Active Learning in Drug Discovery

Main Paper:
Reker, D.; Schneider, P.; Schneider, G.
Chem. Sci. **2016**, *7*, 3919.

Literature Seminar
2016/7/23
D3 Shun-ichiroh Katoh

Topics

Active Learning

Random Forest

Exploration for Protein-Protein Interaction Inhibitors

Topics

✓ **Active Learning**

Random Forest

Exploration for Protein-Protein Interaction Inhibitors

Active Learning

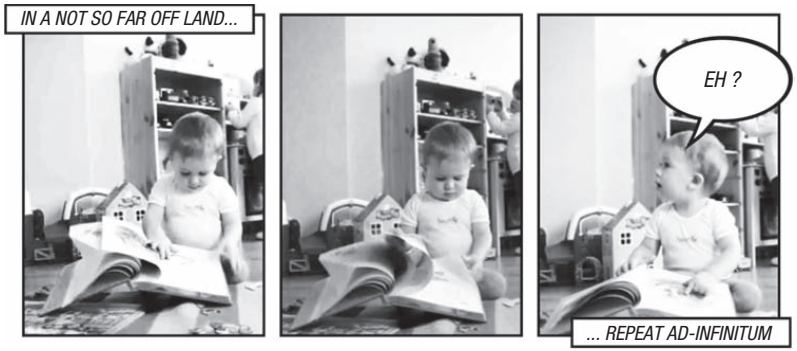
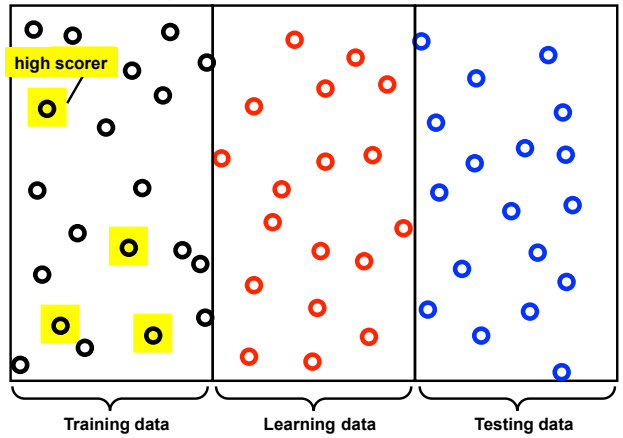


Fig. 1. An example of self-directed learning in everyday life. In the scene, a young child is flipping through the pages of a storybook. At some point, the child comes to a picture she finds interesting and requests the name of the object from the caregiver. A key feature of this example is that the learner herself, as opposed to the parent or teacher, controls the learning sequence through her choices and actions.

Active Learning in Drug Discovery - 1

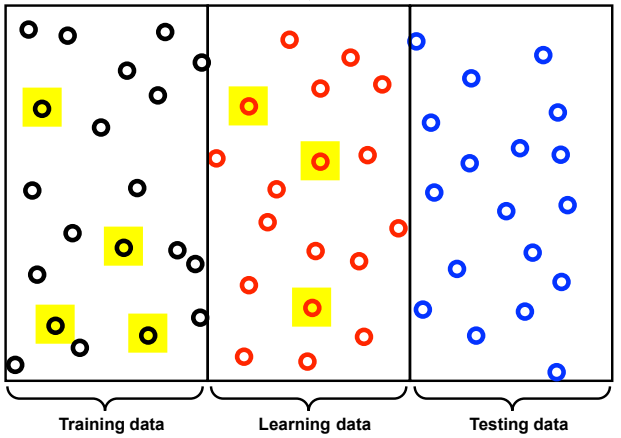
Trains using this compounds



1) Reker, D.; Schneider, G. *Drug Discovery Today* **2015**, *20*, 458.

Active Learning in Drug Discovery - 2

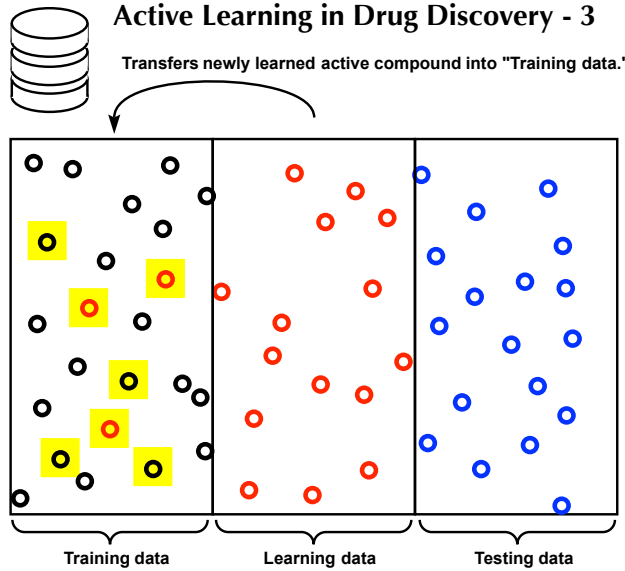
Learns whether the compounds are active or inactive



1) Reker, D.; Schneider, G. *Drug Discovery Today* **2015**, *20*, 458.

Active Learning in Drug Discovery - 3

Transfers newly learned active compound into "Training data."

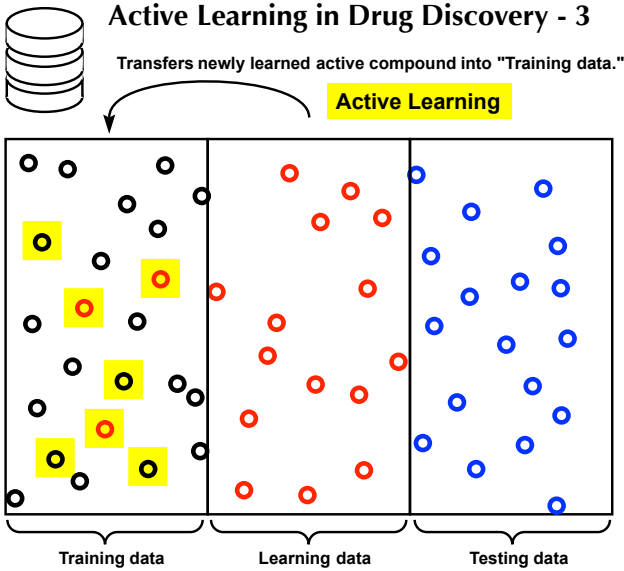


1) Reker, D.; Schneider, G. *Drug Discovery Today* **2015**, *20*, 458.

Active Learning in Drug Discovery - 3

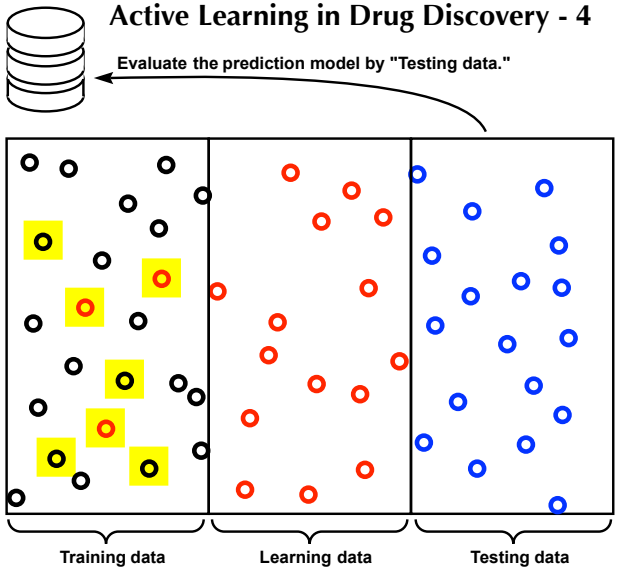
Transfers newly learned active compound into "Training data."

Active Learning



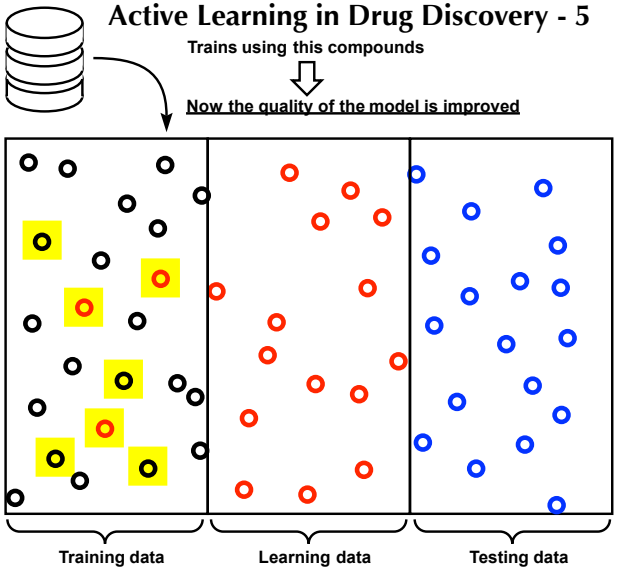
1) Reker, D.; Schneider, G. *Drug Discovery Today* **2015**, *20*, 458.

Active Learning in Drug Discovery - 4



1) Reker, D.; Schneider, G. *Drug Discovery Today* **2015**, *20*, 458.

Active Learning in Drug Discovery - 5



1) Reker, D.; Schneider, G. *Drug Discovery Today* **2015**, *20*, 458.

Topics

Active Learning

✓ Random Forest

Exploration for Protein-Protein Interaction Inhibitors

Regression Tree - Example - 1

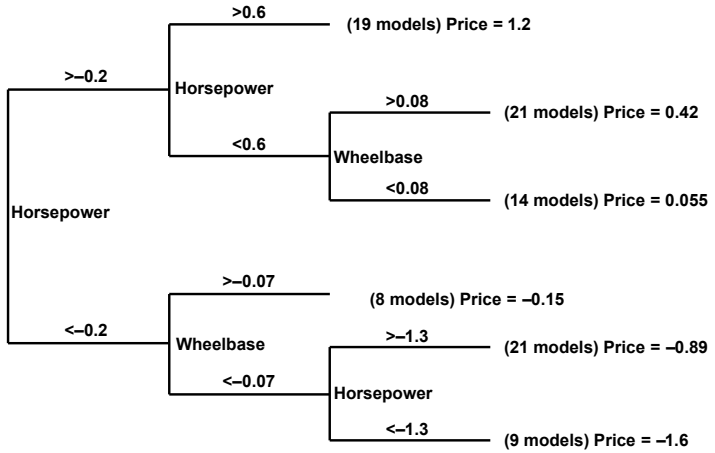


Figure 1: Regression tree for predicting price of 1993-model cars. All features have been standardized to have zero mean and unit variance. Note that the order in which variables are examined depends on the answers to previous questions. The numbers in parentheses at the leaves indicate how many cases (data points) belong to each leaf.

1) http://www.stat.cmu.edu/~cshalizi/350-2006/lecture-10.pdf

Regression Tree - Example - 2

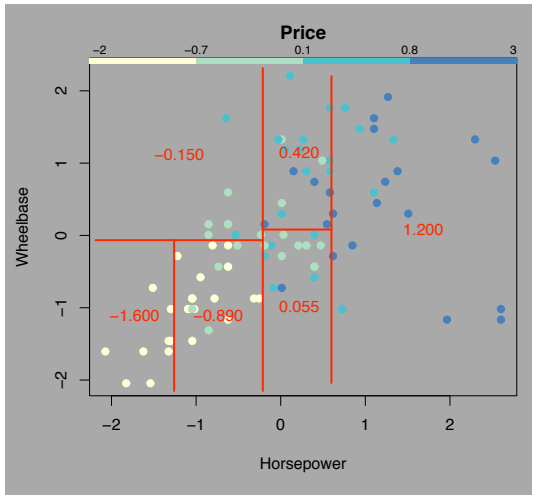
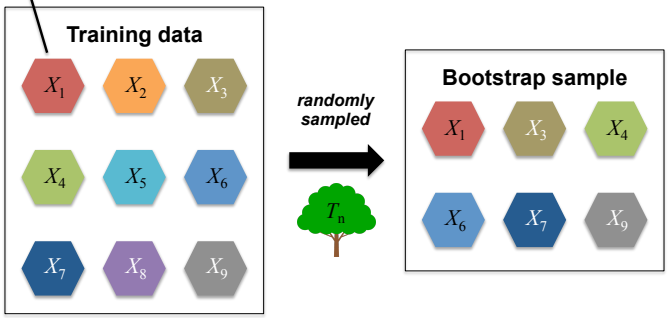


Figure 2:
The partition of the data implied by the regression tree from Figure 1. Notice that all the dividing lines are parallel to the axes, because each internal node checks whether a single variable is above or below a given value.

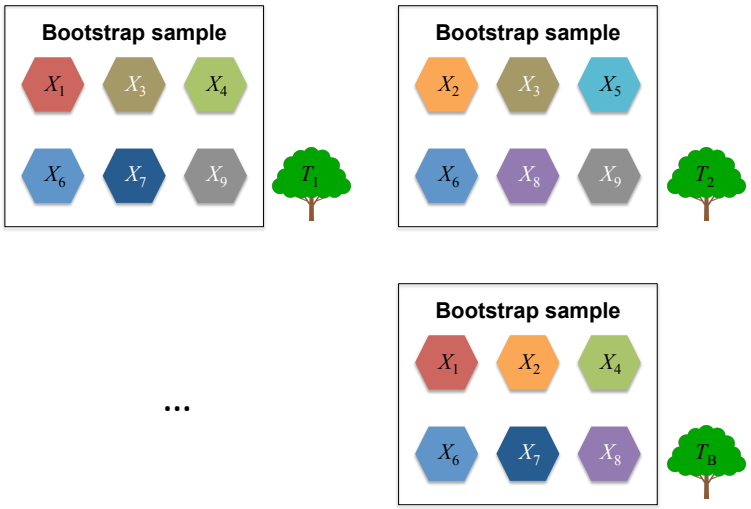
1) http://www.stat.cmu.edu/~cshalizi/350-2006/lecture-10.pdf

Random Forest - 1

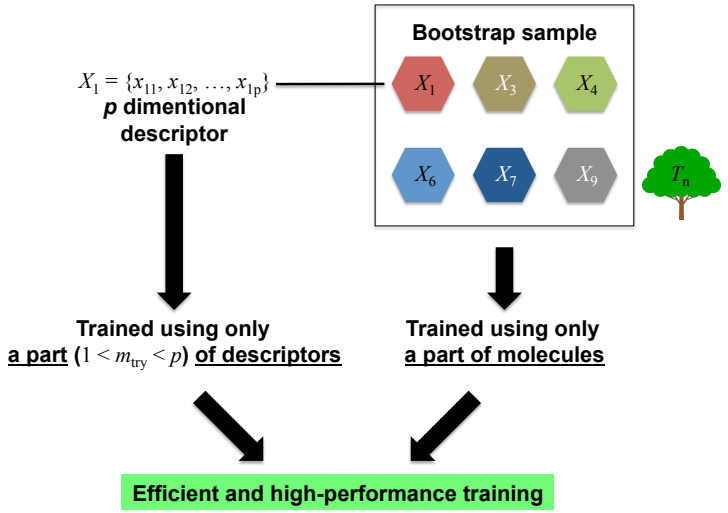
Each compound has descriptors $\{x_1, x_2, \ldots, x_p\}$ and score $\{y\}$



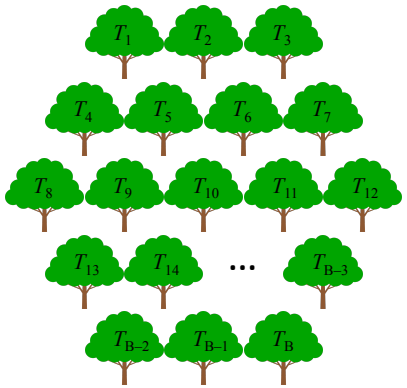
Random Forest - 2



Random Forest - 3



Random Forest - 4



= Random Forest

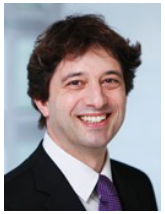
Topics

Active Learning

Random Forest

✓ **Exploration for Protein-Protein Interaction Inhibitors**

Prof. Gisbert Schneider



1965	Born in Fulda, Germany
1991	Freie Universität Berlin, Germany (Prof. Paul Wrede)
1994	Freie Universität Berlin, Germany (Ph.D., Prof. Paul Wrede)
1994-1997 (postdoc)	Benjamin Franklin University Clinic, Berlin; The Massachusetts Institute of Technology in Cambridge, MA; The University of Stockholm, Sweden; The Max-Planck-Institute of Biophysics in Frankfurt, Germany;
1997-2002	*Scientific specialist in industrial research* (Hoffmann - La Roche Ltd.)
2002-2009	*Professor* Goethe University in Frankfurt, Germany
2010-	*Professor* ETH Zürich, Switzerland

CXCR4 and CXCL-12

CXCR4:

- **G-protein-coupled receptor**
- **Involved in a number of hematopoietic and immune systems**
- **Associated with HIV, cancer, rheumatoid arthritis, etc.**

CXCL-12:

- **Endogenous ligand of CXCR4**
- **Form a part of inter-cellular signaling system**

However, It is difficult to find low-molecular-weight inhibitors of protein-protein interactions.
➪ Decided to utilize active learning.

Explorative vs Exploitive - 1

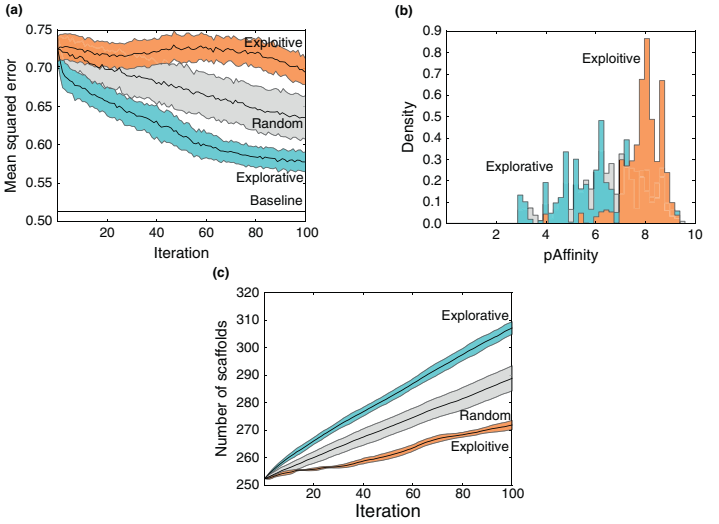
Explorative strategy:

- **Improves the model**
- **Picks wide range of molecule scaffolds**
- **Not always proposes favorable structures**

Exploitive strategy:

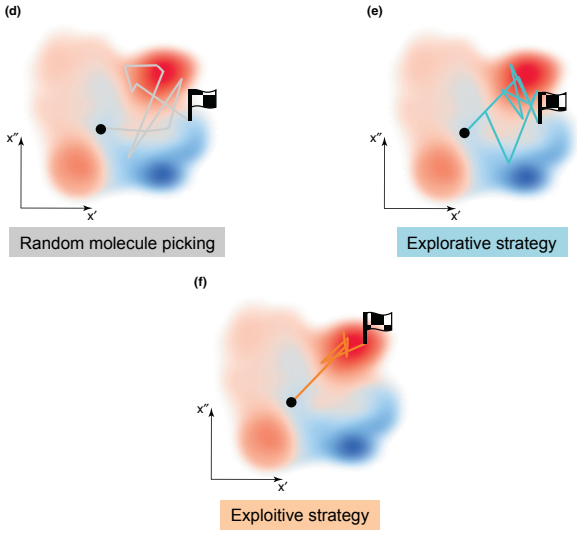
- **Retrieves active compounds**
- **Not always proposes various structures**
- **Sometimes the model decays**

Explorative vs Exploitive - 2[1)]



1) Reker, D.; Schneider, G. *Drug Discovery Today* **2015**, *20*, 458.

Explorative vs Exploitive - 3[1)]



1) Reker, D.; Schneider, G. *Drug Discovery Today* **2015**, *20*, 458.

Explorative Strategy

Enamine HTS compound collection (1465960 compounds)

+

CXCR4 ligand data (IC_{50}, K_i) from ChEMBL19

Training

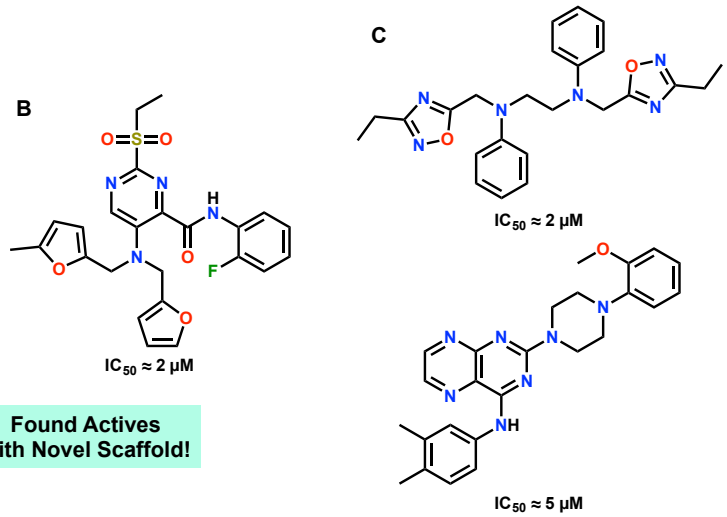
Random Forest

scores the compounds according to
1) affinity (w_1 = 2)
2) variance (w_2 = 1)
3) similarity (w_3 = –1)

Chose the best-score compound and other compounds which the Random Forest deemed "similar" (a batch of 30 compounds)

*** Variance: opposite of "uncertainty"; Similarity: opposite of outlier measure.**

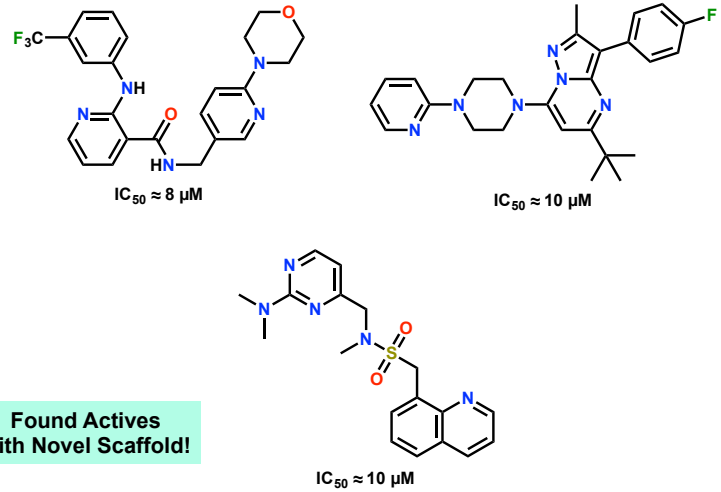
Explorative Strategy - 1st Iteration - 1



Found Actives with Novel Scaffold!

1) Reker, D.; Schneider, P.; Schneider, G. *Chem. Sci.* **2016**, *7*, 3919.

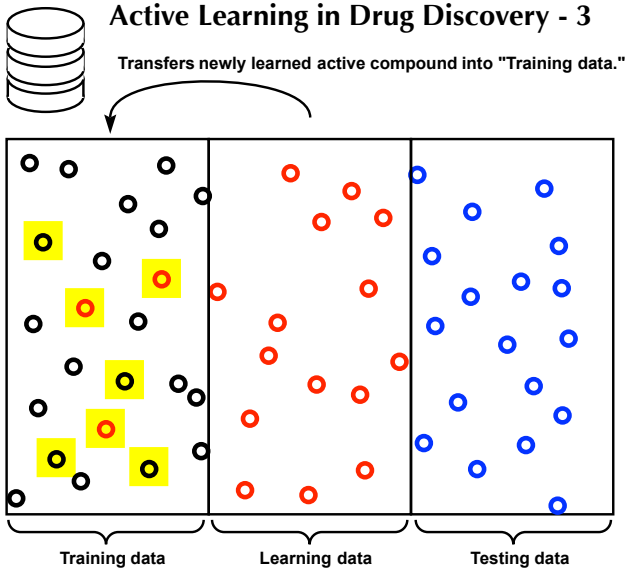
Explorative Strategy - 1st Iteration - 2



Found Actives with Novel Scaffold!

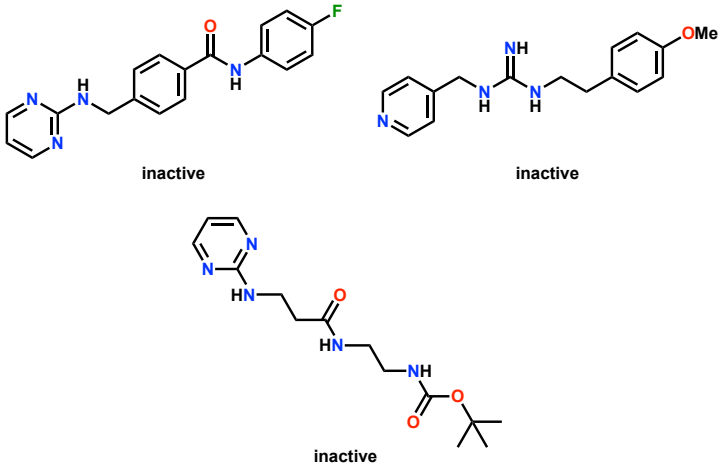
1) Reker, D.; Schneider, P.; Schneider, G. *Chem. Sci.* **2016**, *7*, 3919.

Active Learning in Drug Discovery - 3



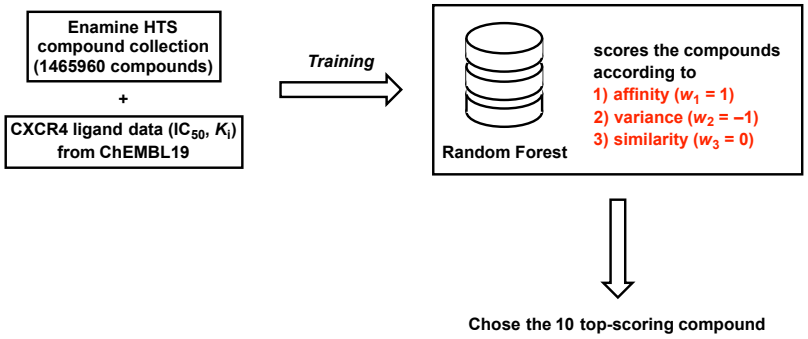
1) Reker, D.; Schneider, G. *Drug Discovery Today* **2015**, *20*, 458.

Explorative Strategy - 2nd Iteration



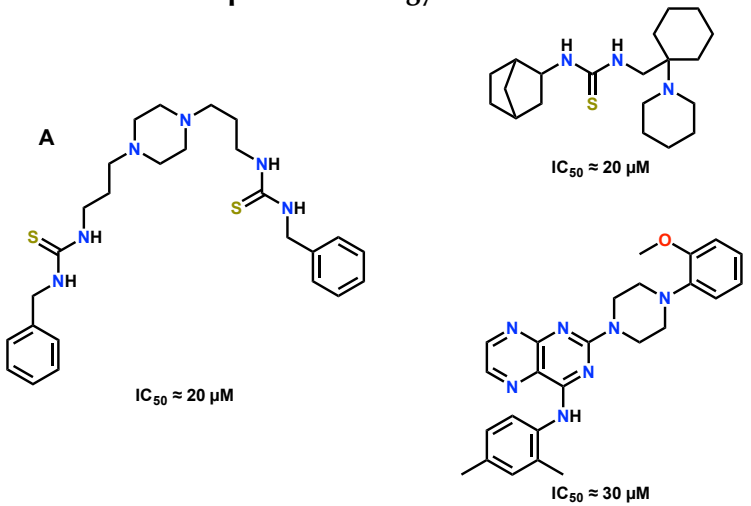
1) Reker, D.; Schneider, G. *Drug Discovery Today* **2015**, *20*, 458.

Exploitive Strategy



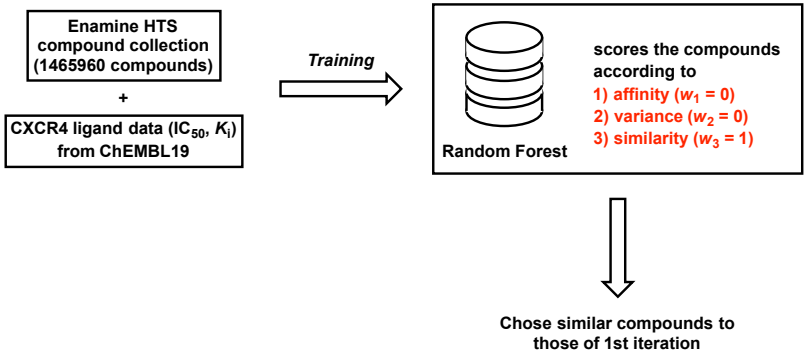
* Variance: opposite of "uncertainty"; Similarity: opposite of outlier measure.

Exploitive Strategy - Results



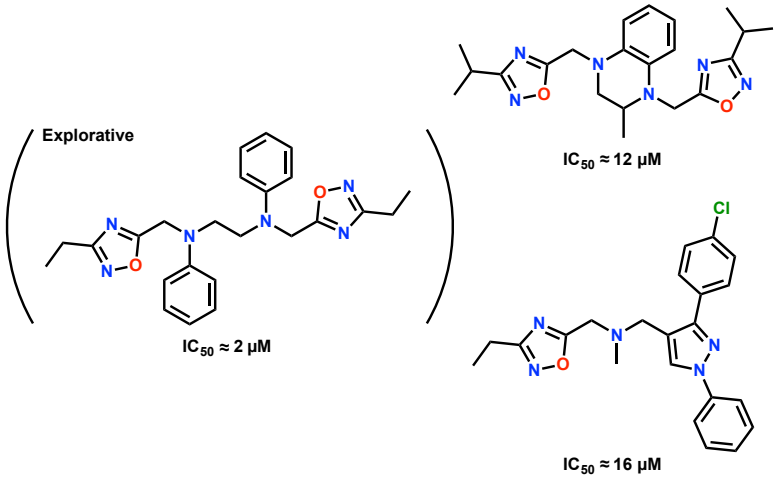
1) Reker, D.; Schneider, P.; Schneider, G. *Chem. Sci.* **2016**, *7*, 3919.

Hit Expansion



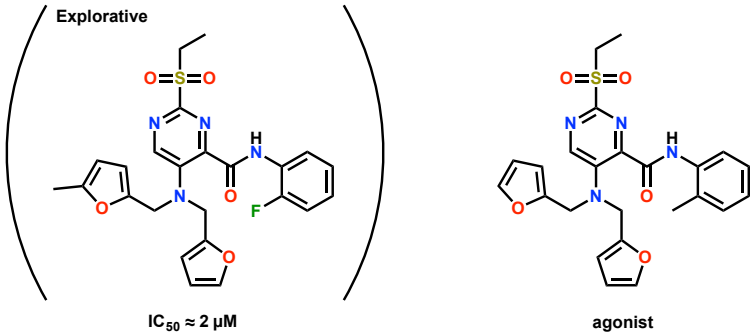
* Variance: opposite of "uncertainty"; Similarity: opposite of outlier measure.

Hit Expansion - Results - 1



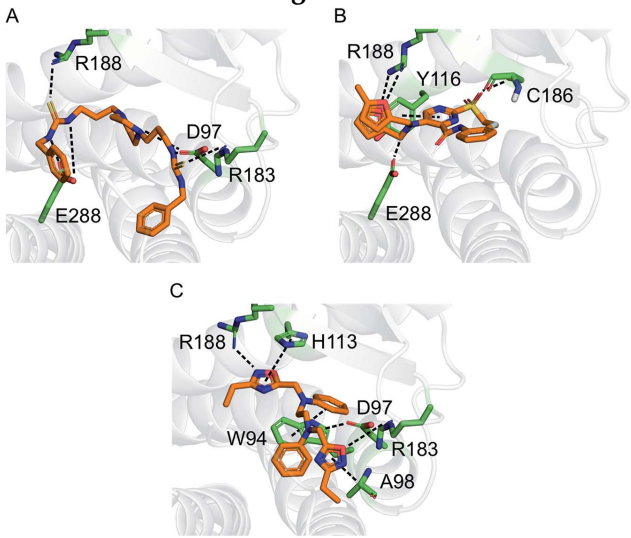
1) Reker, D.; Schneider, P.; Schneider, G. *Chem. Sci.* **2016**, *7*, 3919.

Hit Expansion - Results - 2

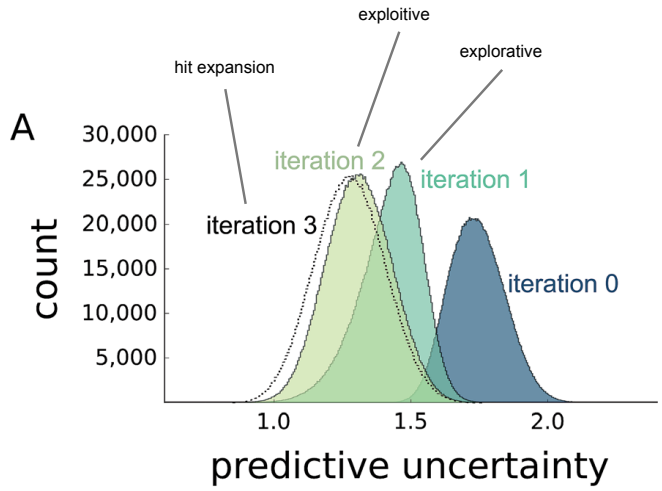


1) Reker, D.; Schneider, P.; Schneider, G. *Chem. Sci.* **2016**, *7*, 3919.

Docking Simulation



Decrease of Uncertainty



Summary

Active Learning for Drug Discovery:

- Needs "Training Data" i.e. substantial precedent experimental results.

- Cannot distinguish agonist and antagonist.

But

• Random forest enables efficient prediction.

• Offers brand new type of lead compounds for difficult target.

Summary

Active Learning for Drug Discovery:

- Needs “Training Data” i.e. substantial precedent experimental results.

- Cannot distinguish agonist and antagonist.

But

• Random forest enables efficient prediction.

• Offers brand new type of lead compounds for difficult target.

• Especially powerful for salvaging false-negatives.

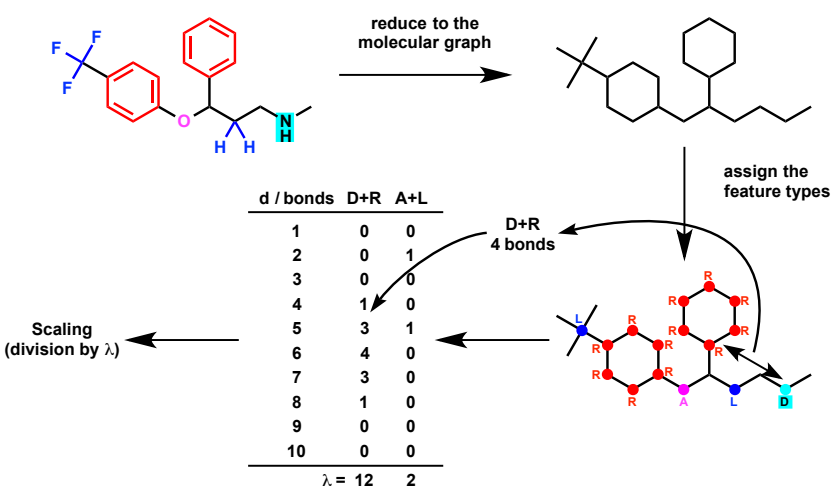
Appendix

Parameter Optimization for Selection

tested parameters			model quality			
w_1	w_2	w_3	MSE	SC	AF	ALC
0.00	0.00	1.00	1.58	90.00	7.37	0.09
0.00	1.00	0.00	1.63	88.00	7.36	0.16
0.00	0.71	0.71	1.62	91.00	7.37	0.18
0.00	0.45	0.89	1.62	91.00	7.37	0.18
0.00	0.24	0.97	1.62	91.00	7.37	0.18
0.00	0.89	0.45	1.63	92.00	7.33	0.17
0.00	0.97	0.24	1.63	91.00	7.36	0.17
1.00	0.00	0.00	0.70	86.00	7.38	0.13
0.71	0.00	0.71	1.04	91.00	7.34	0.13
0.45	0.00	0.89	1.04	91.00	7.34	0.13
0.24	0.00	0.97	1.04	91.00	7.34	0.13
0.71	0.71	0.00	1.64	90.00	7.38	0.17
0.58	0.58	0.58	1.62	91.00	7.40	0.20
0.41	0.41	0.82	1.62	91.00	7.40	0.20
0.24	0.24	0.94	1.62	91.00	7.40	0.20
0.45	0.89	0.00	1.63	92.00	7.43	0.17
0.41	0.82	0.41	1.63	92.00	7.38	0.19
0.33	0.67	0.67	1.62	92.00	7.41	0.19
0.22	0.44	0.87	1.62	92.00	7.41	0.19
0.24	0.97	0.00	1.65	91.00	7.41	0.18
0.24	0.94	0.24	1.63	91.00	7.40	0.21
0.22	0.87	0.44	1.64	91.00	7.40	0.21
0.17	0.70	0.70	1.61	91.00	7.35	0.22
0.89	0.00	0.45	1.04	91.00	7.34	0.13
0.89	0.45	0.00	1.63	90.00	7.47	0.17
0.82	**0.41**	**0.41**	**1.59**	**92.00**	**7.56**	**0.17**
0.67	0.33	0.67	1.59	92.00	7.56	0.17
0.44	0.22	0.87	1.59	92.00	7.56	0.17
0.67	0.67	0.33	1.63	90.00	7.35	0.20
0.44	0.87	0.22	1.66	88.00	7.37	0.18
0.97	0.00	0.24	1.04	91.00	7.34	0.13
0.97	0.24	0.00	1.43	87.00	7.64	0.15
0.94	0.24	0.24	1.49	88.00	7.50	0.15
0.87	0.22	0.44	1.49	88.00	7.50	0.15
0.70	0.17	0.70	1.49	88.00	7.50	0.15
0.87	0.44	0.22	1.59	92.00	7.56	0.17
0.70	0.70	0.17	1.66	91.00	7.40	0.19

Table S1: Parameter optimization for balanced active learning model on CXCR4 time series data. Compounds were sorted according to publication year and the first 33% served as training data. The remaining 66% were randomly split into learning set and external test set. Active learning was run for 50 iterations. The parameters are the weights of the weighted average for the selection function, balancing the influence of the predicted affinity (w1), the uncertainty about that prediction (w2), and the random forest outlier measure calculated for that compound (w3). After the active learning model was trained, we calculated four different evaluation criteria: (i) the reduction of the mean squared error (MSE) on a randomly selected test set (ii) the number of scaffolds investigated (scaffold count SC) (iii) the average affinity of the picked compounds (AF) (iv) the area under the learning curve (ALC). The selected parameter set and the associated model quality is shown in bold.

Principle of CATS descriptor calculation



Random Forest - Best Choice of m_{try} [1)]

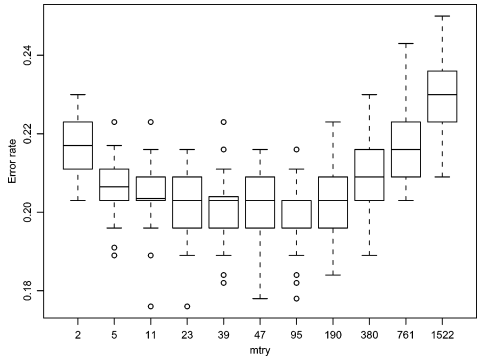
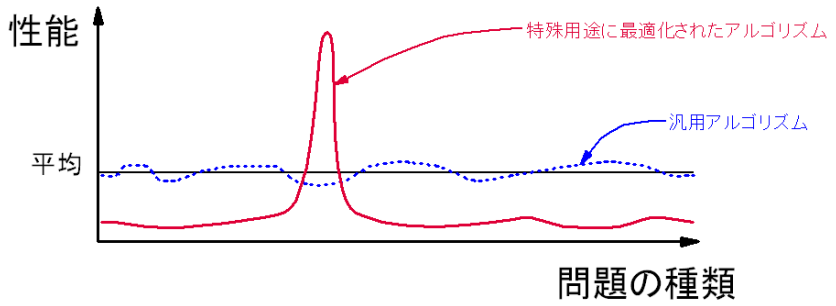


Figure 1. Boxplots of 50 5-fold cross-validation test error rates at various values of m_{try} for the P-gp data set. Horizontal lines inside the boxes are the median error rates. The plot suggests that m_{try} is optimal near 39, the default value, and that the performance is similar for values ranging from 11 to 190.

1) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1947.

No-Free-Lunch Theorem[1)]



1) https://ja.wikipedia.org/wiki/ノーフリーランチ定理