A Tiered Screen Protocol for the Discovery of Structurally Diverse HIV Integrase Inhibitors

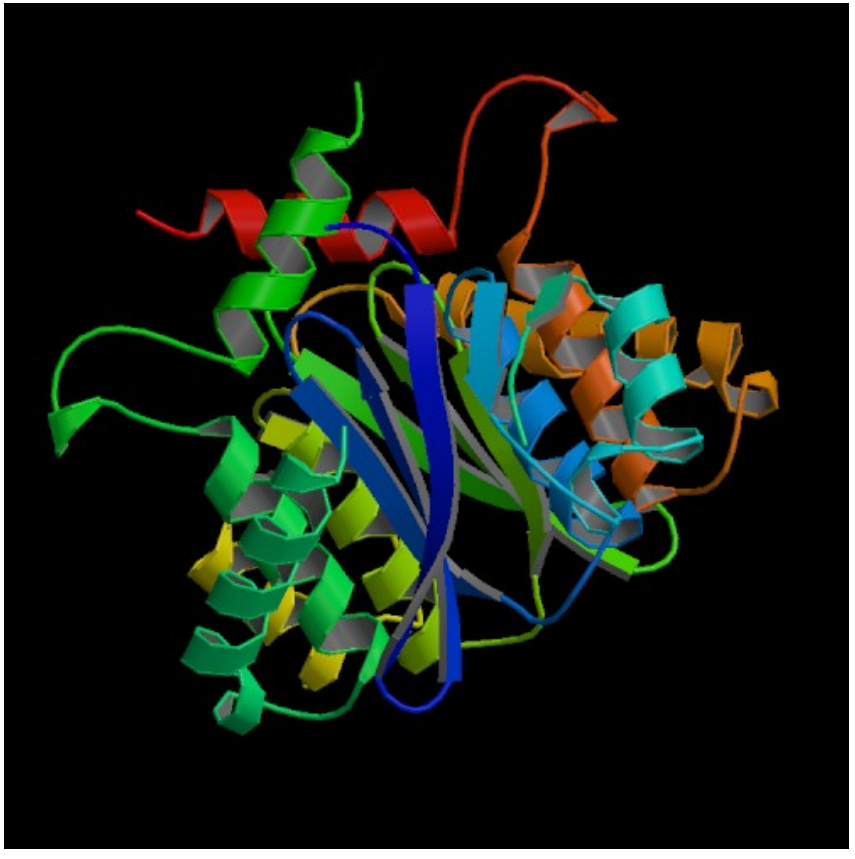
Rajarshi Guha, Debojyoti Dutta, Ting Chen and David J. Wild

School of Informatics
Indiana University
and
Dept. Computational Biology
University of Southern California

28th March, 2007
Chicago

Background

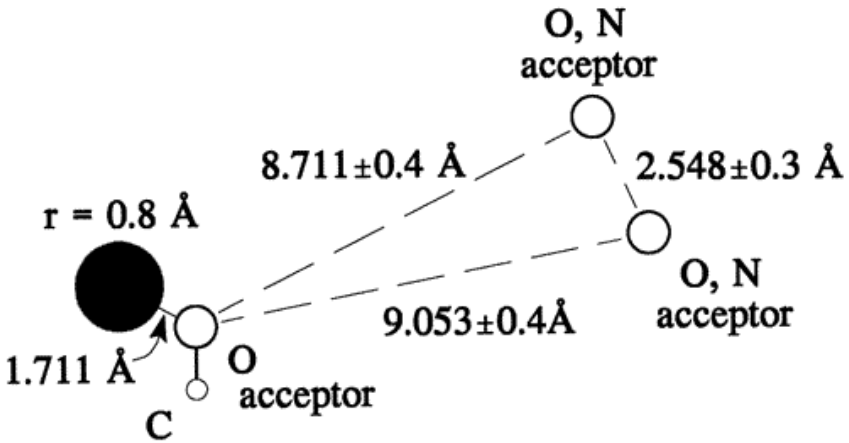
- Most drugs target HIV reverse transcriptase or protease
- HIV integrase is vital for viral replication
- No drugs have been approved for HIV integrase

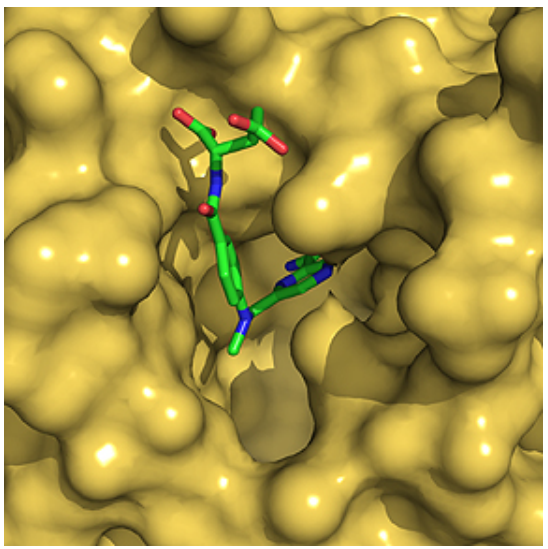


1BIS

Previous Approaches

- Pharmacophores
- Docking
- Disadvantages
  - Requires a reliable receptor structure
  - Computationally intensive
  - Not necessarily diverse





Nicklaus, M.C. et al., *J. Med. Chem.*, **1997**, *40*, 920–929

Nair, V; Chi, G.; Neamati, N.; *J. Med. Chem.*, **2006**, *49*, 445–447

Dayam, R; Sanchez, T.; Neamati, N.; *J. Med. Chem.*, **2005**, *48*, 8009–8015

Goals

- High speed
  - Be able to process large libraries rapidly
  - Avoid docking till required
  - Try and use connectivity information only
- Reliability
  - Use consensus approaches for prediction
  - Use molecular similarity
- Novely
  - Try to obtain a diverse set of hits
  - Try to obtain hits suitable for lead hopping

Datasets & Tools

Training Data

- Curated dataset
- 529 actives
- 395 inactives

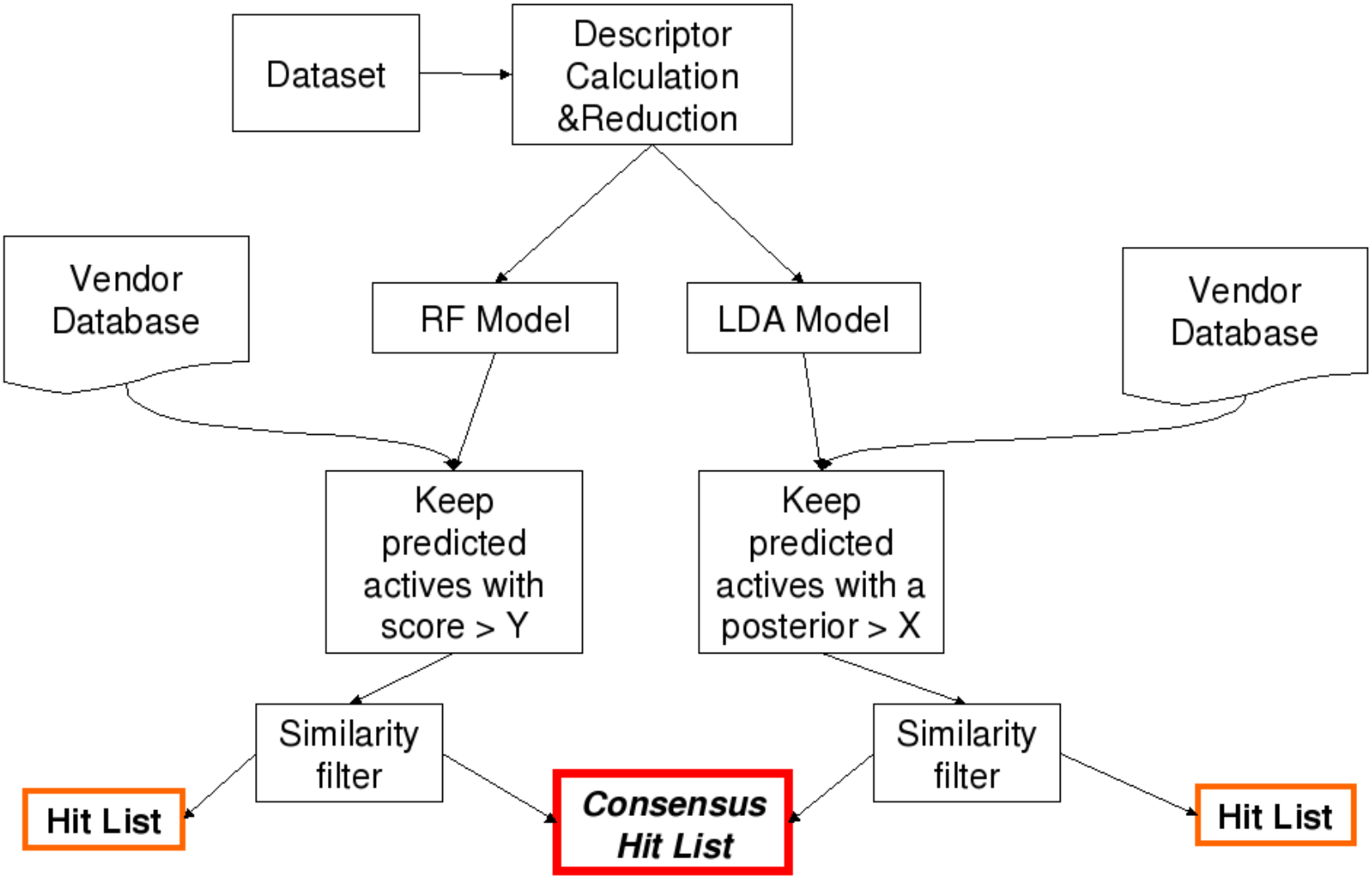
Vendor Database

- 50,000 compounds
- 2D structures

Software

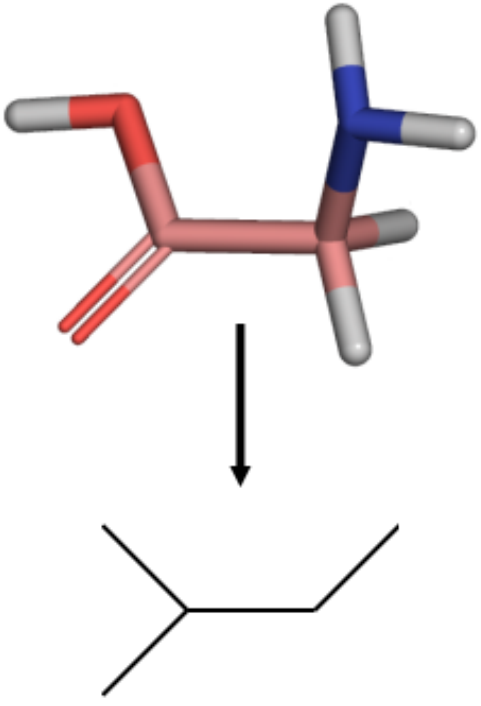
- MOE
- R 2.2.0
- `MASS`, `randomForest`, `fingerprint`

Overview of the Screening Protocol



Descriptor Calculations

- Restricted to topological descriptors
- Calculated 142 descriptors
- Removed low variance and correlated descrirptors
- Size of reduced pool was 45



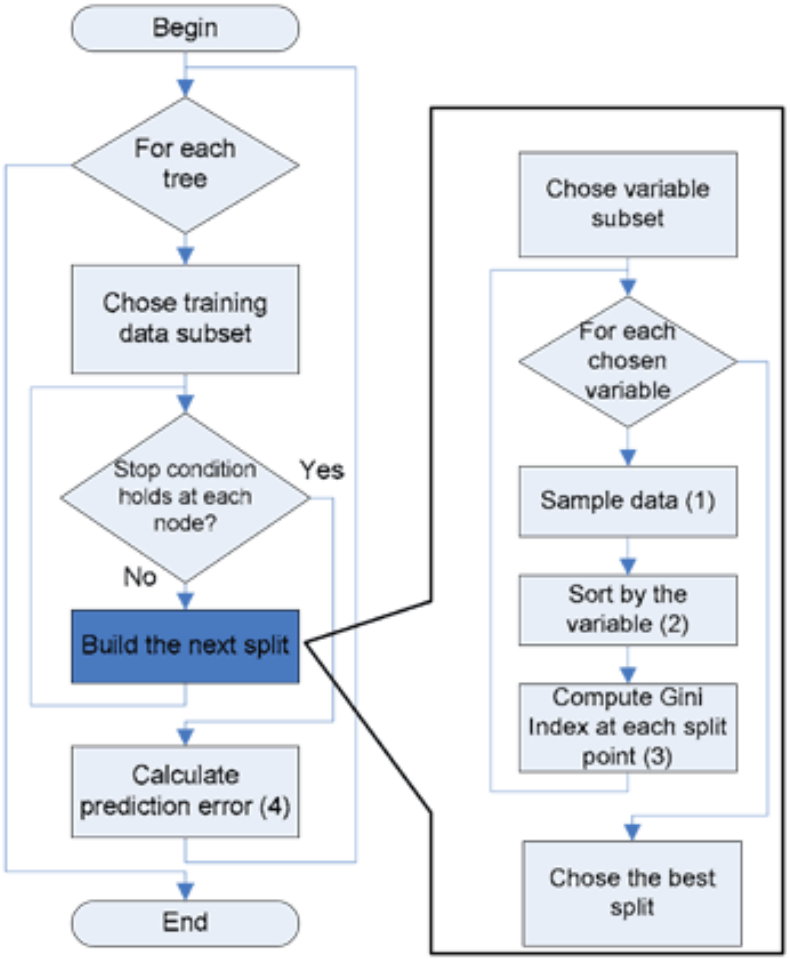
- Only considers 2D connectivity
- Very fast to compute

Predictive Models

- Linear discriminant analysis
- Why?
  - Simple and may be sufficient
- How?
  - Used a genetic algorithm to search for descriptor subsets
  - Finally used a 6-descriptor model
- Accuracy
  - On the whole dataset, 72%
  - On TSET/PSET, 72% / 71%
  - With leave-10%-out, 71%

Predictive Models

- Random Forest
- Why?
  - No feature selection required
  - Does not overfit
  - May capture non-linearities
- How?
  - Used 500 trees
  - Sampled 6 features at each split
- Accuracy
  - On the whole dataset, 72%
  - On TSET/PSET, 75% 70%

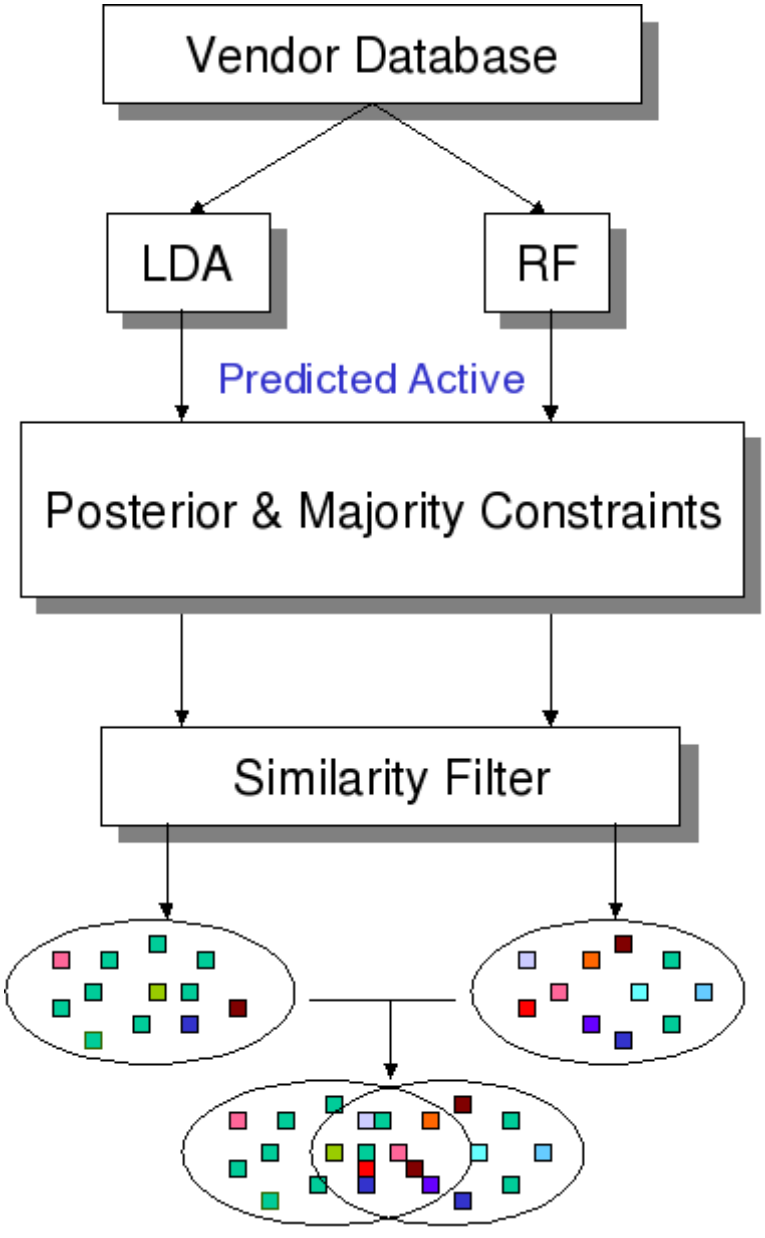


Breiman, L. et al., *Classification and Regresion Trees*, CRC Press, Boca Raton, FL. 1984

Borisov, A. et al., *Intel Technology Journal*, **2005**, *9*

Consensus Predictions

- We first consider compounds predicted active by both models
  - For the LDA model we consider compounds with a posterior $>$ P
  - For the RF model we consider compounds which had a majority $>$ M
- We then apply the similarity filter
- This results in two hit lists
- A final hit list is obtained from the intersection of the hit lists for the individual model



Similarity Filter

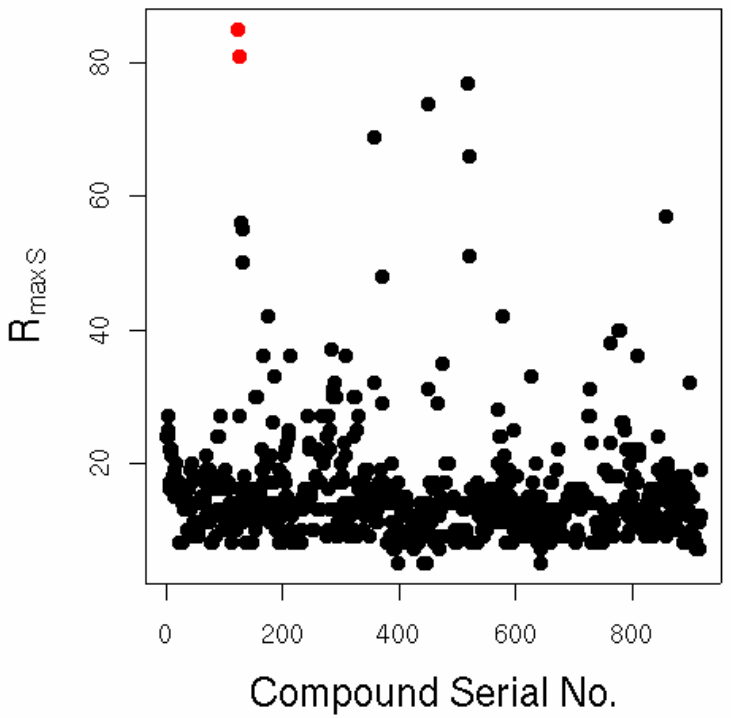
- The goal is to predict actives from the vendor database
- Evaluate 166 bit MACCS fingerprints
- The Tanimoto Similarity between two compounds is defined by

$$S = \frac{N_c}{N_a + N_b - Nc}$$

- For a compound predicted active
  - Calculate average similarity to TSET actives (S_1)
  - Calculate average similarity to TSET inactives (S_2)
  - Select compounds where $S_1 - S_2 \geq \epsilon$

Suggesting Good Leads

- Suggest further prioritization?
  - Look for TSET actives that are not in the bulk of the dataset
  - These compounds may be good starting points for further modifications
  - Evaluate similarity of the hit list compounds to these isolated TSET compounds
- Hit list compounds most similar to the most isolated TSET active may be good leads



Saeh, J.C.; Lyne, P.D.; Takasaki, B.K.; Cosgrove, D.A.; *J. Chem. Inf. Comput. Sci.*, **2005**, *45*, 1122–1133.

Guha, R.; Dutta, D.; Jurs, P.C; Chen, T.; *J. Chem. Inf. Model.*, **2006**, *46*, 1713–1722

Time Considerations

- Model development
  - Descriptor calculation is rapid and a one time event
  - Building individual LDA or RF models is fast
  - Time required to obtain optimal model can be large when a GA is used (partly due to interpreted code)
  - Predictions for 50,000 compounds $<$ 1min
- Similarity calculation is very time consuming
- Detecting spatial outliers is slow
  - Can be improved with approximate NN algorithms

Dutta, D.; Guha, R.; Jurs, P.C.; Chen, T.; *J. Chem. Inf. Model.*, **2006**, *46*, 321–333

Results

- Parameters
  - P, M > 0.7
  - $\epsilon \geq 0.01$
- Of the 34 hits, 7 have a similarity greater than 0.5 with the most outlying TSET active
- We obtained 66 more hits from an ensemble of LDA models

- LDA - 313 hits
- RF - 57 hits
- Consensus - 34 hits

Average similarity $= 0.64$

None were in common with pharmacophore predictions

Results

- Parameters
  - P, M > 0.7
  - $\epsilon \geq 0.01$
- Of the 34 hits, 7 have a similarity greater than 0.5 with the most outlying TSET active
- We obtained 66 more hits from an ensemble of LDA models

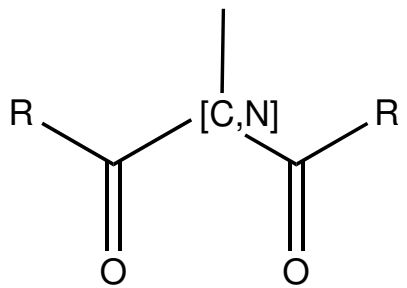
- LDA - 313 hits
- RF - 57 hits
- Consensus - 34 hits

Average similarity $= 0.64$

None were in common with pharmacophore predictions

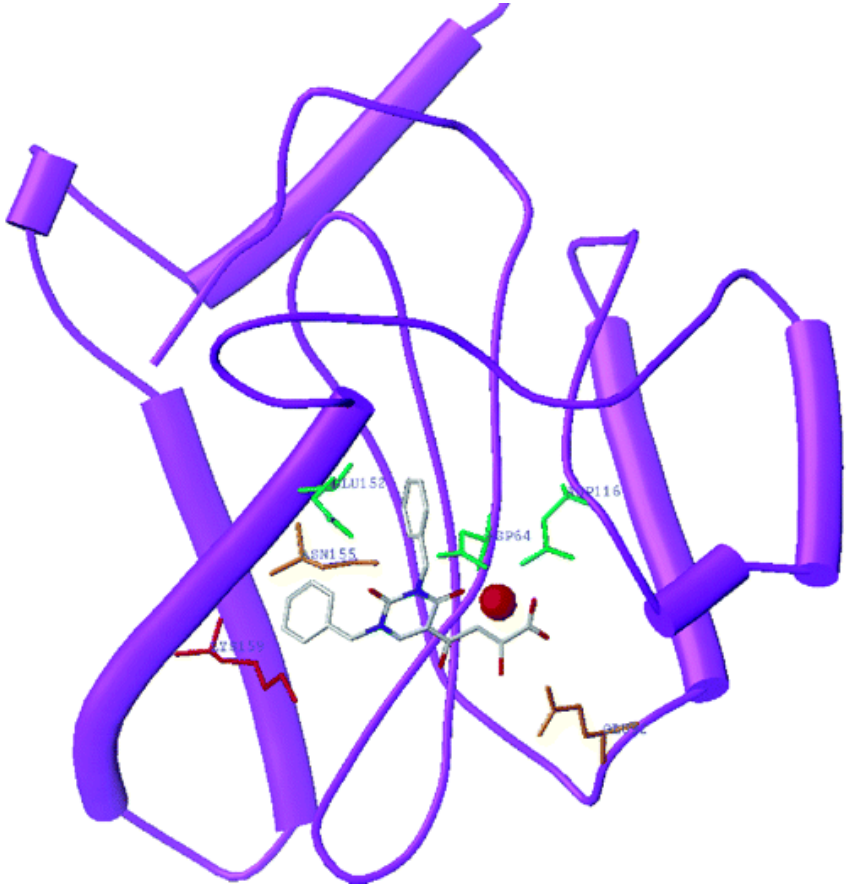
The Search for β diketones

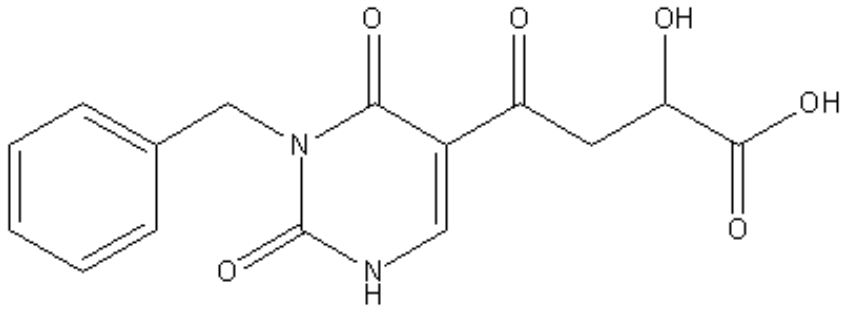
- A recently published inhibitor exhibited a β diketone moiety
- Searched the vendor DB for structures with this feature and predicted them
  - LDA model : 244 actives (posterior > 0.8)
  - RF model : 4 actives (score > 0.85)
  - The intersection set contained 4 compounds
- We missed these hits due to our similarity constraints being too strict



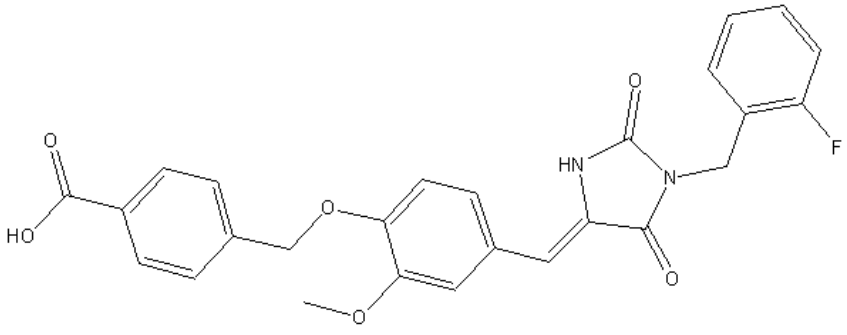
Nair, V.; Chi, G.; Ptak, R.; Neamati, N.; *J. Med. Chem.*, **2006**, *45*, 445–447

Similarity to the Published Compound





Published



Predicted

Future Work

- Investigate similarity to known inhibitors in terms of pharmacophore similarity
- Dock our best hits (may not be conclusive)
- Investigate the distribution of vendor compounds in descriptor space
- Cluster the vendor database and predict representative members of clusters
- Perform assays

Conclusions

- A consensus approach to predictive models allows us to increase confidence in activity predictions
- Prioritization based on similarity is intuitive, but may not work well with homogenous datasets
- The protocol appears to have identified possible leads which await confirmation

Acknowledgements

- Prof. Nouri Neamati
- Dr. Raveendra Dayam