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

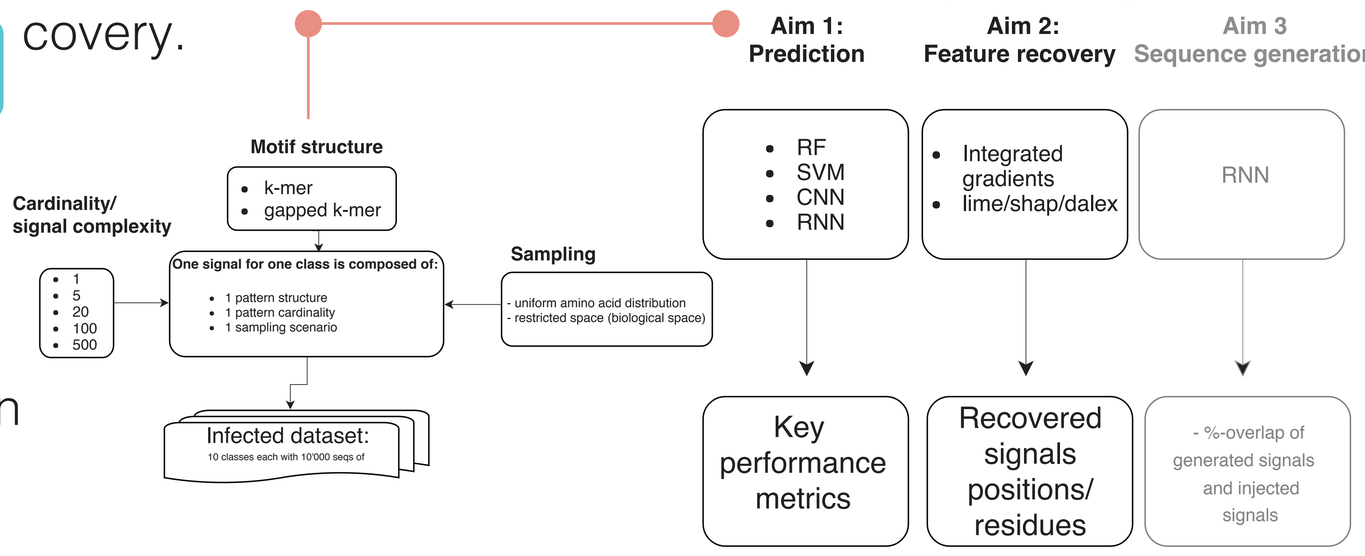
[1] UiO, Department of Immunology, Computational and Systems Immunology, Oslo, Norway
[2] UiO, Institute of Informatics, Biomedical Informatics, Oslo, Norway
[3] ETH Zürich, D-BSSE, Mattenstrasse 26, 4058, Basel, Switzerland

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

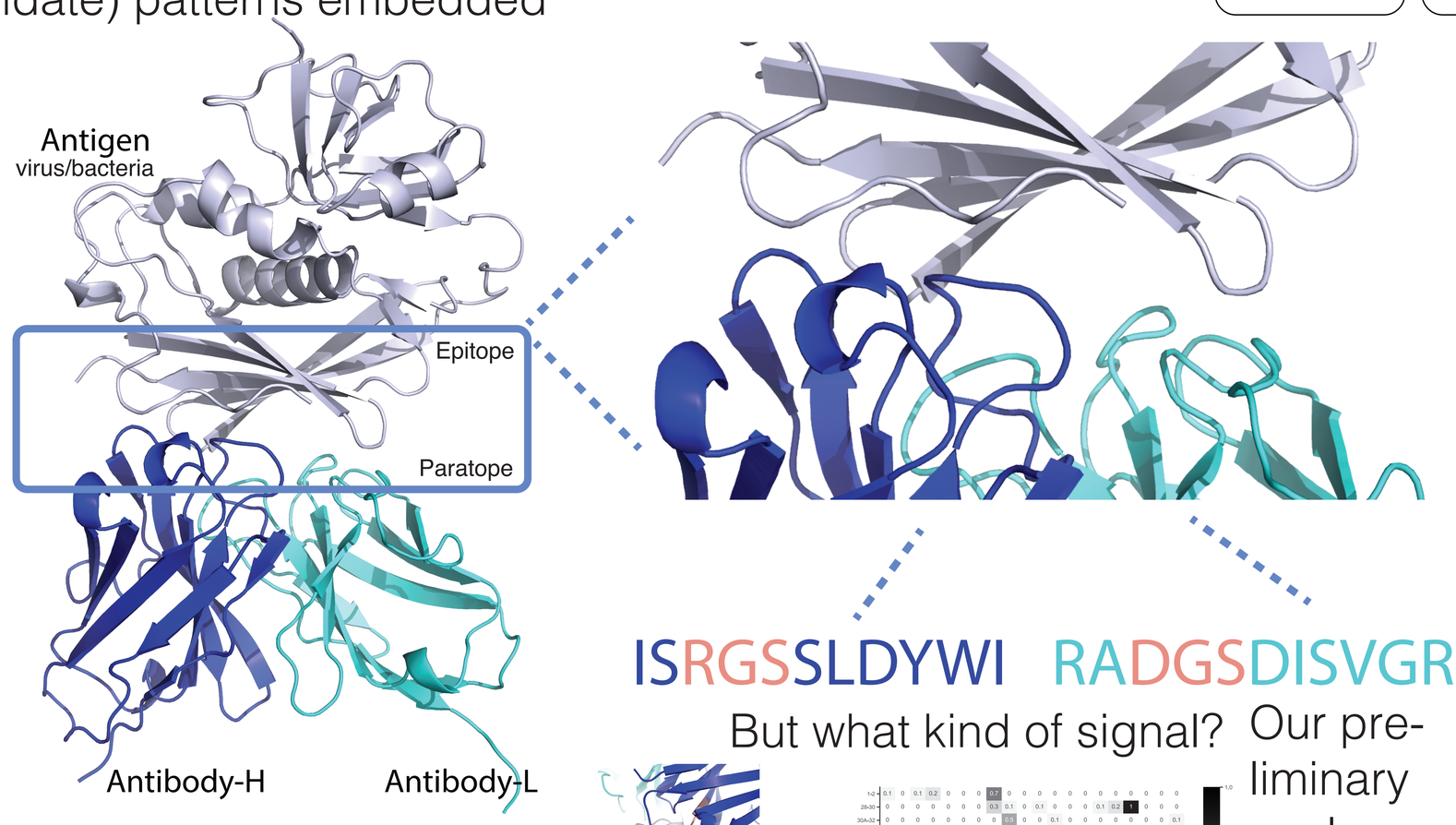
Evaluate a set of machine learning models on synthetic antibody sequences for prediction accuracy and signal recovery.



(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.

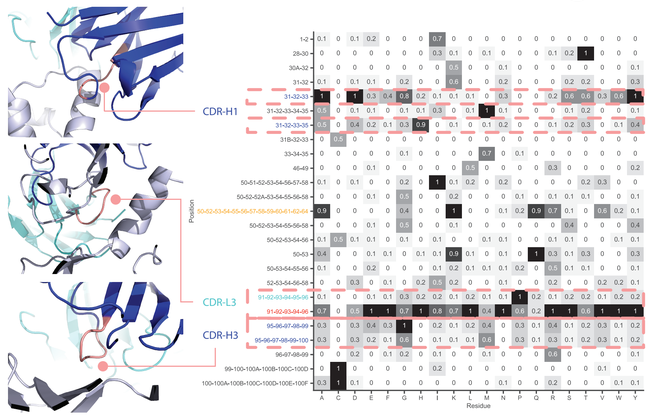
Results

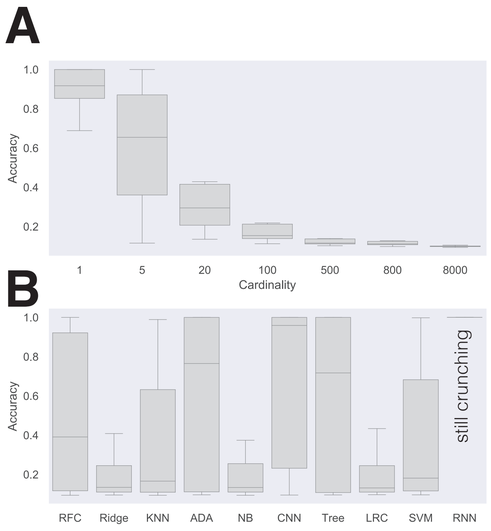
A. Accuracy decays as signals get more complex. **B.** Models with non-linear decision functions are superior to their linear counter parts. **C.** Longer signals (k-mers) yield higher accuracy. **D.** Signals were successfully recovered using local linear approximations. **E.** Signal recovery decays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

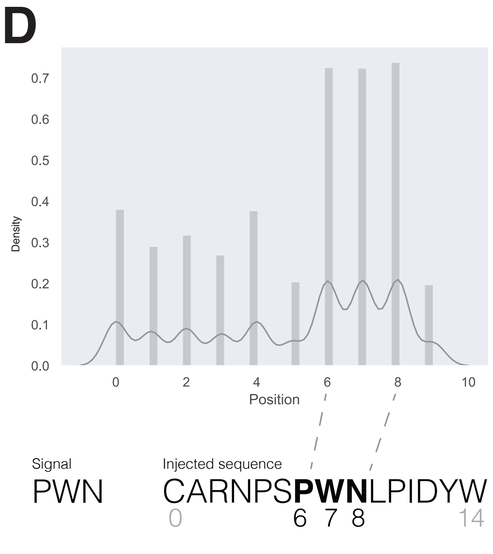


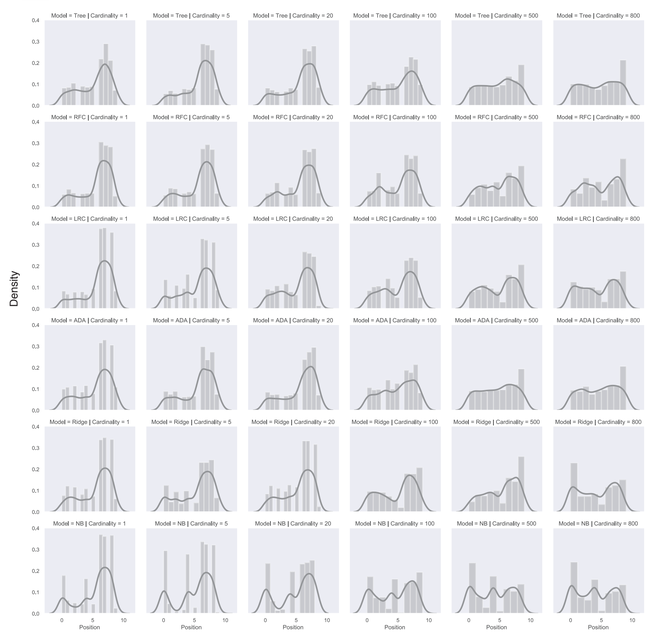
ISRGSSLDYWI RADGSDISVGR

But what kind of signal? Our preliminary analyses suggest that short and continuous (linear) residue-segments mediate the majority of antibody-antigen interactions.



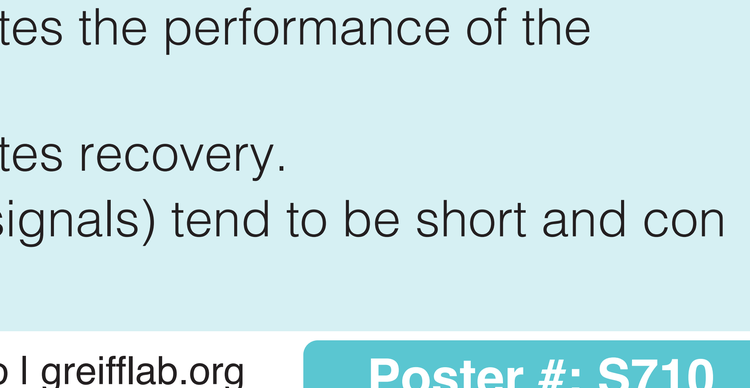






Future works

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 con tinuous.

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

victor.greiff@medisin.uio.no | rahmad.akbar@medisin.uio.no | igor.snapkov@medisin.uio.no | greifflab.org

Poster #: S710