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

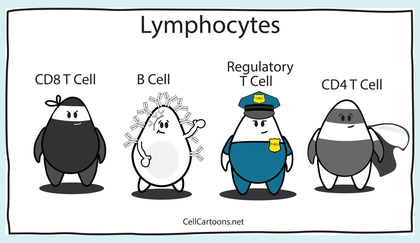
Rahmad Akbar[1], Cédric R. Weber[3], Igor Snapkov[1], Daniel Heinesen[1], Edvard Aksnes[1], Zixuan Liu[1], Milena Pavlovic[1,2], Geir Kjetil Sandve[2], Sai T. Reddy[3], and Victor Greiff[1].

UiO : University of Oslo

ETH Eidgenössische Technische Hochschule Zürich Swiss Federal Institute of Technology Zurich

[1] UiO, Department of Immunology, Computational and Systems Immunology, Oslo, Norway
[2] UiO, Institute of Informatics, Biomedical Informatics, Oslo, Norway
[3] ETH Zürich, D-BSSE, Mattenstrasse 26, 4058, Basel, Switzerland

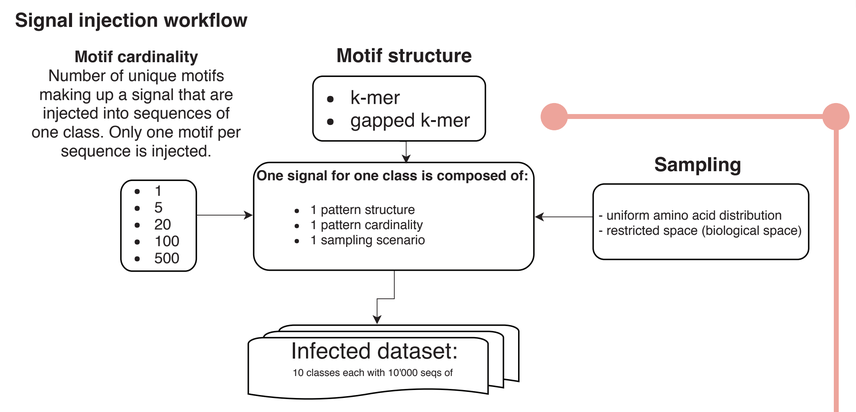
Motivation



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?

Approach and methods

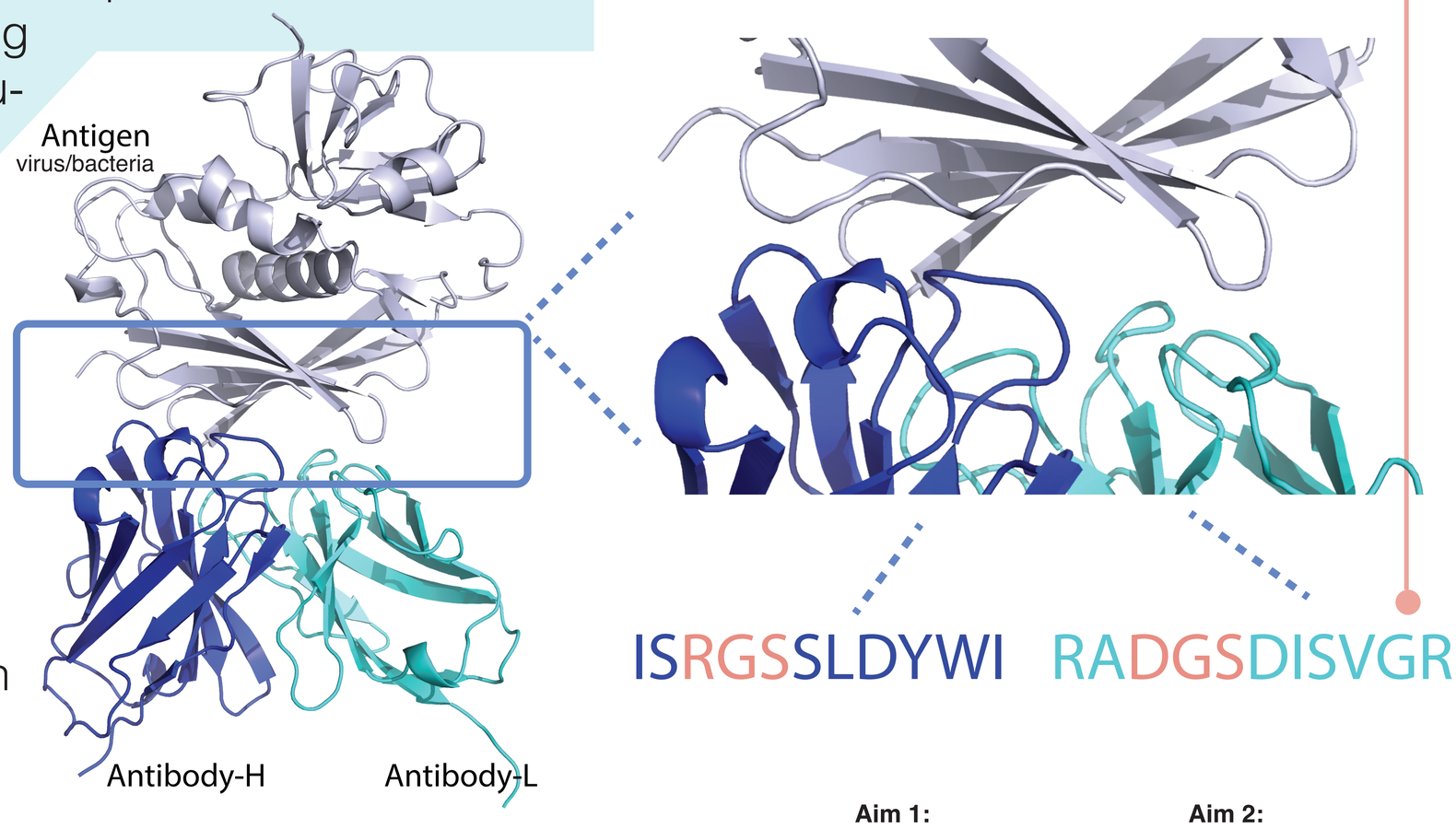
Evaluate a set of machine learning models on synthetic antibody sequences for prediction accuracy and signal recovery.



An excerpt of a dataset with k=3, cardinality=5, class=10.
For brevity, only 10 sequences per class and two classes are shown.

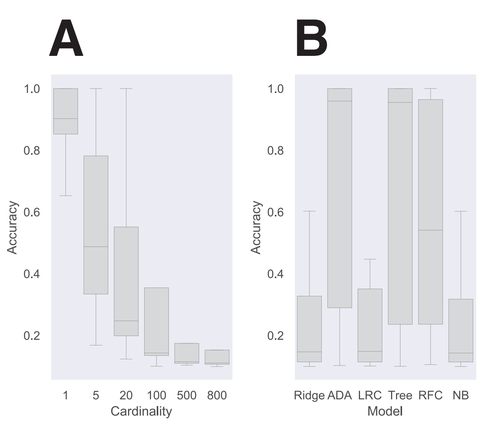
Sequence	Injected sequence	Signal	Class
CARNPSAPQLPIDYW	CARNPS**PWN**LPIDYW	PWN	0
CAKEGYSSGPFFDYW	CAKEGY**PWN**PFFDYW	PWN	0
CAGLLWFRELIADLW	CAGLLW**DEL**LIADLW	DEL	0
CARDASPRNDAFDIW	CARDAS**MAW**DAFDIW	MAW	0
CAREWSVAGTPIDYW	CAREWS**MAW**TPIDYW	MAW	0
CARRSRGVIIGSDYW	CARRSR**PWN**IGSDYW	PWN	0
CARSAPWAGYYPDYW	CARSAP**SQK**YYPDYW	SQK	0
CAKLPSSSWLHFDYW	CAKLPS**MAW**LHFDYW	MAW	0
CARTRGVPSYYFDYW	CARTRG**DEL**YYFDYW	DEL	0
CARVSVQYYYGMDVW	CARVSV**PWN**YGMDVW	PWN	0
CALPLEWTQYYMDVW	CALPLE**NTA**YYMDVW	NTA	1
CARTHVIAAGRFDPW	CARTHV**DWD**GRFDPW	DWD	1
CARGHNWMYNWFDPW	CARGHN**SHS**NWFDPW	SHS	1
CAKGVSDPFDAFDIW	CAKGVS**DWD**DAFDIW	DWD	1
CARHELDRGAMFDYW	CARHEL**SRT**AMFDYW	SRT	1
CAAQLELLWVGFDPW	CAAQLE**YGY**VGFDPW	YGY	1
CARDLKQIVGATYYW	CARDLK**DWD**GATYYW	DWD	1
CARGTAMVIYLFDYW	CARGTA**SRT**YLFDYW	SRT	1
CARHSSGDWGRIDYW	CARHSS**DWD**GRIDYW	DWD	1
CAKDWGPSGSYLTGW	CAKDWG**YGY**SYLTGW	YGY	1

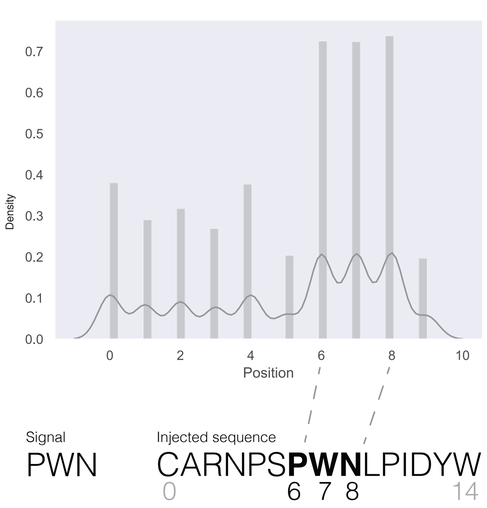
The idea is rather straightforward. (i) Generate synthetic datasets with varying signal complexity. (ii) Feed the datasets to machine learning alogrithms. (iii) Evaluate and recover key performance metrics and features.

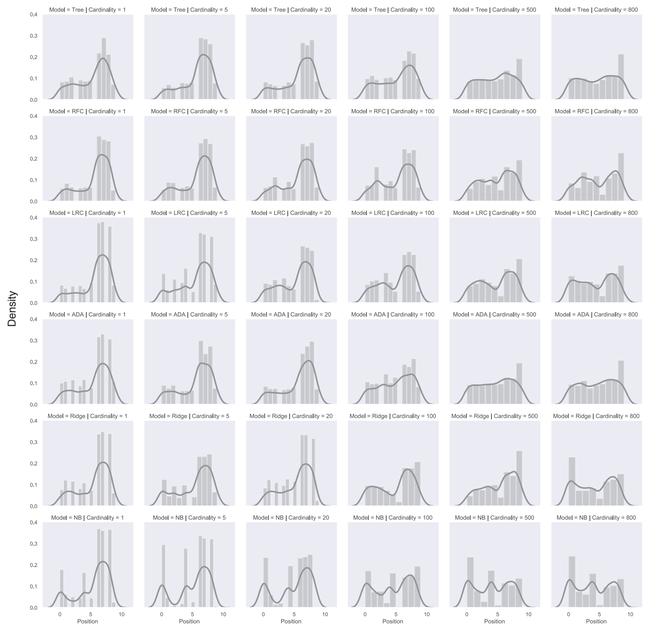


Results

A/C. Accuracy decays as signals get more complex. **B.** Models with non linear decision functions are superior to their linear counter parts. **D.** Signals were succesfully recovered using local linear approximations. **E.** Signal recovery decays as a function of complexity.







Aim 1: Prediction
- RF
- SVM
- CNN
- RNN

→ PredAccplot all different combinations

Aim 2: Feature recovery
- Integrated gradients
- lime/shap/dalex

→ - example plot showing recovered features
- %-overlap of recovered features with injected features

Aim 3 Sequence generation

RNN (check whether signal is reproduced)

→ - %-overlap of generated signals and injected signals

Concluding remarks

We learnt quite a few things:

- Non linear models are superior (at least for this kind of datasets).
- Signal complexity dictates the performance of the models.
- Signal can be recovered quite successfully.
- Signal complexity dictates recovery.

About the authors:

Rahmad (Mat) and Igor are postdocs at the Department of Immunology, UiO. Cédric is a PhD student at ETH, Basel. Milena is a PhD student at UiO. Zixuan, Daniel, and Edvard are a team of highly motivated students at UiO. Victor is a professor at UiO and the group leader. Geir is a professor at UiO. Sai is a professor at ETH, Basel. Together they **nerd** over antibody-antigen interactions and hope to one day crack the key to this [and win nobel prizes along the way 😊]

Reach out to us

GREIFFLAB Immunity | AI | Therapeutics Laboratory for Computational and Systems Immunology

victor.greiff@medisin.uio.no I rahmad.akbar@medisin.uio.no I greifflab.org