UNIVERSITY OF CAMBRIDGE

Maximizing Gain in HTS Screening Using Conformal Prediction

Ulf Norinder, Fredrik Svensson, Avid Afzal and Andreas Bender

fs447@cam.ac.uk

COPA 2017, Stockholm 16th June

Outline

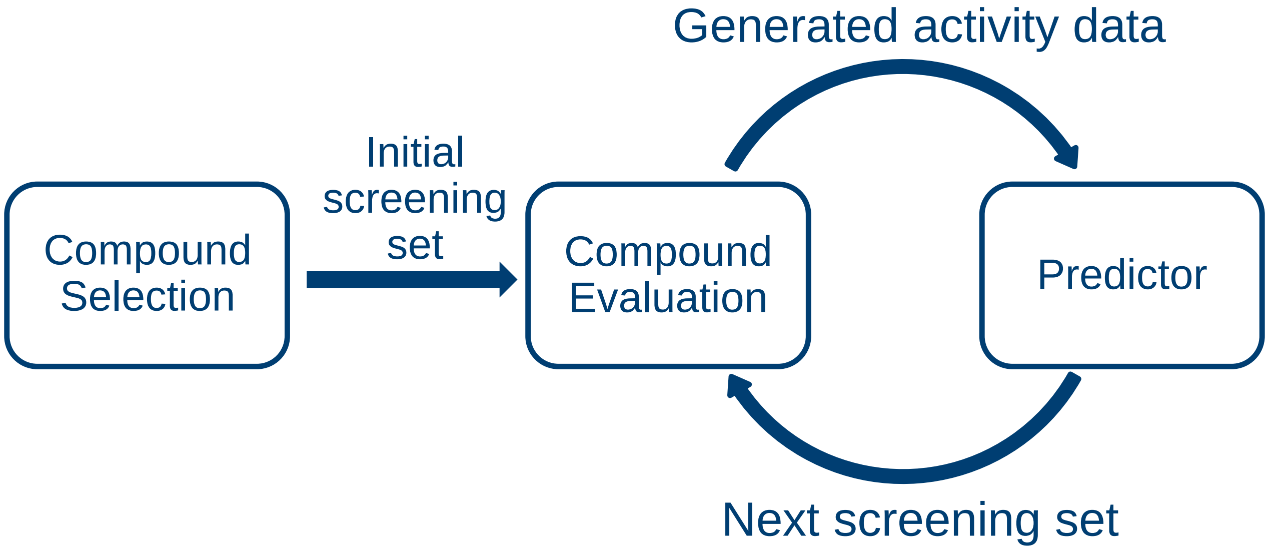
- Introduction to screening
- Datasets
- What confidence do we need for screening?
- Gain-cost function
- Results

High throughput screening

- Screening of large compound collections have been the backbone of early stage drug discovery for many years
  - However, these approaches are often costly
  - New focus on phenotypic assays increases the cost even further

Predicting assay outcomes

- Previous work has shown that iterative screening can improve the efficiency of large scale screening



S. Paricharak et al., **Analysis of Iterative Screening with Stepwise Compound Selection Based on Novartis In-house HTS Data,** *ACS Chem. Biol.*, 2016, 11 (5),1255–1264
F. Svensson et al., **Improving Screening Efficiency through Iterative Screening Using Docking and Conformal Prediction,** *J. Chem. Inf. Model.*, 2017, 57 (3), 439–444

Imbalanced data

- Screening data is often highly imbalanced
- Mondrian conformal predictors have been shown to handle imbalanced data very well

T. Löfström et al. **Bias reduction through conditional conformal prediction**, *Intell. Data Anal.* 2015, 19, 1355–1375
U. Norinder and S. Boyer. **Binary classification of imbalanced datasets using conformal prediction**. *J. Mol. Graph. Model. 2017,* 72, 256-265

Datasets

- Collected from PubChem

PubChem AID	Active	Inactive	%Active	Target/Readout
868	3,545	194,381	1.8	RAM network signalling
1460	1,189	47,025	2.5	tau fibrillization
2314	36,955	295,303	12	Stabilization of luciferase activity
2551	16,638	269,830	5.8	ROR gamma activity

Modelling

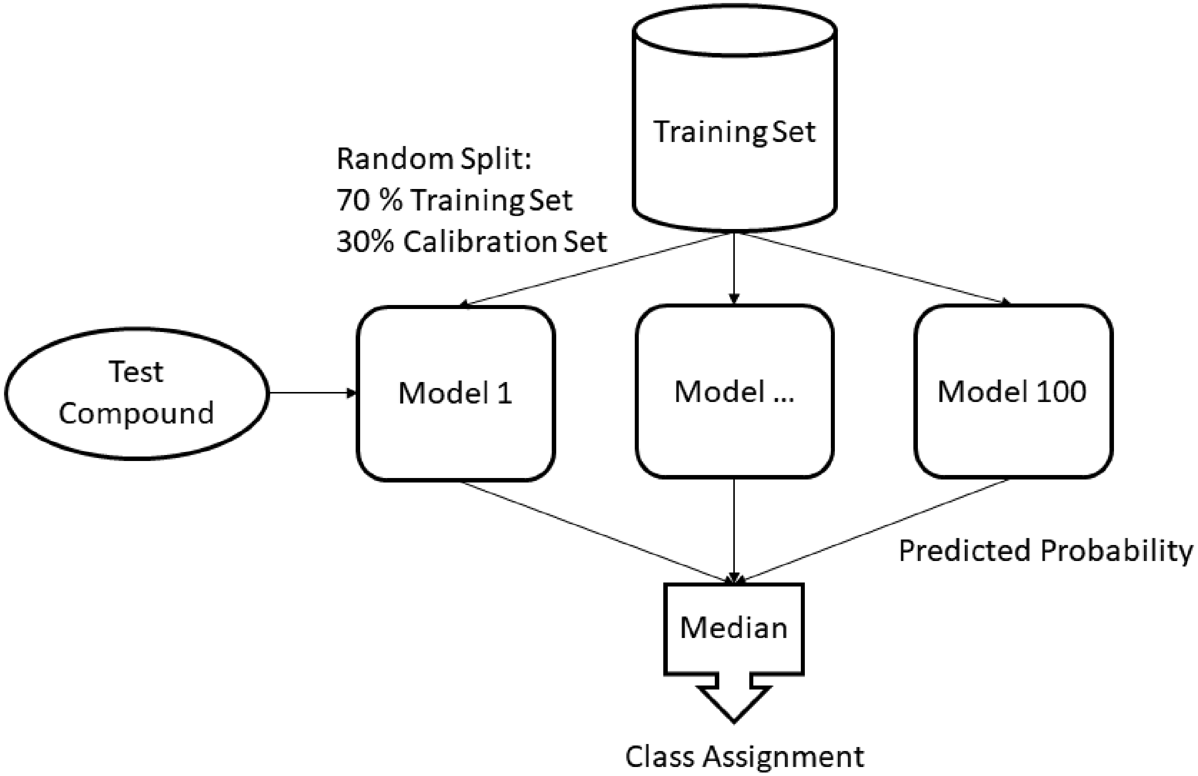
- RDKit molecular descriptors (97) and Morgan fingerprints (4,096 bits) used as features
- Modelling was done using Python, scikit-learn, and the nonconformist package
- Random forests (500 trees) were used as the underlying models
- Default values used for other parameters

Aggregated conformal predictors

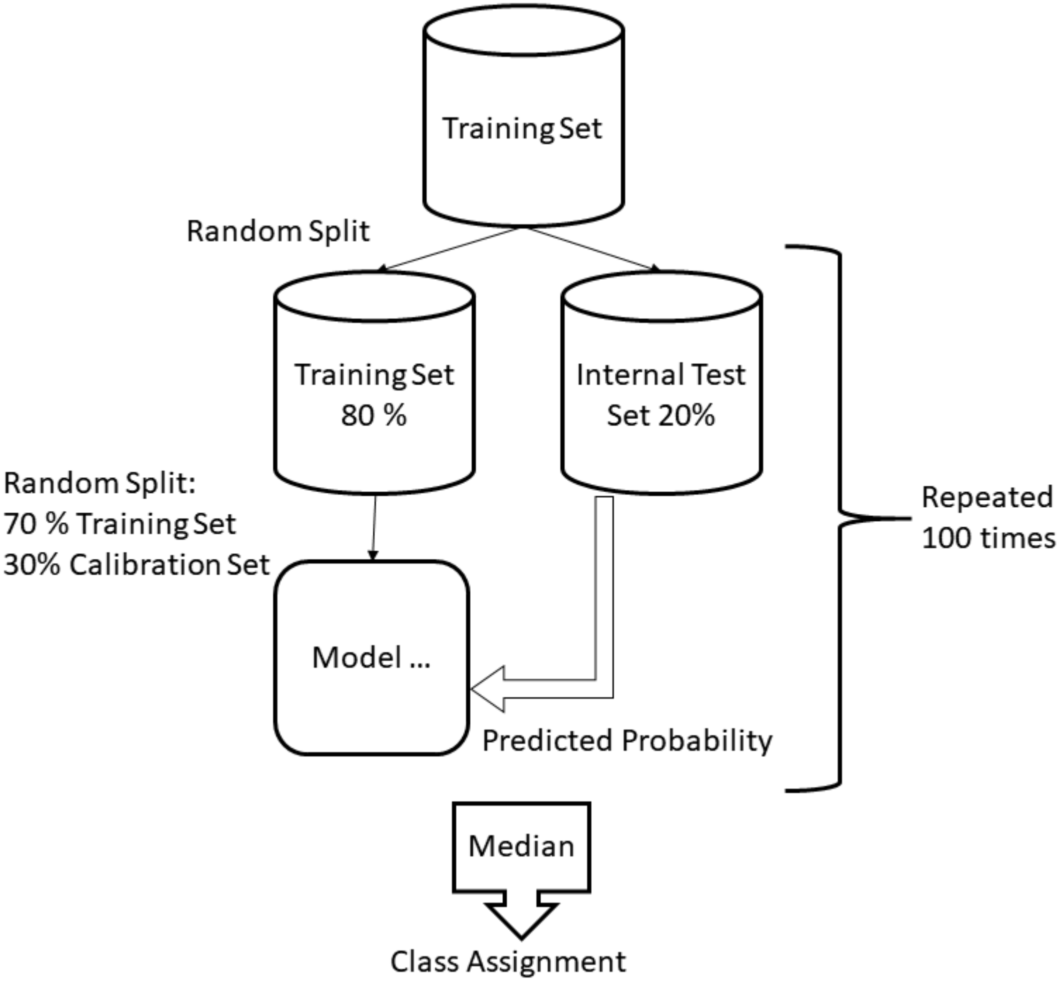
- 100 models with random split for proper training and calibration
  - 70% training, 30% calibration

L. Carlsson, M. Eklund, and U. Norinder. **Aggregated conformal prediction.** In L. Iliadis, I. Maglogiannis, H. Papadopoulos, S. Sioutas, and C. Makris, editors, Artificial Intelligence Applications and Innovations: AIAI 2014 Workshops: CoPA, MHDW, IIVC, and MT4BD, Rhodes, Greece, September 19-21, 2014. Proceedings, pages 231–240, Berlin, Heidelberg, 2014. Springer International Publishing

External validation



Internal validation



Training Set

Random Split

Training Set 80 %

Internal Test Set 20%

Random Split:
70 % Training Set
30% Calibration Set

Model ...

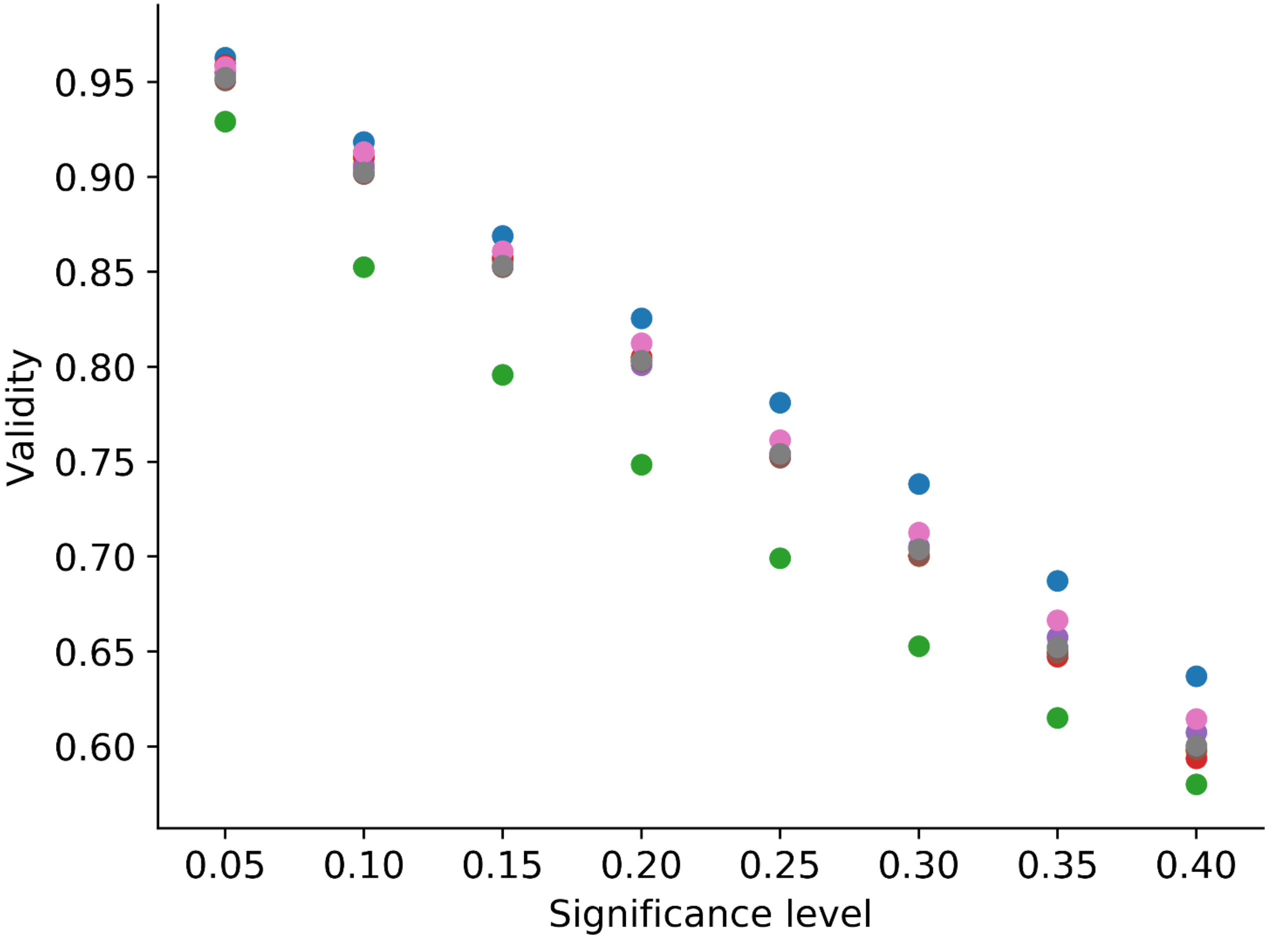
Predicted Probability

Repeated 100 times

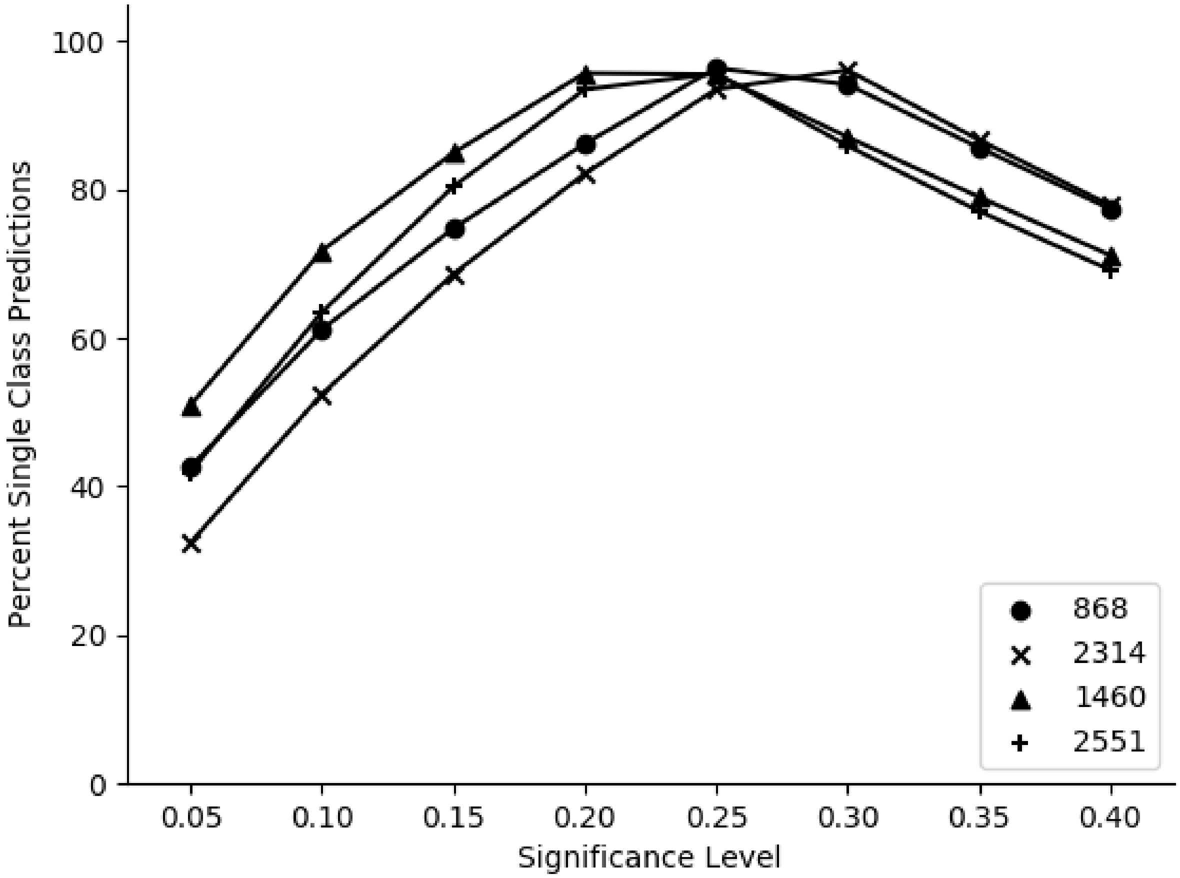
Median

Class Assignment

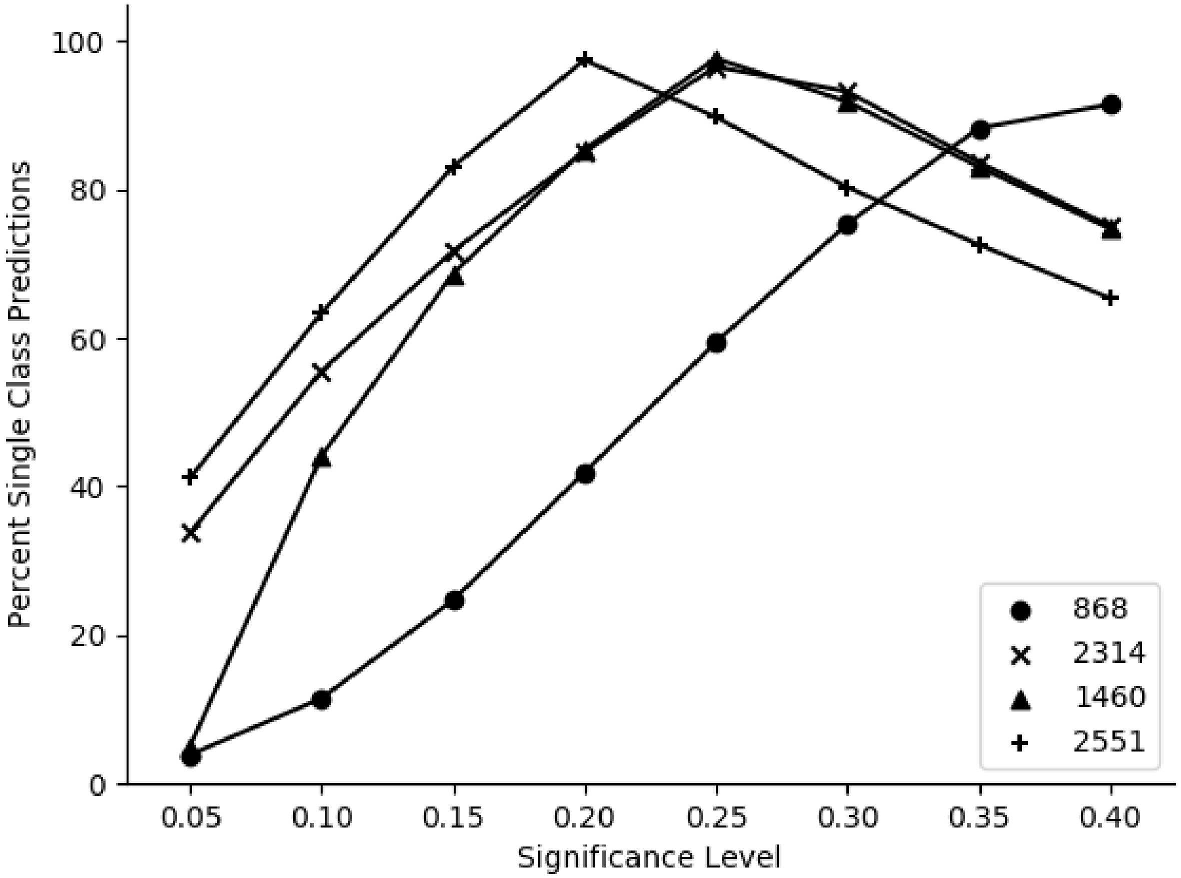
The generated models are valid



Efficiency (physico-chemical descriptors)



Efficiency (fp descriptors)



What certainty do we need?

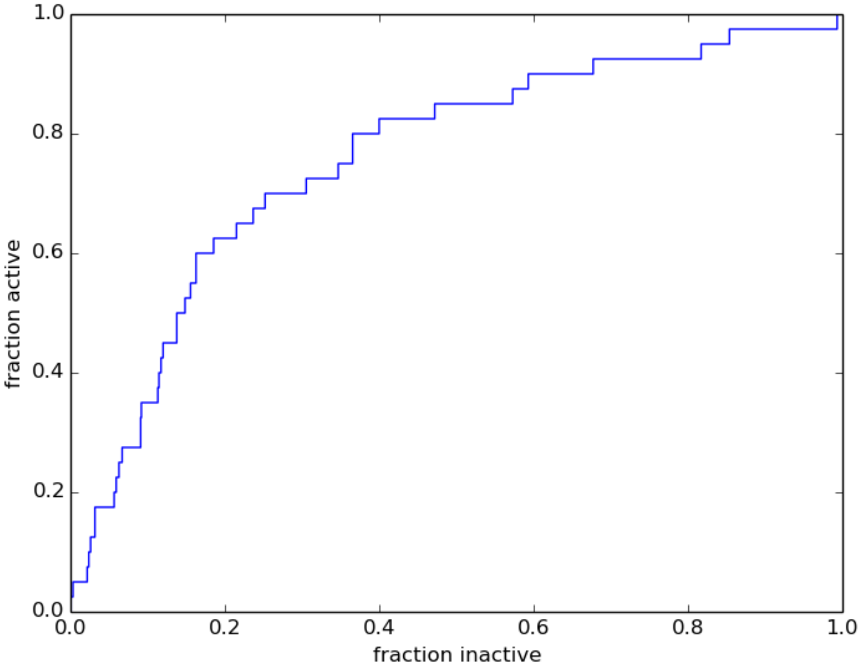
- How can we define the optimal confidence level for screening outcome predictions?

Or more generally:

- What is the optimal number of compounds to screen?

What about traditional performance metrics?

- Enrichment factor and related metrics do not provide an answer to how many compounds to screen



How to decide on the optimal fraction to screen?

Gain Cost of screening

- Rudimentary gain-cost function defined:

$$gain= \sum_{i=1}^{ntra}(gc) - \sum_{i=1}^{ntr}(fc + sdc) + \sum_{i=1}^{ntesta}(gc) - \sum_{i=1}^{ntest}(fc + sdc)$$

where

- gc: gain per hit compound
- fc: compound purchase and handling cost
- sdc: screen dependent cost
- ntr: number of training compounds
- ntra: number of active training compounds
- ntest: number of test set compounds
- ntesta: number of active test set compounds

Assigning cost and gain

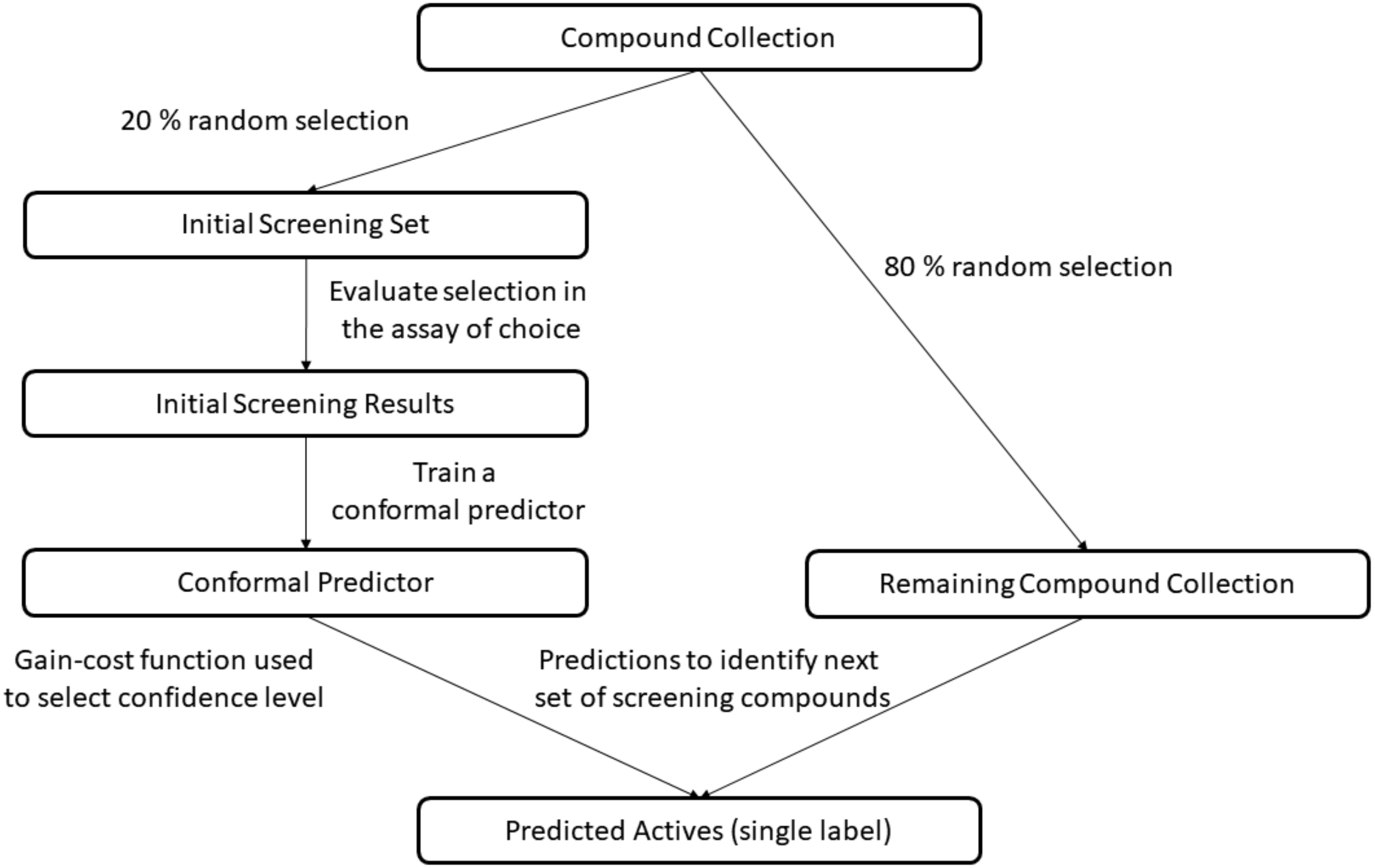
- Based on discussions with screening experts we decided on:
  - A fixed cost of 2 per compound
  - An assay dependent cost of 4, 8, and 12

Thanks to Dr. Anna-Lena Gustavsson, Chemical Biology Consortium Sweden, CBCS, Karolinska Institutet, SciLifeLab, Stockholm, for fruitful discussions on the design of the gain-cost function.

Assigning cost and gain

- We defined a gain that approximately balances the cost for the HTS data
  - This was found to be a gain of 400
- Overall we applied three different cost:gain ratios:
  - 6:400
  - 10:400
  - 14:400

Workflow



Compound Collection

20 % random selection

Initial Screening Set

Evaluate selection in the assay of choice

Initial Screening Results

Train a conformal predictor

Conformal Predictor

80 % random selection

Remaining Compound Collection

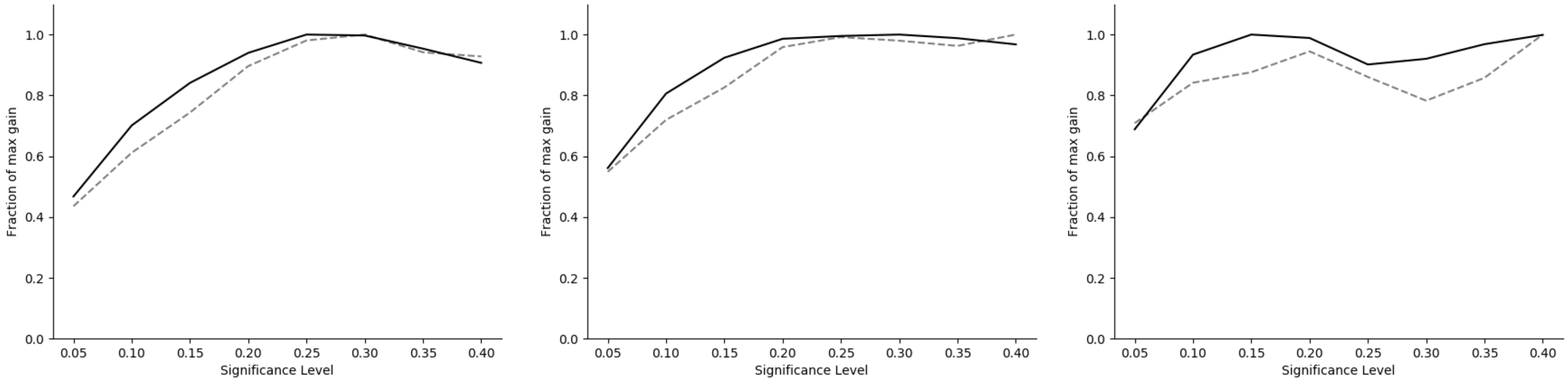
Gain-cost function used to select confidence level

Predictions to identify next set of screening compounds

Predicted Actives (single label)

Gain-Cost evaluation AID868

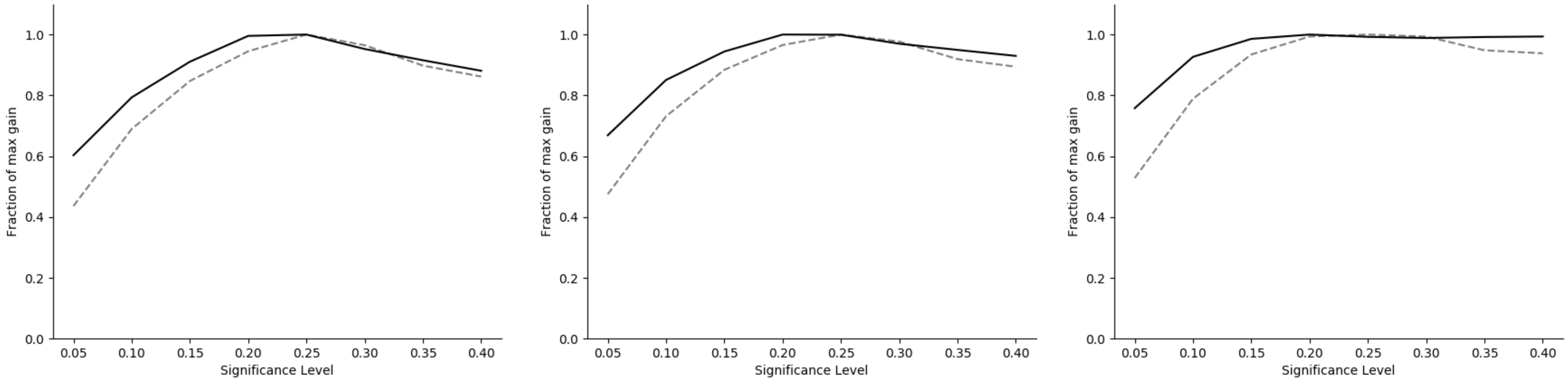
- Physico-chemical descriptors
- Dashed line internal validation, solid line test data



Cost: 6 10 14

Gain-Cost evaluation AID1460

- Physico-chemical descriptors
- Dashed line internal validation, solid line test data

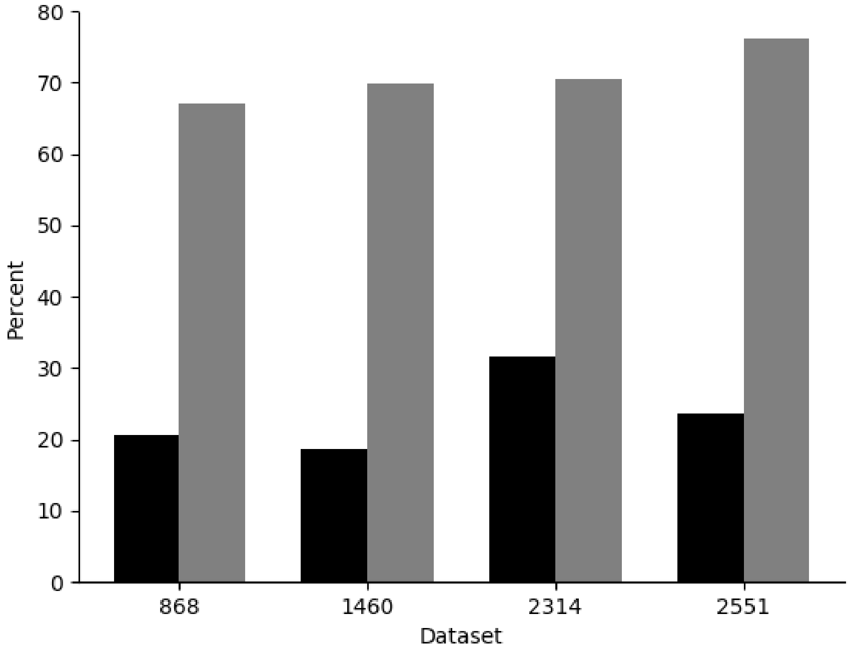


Cost: 6 10 14

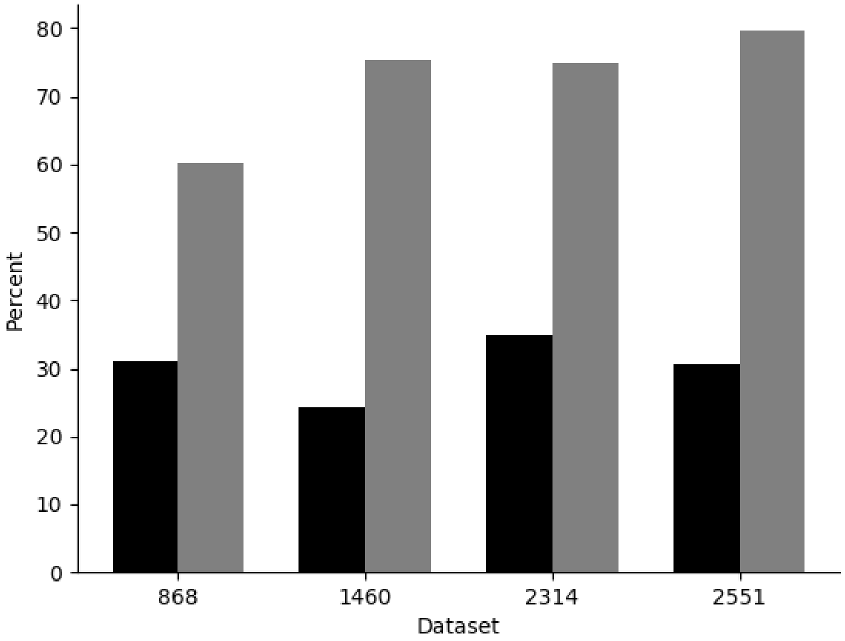
Train test correspondence

- The internal validation of the training data was very successful in identifying the test set optimum
  - 7/12 identified the optimal confidence level
  - For the remaining the average deviation from maximum gain was 1% (physio-chemical descriptors)
  - Fore some datasets the overall gain from the whole screening set is greater than screening the predicted actives
    - Also these cases were correctly predicted by the internal validation

Percent screened and percent actives fund



Physico-chemical



Fingerprints

Black bars = percent screened
Grey bars = percent actives found

Conclusions

- A gain-cost was used to find the optimal significance level for activity prediction in a HTS setting
- Evaluation on the training data was highly indicative of the result using the test data

Future work

- Expand the result to additional datasets
    - 8 more datasets underway
- Evaluate more complex gain-cost functions

Acknowledgments

UNIVERSITY OF CAMBRIDGE

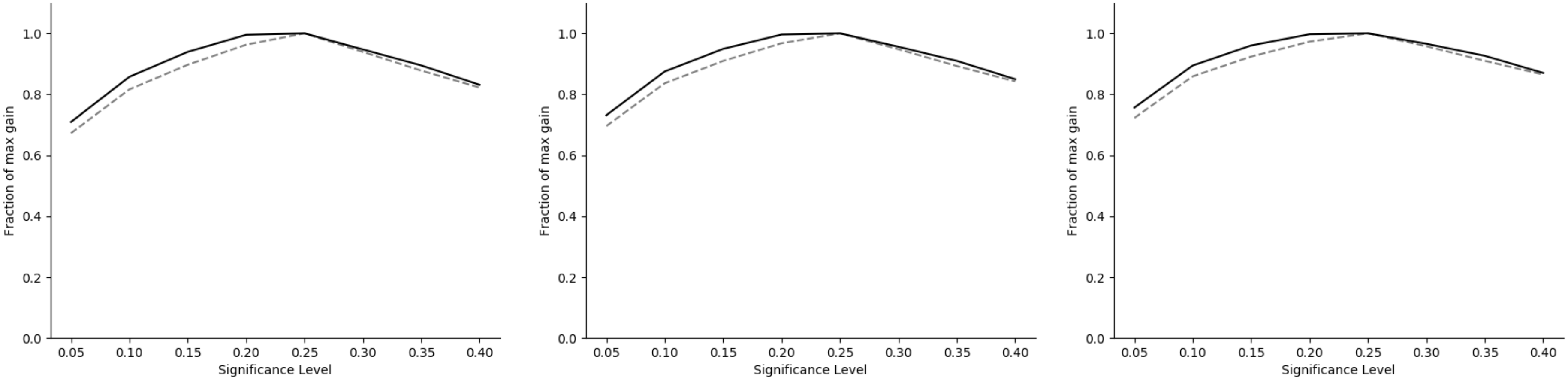
Dr Andreas Bender
Dr Avid Afzal

SWETOX Swedish Toxicology Sciences Research Center

Dr Ulf Norinder

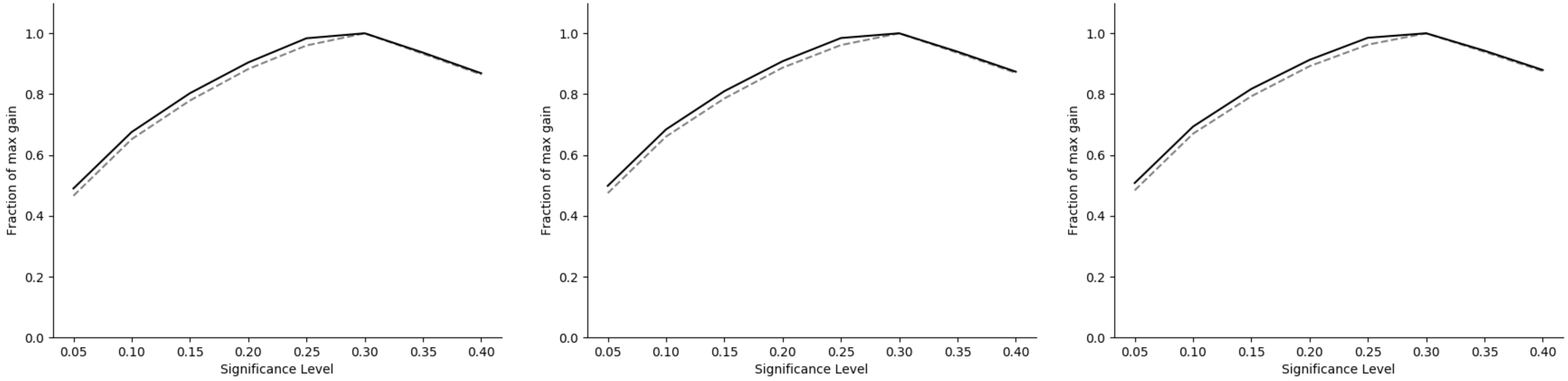
IOTA Pharmaceuticals

Gain-Cost 2551 (physico-chemical descriptors)



Cost: 6 10 14

Gain-Cost 2314 (physico-chemical descriptors)



Cost: 6 10 14