GEORGETOWN UNIVERSITY

Capstone Project at Argentys Informatics

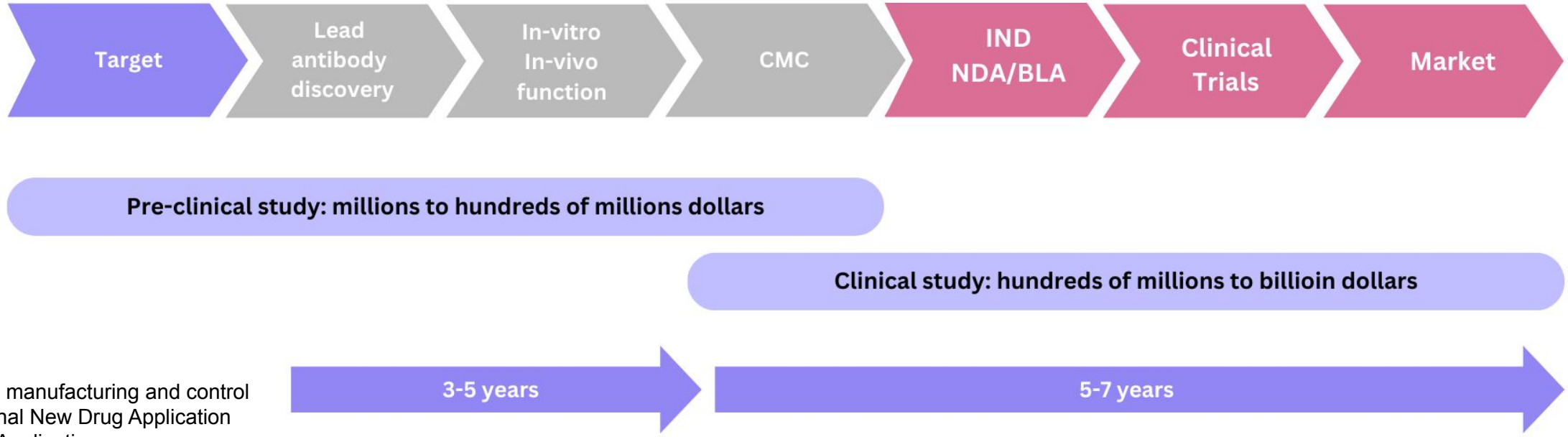
Antibody Design: Antigen-Antibody Binding Prediction

by Aizhan Uteubayeva

Table of content

Introduction: Antibody Therapy

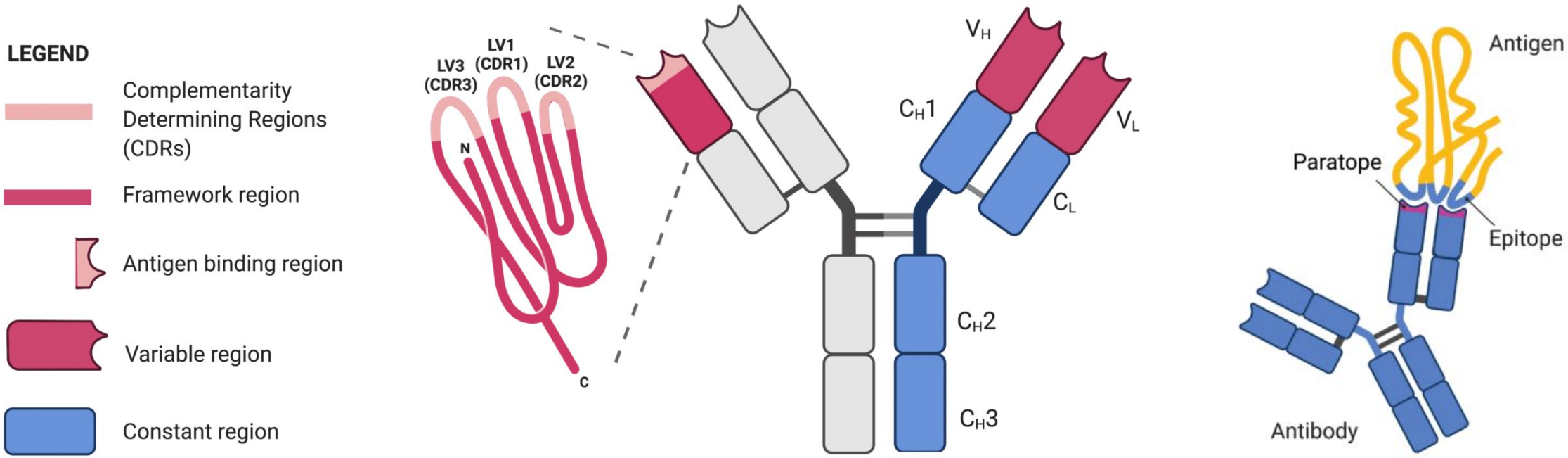
- treatment of cancer, autoimmune disorder, infectious disease, etc.
- revenue **>$140 billion** by 2025



CMC: Chemistry, manufacturing and control
IND: Investigational New Drug Application
NDA: New Drug Application
BLA: Biologics License Application

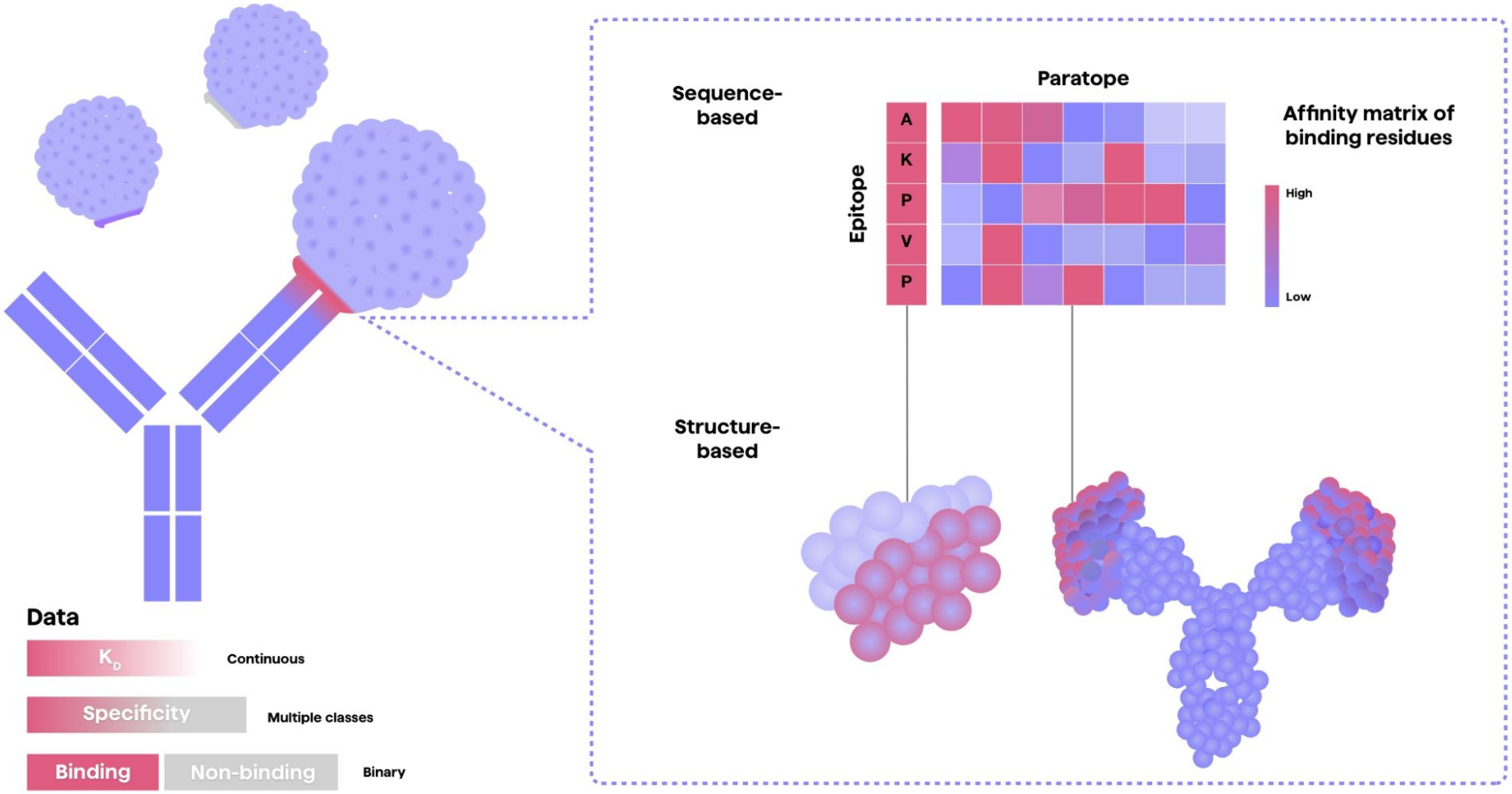
(Kaplon *et al.*, 2020; Ecker *et al.*, 2015; Reichert, 2012)

Introduction: Antibody and Antigen Structure

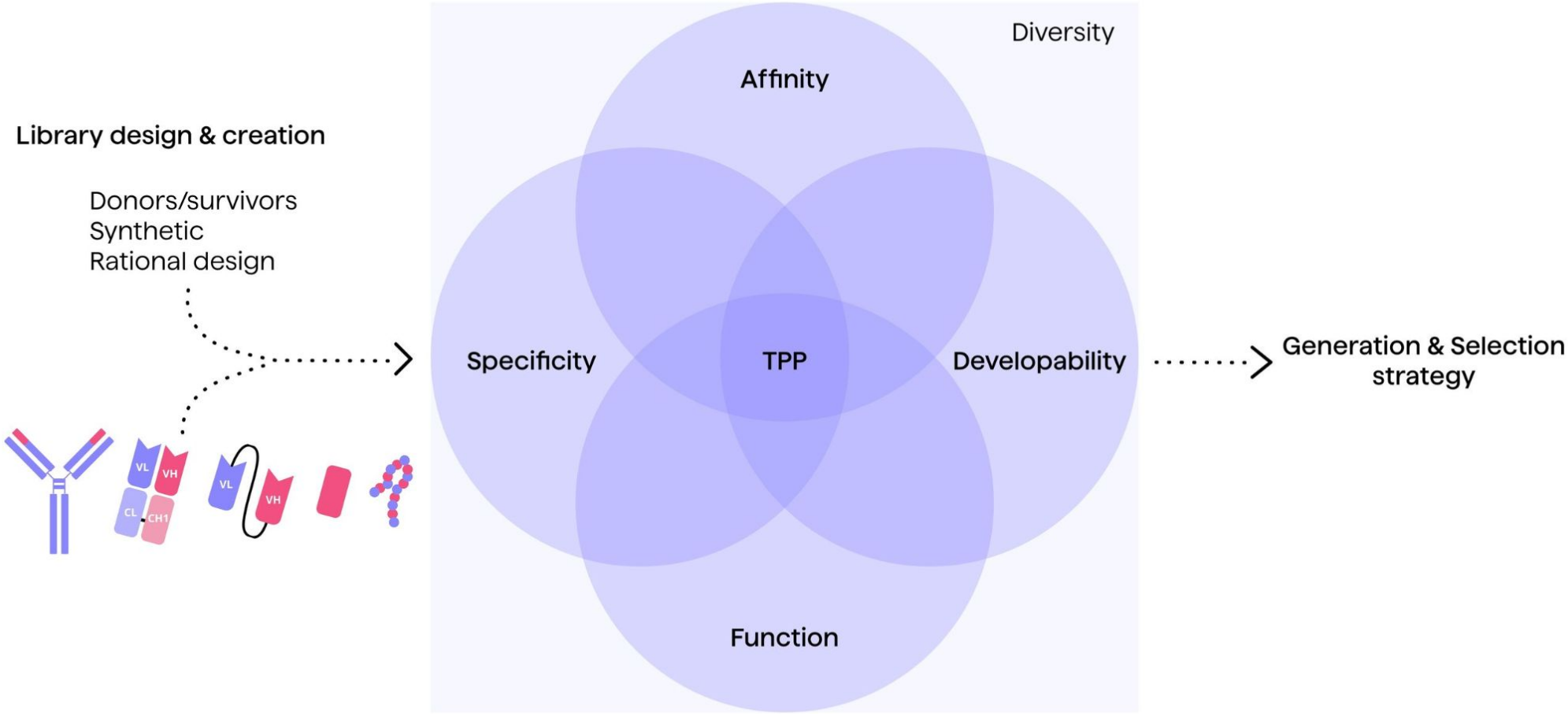


(Kulytė *et al.*, 2024)

Introduction: ML in Antibody Design

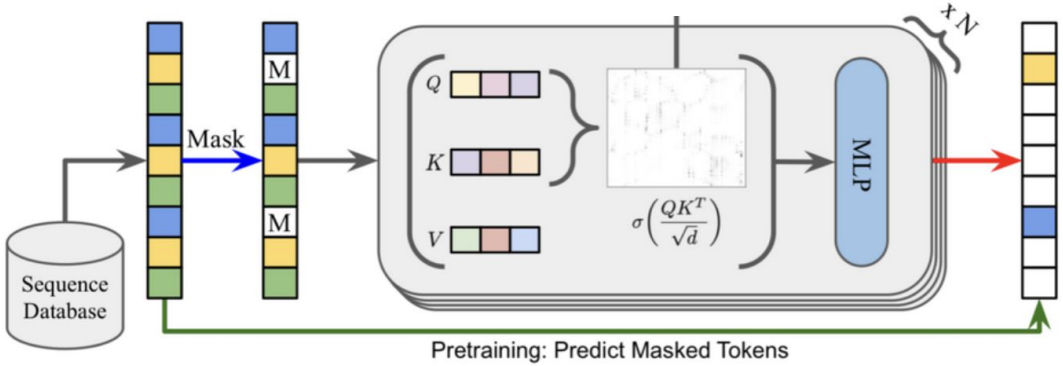


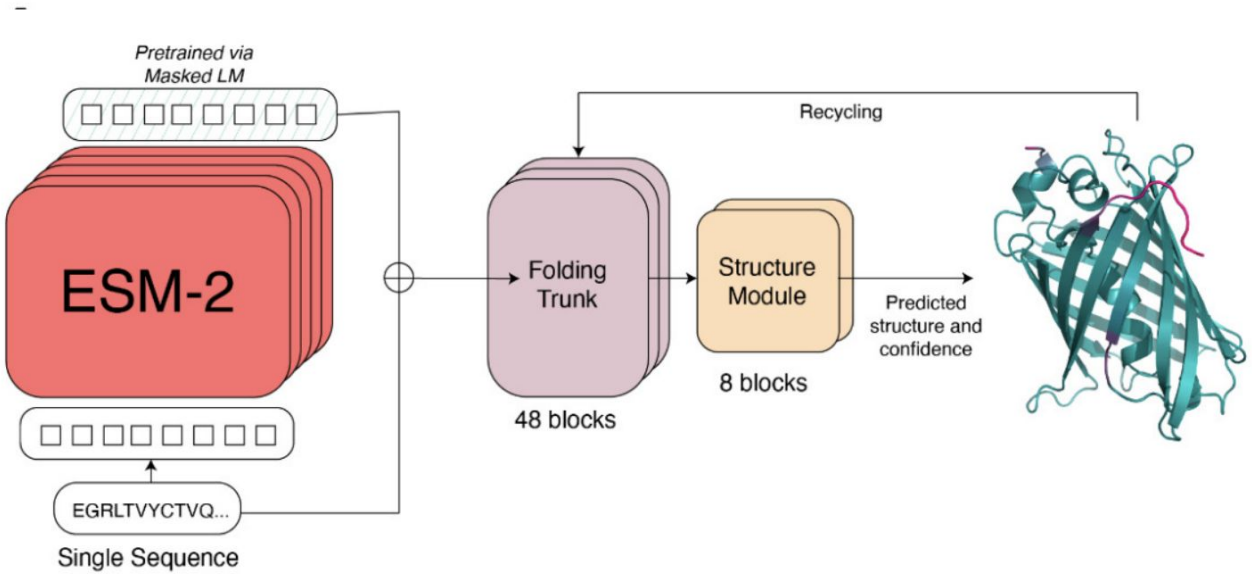
Introduction: Antibody Libraries



Introduction: ESM2

- state-of-the-art protein model trained on a masked language modelling (MLM)
- trained with ~65 million unique sequences
- publicly available on Hugging Face





(Lin *et al.*, 2023)

Introduction: ESM2 evaluation

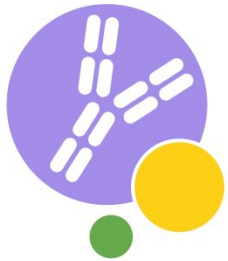
Model	# Params	# Updates	Validation Perplexity	LR P@L	LR P@L/5	CASP14	CAMEO
ESM-2	8M	270K	10.45	0.16	0.28	0.37	0.48
	35M	270K	9.12	0.29	0.49	0.41	0.56
	150M	270K	8.00	0.42	0.68	0.47	0.63
	650M	270K	7.23	0.50	0.77	0.51	0.68
	3B	270K	6.73	0.53	0.80	0.51	0.71
	8M	500K	10.33	0.17	0.29	0.37	0.48
	35M	500K	8.95	0.30	0.51	0.41	0.56
	150M	500K	7.75	0.44	0.70	0.49	0.65
	650M	500K	6.95	0.52	0.79	0.51	0.70
	3B	500K	6.49	**0.54**	0.81	0.52	**0.72**
	15B	270K	**6.37**	**0.54**	**0.82**	**0.55**	**0.72**
ESM-1b	650M	—	—	0.41	0.66	0.42	0.64
Prot-T5-XL (UR50) (21)	3B	—	—	0.48	0.72	0.50	0.69
Prot-T5-XL (BFD) (21)	3B	—	—	0.36	0.58	0.46	0.63
CARP (24)	640M	—	—	—	—	0.42	0.59

Project Goal

As part of the *in silico de novo* antibody design project, we aim to be able to predict antigen-antibody binding solely based on the sequence information

Method: Data Collection

- public database documenting all published/patented antibodies and nanobodies able to bind to coronaviruses, including SARS-CoV2, SARS-CoV1, and MERS-CoV
- **12,916 entries**



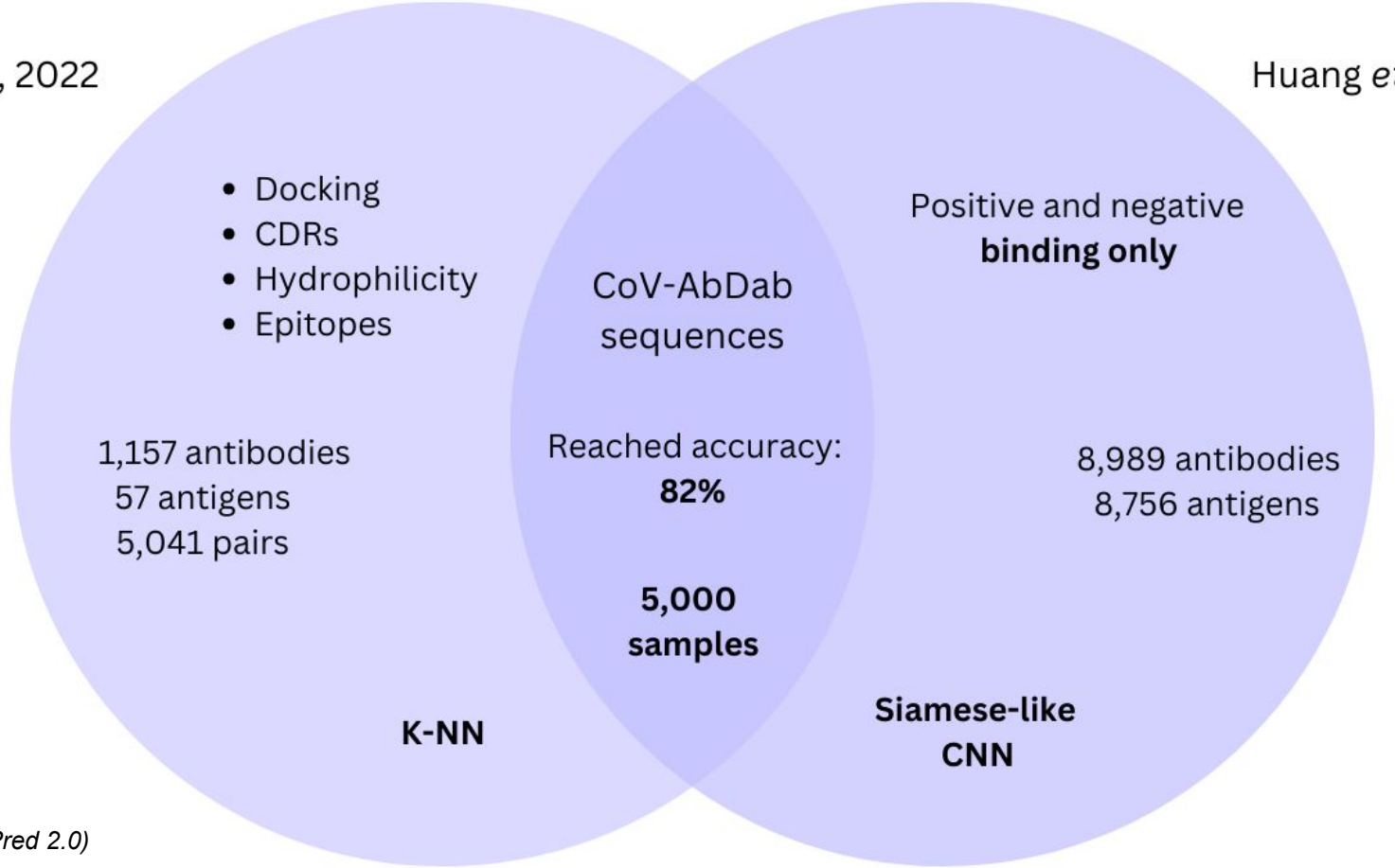
CoV-AbDab

The Coronavirus Antibody Database

opig

UNIVERSITY OF OXFORD

Method: Datasets



Applied tools for feature extraction:
ImMunoGeneTics
Immune Epitope Database (BepiPred 2.0)
Bachem peptide calculator

Method: Feature Extraction

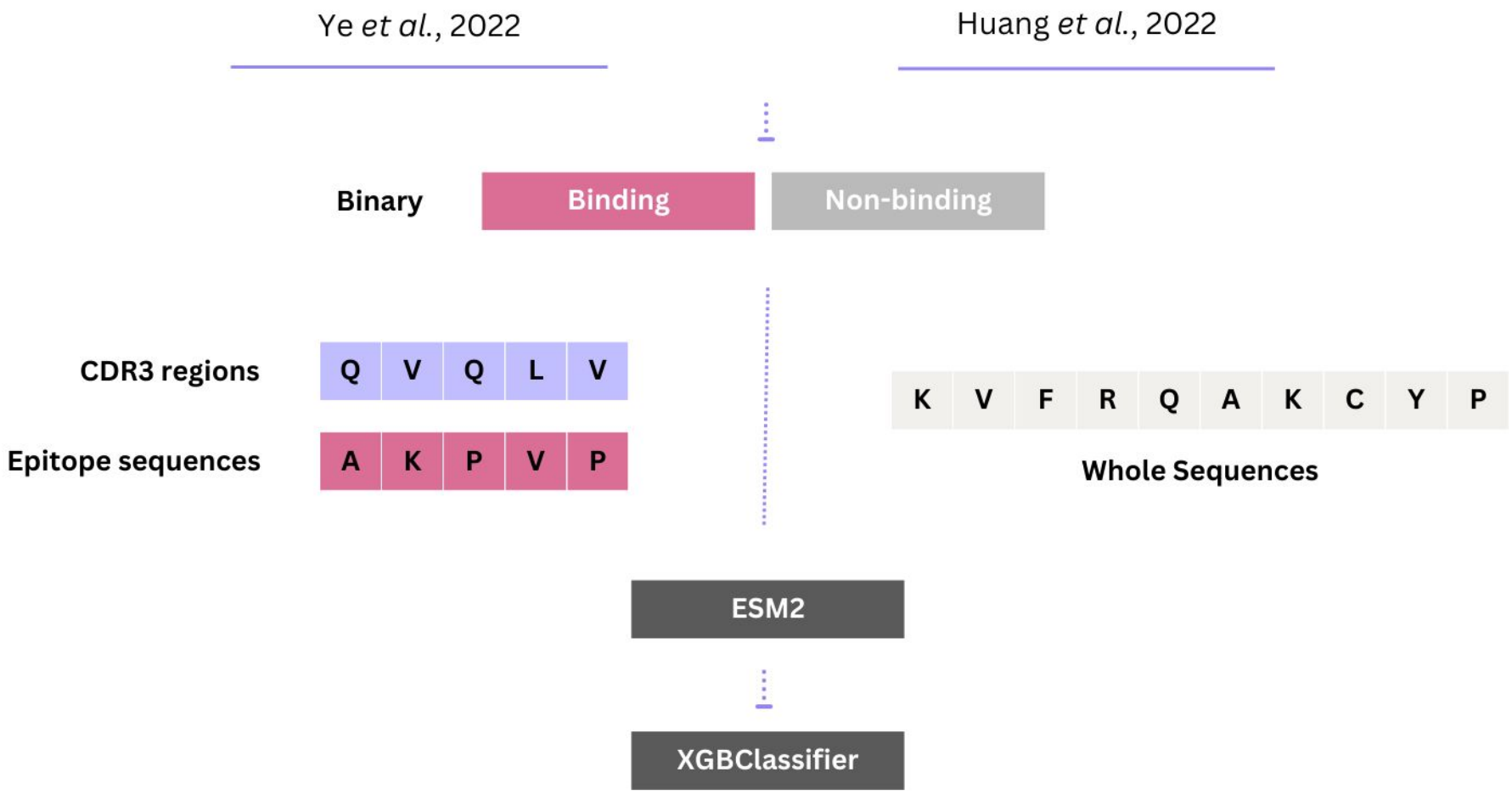
Checkpoint name	Num layers	Num parameters
esm2_t48_15B_UR50D	48	15B
esm2_t36_3B_UR50D	36	3B
esm2_t33_650M_UR50D	33	650M
esm2_t30_150M_UR50D	30	150M
esm2_t12_35M_UR50D	12	35M
esm2_t6_8M_UR50D	6	8M

```
Tokenized Input:
{'input_ids': tensor([[ 0,  9,  7, 16,  4,  7,  9,  8,  6,  6,  6,  7,  7, 17, 14,  6,  6,  8,
          4, 10,  4,  8, 23,  5,  6,  8,  6, 18, 11, 18,  8, 13, 19, 19, 20,  6,
         22, 12, 10, 16,  5, 14,  6, 15,  6,  4,  9,  7,  7,  8, 19, 12,  8, 11,
         11,  6,  8, 19, 12, 15, 13,  5, 13,  8,  7, 15,  6, 10, 18, 11, 12,  8,
         10, 13, 17,  5, 15, 17,  8,  7, 19,  4, 16, 20, 17,  8,  4, 10,  5,  9,
         13, 11,  5,  7, 19, 19, 23,  5, 10, 20,  6, 14, 19,  6,  8,  6,  8, 18,
         13, 19, 22,  6, 16,  6, 11,  4,  7, 11,  7,  8,  8,  2]]), 'attention_mask': tensor([[1,
         1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
         1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
         1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
         1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
         1, 1]])}

Embeddings:
tensor([[[-0.0500,  0.8885,  0.0392,  ...,  1.0270, -0.0238, -0.5494],
         [-0.0302, -0.1456,  0.0300,  ...,  0.5343,  0.0169, -0.1203],
         [-0.5252, -0.2263,  0.4586,  ...,  0.7090, -0.0535,  0.2848],
         ...,
         [-0.2821, -0.6369, -0.2422,  ...,  0.2596,  0.1478, -0.1605],
         [-0.1095, -0.7489, -0.4198,  ...,  0.1201, -0.0409, -0.2009],
         [ 0.0416, -0.3325,  0.0533,  ...,  0.3160, -0.8324, -0.2321]]])
```

- amino acid → tokens → embeddings

Method: Modelling



Results: 89% accuracy and improved class distinction

Metric	Ye *et al* (CDR3 + Epitopes)	Huang *et al* (whole sequences)
Sample Size	5041	11019
Accuracy	**0.8961**	**0.8982**
Precision (Class 0)	0.79	0.76
Precision (Class 1)	**0.97**	**0.91**
Recall (Class 0)	0.96	*0.5*
Recall (Class 1)	*0.86*	0.97
F1-Score (Class 0)	0.86	0.6
F1-Score (Class 1)	0.92	0.94
AUROC	**0.9673**	**0.9093**

- ➔ high precision for both
- ➔ identification of false negatives vs. missing true positives

Results: 89% accuracy and improved class distinction

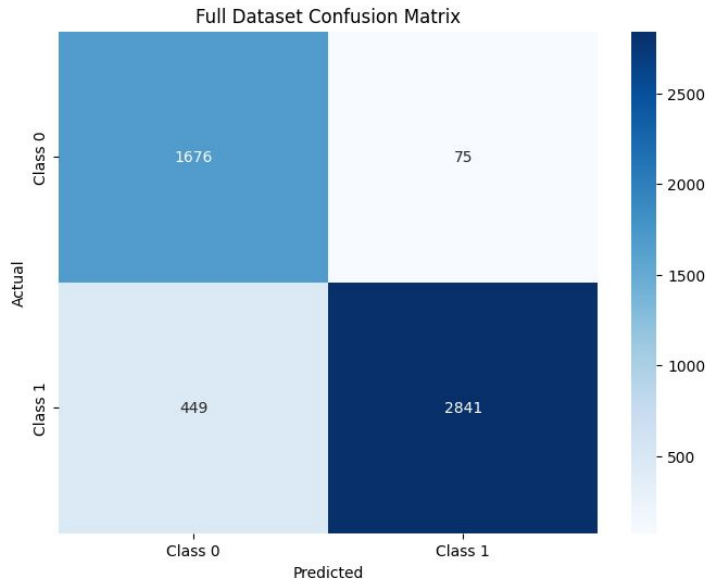
Metric	Ye *et al* (RF Test)	Ye *et al* (RF Full)	Ye *et al* (XGB Test)	Ye *et al* (XGB Test)
Sample Size	701	5041	701	5041
Accuracy	0.8474	0.8939	0.8359	0.8961
Precision (Class 0)	0.84	0.78	0.85	0.79
Precision (Class 1)	0.86	0.98	0.82	0.97
Recall (Class 0)	0.89	0.97	0.85	0.96
Recall (Class 1)	0.8	0.86	0.82	0.86
F1-Score (Class 0)	0.86	0.86	0.85	0.86
F1-Score (Class 1)	0.83	0.91	0.82	0.92
AUROC	0.9176	0.9707	0.9151	0.9673

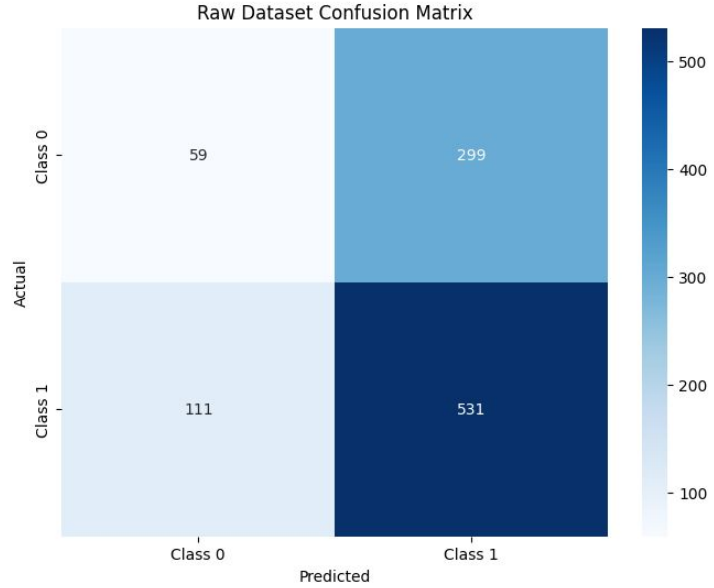
- ➔ high precision for both
- ➔ identification of false negatives vs. missing true positives

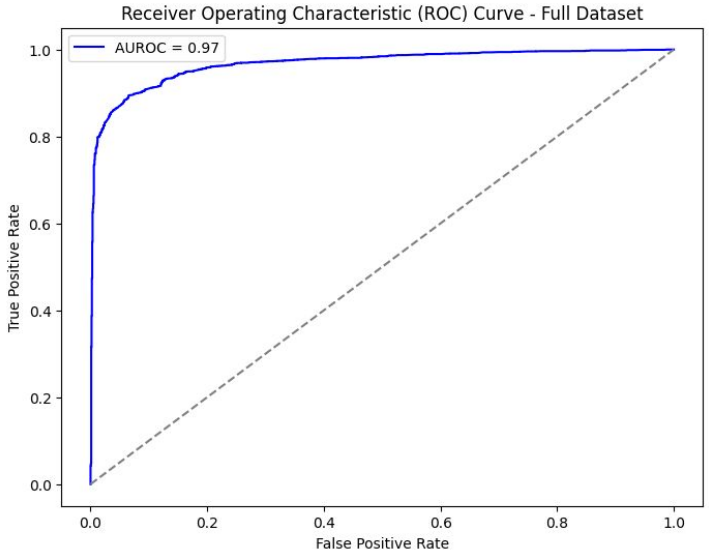
Results:

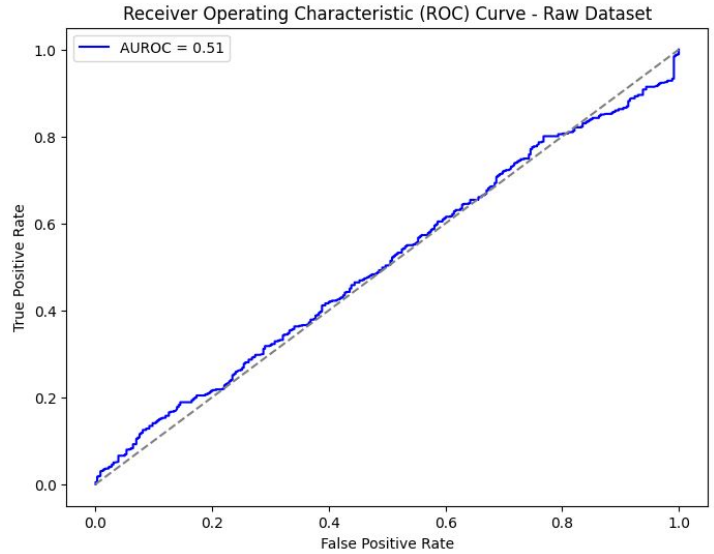
Sample size: 5,000

Accuracy for
Whole Sequences = **0.59**









Results: EpiDope

- prediction model for epitope extraction
- computationally heavy
- better performance then existing tools
- easily available

```
Reading input fasta.
Deep Neural Network model summary:
Model: "model_1"
```

Layer (type)	Output Shape	Param #	Connected to
input_2 (InputLayer)	(None, 49)	0	
input_1 (InputLayer)	(None, 49, 1024)	0	
embedding_1 (Embedding)	(None, 49, 10)	270	input_2[0][0]
bidirectional_1 (Bidirectional)	(None, 49, 10)	41200	input_1[0][0]
bidirectional_2 (Bidirectional)	(None, 49, 20)	1680	embedding_1[0][0]
dense_1 (Dense)	(None, 49, 10)	110	bidirectional_1[0][0]
dense_2 (Dense)	(None, 49, 10)	210	bidirectional_2[0][0]
leaky_re_lu_1 (LeakyReLU)	(None, 49, 10)	0	dense_1[0][0]
leaky_re_lu_2 (LeakyReLU)	(None, 49, 10)	0	dense_2[0][0]
flatten_1 (Flatten)	(None, 490)	0	leaky_re_lu_1[0][0]
flatten_2 (Flatten)	(None, 490)	0	leaky_re_lu_2[0][0]
concatenate_1 (Concatenate)	(None, 980)	0	flatten_1[0][0] flatten_2[0][0]
dense_3 (Dense)	(None, 10)	9810	concatenate_1[0][0]
leaky_re_lu_3 (LeakyReLU)	(None, 10)	0	dense_3[0][0]
dense_4 (Dense)	(None, 2)	22	leaky_re_lu_3[0][0]

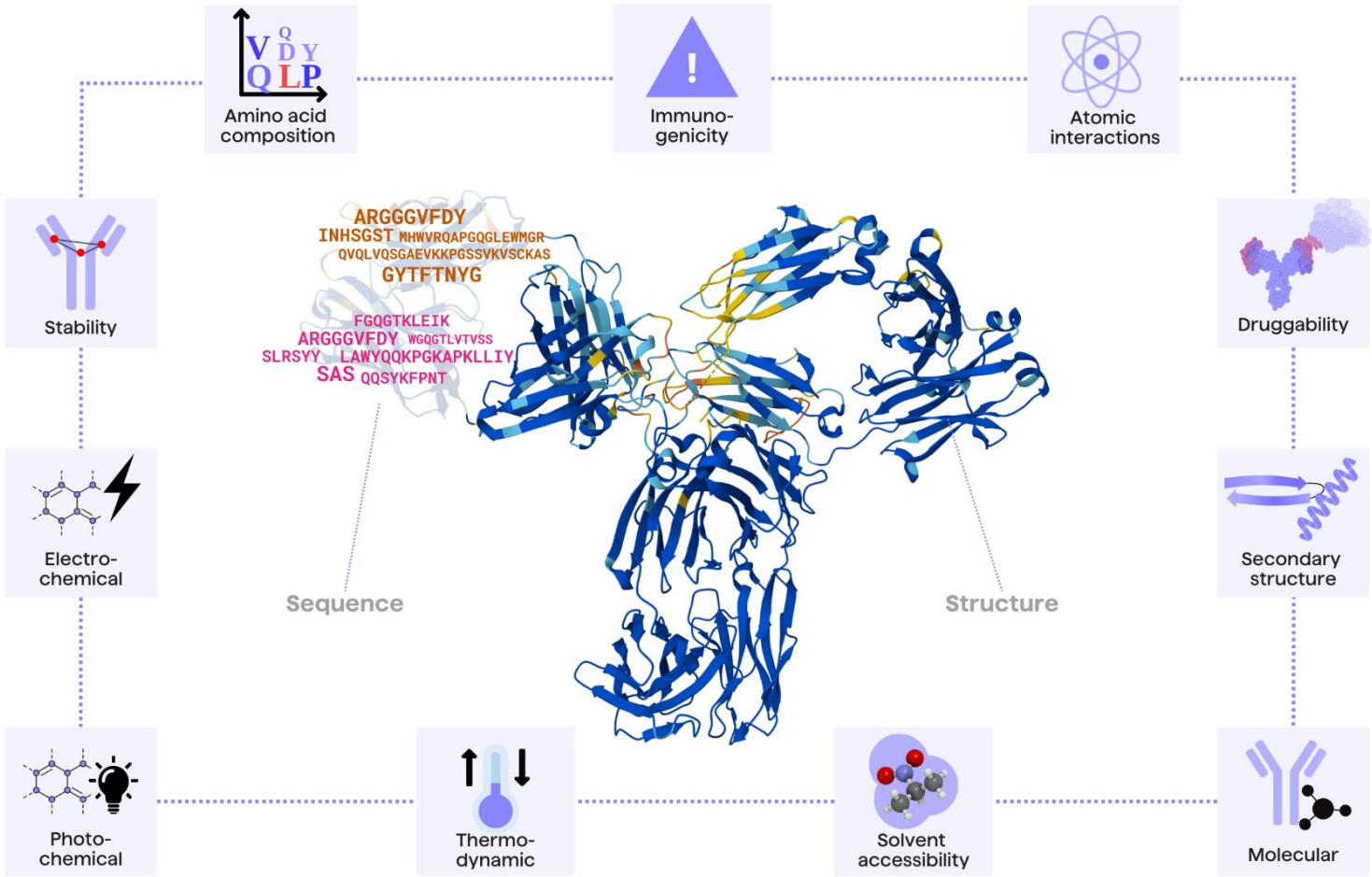
```
Total params: 53,302
Trainable params: 53,302
Non-trainable params: 0
```

Results: Interface Prototype

Discussion

- ➔ Open-source database, platform and tools
- ➔ Achieved 89% accuracy and AUROC > 0.90 for both datasets
- ➔ Choosing CDR3 regions and epitopes has shown to improve class distinction and requires half the dataset size for same level of performance
- ➔ Additional tools exist to extract CDR3 regions and epitopes but require computational resources - AlphaFold3

Future work: Developability



References

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., Rives, A., 2023. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science 379, 1123–1130. https://doi.org/10.1126/science.ade2574

Raybould, M.I.J., Kovaltsuk, A., Marks, C., Deane, C.M., 2021. CoV-AbDab: the coronavirus antibody database. Bioinformatics 37, 734–735. https://doi.org/10.1093/bioinformatics/btaa739

Ecker, D.M., Jones, S.D., Levine, H.L., 2015. The therapeutic monoclonal antibody market. MAbs 7, 9–14. https://doi.org/10.4161/19420862.2015.989042

Ye, C., Hu, W., Gaeta, B., 2022. Prediction of Antibody-Antigen Binding via Machine Learning: Development of Data Sets and Evaluation of Methods. JMIR Bioinform Biotechnol 3, e29404. https://doi.org/10.2196/29404

Huang, Y., Zhang, Z., Zhou, Y., 2022. AbAgIntPre: A deep learning method for predicting antibody-antigen interactions based on sequence information. Front. Immunol. 13. https://doi.org/10.3389/fimmu.2022.1053617

Collatz, M., Mock, F., Barth, E., Hölzer, M., Sachse, K., Marz, M., 2021. EpiDope: a deep neural network for linear B-cell epitope prediction. Bioinformatics 37, 448–455. https://doi.org/10.1093/bioinformatics/btaa773

Thank you for your attention!