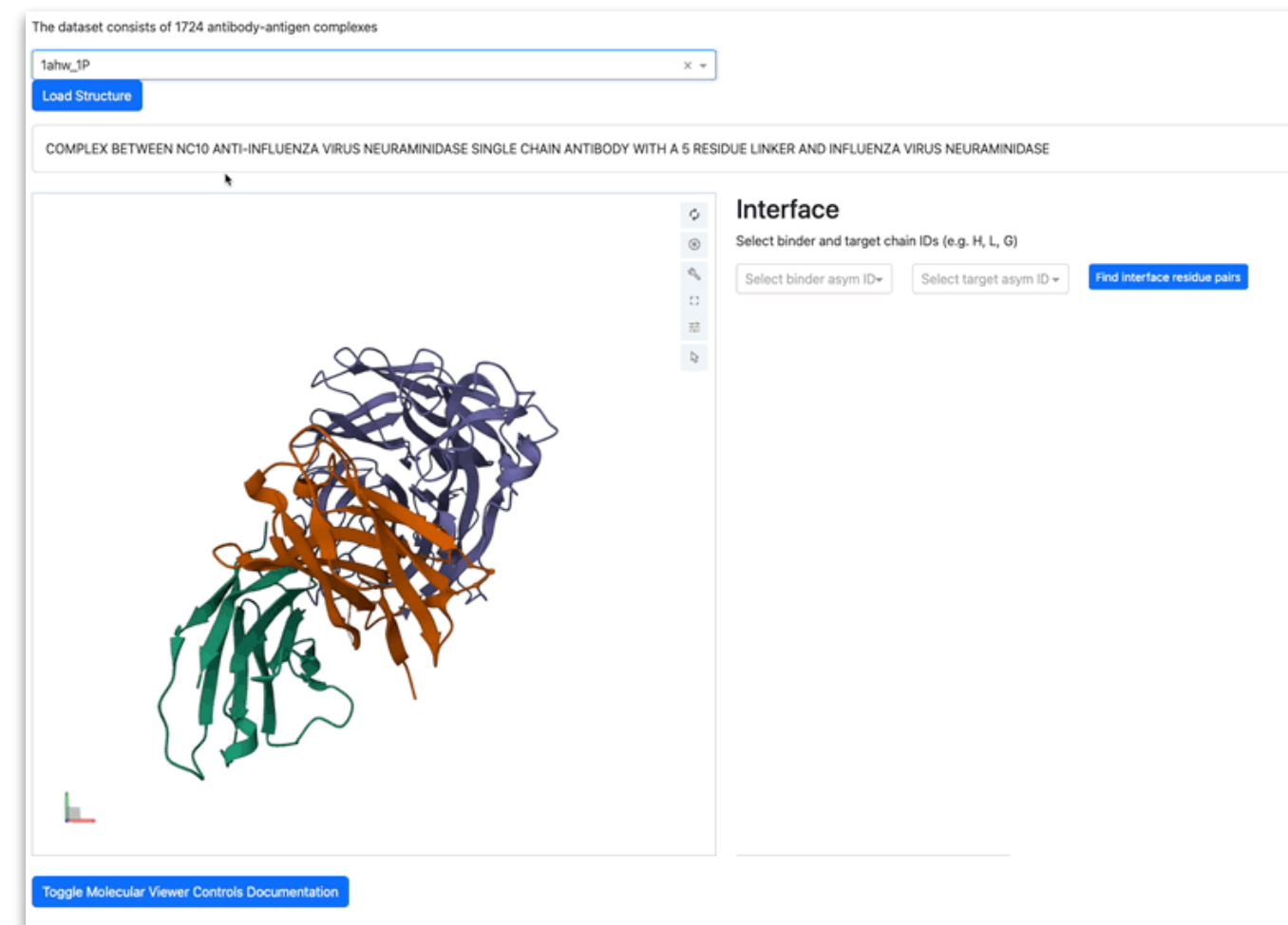
AsEP Dataset

DOI 10.5281/zenodo.11495514 License MIT Python 3.10 PyTorch-Geometric 2.5.3 PyTorch 2.1.1

Antibody-specific Epitope Prediction (AsEP) Dataset. This dataset is used in the manuscript [AsEP: Benchmarking Deep Learning Methods for Antibody-specific Epitope Prediction](#) (submitted to NeurIPS 2024 Datasets and Benchmarks).

The raw dataset can be downloaded from [Zenodo](#).

- [AsEP Dataset](#)
 - [Structure viewer](#)
 - [Dataset Python Interface \(asep \)](#)
 - [Installation](#)
 - [devcontainer](#)
 - [conda environment](#)
 - [Download dataset](#)
 - [Data Loader](#)
 - [Data Split](#)
 - [Evaluation](#)
 - [Benchmark Performance](#)
 - [Epitope Ratio](#)
 - [Epitope Group](#)



```
from asep.data.asepv1_dataset import AsEPv1Dataset, EmbeddingConfig

# one-hot encoding
config = EmbeddingConfig(node_feat_type="one-hot")
asepv1_dataset = AsEPv1Dataset(
    root="/path/to/asep/download/folder", # replace with the path to the parent folder of dow
    name="AsEP",
    embedding_config=config,
)

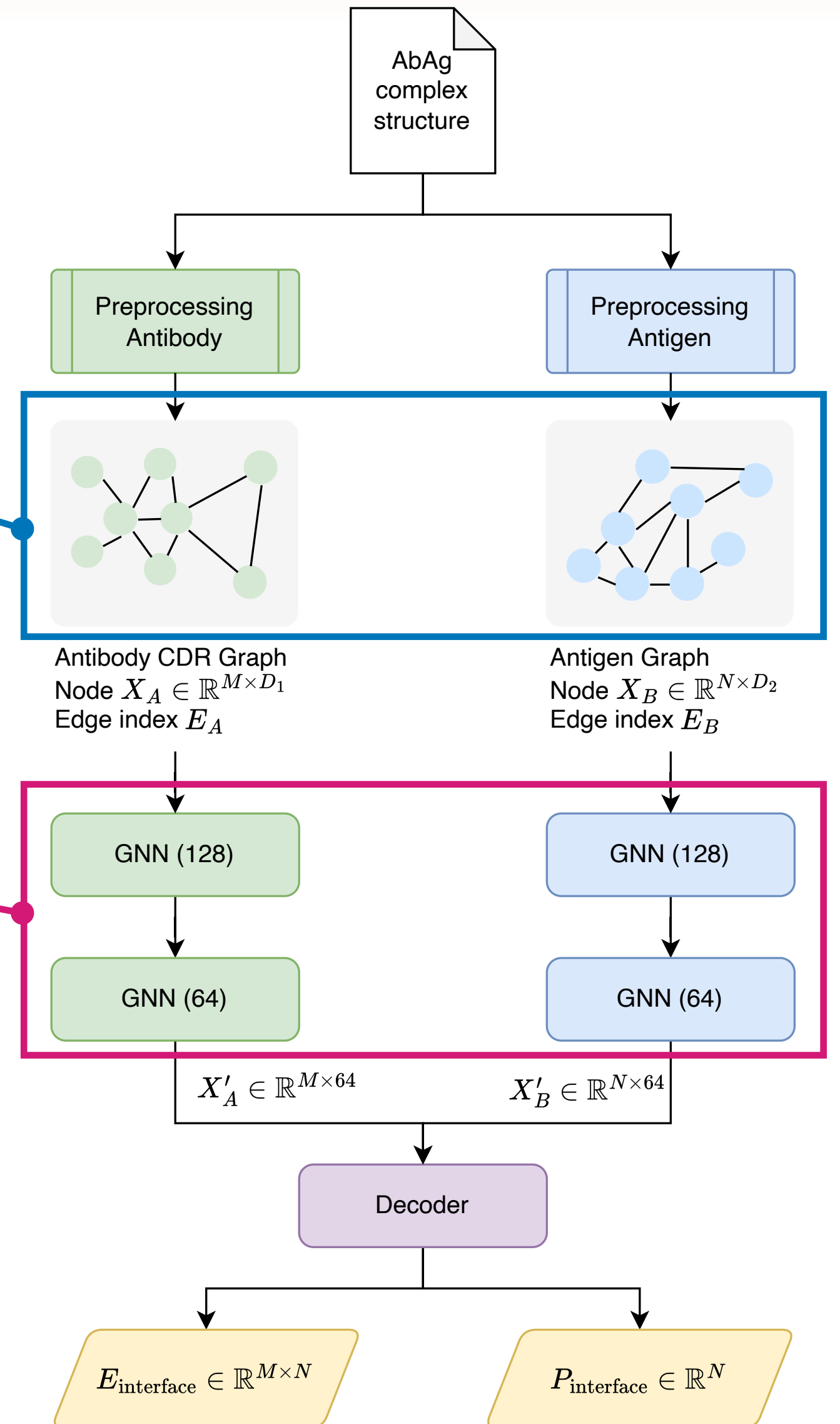
# pre-calculated embeddings with AntiBERTy (via igfold) and ESM2
config = EmbeddingConfig(
    node_feat_type='pre_cal',
    ab={"embedding_model": "igfold"}, # change this "esm2" for ESM2 embeddings
    ag={"embedding_model": "esm2"},
)
asepv1_dataset = AsEPv1Dataset(
    root="/path/to/asep/download/folder", # replace with the path to the parent folder of do
    name="AsEP",
    embedding_config=config,
)

# get i-th graph pair and node labels
i = 0
graph_pair = asepv1_dataset[i]
node_labels_b = graph_pair.y_b # antibody graph node labels (1 => interface nodes)
node_labels_g = graph_pair.y_g # antigen graph node labels (1 => interface nodes)

# bipartite graph edges
edge_index_bg = graph_pair.edge_index_bg # bipartite graph edge indices between the antibody
```


A hybrid method leveraging PLMs & GNNs

- Protein Language Models (PLMs)
 - AntiBERTy (Antibody only)
 - ESM2-35M & ESM2-650M
- Graph Neural Networks (GNNs)
 - Graph Convolutional Network (GCN)
 - Graph Attention Network (GAT)
 - GraphSAGE (**S**Ample and aggre**G**at**E**)



■ Benchmarking Performance

Hybrid method works better than existing methods

Table 1: Performance on test set from dataset split by epitope to antigen surface ratio and epitope groups.

(a) Performance on dataset split by epitope to antigen surface ratio.

Method	MCC	Precision	Recall	AUCROC	F1
WALLE	0.305 (0.023)	0.308 (0.019)	0.516 (0.028)	0.695 (0.015)	0.357 (0.021)
EpiPred	0.029 (0.018)	0.122 (0.014)	0.180 (0.019)	—	0.142 (0.016)
ESMFold	0.028 (0.010)	0.137 (0.019)	0.043 (0.006)	—	0.060 (0.008)
ESMBind	0.016 (0.008)	0.106 (0.012)	0.121 (0.014)	0.506 (0.004)	0.090 (0.009)
MaSIF-site	0.037 (0.012)	0.125 (0.015)	0.183 (0.017)	—	0.114 (0.011)

MCC: Matthews Correlation Coefficient; **AUCROC**: Area Under the Receiver Operating Characteristic Curve; **F1**: F1 score. Standard errors are included in the parentheses. We omitted the results of EpiPred, ESMFold and MaSIF-site for AUCROC. For EpiPred and ESMFold, the interface residues are determined from the predicted structures by these methods such that the predicted values are binary and not comparable to other methods; As for MaSIF-site, it outputs the probability of mesh vertices instead of node probabilities and epitopes are determined as residues close to mesh vertices with probability greater than 0.7.

$$\text{MCC} = \frac{(TP \times TN - FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

AbAb Ablation studies

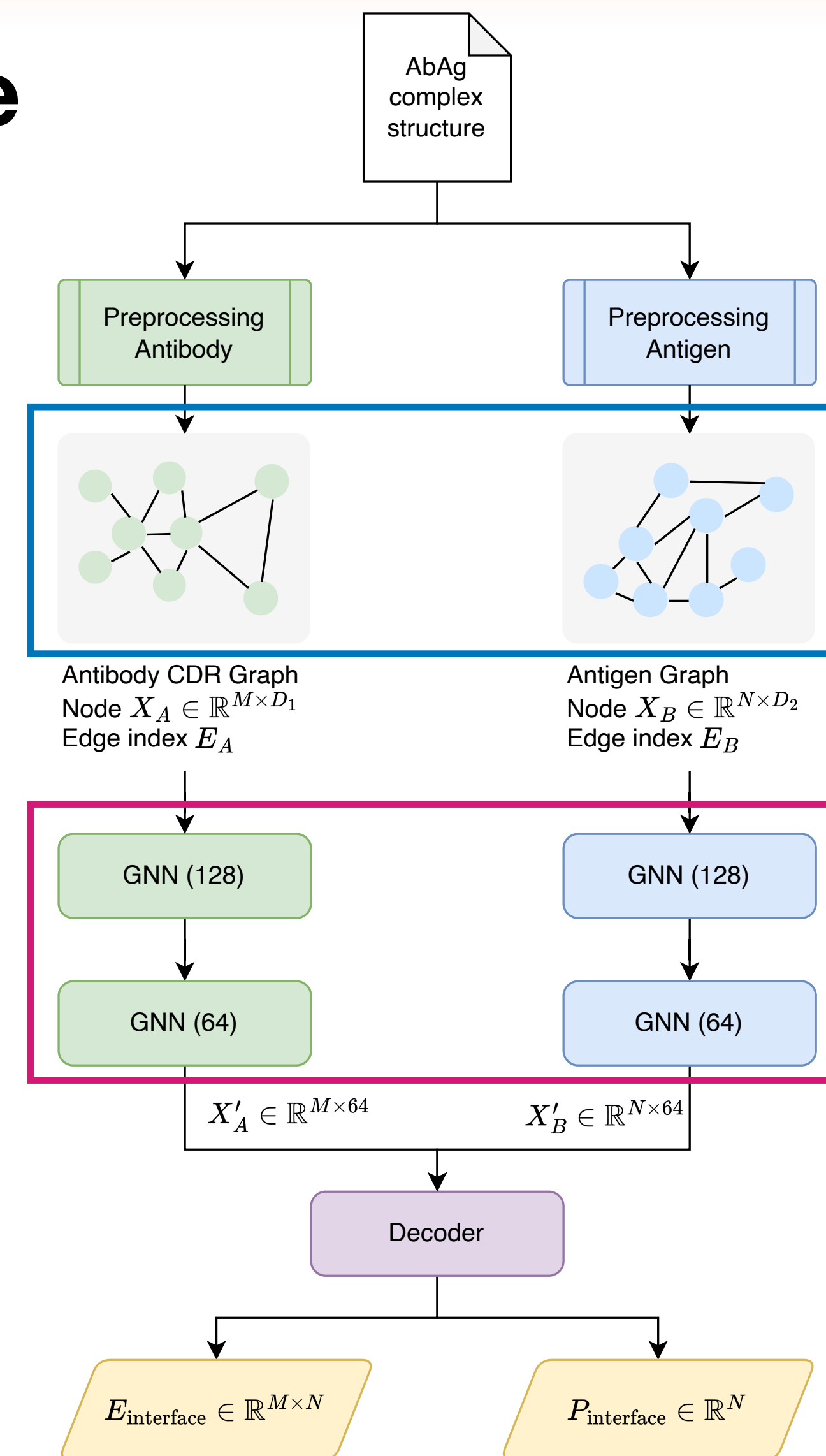
Both PLMs and GNNs contribute to performance

Method	Encoding	MCC
WALLE	Both	0.264 (0.021)
WALLE-L	Both	0.159 (0.016)
WALLE	ESM2	0.196 (0.021)
WALLE-L	ESM2	0.145 (0.014)
WALLE	One-hot	0.097 (0.009)
WALLE	BLOSUM	0.085 (0.010)

WALLE-L: replace GNN with linear layers
Both: AntiBERTy + ESM2-35M
ESM2: ESM2-35M

Replace node embeddings

Replace with Linear layers



III Ablation studies

Both PLMs and GNNs contribute to performance

Method	Encoding	MCC
WALLE	Both	0.264 (0.021)
WALLE-L	Both	0.159 (0.016)
WALLE	ESM2	0.196 (0.021)
WALLE-L	ESM2	0.145 (0.014)
WALLE	One-hot	0.097 (0.009)
WALLE	BLOSUM	0.085 (0.010)

Graph topology, i.e. residue neighborhood, contributes to performance

Meaningful node embeddings, i.e. from PLMs contribute performance

WALLE-L: replace GNN with linear layers

Both: AntiBERTy + ESM2-35M

ESM2: ESM2-35M

■ Benchmarking Performance

Generalizing to novel epitopes needs improvement

(a) Performance on dataset split by epitope to antigen surface ratio.

Method	MCC	Precision	Recall	AUCROC	F1
WALLE	0.305 (0.023)	0.308 (0.019)	0.516 (0.028)	0.695 (0.015)	0.357 (0.021)
EpiPred	0.029 (0.018)	0.122 (0.014)	0.180 (0.019)	—	0.142 (0.016)
ESMFold	0.028 (0.010)	0.137 (0.019)	0.043 (0.006)	—	0.060 (0.008)
ESMBind	0.016 (0.008)	0.106 (0.012)	0.121 (0.014)	0.506 (0.004)	0.090 (0.009)
MaSIF-site	0.037 (0.012)	0.125 (0.015)	0.183 (0.017)	—	0.114 (0.011)

(b) Performance on dataset split by epitope groups.

Method	MCC	Precision	Recall	AUCROC	F1
WALLE	0.152 (0.019)	0.207 (0.020)	0.299 (0.025)	0.596 (0.012)	0.204 (0.018)
EpiPred	-0.006 (0.015)	0.089 (0.011)	0.158 (0.019)	—	0.112 (0.014)
ESMFold	0.018 (0.010)	0.113 (0.019)	0.034 (0.007)	—	0.046 (0.009)
ESMBind	0.002 (0.008)	0.082 (0.011)	0.076 (0.011)	0.500 (0.004)	0.064 (0.008)
MaSIF-site	0.046 (0.014)	0.164 (0.020)	0.174 (0.015)	—	0.128 (0.012)

Summary

- Epitopes are important for antibody development
- Existing methods are either trained on a small dataset (less than 1K) or do not consider antibodies in prediction
- We proposed a new dataset with a maintenance plan to enrich novel antibody types and general protein-protein complexes
- We benchmarked representative methods and a hybrid method leveraging both PLMs and GNNs, which showed promising performance (3-10X better than existing methods)
- Further development will focus on improving generalizability to unseen epitopes