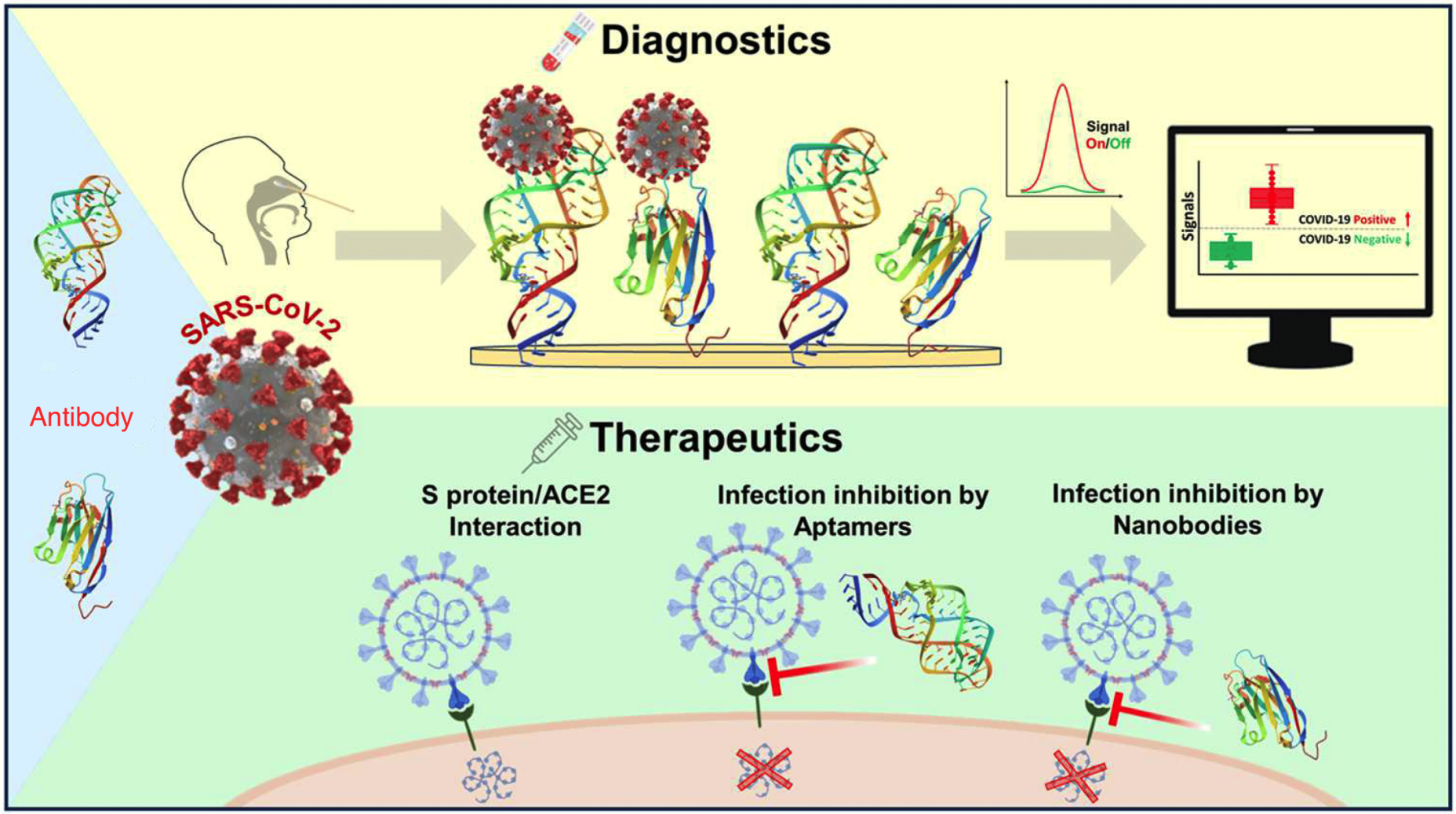
Exploring the Impact of Embedding Methods on Graph-based Antibody-aware Epitope Prediction

Mansoor Ahmed

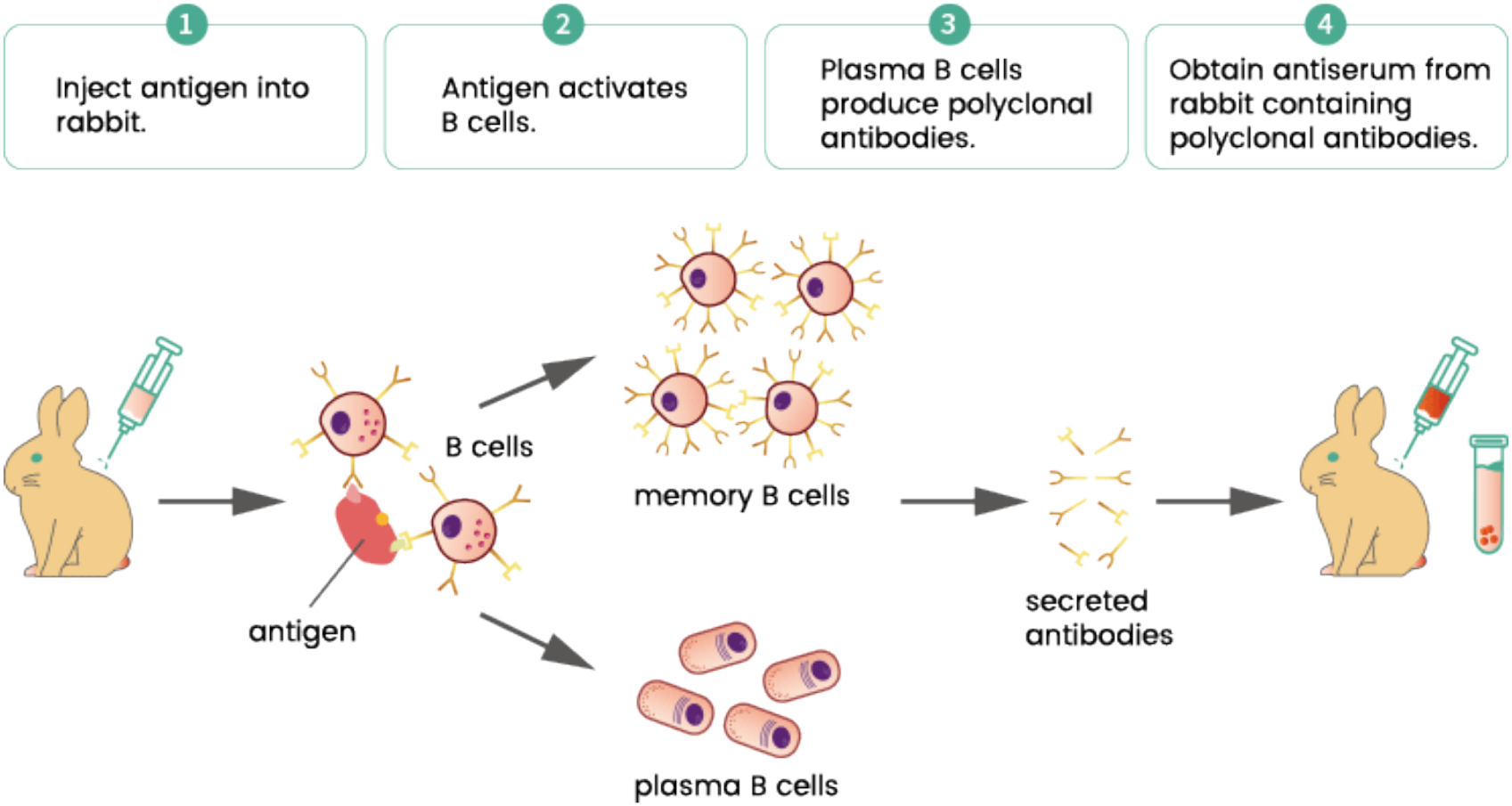
Georgia State University
ICCABS 2025

Why are we interested in Antibodies?



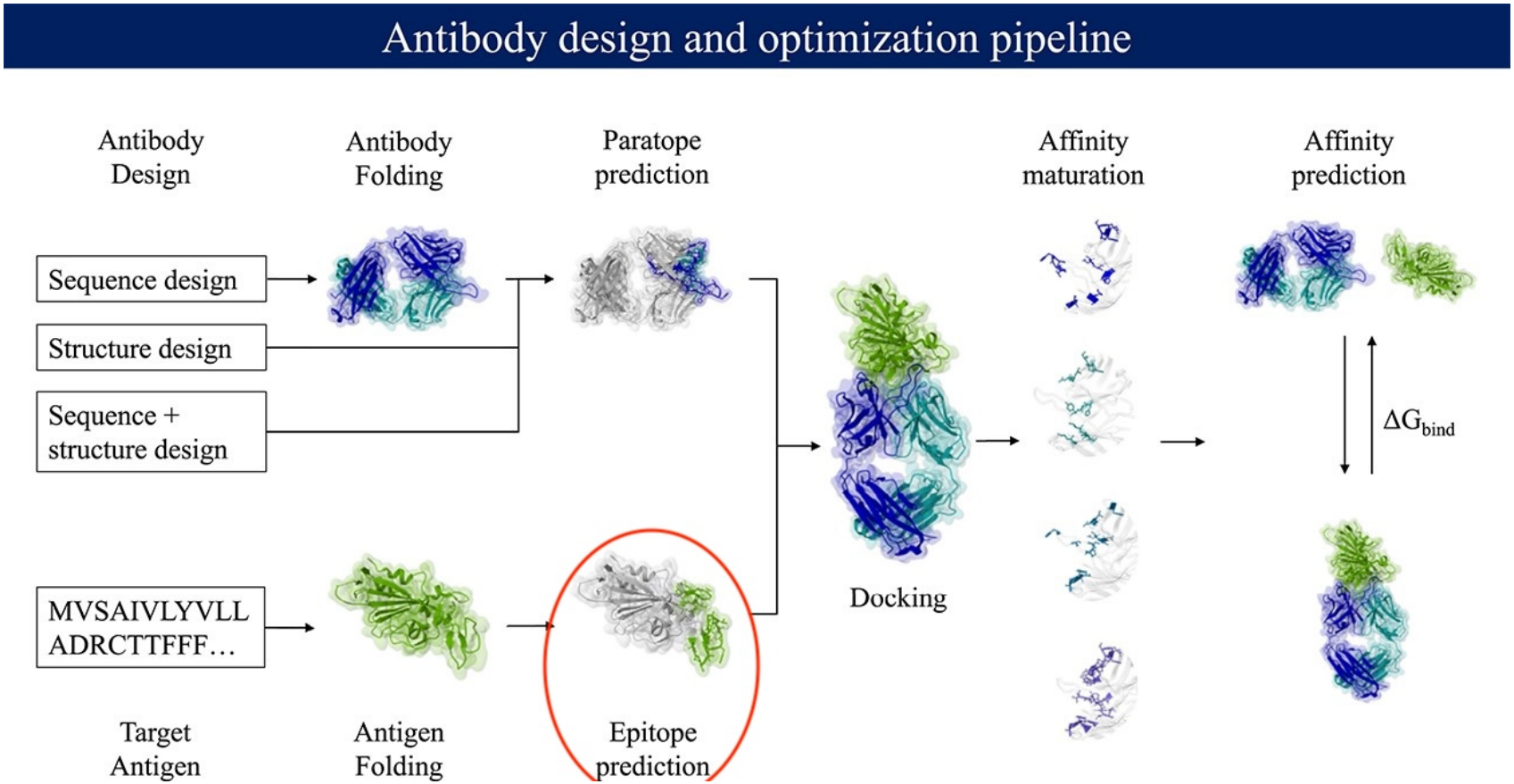
(Source: Park et al. [2024])

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



(Source: Lumen Learning)

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



Source: Joubbi et al. [2024]

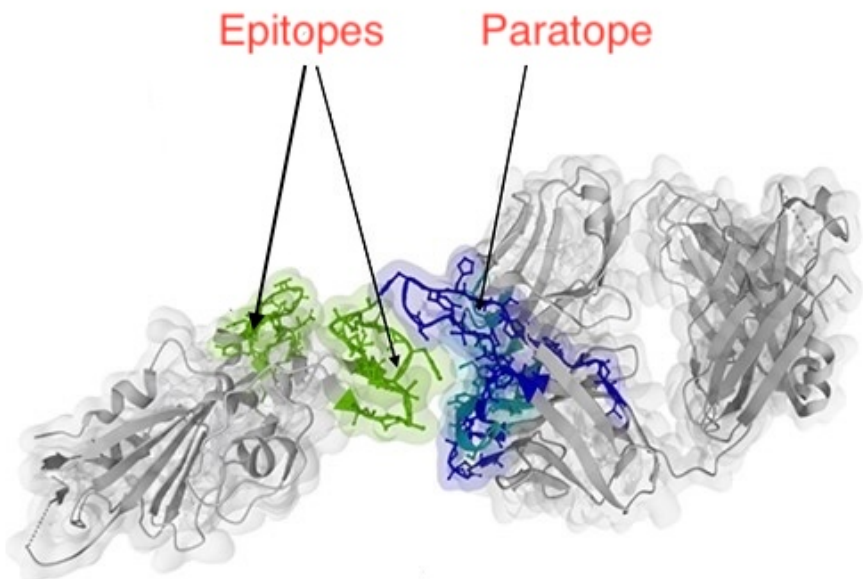
Some Definitions

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

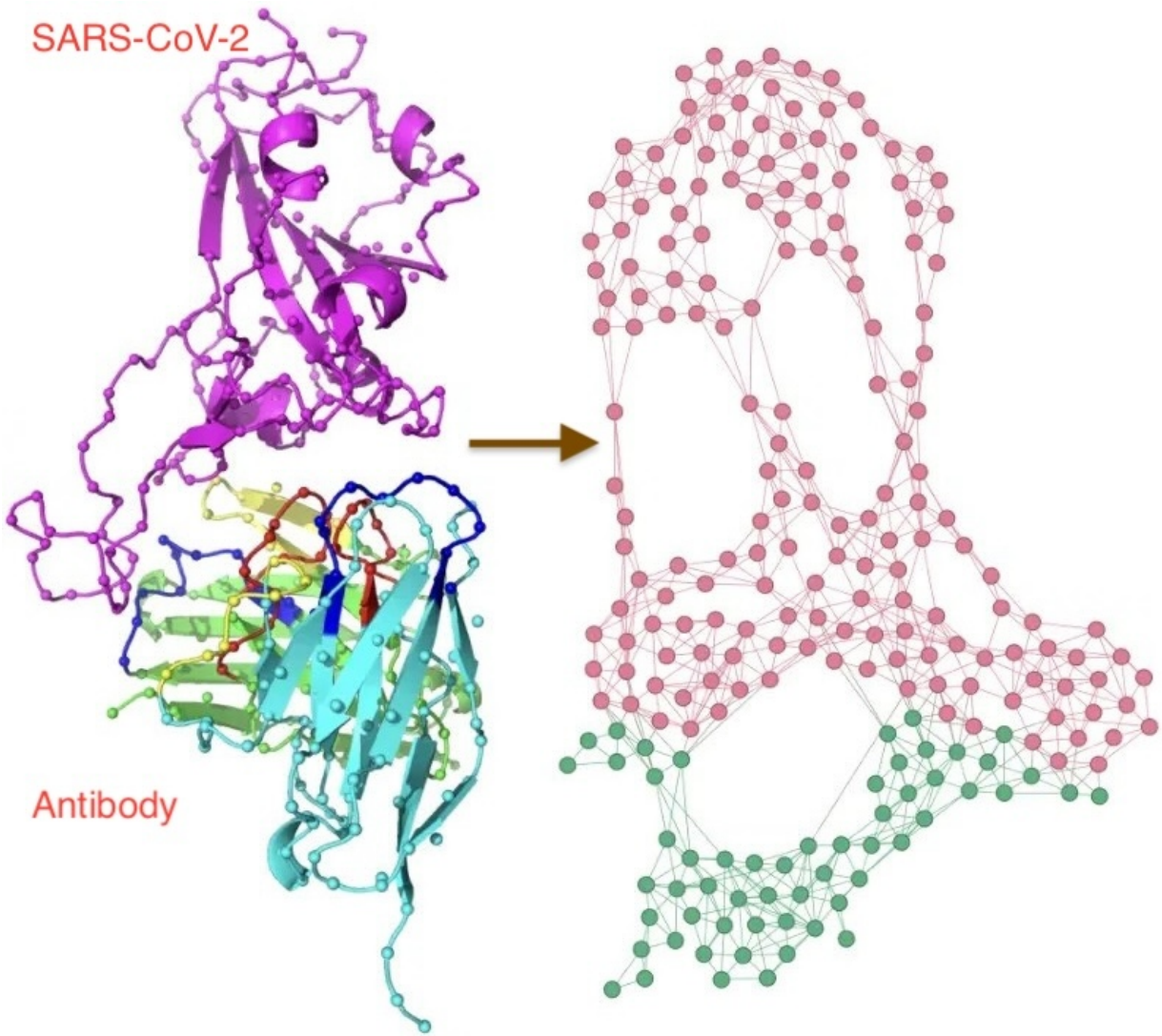


Challenges – Research Gap

- Non-conformational/sequential – sequence-based approaches fail to capture the spatial relationship
- Multiple epitopes on a single antigen
- Limited datasets

Solution: work with antigen structures coupled with frontier deep learning methods for accurate epitope prediction problem

Need Graph-based Representation for 3D Models



Source: Liu et al. [2024]

Problem Formulation

- **Input**:
  - Two disjoint graphs:
    - **Antibody Graph** $G_A = (V_A, E_A)$: CDR residues from heavy and light chains.
    - **Antigen Graph** $G_B = (V_B, E_B)$: Surface residues of the antigen.
  - Adjacency matrices E_A and E_B based on residue proximity (distance $<$ 4.5Å).
- **Encoding**: Encode each residue into a vector
- **Output**: Nodes and edges between interacting antibody and antigen residues.

Given an antibody-antigen graph pair, predict the binding nodes on the antigen – essentially a binary classification problem

Task 1: Epitope Prediction

- **Definition**: Epitopes are antigen residues in contact with the antibody (distance < 4.5Å).
- **Task**: Binary classification of antigen nodes:
  - Label $= 1$: Epitope residue.
  - Label $= 0$: Non-epitope residue.
- **Formulation**:

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

Task 2: Bipartite Link Prediction

- **Definition**: Predict interactions between antibody and antigen residues.
- **Task**: Binary classification of edges in a bipartite graph:
  - Label $= 1$: Residues are in contact (distance < 4.5Å).
  - Label $= 0$: Residues are not in contact.
- **Formulation**:

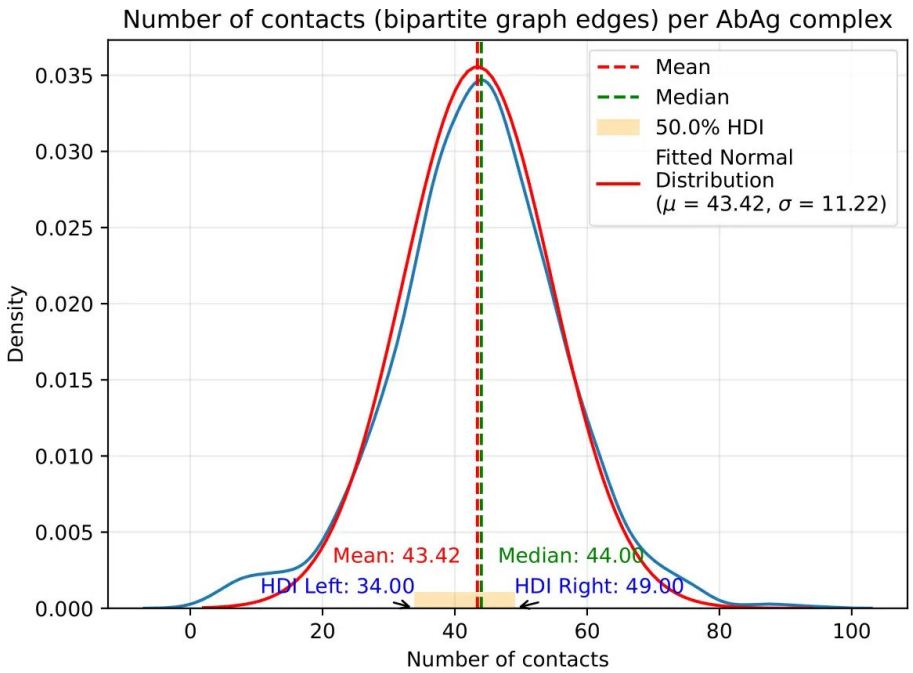
$$g(v_a, v_b; K_{m,n}) = \begin{cases} 1 & \text{if } v_a \text{ and } v_b \text{ are in contact,} \\ 0 & \text{otherwise.} \end{cases}$$

AsEP Benchmark Dataset

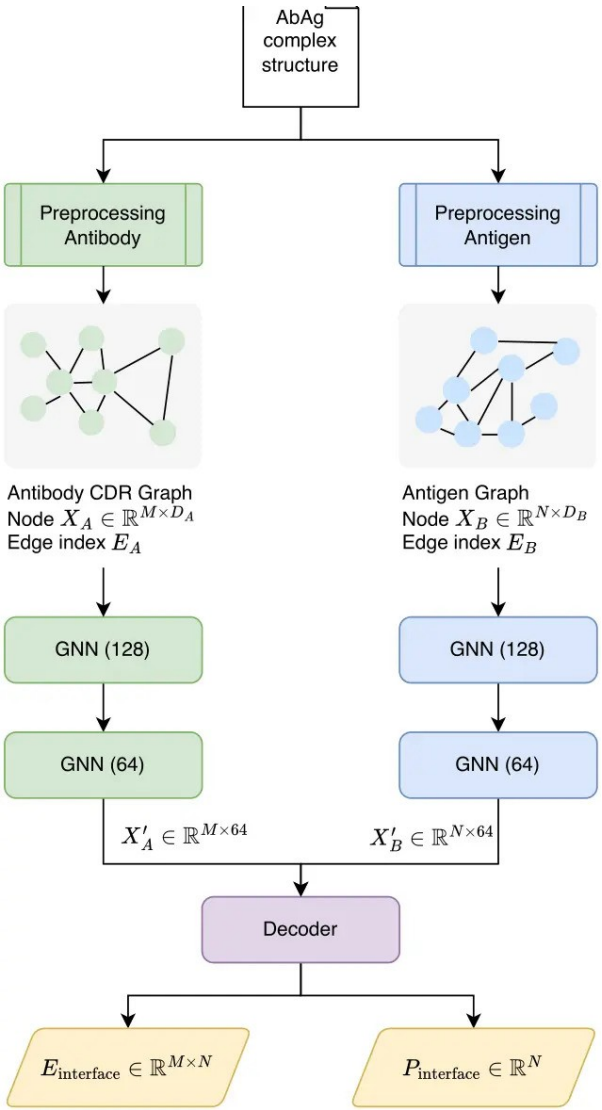
- Currated by Liu et al. [2024] using Antibody Database (AbDb) and Protein Data Bank (PDB)
- **Size**: 1,723 antibody-antigen complexes.
- **Features**:
  - Clustered epitope groups.
  - Pre-built graph representations.
  - Customizable embeddings (AntiBERTy, ESM2, ProtBERT).
- **Split**: Random split by epitope-to-antigen surface ratio and epitope groups.

Exploratory Data Analysis (EDA)

- **Distribution of Epitope Residues**:
  - Mean epitope residues: 14.6 $\pm$ 4.9.
  - Antigen surface residues: Hundreds.
- **Contact Distribution**:
  - Mean residue-residue contacts: 43.42.
  - Standard deviation: 11.22.
- **Epitope Groups**:
  - 641 unique antigens.
  - 973 epitope groups.

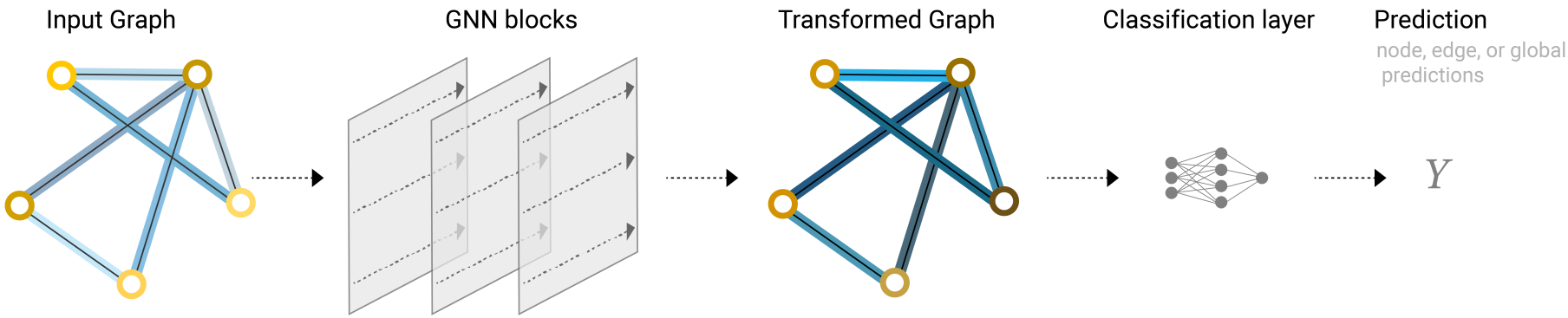


Experimental Setup



Source: Liu et al. [2024]

Graph Neural Networks – a Recap



Source: Distill

Embedding Methods – Protein Language Models (PLM)

1. **AntiBERTy**[1]: an antibody-specific transformer language model pre-trained on 558M natural antibody sequences producing embeddings of size $L \times 512$
2. **ESM2**[2]: SOTA general-purpose PLM – used `esm2_t33_650M_UR50D` producing embeddings of size $L \times 1280$
3. **ESM-IF1**[2]: Inverse folding model to design sequences for given structures – used `esm_if1_gvp4_t16_142M_UR50` producing embeddings of size $L \times 512$ given PDB structures
4. **ProtBERT**[3]: producing embeddings of size $L \times 1024$
5. Classical embeddings methods such as **BLOSUM62** and **One-Hot Encoding**

[1] https://github.com/jeffreyruffolo/AntiBERTy.git
[2] https://github.com/facebookresearch/esm.git
[3] https://github.com/agemagician/ProtTrans.git

Results: Epitope Node Prediction – Epitope Ratio Split

Algorithm	Embedding (Ab/Ag)	MCC	AUPRC	AUC-ROC	Precision	Recall	F1	BACC
GCN	AntiBERTy/ESM-IF	0.000	0.101	0.500	0.074	0.211	0.110	0.510
	AntiBERTy/ESM2	0.245	0.226	0.649	0.247	0.488	0.328	0.690
	AntiBERTy/ProtBERT	**0.263**	**0.251**	**0.650**	**0.281**	0.457	**0.348**	0.686
	BLOSUM62/BLOSUM62	0.048	0.113	0.541	0.081	0.465	0.139	0.541
	ESM-IF/ESM-IF	0.000	0.101	0.500	0.068	1.000	0.127	0.500
	ESM-IF/ESM2	0.219	0.210	0.633	0.243	0.466	0.320	0.680
	ESM-IF/ProtBERT	0.232	0.225	0.640	0.253	0.459	0.326	0.680
	ESM2/ESM-IF	0.000	0.101	0.500	0.042	0.094	0.058	0.468
	ESM2/ESM2	0.228	0.211	0.643	0.231	0.504	0.317	0.691
	ESM2/ProtBERT	0.228	0.205	0.655	0.218	0.553	0.313	**0.705**
	ProtBERT/ESM-IF	0.000	0.101	0.500	0.068	**0.956**	0.127	0.501
	ProtBERT/ESM2	0.222	0.215	0.631	0.249	0.441	0.319	0.672
	ProtBERT/ProtBERT	0.236	0.225	0.642	0.259	0.461	0.332	0.682
	One-Hot/One-Hot	0.049	0.113	0.543	0.079	0.533	0.138	0.542
GCN-L	AntiBERTy/ESM-IF	0.000	0.101	0.500	0.074	0.242	0.113	0.510
	AntiBERTy/ESM2	**0.256**	0.247	**0.650**	0.264	0.468	0.338	**0.687**
	AntiBERTy/ProtBERT	0.252	**0.251**	0.642	**0.283**	0.426	**0.340**	0.674
	BLOSUM62/BLOSUM62	0.048	0.113	0.541	0.082	0.457	0.140	0.543
	ESM-IF/ESM-IF	0.000	0.101	0.500	0.068	**1.000**	0.127	0.500
	ESM-IF/ESM2	0.000	0.101	0.500	0.068	**1.000**	0.127	0.500
	ESM-IF/ProtBERT	0.219	0.218	0.634	0.240	0.472	0.318	0.682
	ESM2/ESM-IF	0.000	0.101	0.500	0.045	0.146	0.069	0.461
	ESM2/ESM2	0.220	0.203	0.641	0.222	**0.521**	0.311	0.694
	ESM2/ProtBERT	0.244	0.224	0.653	0.244	0.511	0.330	0.698
	ProtBERT/ESM-IF	0.000	0.101	0.500	0.056	0.293	0.094	0.466
	ProtBERT/ESM2	0.234	0.217	0.640	0.253	0.470	0.329	0.684
	ProtBERT/ProtBERT	0.216	0.204	0.634	0.238	0.480	0.318	0.684
	One-Hot/One-Hot	0.045	0.112	0.539	0.078	0.509	0.136	0.537

Results: Bipartite Link Prediction – Epitope Ratio Split

Algorithm	Embedding (Ab/Ag)	MCC	AUPRC	AUC-ROC	Precision	Recall	F1	BACC
GCN	AntiBERTy/ESM-IF	0.000	0.013	0.752	0.000	0.000	0.000	0.500
	AntiBERTy/ESM2	0.114	0.086	**0.852**	0.058	0.298	0.098	0.643
	AntiBERTy/ProtBERT	**0.126**	**0.100**	0.840	**0.063**	0.302	**0.104**	0.645
	BLOSUM62/BLOSUM62	0.000	0.011	0.688	0.000	0.000	0.000	0.500
	ESM-IF/ESM-IF	0.000	0.005	0.496	0.000	0.000	0.000	0.500
	ESM-IF/ESM2	0.036	0.022	0.723	0.027	0.128	0.045	0.558
	ESM-IF/ProtBERT	0.042	0.023	0.713	0.028	0.164	0.047	0.574
	ESM2/ESM-IF	0.000	0.013	0.761	0.000	0.000	0.000	0.500
	ESM2/ESM2	0.100	0.064	0.843	0.048	0.308	0.083	0.646
	ESM2/ProtBERT	0.102	0.070	0.833	0.039	**0.373**	0.070	**0.674**
	ProtBERT/ESM-IF	0.000	0.013	0.763	0.000	0.000	0.000	0.500
	ProtBERT/ESM2	0.099	0.065	0.840	0.054	0.247	0.089	0.618
	ProtBERT/ProtBERT	0.113	0.078	0.827	0.053	0.312	0.090	0.649
	One-Hot/One-Hot	0.000	0.011	0.687	0.000	0.000	0.000	0.500
GCN-L	AntiBERTy/ESM-IF	0.000	0.012	0.752	0.000	0.000	0.000	0.500
	AntiBERTy/ESM2	**0.121**	0.089	0.843	0.057	0.315	0.096	0.651
	AntiBERTy/ProtBERT	0.115	**0.090**	0.833	**0.059**	0.297	**0.098**	0.642
	BLOSUM62/BLOSUM62	0.000	0.011	0.684	0.000	0.000	0.000	0.500
	ESM-IF/ESM-IF	0.000	0.004	0.504	0.000	0.000	0.000	0.500
	ESM-IF/ESM2	0.001	0.014	0.676	0.023	0.003	0.005	0.501
	ESM-IF/ProtBERT	0.045	0.023	0.708	0.027	0.178	0.046	0.580
	ESM2/ESM-IF	0.000	0.013	0.761	0.000	0.000	0.000	0.500
	ESM2/ESM2	0.091	0.054	0.845	0.045	0.278	0.077	0.631
	ESM2/ProtBERT	0.109	0.083	0.838	0.046	**0.344**	0.081	**0.662**
	ProtBERT/ESM-IF	0.000	0.013	0.765	0.000	0.000	0.000	0.500
	ProtBERT/ESM2	0.105	0.072	**0.850**	0.058	0.274	0.096	0.631
	ProtBERT/ProtBERT	0.098	0.067	0.835	0.049	0.288	0.083	0.637
	One-Hot/One-Hot	0.000	0.011	0.680	0.000	0.000	0.000	0.500

Results: Epitope Node Prediction – Epitope Group Split

Algorithm	Embedding (Ab/Ag)	MCC	AUPRC	AUC-ROC	Precision	Recall	F1	BACC
GCN	AntiBERTy/ESM-IF	0.000	0.098	0.500	0.076	0.275	0.119	0.521
	AntiBERTy/ESM2	0.112	0.149	0.566	0.158	0.264	0.197	0.583
	AntiBERTy/ProtBERT	0.112	**0.153**	0.564	0.167	0.240	0.197	0.578
	BLOSUM62/BLOSUM62	0.045	0.112	0.537	0.084	0.412	0.139	0.541
	ESM-IF/ESM-IF	0.000	0.098	0.500	0.065	1.000	0.122	0.500
	ESM-IF/ESM2	0.117	0.148	**0.573**	0.165	0.291	0.211	**0.594**
	ESM-IF/ProtBERT	0.107	0.152	0.560	0.162	0.231	0.190	0.574
	ESM2/ESM-IF	0.000	0.098	0.500	0.063	0.025	0.036	0.500
	ESM2/ESM2	0.112	0.151	0.566	0.157	0.252	0.193	0.579
	ESM2/ProtBERT	**0.117**	0.152	0.567	**0.181**	0.255	**0.212**	0.587
	ProtBERT/ESM-IF	0.000	0.098	0.500	0.061	0.445	0.108	0.485
	ProtBERT/ESM2	0.098	0.145	0.551	0.175	0.188	0.181	0.563
	ProtBERT/ProtBERT	0.094	0.144	0.552	0.172	0.195	0.183	0.565
	One-Hot/One-Hot	0.044	0.108	0.542	0.076	**0.556**	0.134	0.543
GCN-L	AntiBERTy/ESM-IF	0.000	0.098	0.500	0.079	0.289	0.124	0.527
	AntiBERTy/ESM2	0.123	0.147	0.569	0.162	0.280	0.205	0.590
	AntiBERTy/ProtBERT	0.112	0.153	0.564	0.167	0.240	0.197	0.578
	BLOSUM62/BLOSUM62	0.050	0.113	0.543	0.080	0.490	0.138	0.540
	ESM-IF/ESM-IF	0.000	0.098	0.500	0.065	1.000	0.122	0.500
	ESM-IF/ESM2	0.000	0.098	0.500	0.060	0.535	0.109	0.478
	ESM-IF/ProtBERT	0.099	0.150	0.557	0.155	0.239	0.188	0.574
	ESM2/ESM-IF	0.000	0.098	0.500	0.053	0.183	0.082	0.478
	ESM2/ESM2	0.06	0.071	0.43	0.032	0.024	0.074	0.038
	ESM2/ProtBERT	0.095	0.144	0.549	0.173	0.186	0.179	0.562
	ProtBERT/ESM-IF	0.000	0.098	0.500	0.060	0.535	0.109	0.478
	ProtBERT/ESM2	0.095	0.144	0.549	0.173	0.186	0.179	0.562
	ProtBERT/ProtBERT	0.115	0.151	0.564	0.183	0.234	0.205	0.581
	One-Hot/One-Hot	0.046	0.109	0.544	0.076	0.542	0.134	0.543

Results: Bipartite Link Prediction – Epitope Group Split

Algorithm	Embedding (Ab/Ag)	MCC	AUPRC	AUC-ROC	Precision	Recall	F1	BACC
GCN	AntiBERTy/ESM-IF	0.000	0.012	0.733	0.000	0.000	0.000	0.500
	AntiBERTy/ESM2	**0.054**	**0.038**	**0.796**	0.031	0.124	0.050	0.557
	AntiBERTy/ProtBERT	0.050	0.034	0.779	0.027	0.128	0.045	0.558
	BLOSUM62/BLOSUM62	0.000	0.011	0.667	0.000	0.000	0.000	0.500
	ESM-IF/ESM-IF	0.000	0.004	0.510	0.000	0.000	0.000	0.500
	ESM-IF/ESM2	0.017	0.013	0.648	0.015	0.060	0.024	0.525
	ESM-IF/ProtBERT	0.015	0.012	0.644	0.012	0.071	0.021	0.528
	ESM2/ESM-IF	0.000	0.013	0.745	0.000	0.000	0.000	0.500
	ESM2/ESM2	0.051	0.032	0.779	0.027	**0.136**	0.045	**0.561**
	ESM2/ProtBERT	0.046	0.031	0.789	**0.034**	0.111	**0.052**	0.551
	ProtBERT/ESM-IF	0.000	0.014	0.750	0.000	0.000	0.000	0.500
	ProtBERT/ESM2	0.045	0.034	0.770	0.029	0.100	0.045	0.546
	ProtBERT/ProtBERT	0.043	0.028	0.768	0.030	0.108	0.047	0.549
	One-Hot/One-Hot	0.000	0.011	0.668	0.000	0.000	0.000	0.500
GCN-L	AntiBERTy/ESM-IF	0.000	0.012	0.731	0.000	0.000	0.000	0.500
	AntiBERTy/ESM2	0.057	0.037	0.807	0.037	0.121	0.056	0.556
	AntiBERTy/ProtBERT	0.050	0.034	0.779	0.027	0.128	0.045	0.558
	BLOSUM62/BLOSUM62	0.000	0.011	0.686	0.000	0.000	0.000	0.500
	ESM-IF/ESM-IF	0.000	0.004	0.490	0.000	0.000	0.000	0.500
	ESM-IF/ESM2	0.000	0.011	0.686	0.000	0.000	0.000	0.500
	ESM-IF/ProtBERT	0.020	0.013	0.641	0.014	0.083	0.024	0.534
	ESM2/ESM-IF	0.000	0.013	0.743	0.000	0.000	0.000	0.500
	ESM2/ESM2	0.045	0.028	0.767	0.029	0.106	0.046	0.548
	ESM2/ProtBERT	0.045	0.028	0.767	0.029	0.106	0.046	0.548
	ProtBERT/ESM-IF	0.000	0.014	0.749	0.000	0.000	0.000	0.500
	ProtBERT/ESM2	0.000	0.011	0.686	0.000	0.000	0.000	0.500
	ProtBERT/ProtBERT	0.053	0.033	0.771	0.031	0.129	0.050	0.559
	One-Hot/One-Hot	0.000	0.012	0.669	0.000	0.000	0.000	0.500

Results: Epitope Node Prediction

Table: Performance on Dataset Split by Epitope-to-Antigen Surface Ratio

Algorithm	**MCC**	**Precision**	**Recall**	**AUCROC**	**F1**
WALLE++	**0.263**	**0.281**	**0.457**	**0.650**	**0.348**
WALLE	0.210	0.235	0.258	0.635	0.145
EpiPred	0.029	0.122	0.142	—	0.112
ESMFold	0.028	0.137	0.060	—	0.046
ESMBind	0.016	0.106	0.090	0.506	0.064
MaSIF-site	0.037	0.125	0.114	—	0.128

Comparison with State-of-the-Art

Table: Benchmark Result on epitope3D Test Set

Method	**F1**	**MCC**	**BACC**	**AUC-ROC**
SEPPA 3	0.14	0.02	0.52	0.52
DiscoTope-2.0	0.11	-0.01	0.50	0.49
ElliPro	0.11	-0.06	0.44	0.44
BepiPred-2.0	0.15	0.04	0.55	0.54
epitope3D	0.02	-0.02	0.49	0.49
Bepipred-3.0	0.19	0.08	0.57	0.71
DiscoTope-3.0	0.20	0.09	0.57	0.71
GraphBepi	0.28	0.16	0.62	0.64
EpiGraph	0.29	0.19	0.62	0.73
WALLE++	**0.348**	**0.263**	**0.686**	**0.650**

Future Work

- **Expand Dataset**: Include novel antibodies (e.g., nanobodies).
- **Integrarte Embedding Methods**: Integrate PLM-based embedding methods to preserve meaningful features.
- **Frontier Deep Learning Models**: Explore DL methods such as Graph Attention Networks (GANs)
- **Enhance Model**: Incorporate more edge features (e.g., inter-residue distances).
- **Applications**: Accelerate antibody design for therapeutic development.

References

Sara Joubbi, Alessio Micheli, Paolo Milazzo, Giuseppe Maccari, Giorgio Ciano, Dario Cardamone, and Duccio Medini. Antibody design using deep learning: from sequence and structure design to affinity maturation. *Briefings in Bioinformatics*, 25(4): bbae307, 2024.

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.

Questions & Suggestions!!