

# Benchmarking machine learning methods for pattern prediction and recovery in antibody sequences

Rahmad Akbar<sup>1</sup>, Cédric R. Weber<sup>3</sup>, Igor Snapkov<sup>1</sup>, Daniel Heinesen<sup>1</sup>, Edvard Aksnes<sup>1</sup>, Zixuan Liu<sup>1</sup>, Milena Pavlovic<sup>1,2</sup>, Geir Kjetil Sandve<sup>2</sup>, Sai T. Reddy<sup>3</sup>, and Victor Greiff<sup>1</sup>.



UiO : University of Oslo



Eidgenössische Technische Hochschule Zürich  
Swiss Federal Institute of Technology Zurich

<sup>1</sup> UiO, Department of Immunology, Computational and Systems Immunology, Oslo, Norway

<sup>2</sup> UiO, Institute of Informatics, Biomedical Informatics, Oslo, Norway  
<sup>3</sup>

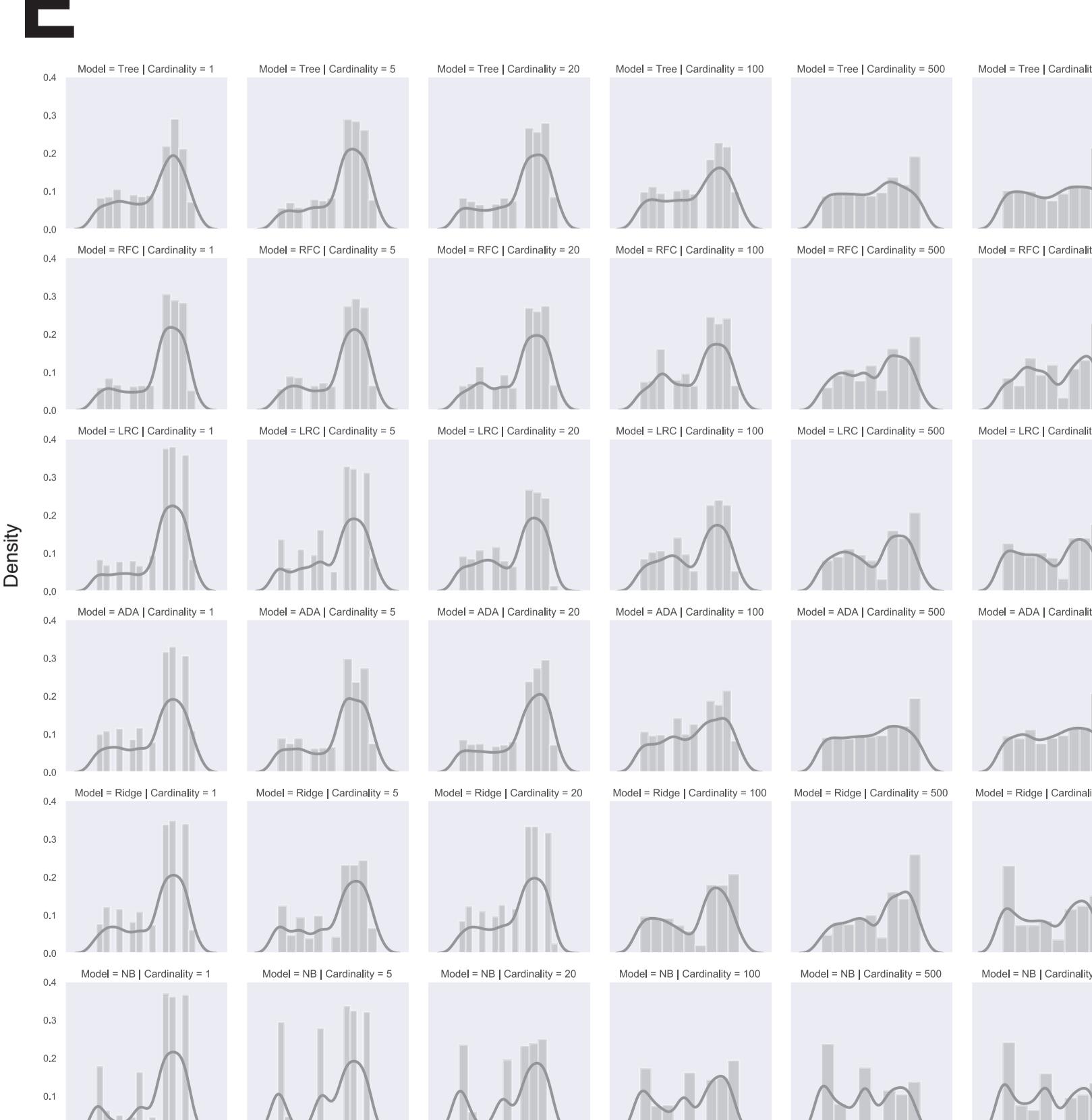
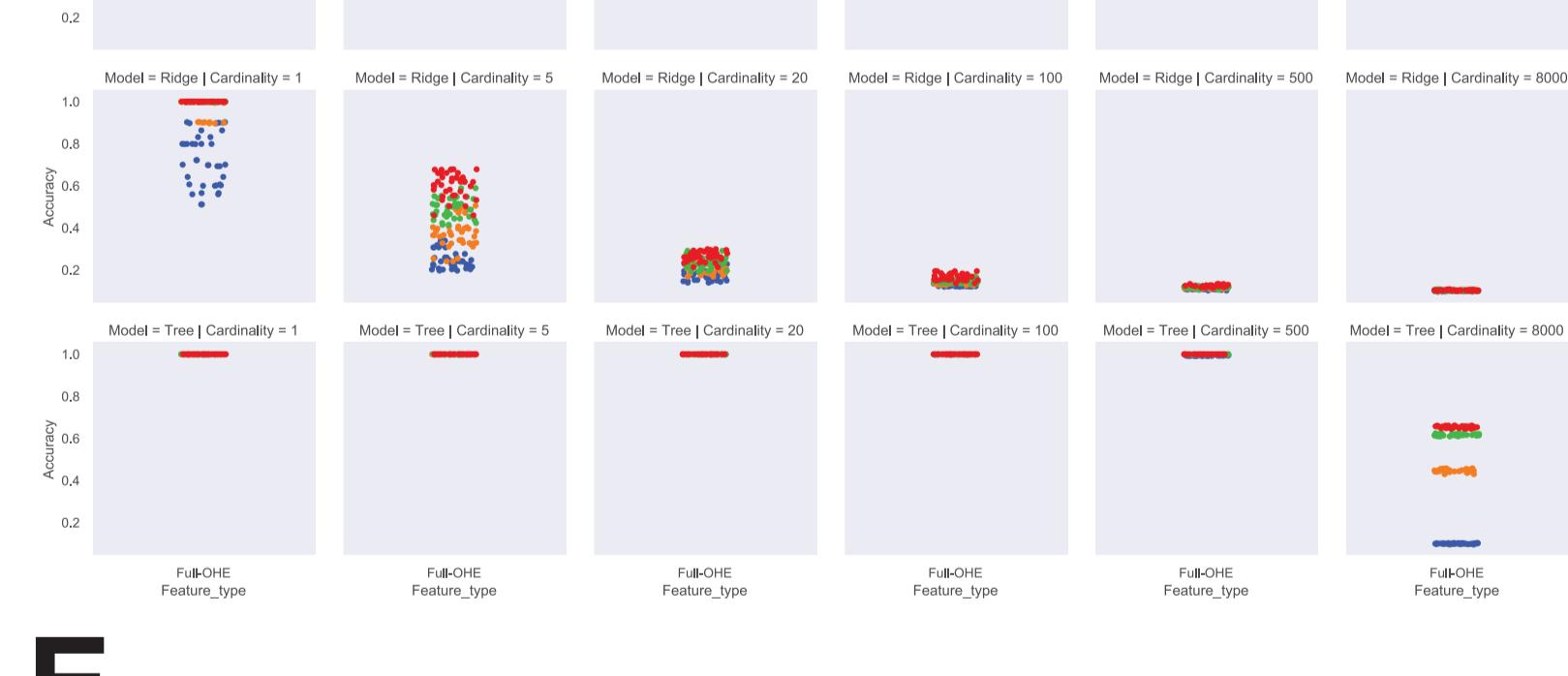
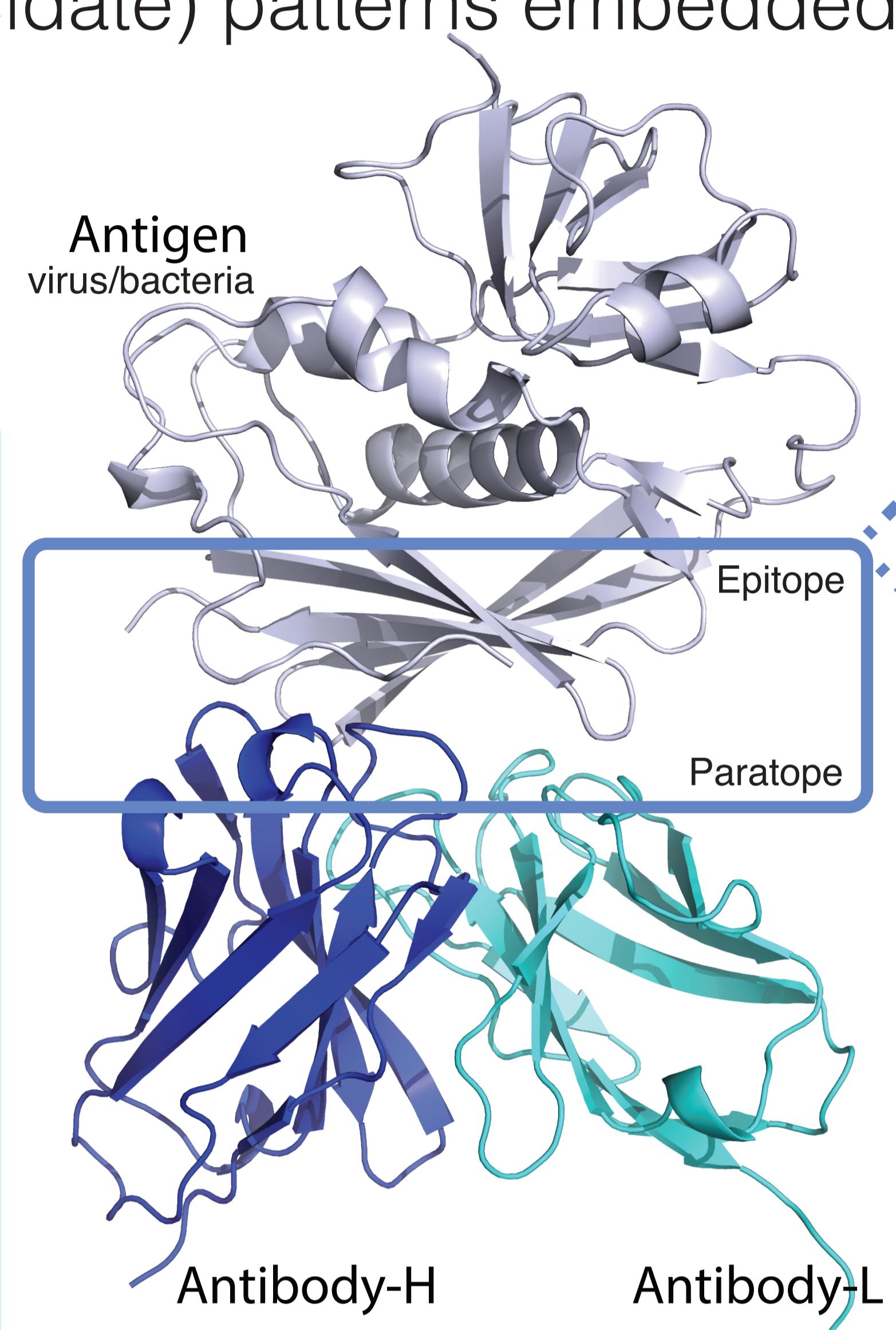
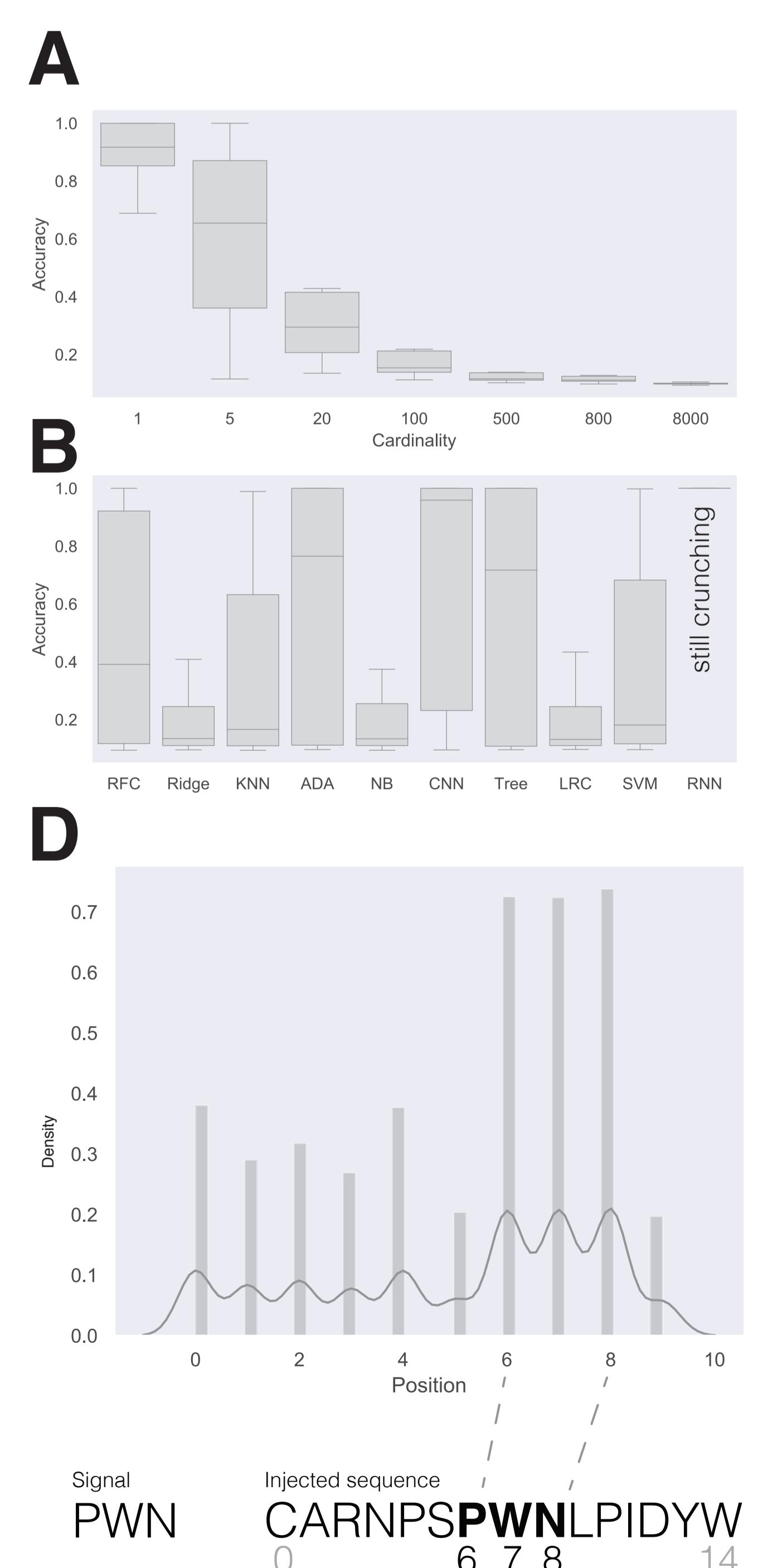
<sup>3</sup> ETH Zürich, D-BSSE, Mattenstrasse 26, 4058, Basel, Switzerland

# Motivation

Six out of ten top selling drugs are antibodies. Hence, accurate prediction of antibody specificity from the antibody sequence alone is of paramount importance for the conception of next-generation antibody therapeutics. Can machine learning help in such predictions? To what extent machine learning algorithms recover (elucidate) patterns embedded in antibody sequences?

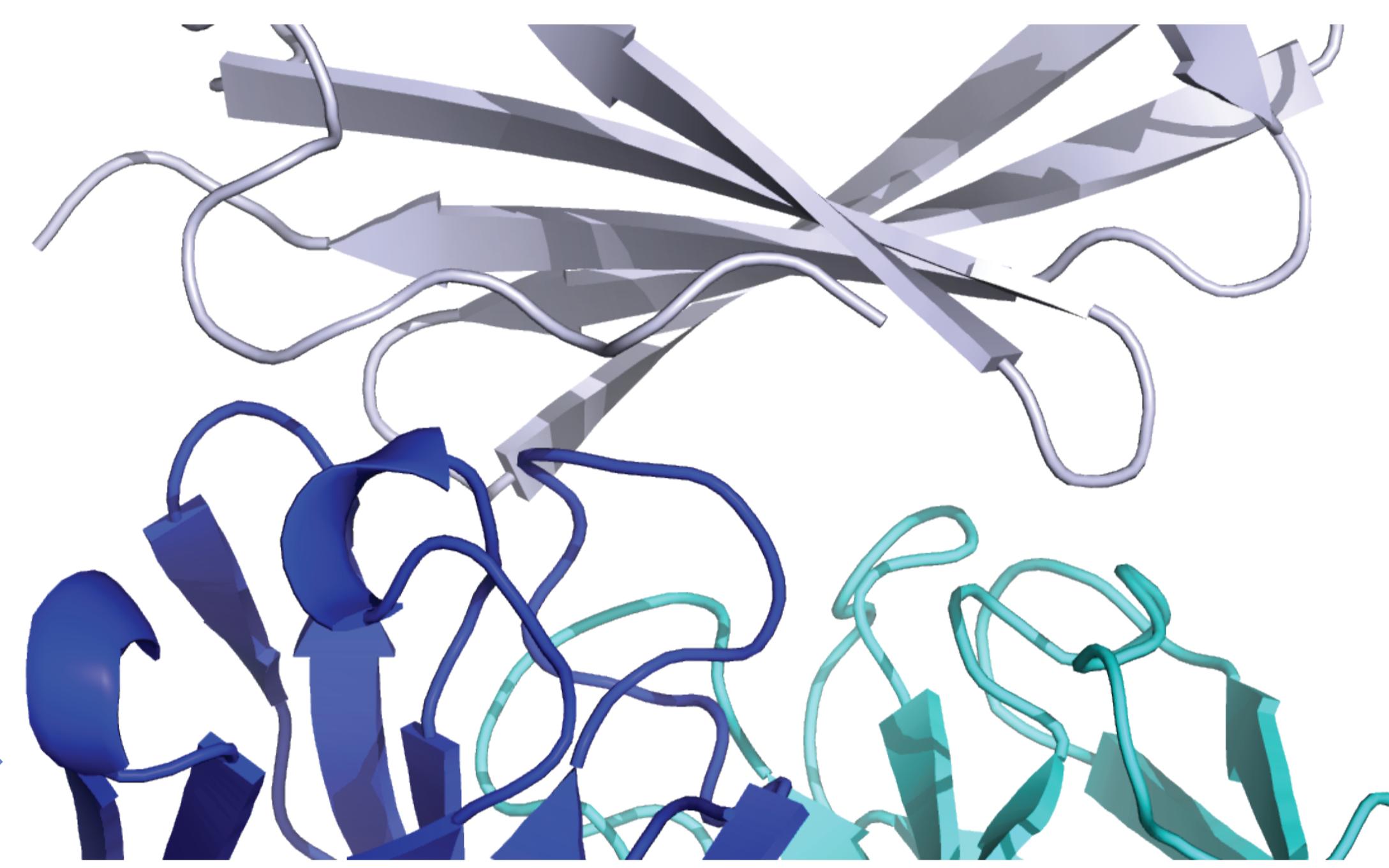
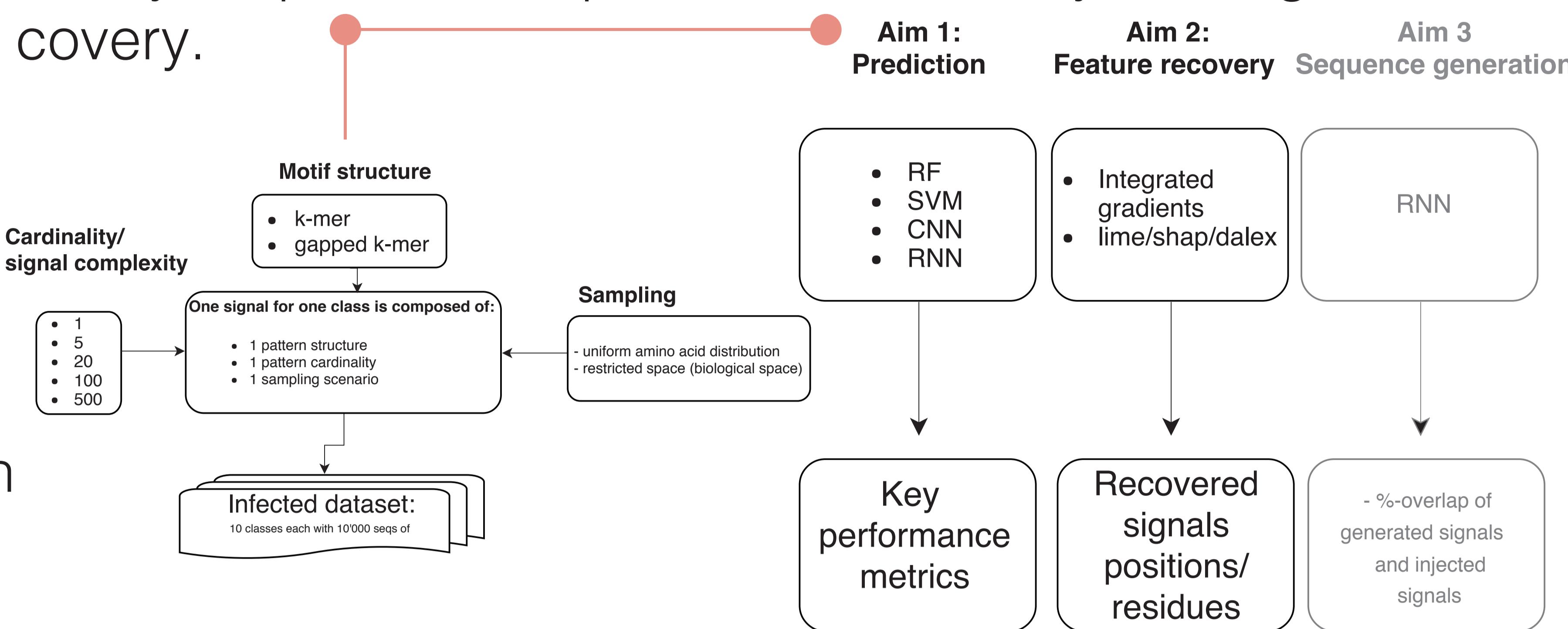
# Results

**A.** Accuracy decays as signals get more complex. **B.** Models with non-linear decision functions are superior to their linear counterparts. **C.** Longer signals ( $k$ -mers) yield higher accuracy. **D.** Signals were successfully recovered using local linear approximations. **E.** Signal recovery decays as a function of signal complexity.



# Approach and methods

Evaluate a set of machine learning models on synthetic antibody sequences for prediction accuracy and signal recovery.



# ISRGSSLDYWI RADGSDISVGR

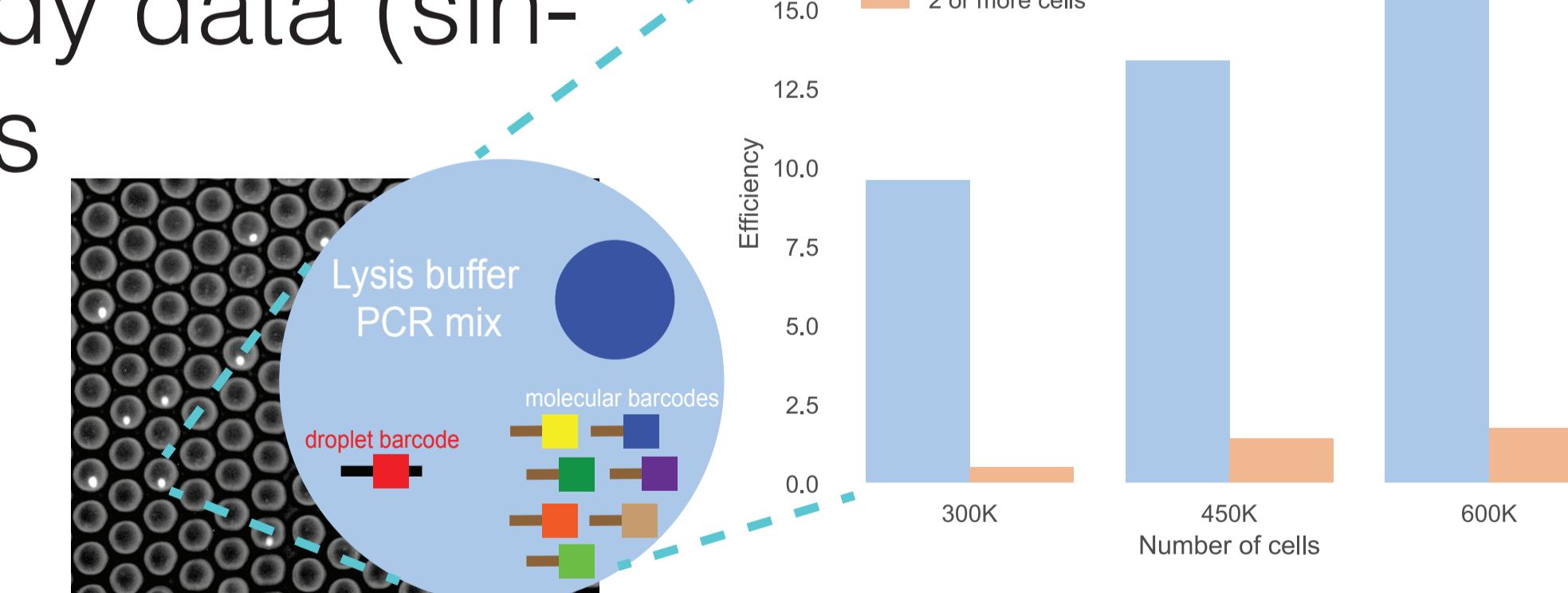
But what kind of signal? Our pre-

The figure displays three ribbon models of antibody structures, specifically focusing on the CDR-H1, CDR-L3, and CDR-H3 regions. Red circles highlight specific residues in each structure. To the right is a heatmap matrix showing interaction scores between residues across different positions. The x-axis represents the residue (A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y) and the y-axis represents the position (1-2, 28-30, 30A-32, 31-32, 31-32-33, 31-32-33-34-35, 31-32-33-35, 31B-32-33, 33-34-35, 46-49, 50-51-52-53-54-55-56-57-58, 50-52-53A-53-54-55-56-58, 50-52-53-54-55-56-58, 50-52-53-54-56, 50-53, 50-53-54-55-56, 52-53-54-56-58, 91-92-93-94-95-96, 91-92-93-94-96, 95-96-97-98-99-100, 96-97-98-99, 99-100-100A-100B-100C-100D-100E-100F). The color scale on the right indicates interaction strength from 0.0 (white) to 1.0 (black).

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
1-2	0.1	0	0.1	0.2	0	0	0	<b>0.7</b>	0	0	0	0	0	0	0	0	0	0	0	
28-30	0	0	0	0	0	0	0	0.3	0.1	0	0.1	0	0	0.1	<b>0.2</b>	<b>1</b>	0	0	0	
30A-32	0	0	0	0	0	0	0	0	<b>0.5</b>	0	0	0.1	0	0	0	0	0	0	0.1	
31-32	0.1	0	0.1	0.1	0	0.2	0	0	0.6	0	0	0.2	0	0	0	0.2	0.2	0	0	0.3
31-32-33	<b>1</b>	0	1	0.3	0.4	0.8	0.2	0.1	0.1	0.1	0	0.3	0	0	0.2	0.6	0.6	0.3	0.6	<b>1</b>
31-32-33-34-35	0.5	0	0.1	0.1	0	0.1	0.1	0.3	0	0.1	1	0.1	0	0	0.1	0.2	0	0	0	0.1
31-32-33-35	0.5	0	0.4	0.2	0.1	0.3	0.9	0	0.1	0	0.1	0.2	0	0	0.1	0.2	0.3	0	0.1	0.4
31B-32-33	0	0.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
33-34-35	0	0	0	0	0	0.1	0	0	0	0	0.7	0	0.1	0	0	0	0	0	0	0
46-49	0	0	0	0.1	0	0	0	0	0	0.5	0	0	0	0.3	0	0	0	0	0.2	
50-51-52-53-54-55-56-57-58	0.1	0	0.1	0	0	0.2	0	<b>1</b>	0.1	0.2	0	0.2	0	0	0	0.2	0.3	0	0	
50-52-52A-53-54-55-56-58	0	0	0	0.1	0	0.5	0	0	0.2	0	0	0.1	0	0.1	0.1	0.1	0	0	0.1	
50-52-53-54-55-56-58	0.9	0	0	0	0	0.4	0	0	<b>1</b>	0	0	0.1	0.2	<b>0.9</b>	<b>0.7</b>	0.1	0	0.6	0.2	0.1
50-52-53-54-55-56-58	0	0	0	0.1	0	0.5	0	0	0.1	0	0	0	0	0	0.4	0	0	0	0.4	
50-52-53-54-56	0.1	0.5	0.1	0.1	0	0.1	0	0.1	0	0	0.1	0.1	0	0	0.1	0.1	0	0	0.1	0.1
50-53	0.4	0	0	0.1	0.1	0	0.1	0.1	<b>0.9</b>	0.1	0	0.1	0	<b>1</b>	0.3	0.1	0.3	0	0	0.3
50-53-54-55-56	0.1	0	0	0.2	0	0	0	0.1	0.2	0.1	0.1	0.1	0.2	0	0.5	0.1	0.1	0	0	0.1
52-53-54-56-58	0	0	0.3	0	0.1	0	0.2	0.5	0.2	0	0	0.1	0	0	0.2	0	0	0.2	0.2	0.1
91-92-93-94-95-96	0.1	0	0	0.1	0.1	0.3	0.2	0.2	0.1	0.2	0.1	0.1	<b>1</b>	0.2	0.1	0.2	0.2	0.1	0.2	0.2
91-92-93-94-96	0.7	0	0.5	1	1	0.7	1	0.8	0.7	1	0.4	<b>1</b>	0.6	0.2	1	1	0.6	1	1	<b>1</b>
95-96-97-98-99-100	0.3	0	0.3	0.4	0.3	<b>1</b>	0	0.2	0.1	0.2	0.4	0	0.3	0.1	0.4	0.1	0.2	0.3	0.1	0.2
95-96-97-98-99-100	0.3	0	0.3	0.2	0.1	0.6	0	0.1	0.1	0.1	0.6	0	0.3	0.1	0.3	0.1	0.2	0.3	0.1	0.2
96-97-98-99	0	0	0.2	0	0.1	0.2	0	0.1	0	0	0.1	0	0	0.6	0	0	0	0	0	0.1
99-100-100A-100B-100C-100D-100E-100F	0	<b>1</b>	0	0	0	0	0	0.2	0	0	0	0	0	0	0	0	0	0.1	0	0
100-100A-100B-100C-100D-100E-100F	0.3	<b>1</b>	0.1	0	0	0	0	0	0	0	0.1	0	0	0.2	0	0	0	0.1	0.1	0.1

# Future works

Train and evaluate the models on single-cell resolution antibody data (single-cell immunogenomics and single antibody proteomics) for in-silico design of antibodies.



# Concluding remarks

- Non-linear models are superior to models with linear decision boundaries (at least for this kind of datasets).
  - Signal complexity dictates the performance of the models.
  - Signal complexity dictates recovery.
  - Interacting segments (signals) tend to be short and continuous.