INTERPRETABLE MODELS EXPLAIN CURRENT PARATOPE PREDICTION LIMITATIONS

GITHUB
bit.ly/topefind-code

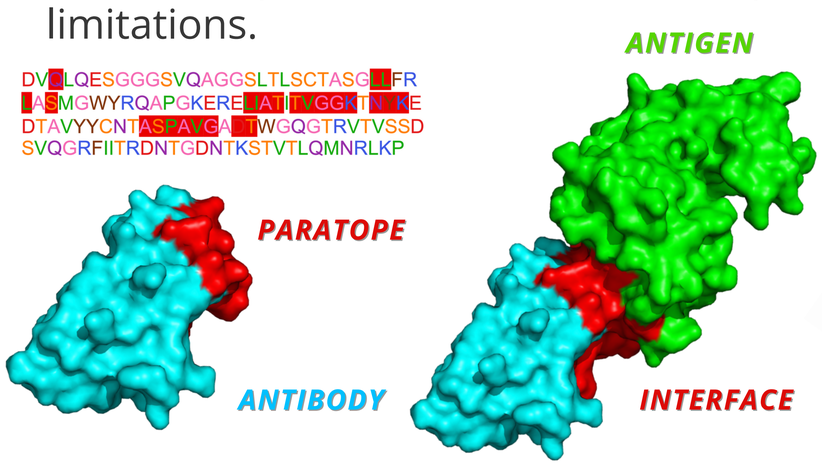
DASHBOARD
bit.ly/topefind

Serban C. Tudosie*, Pedro B.P.S. Reis, Adrien Bitton**, Sven H. Giese**, and Santiago D. Villalba****

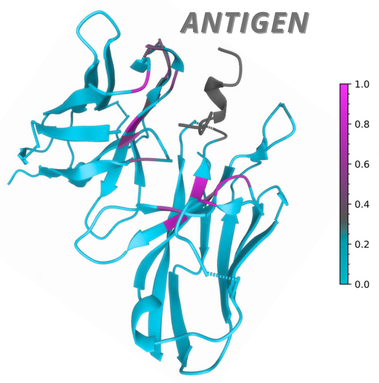
***sc.tudosie@gmail.com ****firstname.surname@bayer.com**

MOTIVATION

- **Paratope prediction** can help engineer more affine and specific therapeutic antibodies.
- Currently, all paratope prediction models face a **performance ceiling**.
- **Dissecting** what models learn and **comparing** deep models with interpretable ones can help shed light on such limitations.



Models errors and differences
(from our visualization tool)



CONTRIBUTIONS

- We **achieve SOTA performance** with either transfer learning from general protein language models (LMs) or models with a combination of interpretable features.
- We **investigate feature importance** with an ablation study over interpretable antibody properties.
- We **provide a tool** to visualize predictions and compare models.

METHODS

DATASET

We use **Expanded Paragraph** [1], a subset of SAbDab [2]. We keep the same train/validation/test split but **filter for data leakage** across models. From the filtered dataset, we derive paratope labels as antibody amino acids that have a heavy atom at a distance less than 4.5Å from any heavy atom of the antigen (or antigens if many bind to the same antibody chain).

MODELS

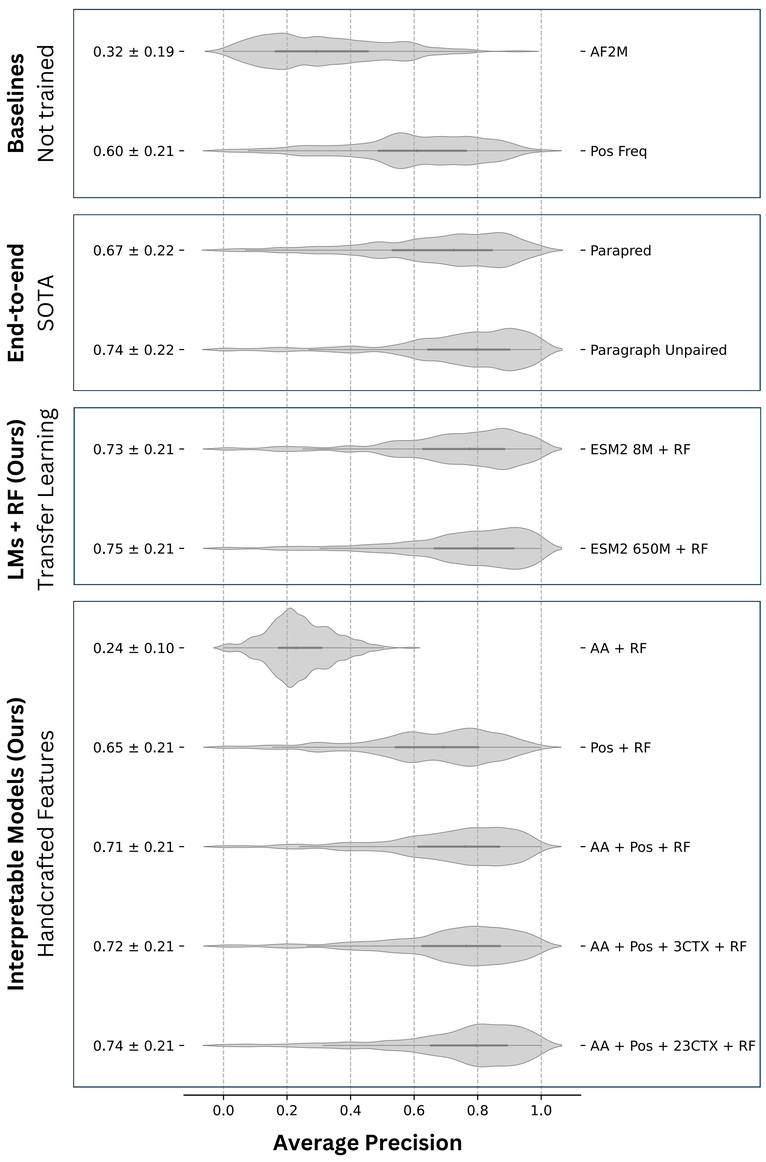
- **Baselines**: 1) Paratope derived from complex prediction with AlphaFold2 Multimer (AF2M); 2) paratope derived from IMGT position prior probability of being paratope (Pos Freq).
- **End-to-end:** Parapred [3] and Paragraph [1] (literature's SOTA).
- **Interpretable models:** Random Forests (RFs) with features such as amino acid physicochemical properties (AA), IMGT numbering (Pos), and context (CTX). Context is the concatenation of features from N-1 nearby residues' features. We also consider a combination of features through concatenation (e.g., AA + Pos).
- **LMs embeddings + RF head:** Generic transfer learning from ESM [4]*. We choose RF based on a model selection procedure.

EVALUATION

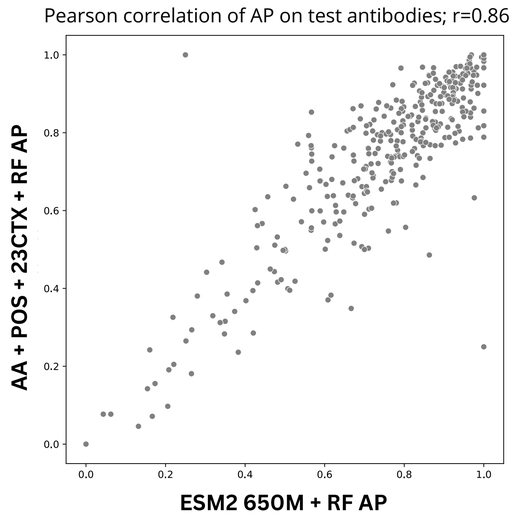
- We use **average precision** (AP) to account for class imbalance.
- We **evaluate model performance for each antibody individually** and report the mean and standard deviation across antibodies. This is in contrast to pooling all residue predictions and computing a single aggregated value, which in this setting is incorrect for ranking metrics like AP.

RESULTS

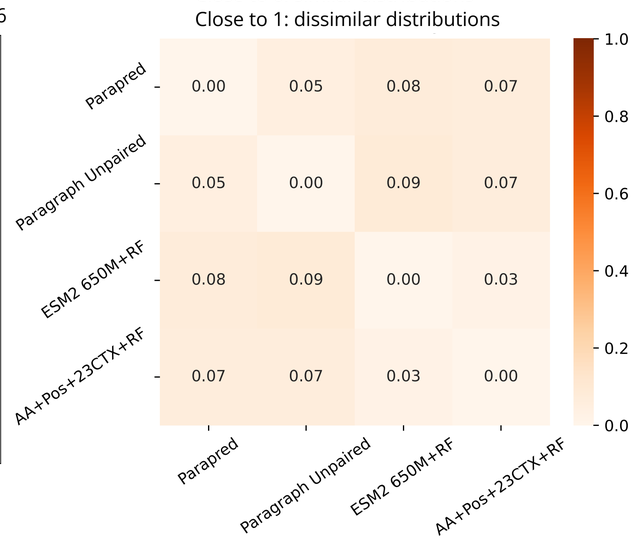
1. Distribution of AP on Expanded Paragraph test set across models



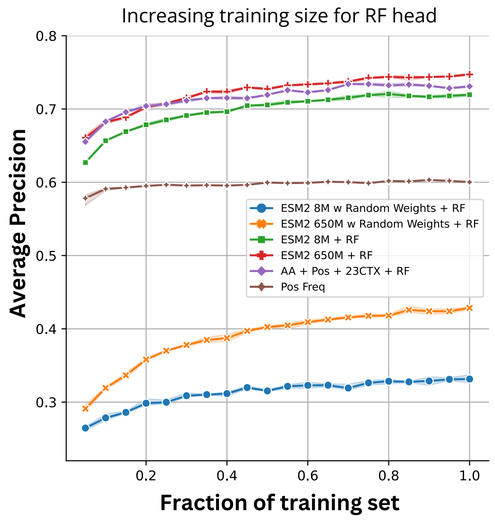
2. Similar performance per antibody between interpretable and deep models



3. Models fail equivalently



4. Similar performance increase with model and dataset size



1&4. Paratope prediction is insensitive to model size. Increasing the size of LMs such as ESM2 does not significantly improve paratope prediction. The insensitivity to capacity implies that models could learn a bias towards underlying data characteristics, such as the abundance of paratopes in certain positions, while ignoring the least common interactions.

1&2. Interpretable models uncover biases and mimic deep counterparts. Amino acid properties, numbering, and context are sufficient to reach the performance (in terms of AP) of more complex counterparts. An ablation study with interpretable models explains how numbering plays a crucial role.

3. Models fail equivalently. Divergence between residue-level distributions of errors is practically zero, highlighting the presence of consistent biases across models.

4. Performance increases similarly with training size. While Pos Freq highlights the baseline threshold performance, the remaining trained models show a similar learning curve, slowly plateauing, with minor sensitivity to model and dataset size. This indicates that performance-explaining biases can be already learned with small samples.

CONCLUSIONS & DISCUSSIONS

- Comparing deep models with interpretable counterparts allows the **dissection of biases and difficulties** in learning to predict paratopes.
- **All current models fail similarly**, given the low value of divergence.
- We propose comparing **residue-level errors** for further model interpretation and **open-source a comprehensive visualisation tool**.
- For future research, we suggest using learning-based methods that **consider biases directly** (e.g., **hard example mining**).
- We also suggest **including the antigen** to account for polyspecificity since antigen-unaware paratope prediction might be generally ill-defined.

ACKNOWLEDGEMENTS

Bayer provided financial support, infrastructure, and the environment for the development of this work. SCT and RossettaCon also partially financed poster finalisation and conference attendance. Open tools and datasets, such as SAbDab, ESM, and Panel, were essential in achieving these results. Ultimately, a deep collaboration between passionate people from different backgrounds made this work possible.

REFERENCES

1. [Chinery et al.(2023)] L Chinery et al. Paragraph—antibody paratope prediction using graph neural networks with minimal feature vectors. Bioinformatics. 39(1):btac732, January 2023. doi: 10.1093/bioinformatics/btac732.
2. [Dunbar et al.(2014)] J Dunbar et al. SAbDab: the structural antibody database. Nucl. Acids Res. 42(D1):D1140–D1146, January 2014. doi: 10.1093/nar/gkt1043.
3. [Liberis et al.(2018)] E Liberis et al. Parapred: antibody paratope prediction using convolutional and recurrent neural networks. Bioinformatics. 34(17):2944–2950, September 2018. doi: 10.1093/bioinformatics/bty305.
4. [Lin et al.(2023)] Z Lin et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science. 379(6637):1123–1130, March 2023. doi: 10.1126/science.ade2574.

* More Language Models are considered in the upcoming full-length article