Performing Fast and Accurate virtual High-Throughput Screening

Chloé-Agathe Azencott, S. Joshua Swamidass, and Pierre Baldi

UCIrvine

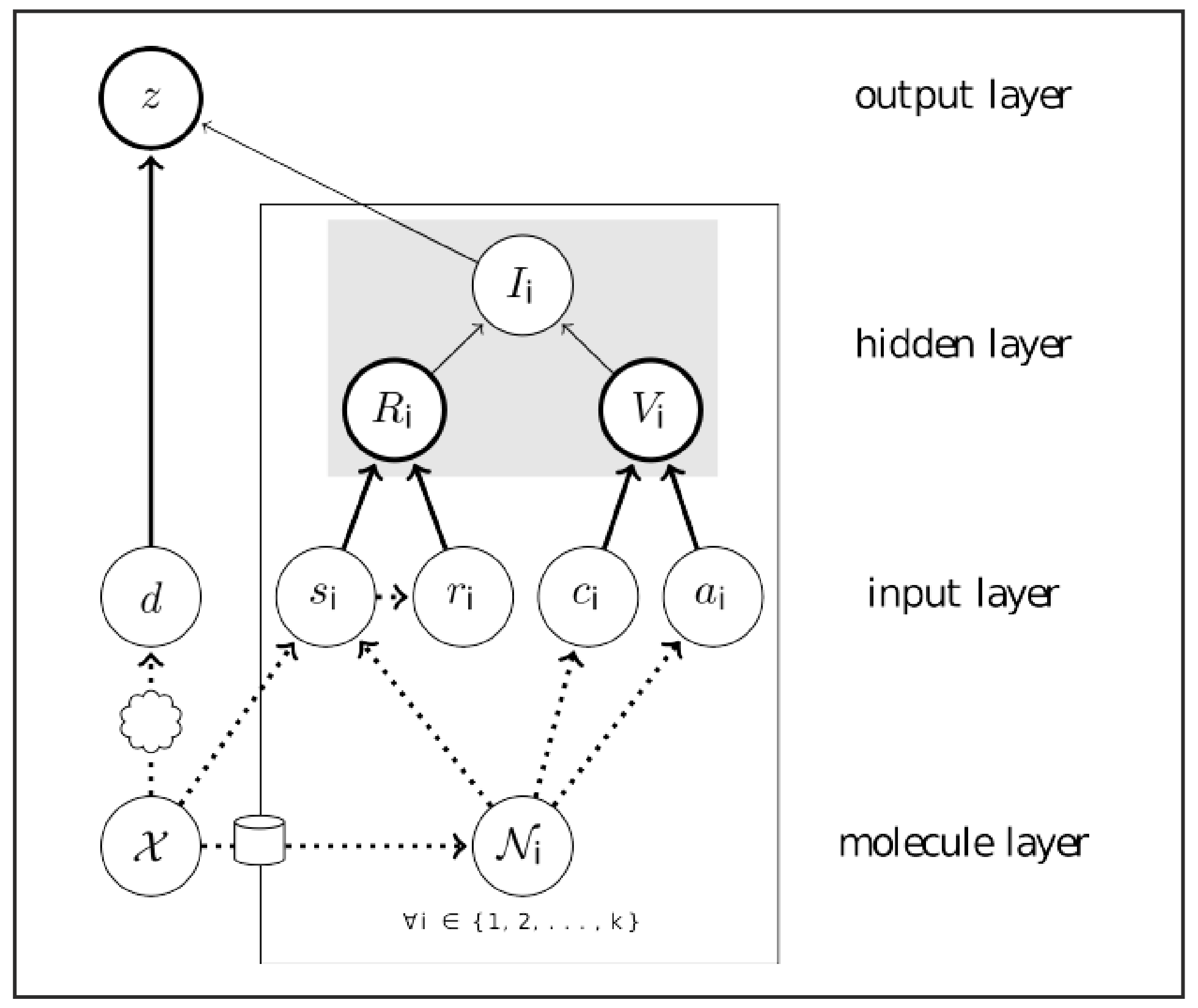
Institute for Genomics and Bioinformatics
Bren School of Information and Computer Sciences

Virtual High-Throughput Screening

Virtual High-Throughput Screening (vHTS) is the cost-effective, in silico complement of experimental High-Throughput Screening (HTS). A vHTS algorithm uses data from HTS experiments to predict the activity of new sets of compounds in silico.
vHTS is most appropriately described as a ranking task, where the goal is to rank compounds such that active ones are close to the top of the prediction-sorted list as possible.
Moreover, being able to assess the performance and evaluate how many hits are retrieved in a fraction of the prediction-sorted list is a major asset for a vHTS algorithm.

Influence Relevance Voter (IRV)

The k-Nearest Neighbors algorithm can be applied to chemical data, but does not perform optimally. The IRV uses a neural network architecture to learn how to best combine information from the nearest structural neighbors contained in the training set.
We compute nearest neighbors of chemicals using a standard MinMax similarity on structural fingerprints.



Benchmarked Performance

IJCNN07 Challenge HIV data: train on 4,229 compounds (149 actives), test on 38,449 compounds (1,354 actives).

McMaster 2005 DHFR data: train on 49,995 compounds (66 actives), test on 50,000 compounds (94 actives).

	BER	AUC
IJCNN07	0.283	0.771
SVM	0.269	0.764
IRV	0.271	0.762

HIV data (IJCNN07 Challenge)

	EF1%	EF5%
McMaster	0.02	0.14
SVM	0.01	0.04
IRV	0.03	0.14

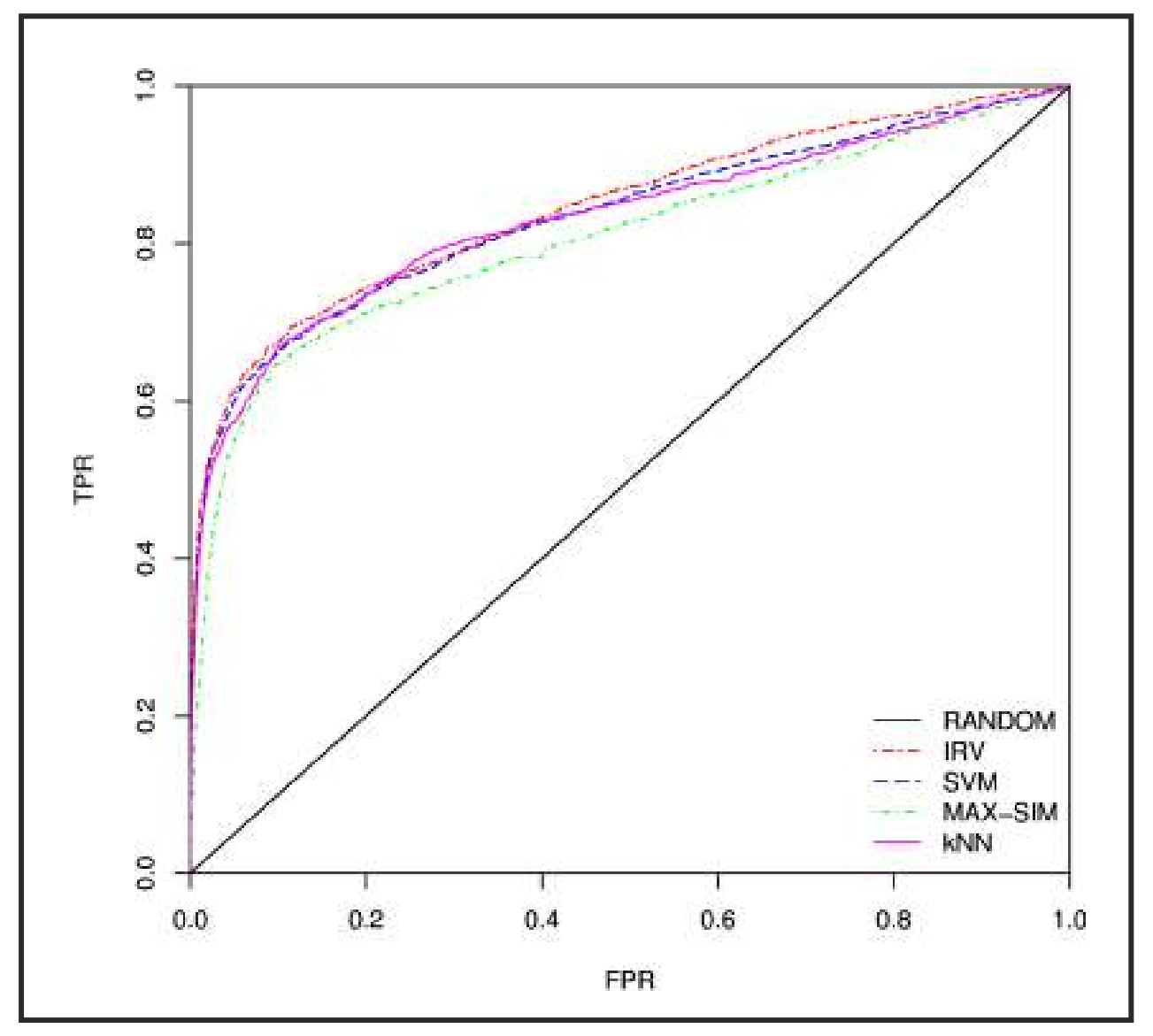
DHFR data (McMaster Challenge)

Early Recognition

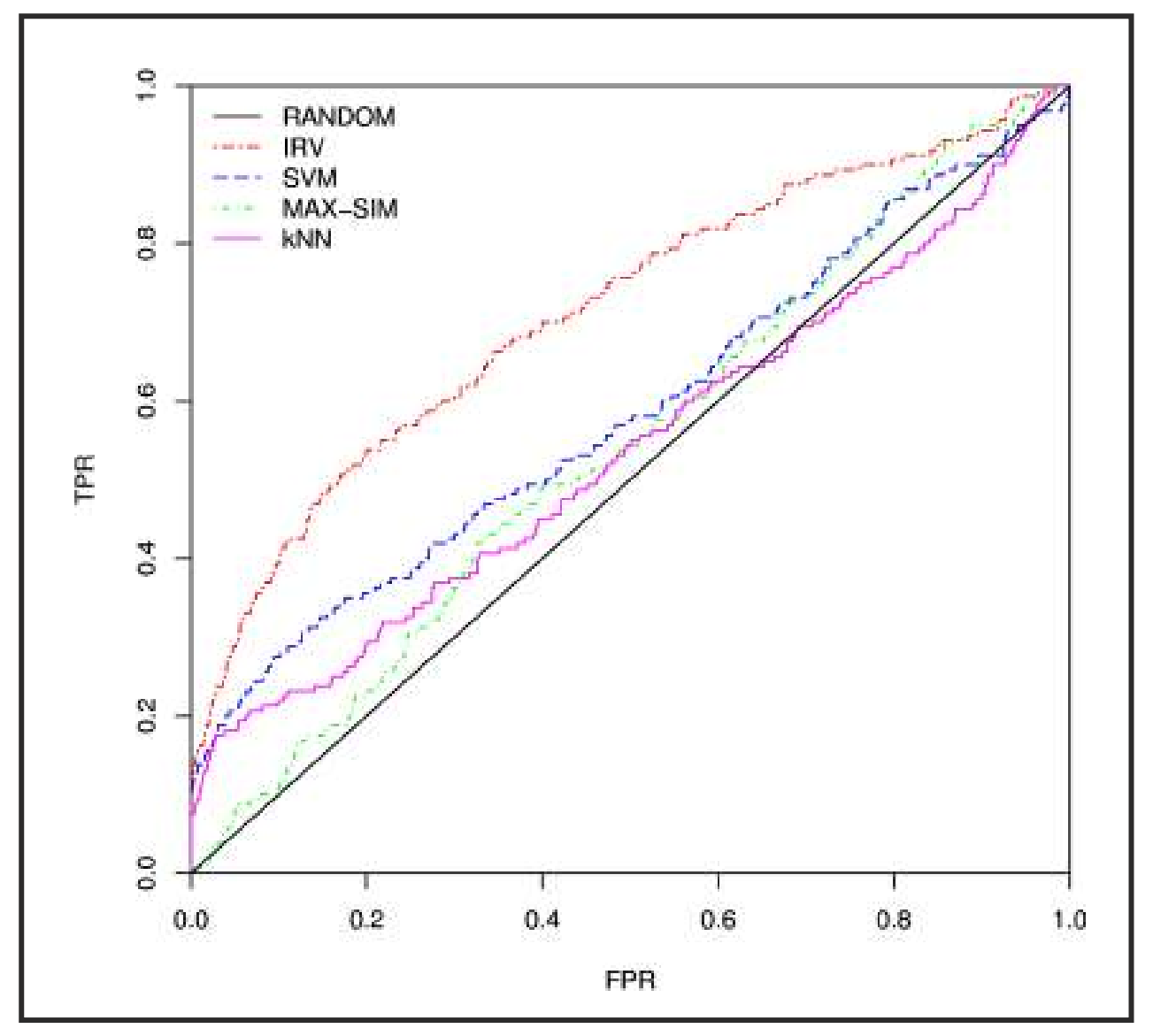
To measure vHTS performance, we need to quantify the ability of a method to rank active compounds early at the top of the prediction-sorted test data.

CROC Curves

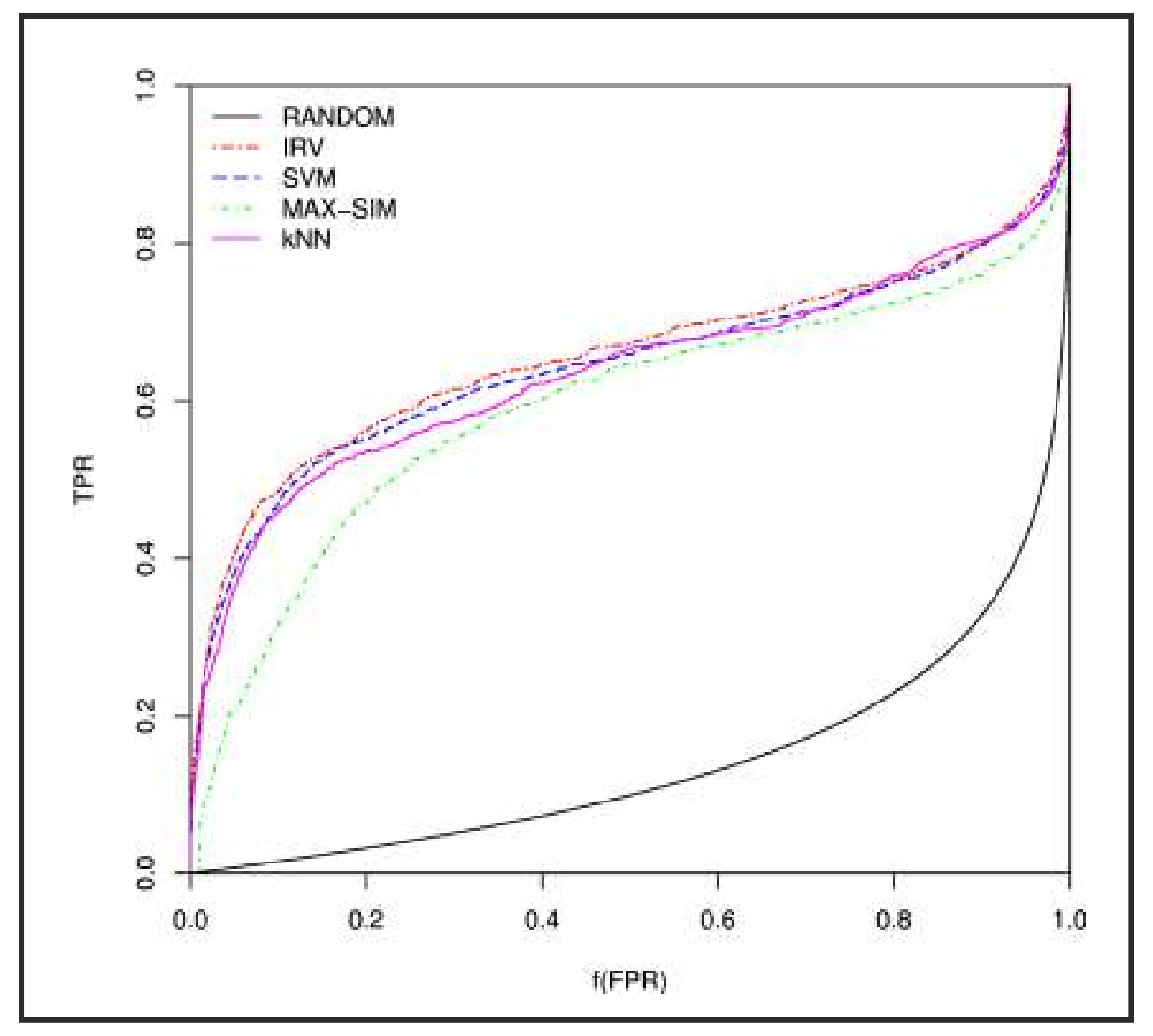
To better assess the results of vHTS experiments, we propose to replace traditional ROC curves with CROC curves, where an exponential transform of parameter α has been applied to emphasize the importance of the early portion of the curve.



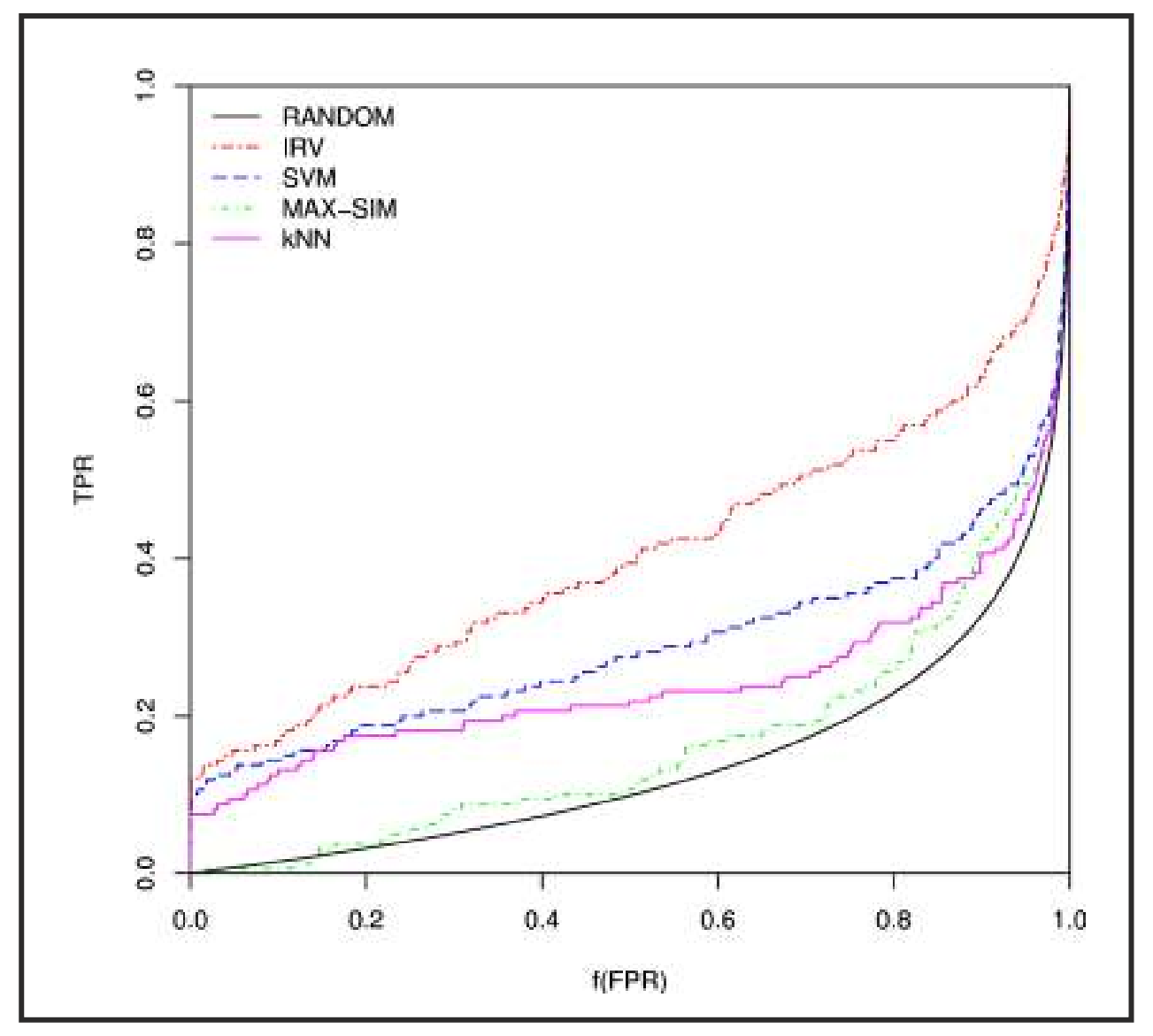
ROC on the HIV data



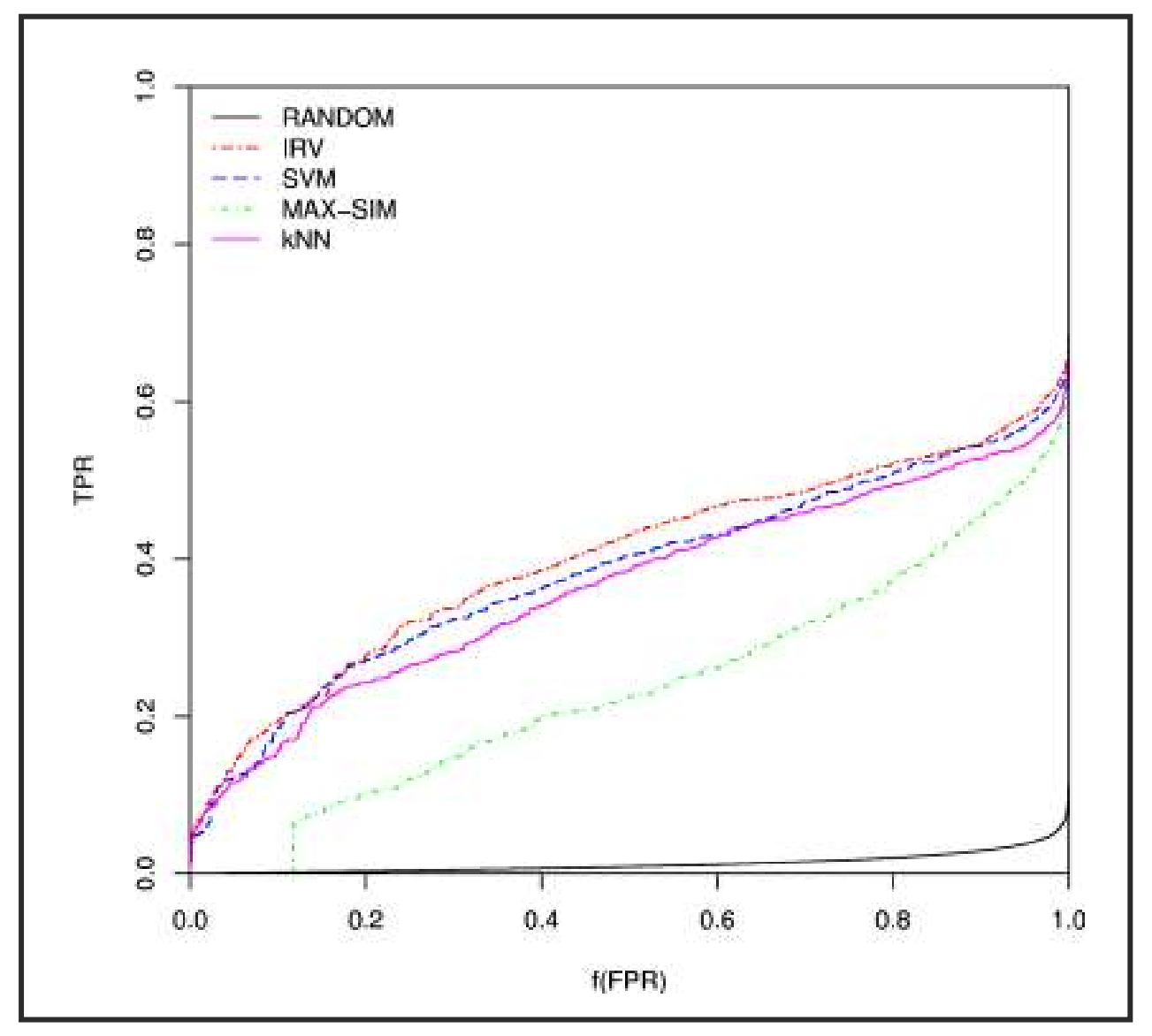
ROC on the DHFR data



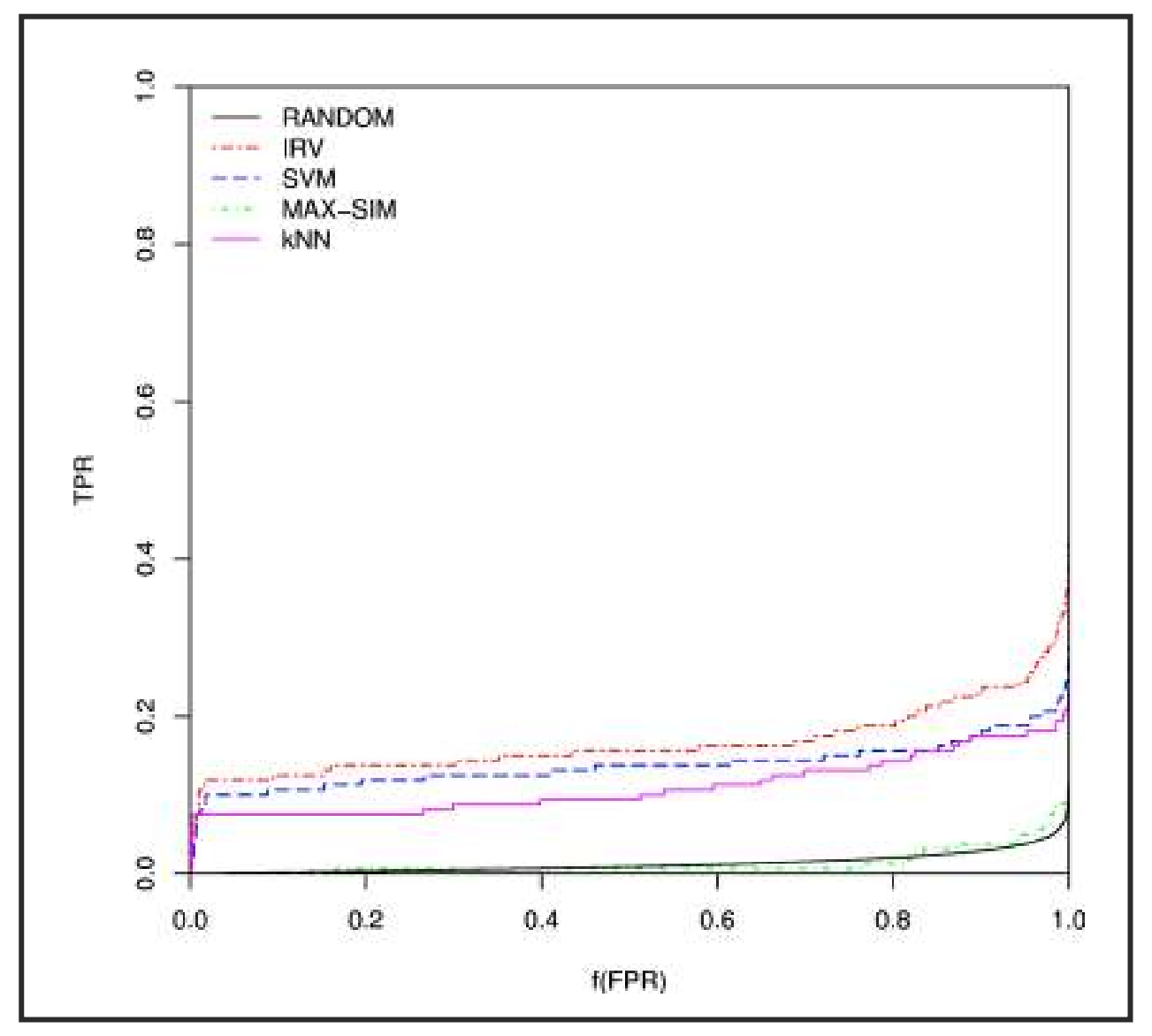
CROC (α=7) on the HIV data



CROC (α=7) on the DHFR data



CROC (α=80) on the HIV data



CROC (α=80) on the DHFR data

	α=7	α=80
SVM	0.644	0.310
kNN	0.638	0.365
IRV	0.656	0.400

Area under CROC on the HIV data

	α=7	α=80
SVM	0.290	0.138
kNN	0.267	0.115
IRV	0.398	0.154

Area under CROC on the DHFR data

Statistical Significance

We use the permutation test described by Zhao et al. [A STATISTICAL FRAMEWORK TO EVALUATE VIRTUAL SCREENING. BMC bioinformatics, 2009] to assess the significance of the observed difference in performance between two methods. We pool the ranks of the actives from both methods and repeatedly partition them at random into equally sized sets of ranks. The p-value is computed as the percentage of sampled differences in performance that are greater than the observed difference.

	SVM vs. IRV	kNN vs. IRV
ROC	0.094	0.094
pROC	0.016	0.025
CROC, α=7	0.001	0.055
CROC, α=80	0.002	0.010

Statistical significance of the difference in performance on the HIV data

Conclusion

We proposed a new vHTS algorithm, the IRV, with the following advantages: (1) the algorithm is suitable for early recognition and achieves state-of-the-art performance; (2) the training time is very short; (3) the risk of overfitting is minimal, due to the small number of free parameters.

Moreover, we proposed a new visualization method, the CROC curve, to better assess the results of vHTS experiments. Our data suggests that the area under the CROC curve has a better statistical power than other commonly used early recognition metrics.

Further Information

S. Joshua Swamidass, Chloé-Agathe Azencott, Ting-Wan Lin, Hugo Gramajo, Sheryl Tsai, and Pierre Baldi. THE INFLUENCE RELEVANCE VOTER: AN ACCURATE AND INTERPRETABLE VIRTUAL HIGH THROUGHPUT SCREENING METHOD, J. Chem. Inf. Model., March 2009. DOI: 10.1021/ci8004379.

S. Joshua Swamidass, Chloé-Agathe Azencott, Kenneth Daily, and Pierre Baldi. A CROC THAT ROC: MEASURING, VISUALIZING, AND OPTIMIZING EARLY RETRIEVAL. Unpublished draft.

Acknowledgements

OpenBabel and OpenEye Scientific Software Academic Licences; SVMTorch implementation; ROCR package for R.

NIH/NLM Biomedical Informatics Training Grant; NSF Grants 0321290 and 0513376; Microsoft Research Award to PB; IBM PhD Fellowship to CA.