De-novo predictions of binding positions within protein domains

Anat Etzion-Fuchs, **David A. Todd** and **Mona Singh**
{anatf, dtodd, mona} @ Princeton.edu
PRINCETON UNIVERSITY LEWIS-SIGELR INSTITUTE FOR INTEGRATIVE GENOMICS

1 Abstract

We present **dSPRINT: d**omain **S**equence-based **PR**ediction of **INT**eraction sites, an ensemble of machine learning classifiers using a novel stacking architecture, that predict binding positions within protein domains.

2 Background

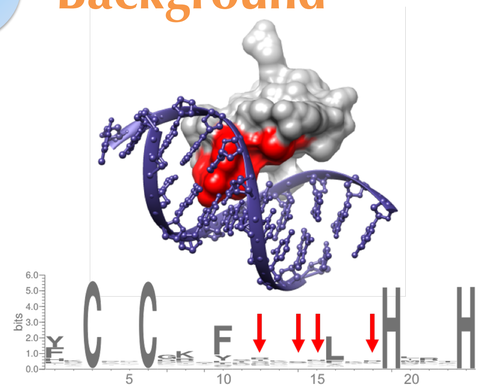


Fig 1: SDPs exemplified in the Cys_2-His_2 Zinc finger domain. Known[1] binding positions are colored in red on the domain surface[2], and pointed by red arrows on the domain sequence logo[3]. They are critical for the DNA-binding specificity, and are not conserved.

3 Methods: pipeline overview

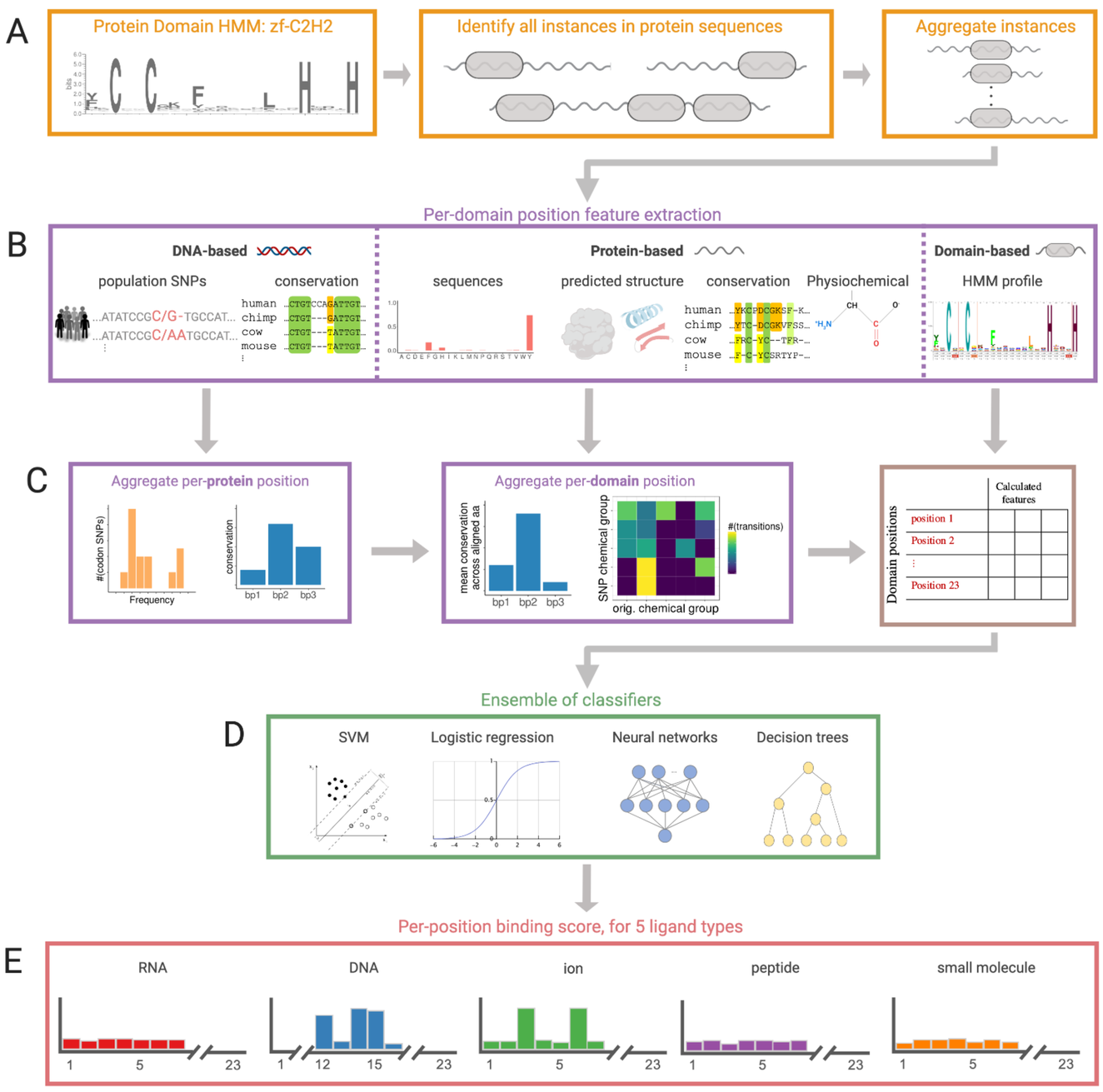


Fig 2: dSPRINT workflow for domain-centered per-position prediction.

4 Methods: ML stacked architecture

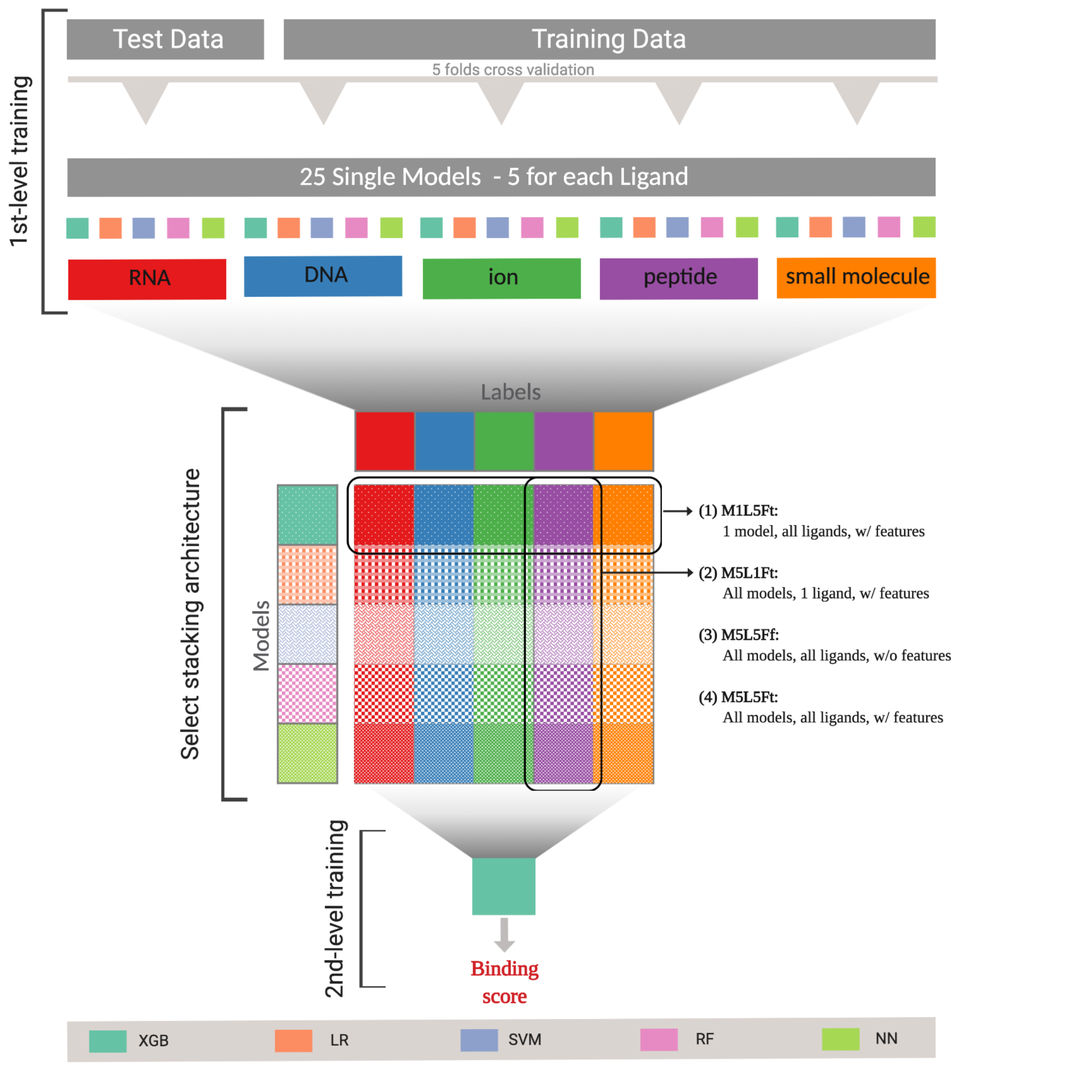
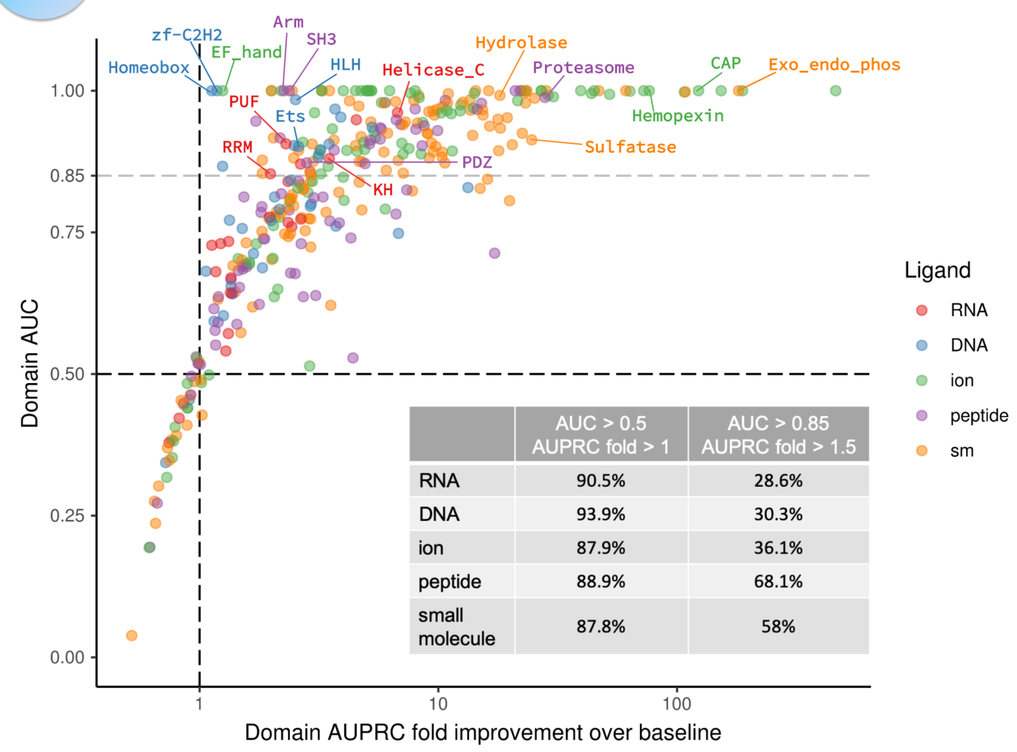


Fig 3: Ligand-combined classifier stacking architecture. Five base-models are trained for each ligand in 5-fold cross validation. This results in 25 base-models that are used in combinations as illustrated in the colorful grid. The chosen stacking architecture is used as input to a meta-classifier in the 2nd stacking level.

5 Results: per-domain evaluation



	AUC > 0.5 AUPRC fold > 1	AUC > 0.85 AUPRC fold > 1.5
RNA	90.5%	28.6%
DNA	93.9%	30.3%
ion	87.9%	36.1%
peptide	88.9%	68.1%
small molecule	87.8%	58%

Fig 5: Performance evaluation on ligand-binding domains. The table represents domain-ligand pairs with performance exceeding that of the random baselines (the dashed lines) of AUC=0.5 and AUPRC fold ratio=1.

6 Results: global evaluation

	Number of most reliable predictions chosen per domain	Number of domains with at least one Correct prediction	Fraction of domains With correct predictions
RNA	1	13	62%
	3	16	76%
	5	18	86%
DNA	1	22	67%
	3	27	82%
	5	29	88%
ion	1	47	52%
	3	60	66%
	5	64	70%
peptide	1	29	40%
	3	44	61%
	5	48	67%
small molecule	1	72	55%
	3	96	73%
	5	104	79%

Table 1: The most reliable prediction(s) in each binding domain.

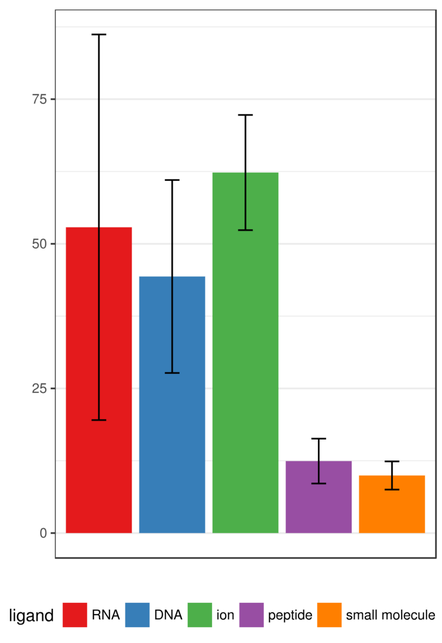


Fig 6: Ligand AUPRC fold improvement. The ratio of the AUPRC to a baseline corresponding to the fraction of binding positions of that ligand at that CV fold.

7 Significance

Systematic identification of ligand-binding residues would have a farther-reaching applications:

- Identify the functional impact of coding variants
- Explore the variation in protein interaction network
- Characterize mutations' effect in the context of disease
- Suggest molecular targets for therapeutic intervention

References

[1] Wolfe et al. Ann. Rev. Biochem. (2001)
[2] Pettersen et al. J. Comput. Chem. (2004)
[3] Crooks et al. Genome Res. (2004)
[4] Kobren and Singh, NAR (2018)

Acknowledgements

We would like to thank Shilpa N. Kobren, and the Singh lab for their helpful insights. Funded by NIH GM076275 (MS) , T32 HG003284 (AE), and CA208148 (MS).