Debiasing Algorithms and Generalization in Protein-Ligand Binding

Vikram Sundar and Lucy Colwell
Department of Chemistry, University of Cambridge

Introduction

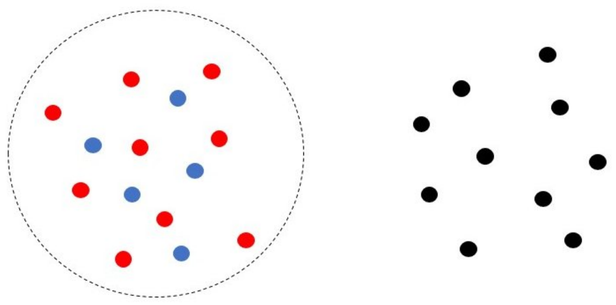
- Many machine learning (ML) approaches to protein/ligand binding have achieved outstanding success on benchmark datasets. [1]
- Unclear whether performance indicates generalizability or overfitting to training data. [2, 3]
- Chemical space is naturally clustered and our datasets are further clustered due to biases from experimentalists. [2, 3]

Dataset

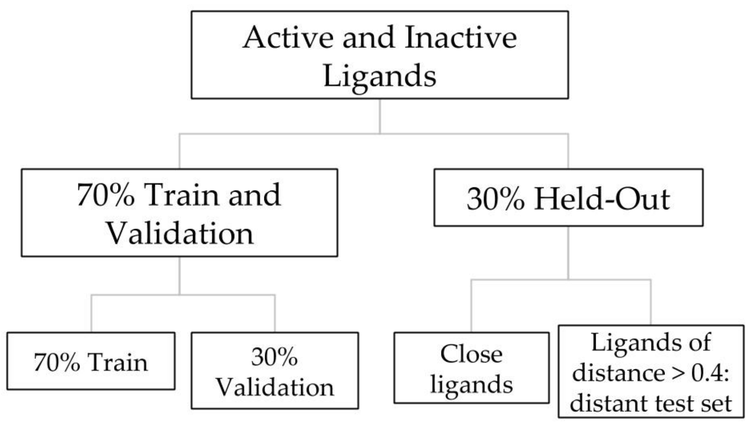
- Active ligands acquired from ChEMBL. Activity threshold of 1 µM.
- Inactive ligands acquired from PubChem. Further inactives randomly drawn until an even split of actives and inactives was achieved.
- ECFP6 fingerprints with 2048 bits used as feature set. Tanimoto similarity as distance metric.
- Tested 189 targets with at least 500 active ligands.

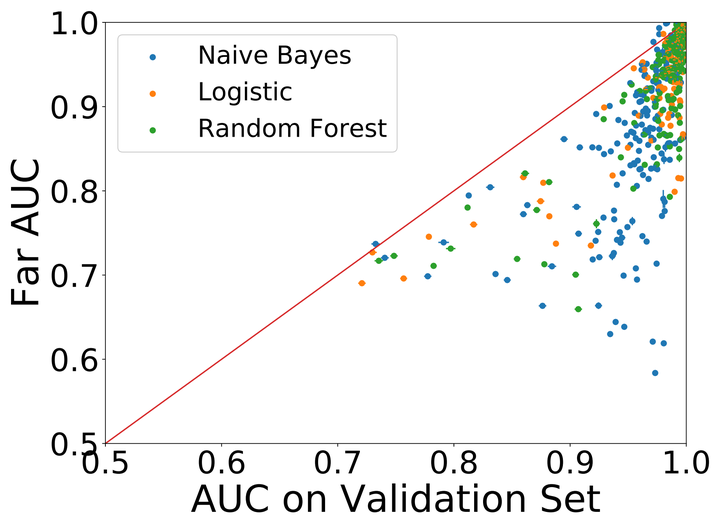
Far AUC

- To measure generalizability, hold out a distant test set to measure algorithm performance.



Training/Validation Sets Held-Out Test Set

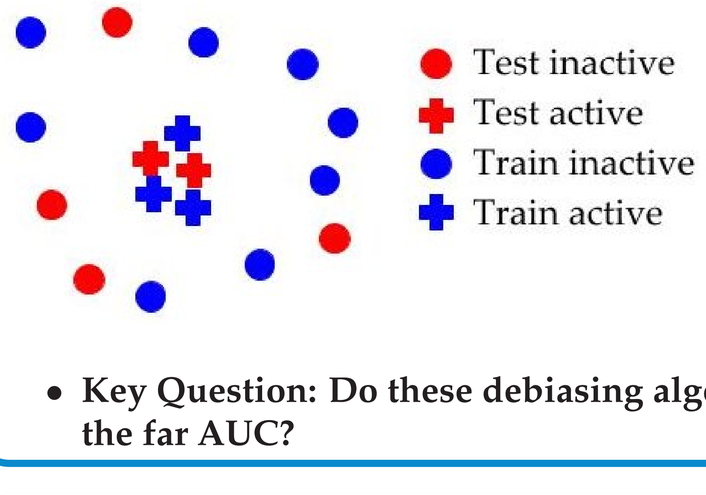




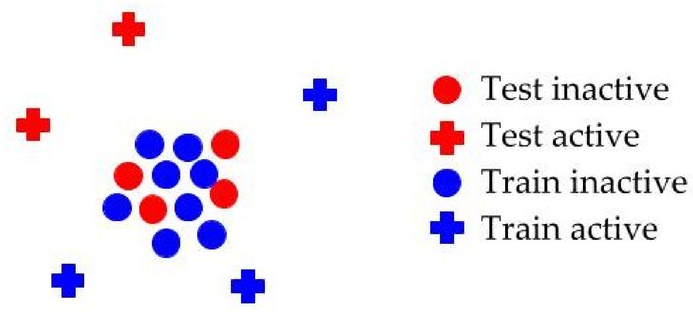
- All models show poor generalization.
- Validation set AUC may not be representative of true performance of model.

Debiasing Algorithms

- Goal of data bias and debiasing algorithm is to prevent models from memorizing and improve their generalization ability.
- Maximum Unbiased Validation (MUV) requires active ligands be uniformly embedded within inactive ligands. [2] This set is MUV-biased:

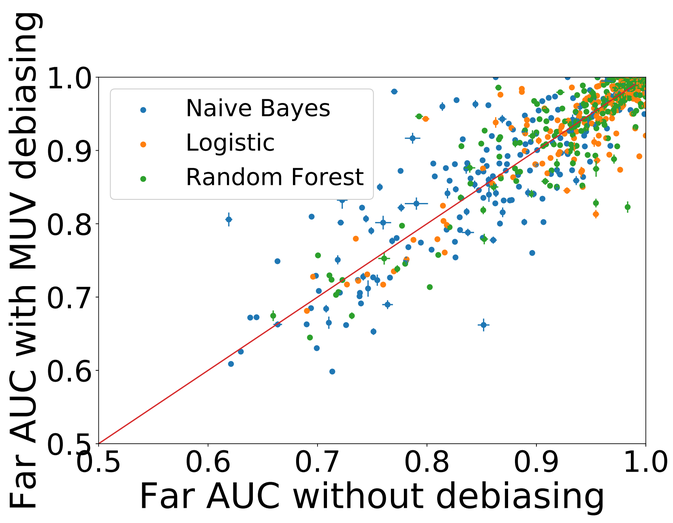


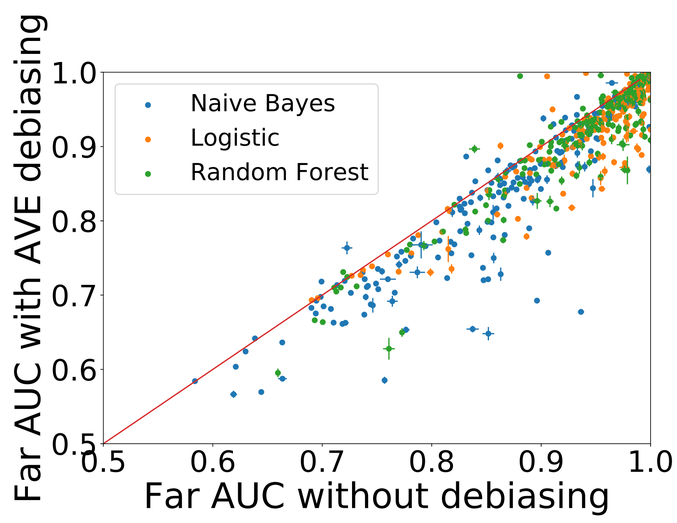
- Asymmetric Validation Embedding (AVE) prevents inactives from being clustered. [3] This set is AVE-biased:



- We use genetic algorithms to debias a given train/validation split. [2, 3]
- **Key Question: Do these debiasing algorithms improve generalization ability as measured by the far AUC?**

Results

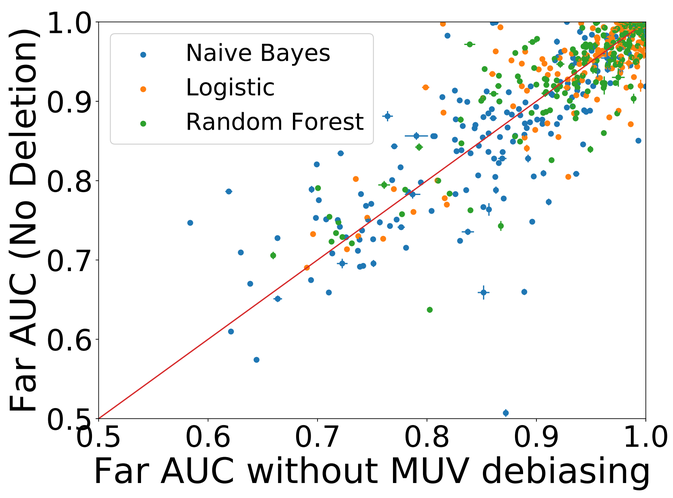




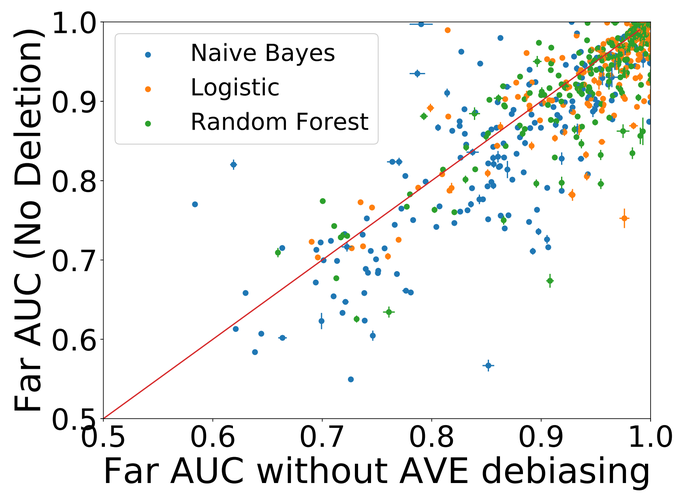
- Debiasing algorithms do not consistently improve, and often worsen, the ability of models to generalize.
- Independent of number of active ligands, number of inactive ligands (or random decoys added), and degree of debiasing.

Avoiding Data Deletion

- Debiasing algorithms allow for deletion of data from both training and validation set.



- Still possible to partially debias datasets without deleting data.



- Slight improvement over the results while allowing deletion.
- Still no consistent improvement over performance without debiasing.

Conclusions

- Debiasing algorithms cannot distinguish between signal and bias and remove relevant information from the training data.
- Some clustering among actives expected in protein/ligand binding. Need to distinguish between this and artificial clustering.
- Generalization in ML is hard. Better approach: understand domains where our models are applicable.

Acknowledgments

I acknowledge the support of the Winston Churchill Foundation of the USA.

References

[1] Lucy J. Colwell. Statistical and machine learning approaches to predicting proteinâĂŞligand interactions. *Current Opinion in Structural Biology*, 49:123–128, 2018.

[2] Sebastian G. Rohrer and Knut Baumann. Maximum unbiased validation (MUV) data sets for virtual screening based on PubChem bioactivity data. *Journal of Chemical Information and Modeling*, 2009.

[3] Izhar Wallach and Abraham Heifets. Most Ligand-Based Classification Benchmarks Reward Memorization Rather than Generalization. *Journal of Chemical Information and Modeling*, 2018.