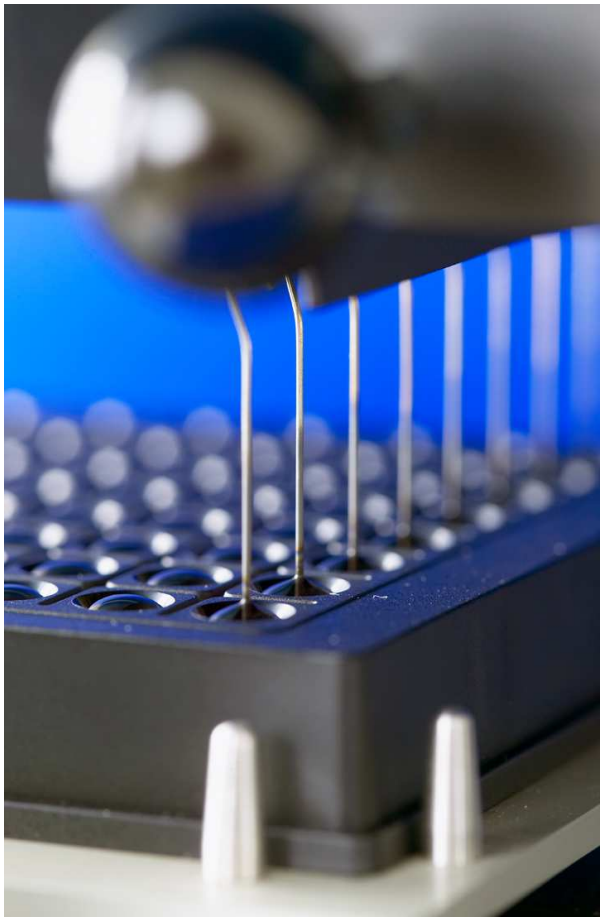
# Performing Fast and Accurate Virtual High-Throughput Screening

Chloé-Agathe Azencott

Institute for Genomics and Bioinformatics
Bren School of Information and Computer Sciences
University of California, Irvine
Joint work with S. Joshua Swamidass (Washington University)
and Pierre Baldi (UC Irvine)

# High-Throughput Screening

- Assay a large library of potential drugs against their target
- Very costly

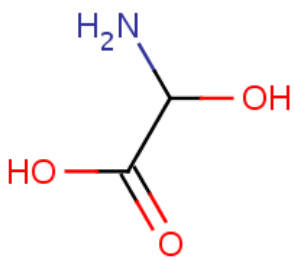


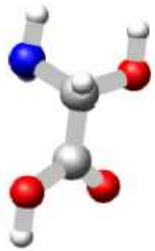
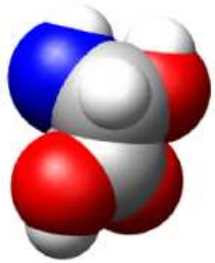
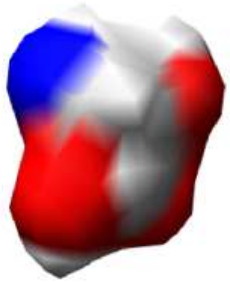
# Virtual High-Throughput Screening

- Cost-effective, in silico complement of experimental HTS
- Predict the activity of new compounds

# Representing Chemicals *in silico*

- Finding molecular descriptors that encapsulate the necessary information is hard
- Expert knowledge may be incomplete
- We need a representation that is *"complete"* enough.



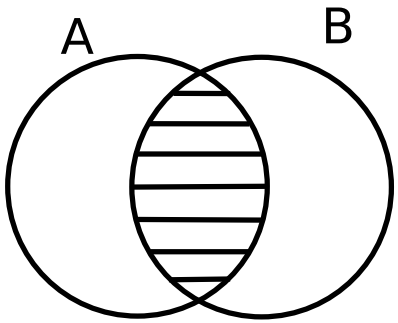




# Feature Vectors based on Structural Information

- Define *feature vectors* that record the presence/absence (or number of occurrences) of particular substructures in a given structure
- $$\phi(A) = (\phi_s(A))_{s \text{ substructure}}$$

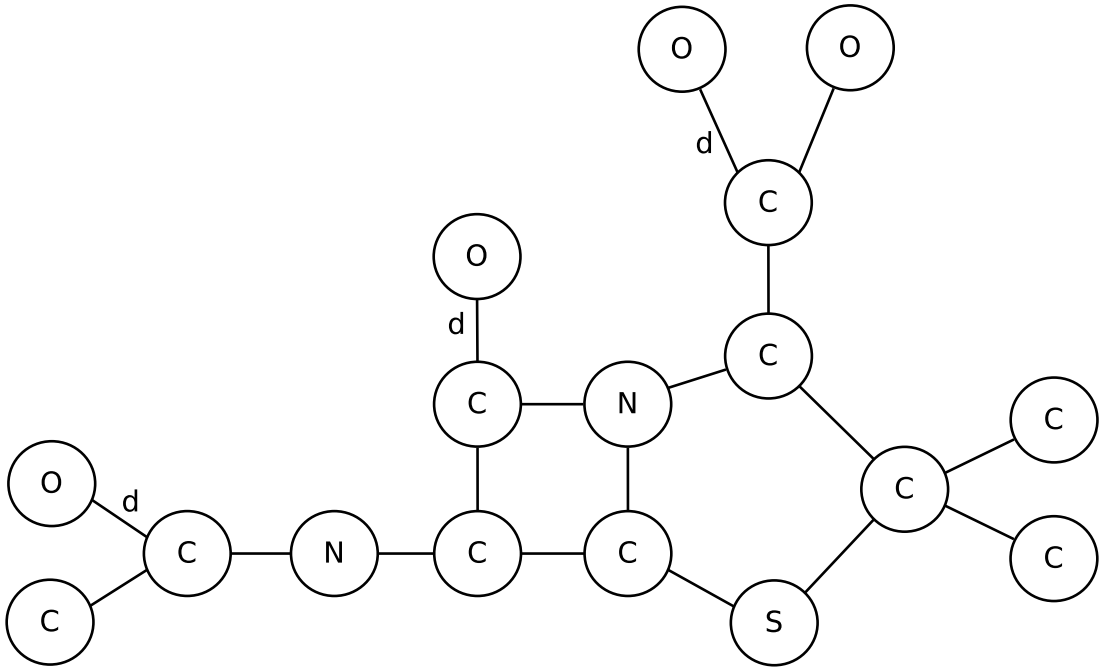
  where
  $$\phi_s(A) = \begin{cases} 1 & \text{if } s \text{ occurs in } A \\ 0 & \text{otherwise} \end{cases}$$
- Extension of traditional chemical fingerprints

# Tanimoto and MinMax Similarities




- Tanimoto (binary setting): $k(A,B) = \frac{A \cap B}{A \cup B}$
- MinMax (counts setting): $\frac{\sum_{i=1}^{N} \min(A_i, B_i)}{\sum_{i=1}^{N} \max(A_i, B_i)}$
- Kernel $\rightarrow$ SVMs

# Molecular Graph



# Two-Dimensional Substructures

- Labeled sub-paths (walks)

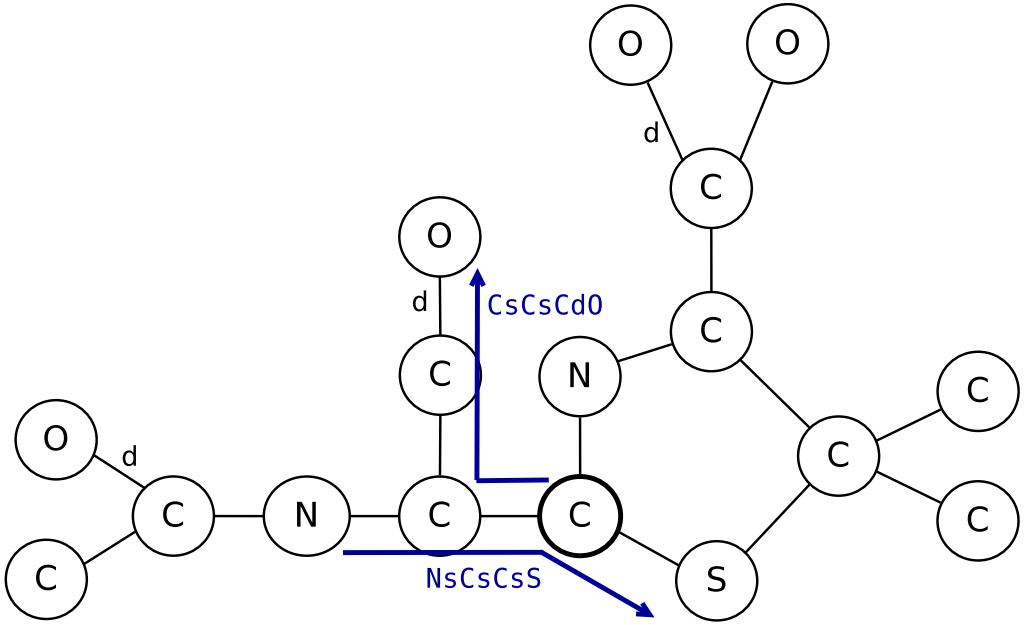



Figure: Some sub-paths of depth 3

# Two-Dimensional Substructures II

- Labeled sub-trees - Extended-Connectivity (or Circular) features

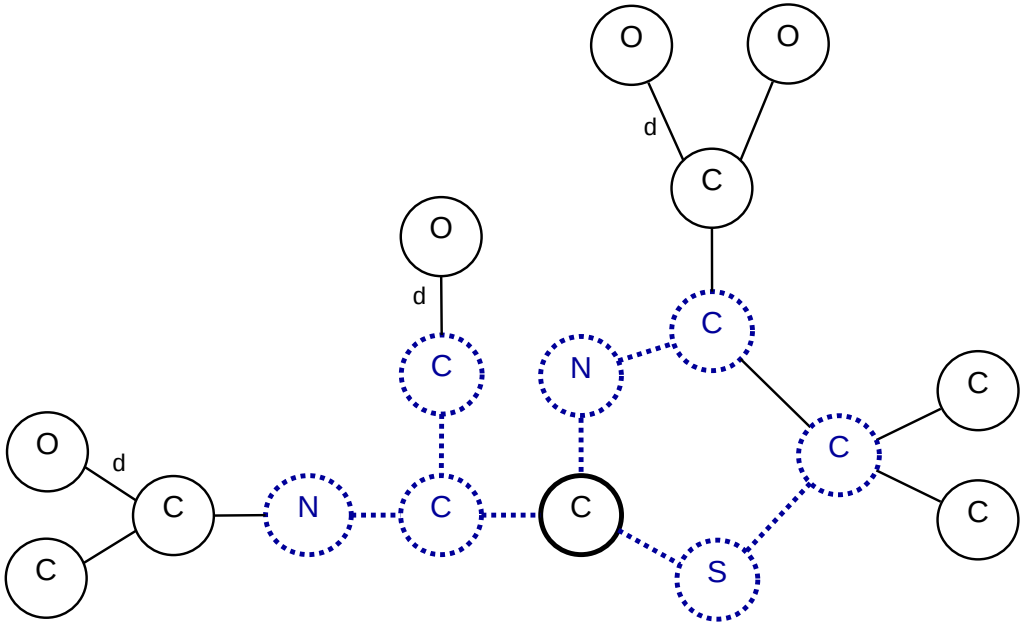
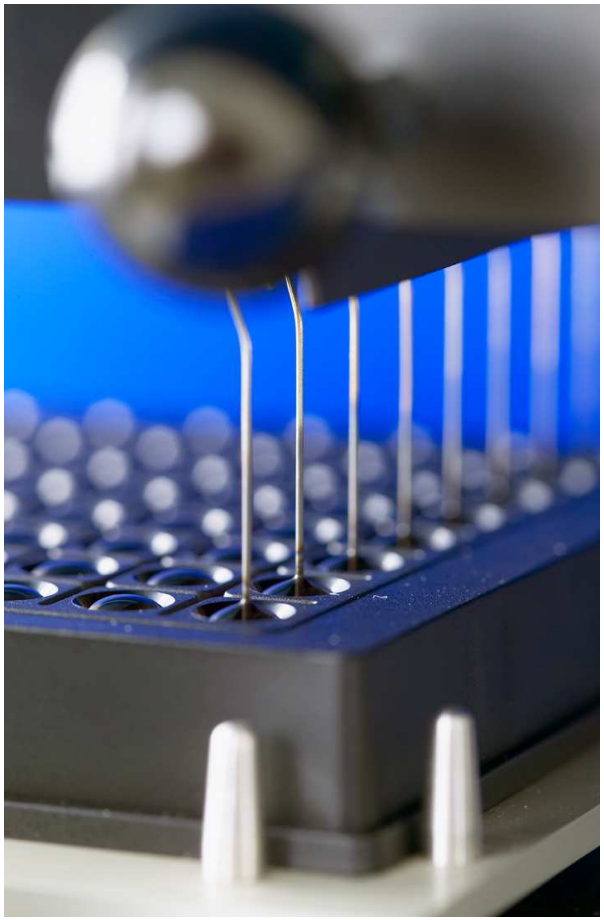



Figure: Example of a circular substructure of depth 2

# High-Throughput Screening

Back to vHTS...



# State-of-the-art

- Max-Sim

$$f(X) = max_{A_i \in \mathcal{A}} k(X, A_i)$$

- kNN

$$f(X) = \frac{|\mathcal{A} \cap \mathcal{N}(X)|}{k}$$

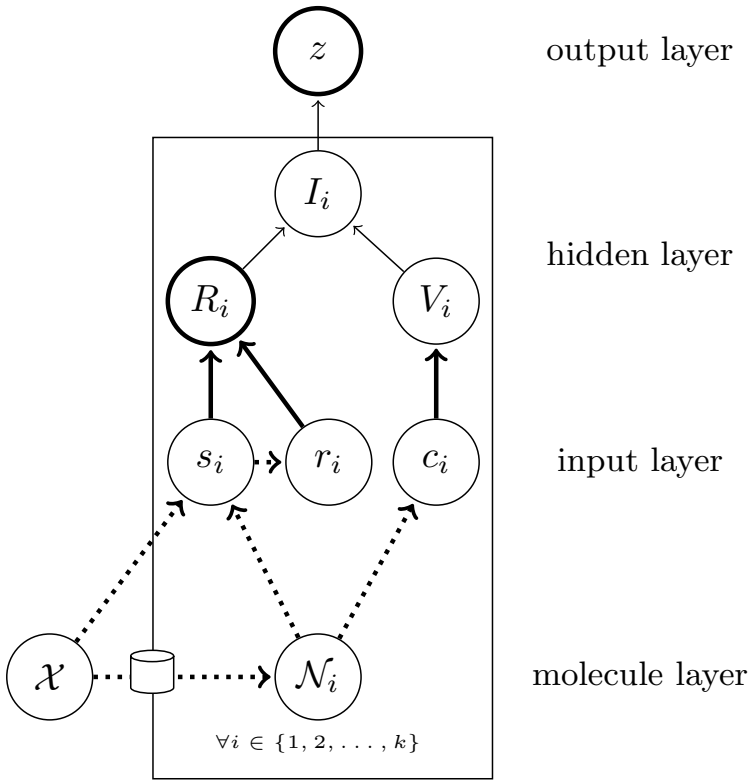
- SVM

$$f(X) = \sum_{i=1}^{n} a_i k(X, X_i) + b$$

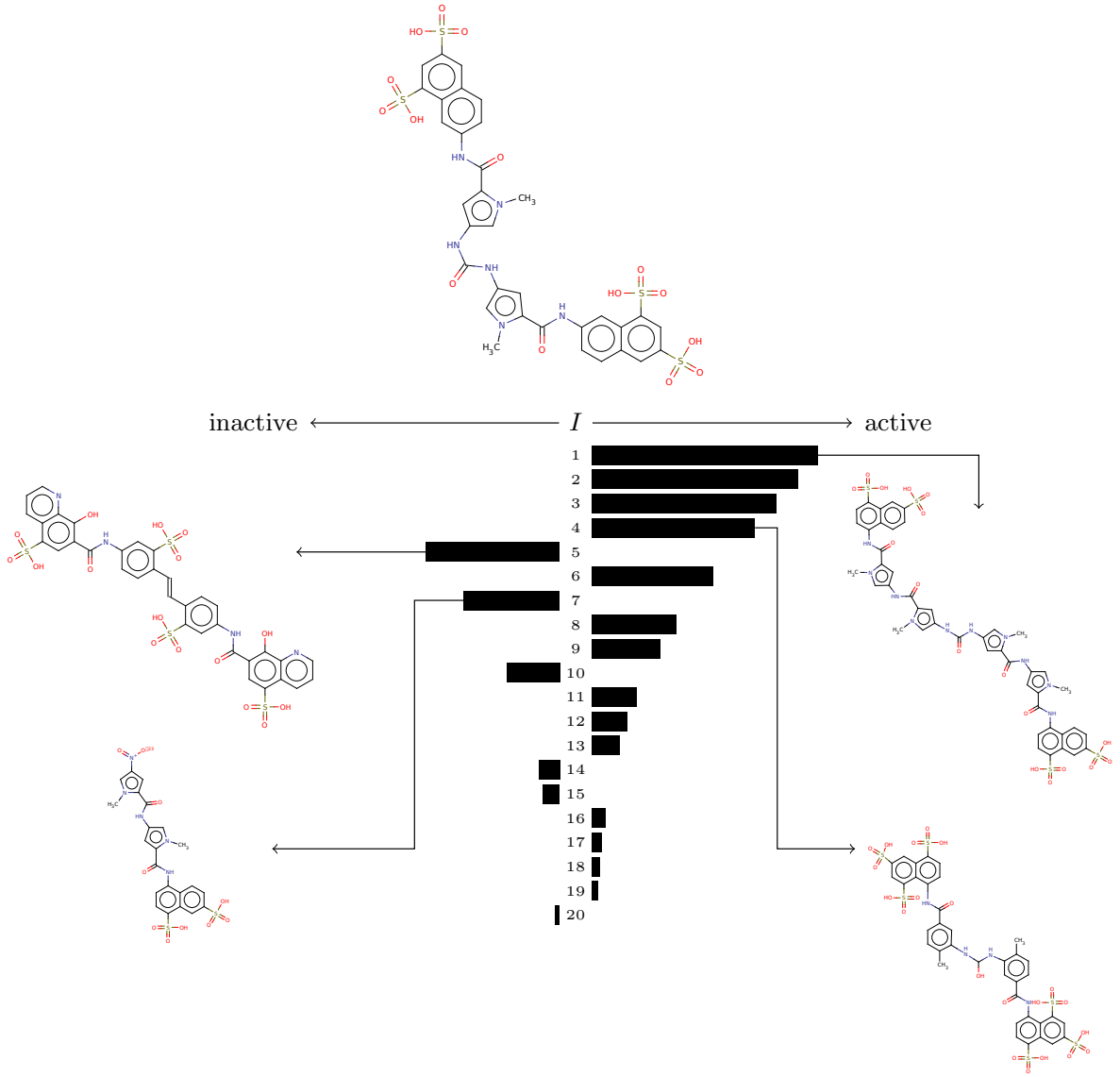
# The Influence Relevance Voter

(Joint work with Josh Swamidass)

- k-Nearest Neighbors
- Neural network on top
- Similarity: MinMax on 2D structural fingerprints



# Interpretability



# Benchmarked Performance

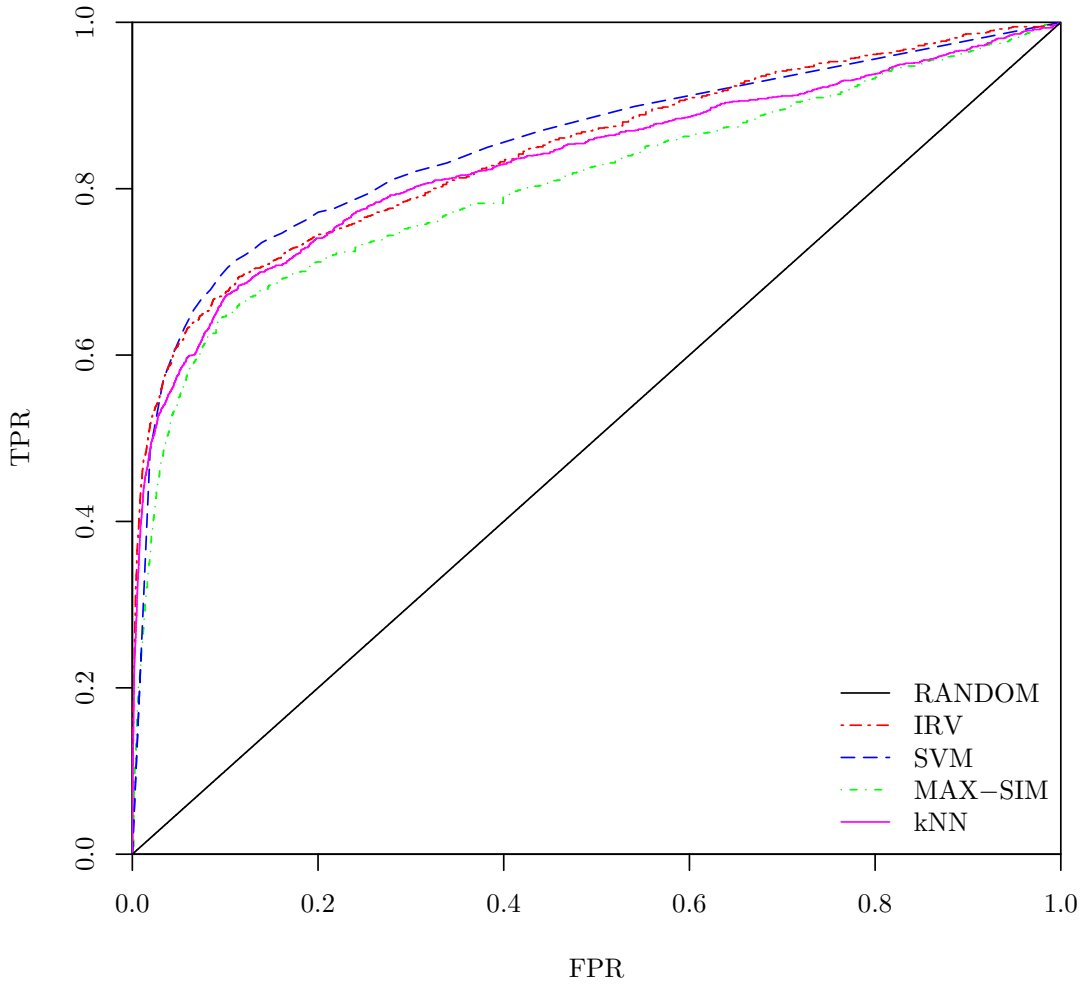
HIV data (IJCNN07 Challenge):

- train: 4,229 compounds (149 actives)
- test: 38,449 compounds (1,354 actives)

| | BER | AUC |
|---|---|---|
| IJCNN07 | 0.283 | 0.771 |
| SVM | 0.269 | 0.764 |
| IRV | 0.271 | 0.762 |
| kNN | 0.276 | 0.747 |
| MAXSIM | 0.283 | 0.739 |

# Early Recognition (HIV)

## ROC Curves



# Early Recognition

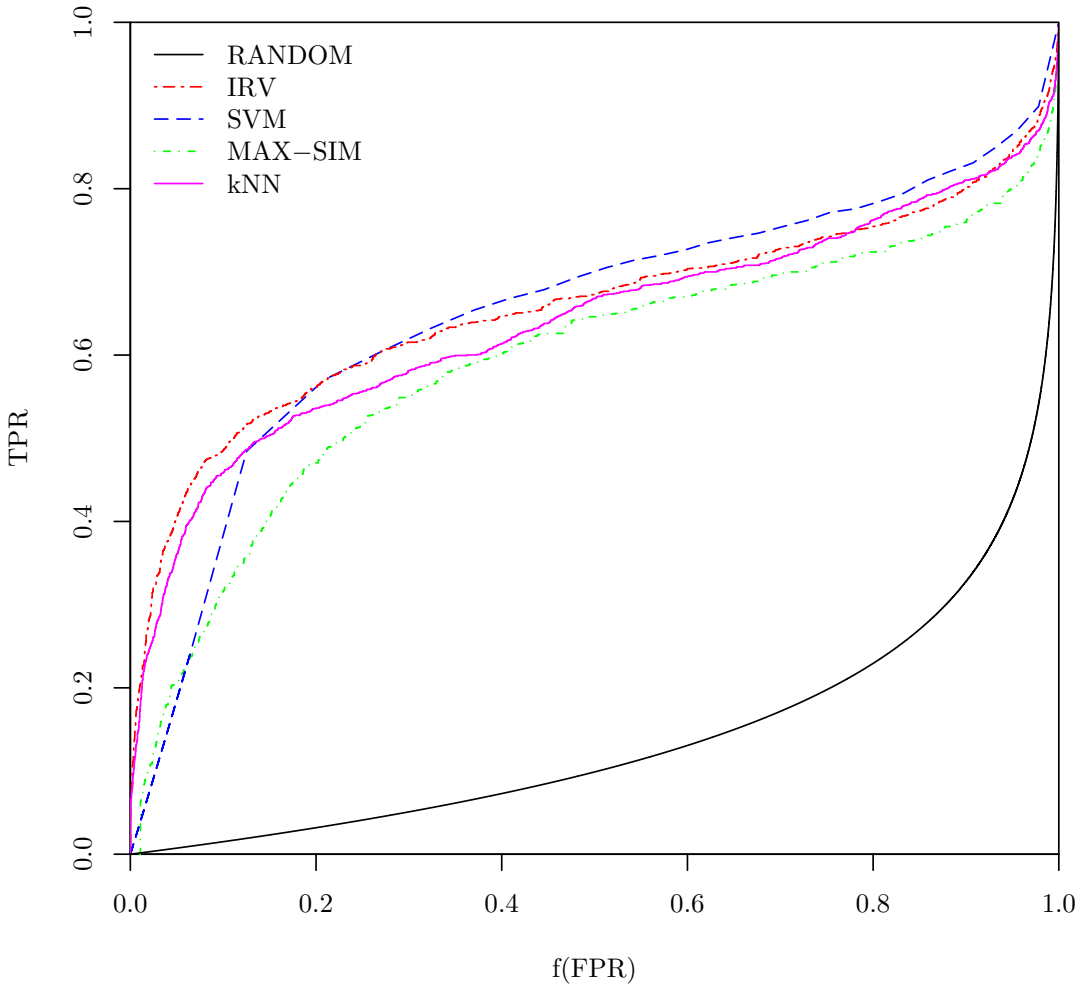
- Rank active compounds early
- $\rightarrow$ Magnify corresponding portion of the curve
  $f : [0, 1] \rightarrow [0, 1]$, continuous, concave down

$$\frac{1 - e^{-\alpha x}}{1 - e^{-\alpha}} \qquad x^{1/(\alpha+1)} \qquad \frac{log(1 + \alpha x)}{log(1 + \alpha)}$$

- $\rightarrow$ Concentrated ROC (CROC) curves

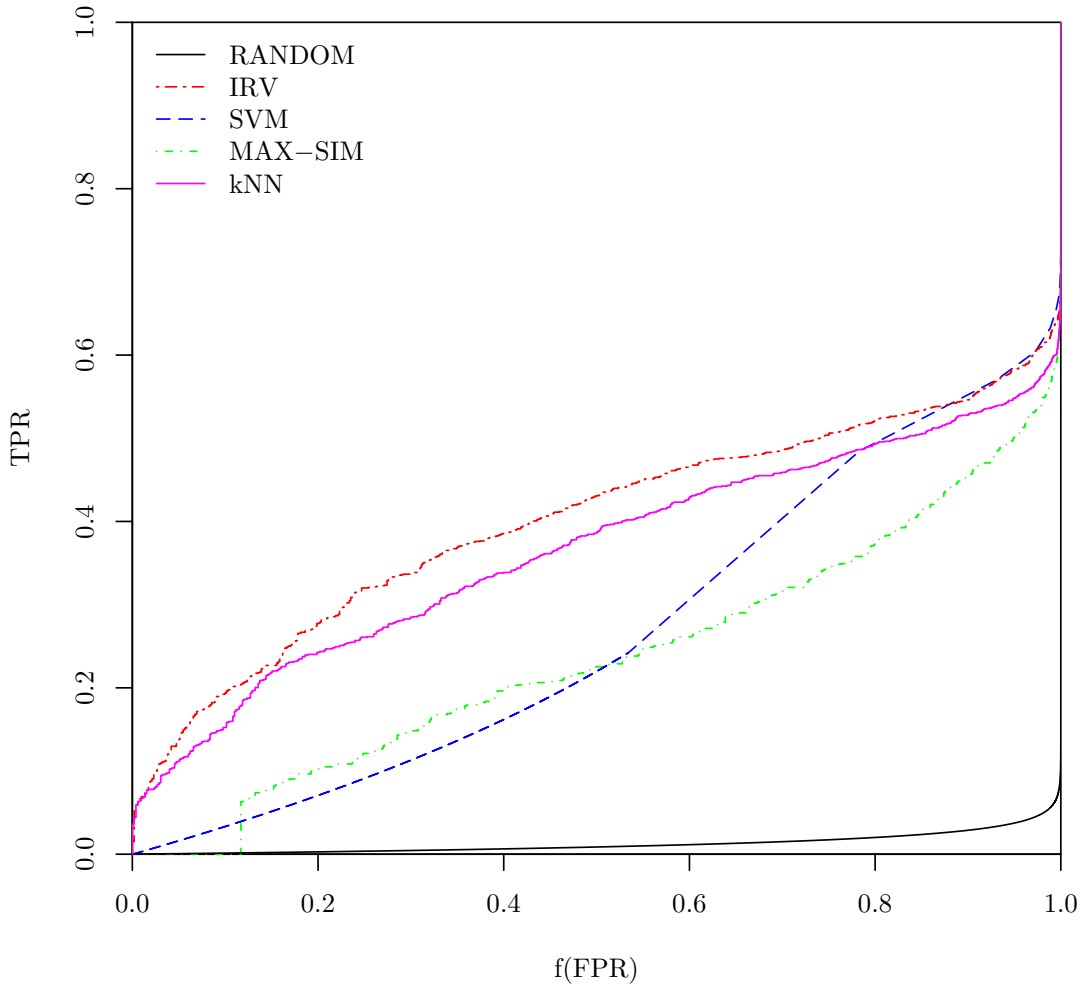
# Early Recognition (HIV)

## CROC Curves ($\alpha = 7$)



# Early Recognition (HIV)

## CROC Curves ($\alpha = 80$)



# Early Recognition (HIV)

| | ROC | CROC ($\alpha = 7$) | CROC ($\alpha = 80$) |
|---|---|---|---|
| SVM | 0.852 (± 0.004) | 0.644 (± 0.003) | 0.310 (± 0.006) |
| kNN | 0.832 (± 0.003) | 0.638 (± 0.004) | 0.365 (± 0.005) |
| IRV | 0.845 (± 0.002) | 0.656 (± 0.003) | 0.400 (± 0.004) |

# Statistical Significance

- Significance of the observed difference in CROC performance?
- Permutation test:
  - pool the ranks of the actives from both methods
  - repeatedly partition them at random into equally sized sets of ranks
  - p-value = fraction of sampled differences in performance greater than the observed difference

| | SVM vs. IRV | kNN vs. IRV |
|---|---|---|
| ROC | 0.094 | 0.094 |
| pROC | 0.016 | 0.025 |
| CROC ($\alpha = 7$) | 0.001 | 0.055 |
| CROC ($\alpha = 80$) | 0.002 | 0.010 |

# Maximizing Early Retrieval

- After each epoch, reweight the training examples
- Give higher weights to higher ranked examples
- Initialize at 1/N
- Weight Update Schemes:

$$w_{t+1}(r) = Ce^{-\gamma r(t)} \qquad w_{t+1}(r) = \frac{C}{r(t)^{\gamma+1}}$$

- Multiply by density of the scores in the neighborhood

$$w_{t+1}(r) = Ce^{-\gamma r(t)} g(r) \qquad w_{t+1}(r) = \frac{C}{r(t)^{\gamma+1}} g(r)$$

- Increase $\gamma$ at each epoch $\rightarrow$ avoid getting stuck in local minima by starting with low values of $\gamma$

# Maximizing Early Retrieval

- Exponential weighting scheme: performance about the same, but convergence sped up (from 100 to 4-8)
- Power weighting scheme

| | AUC[CROC] ($\alpha = 7$) | | |
|---|---|---|---|
| | $w_r = 1$ | $w_r = 1/r$ | $w_r = g(r)/r$ |
| 1 iter | 0.569 | 0.598 | 0.611 |
| 3 iters | 0.641 | 0.646 | 0.652 |
| convergence | 0.660 | 0.660 | 0.661 |

# Conclusion

- New vHTS algorithm, the IRV
  - Suitable for early recognition
  - Achieves state-of-the-art performance
  - Interpretable underlying inferences
  - Probabilistic semantic of the output predictions
  - Short training time
- New early recognition evaluation method, the CROC
  - CROC curve (visualization)
  - Area under the CROC curve (performance measure)
  - CROC optimization

# Future Directions

- Apply the IRV to non-biological data (e.g. information retrieval)
- Incorporate more information in the IRV
- Learning algorithms that optimize the area under the CROC curve


S. J. Swamidass, C.-A. Azencott, T. -W. Lin, H. Gramajo, S. Tsai, and P. Baldi.
The Influence Relevance Voter: An Accurate And Interpretable Virtual High-Throughput Screening Method,
*J. Chem. Inf. Model*, March 2009


S. J. Swamidass, C.-A. Azencott, K. Daily, and P. Baldi.
A CROC Stronger than ROC: Measuring, Visualizing, and Optimizing Early Retrieval,
*Bioinformatics*, accepted with minor revisions

Acknowledgements

- Dr. S. Joshua Swamidass and Prof. Pierre Baldi
- 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.

Thank you!