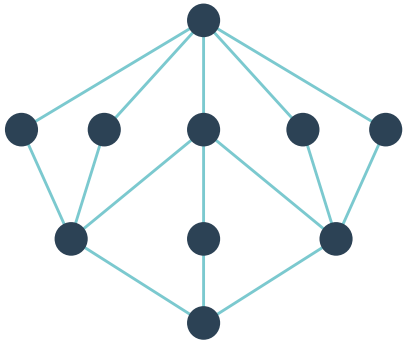
EVARISTE TECHNOLOGIES

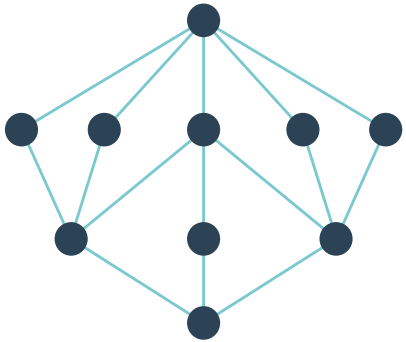


Multiparameter optimisation using RDKit and scipy: what's the chance of success?

Nicholas Firth - CSO

RDKit UGM 19th September 2017

EVARISTE
TECHNOLOGIES



Quantitative Drug Design @ Evariste

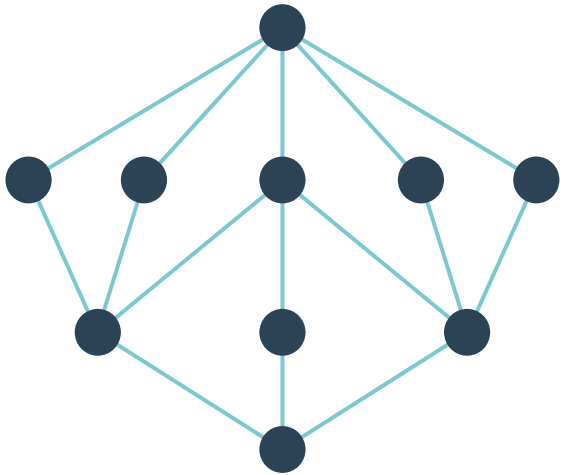
Nicholas Firth - CSO

RDKit UGM 19th September 2017

OVERVIEW

EVARISTE TECHNOLOGIES

- Evariste
- Prediction
- Design
- Selection



EVARISTE
TECHNOLOGIES

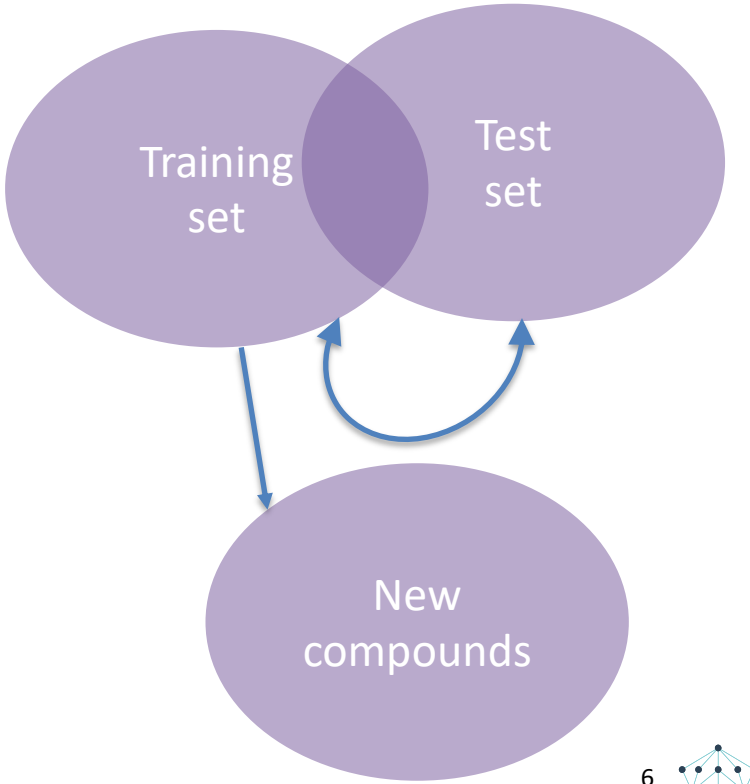
- Founded by Oliver Watson
- Isidro Cortes and I joined
- QSAR modelling
- *de novo* design (generative chemistry)
- Selecting optimal subsets

COMPANY THESIS

EVARISTE TECHNOLOGIES

We can increase the speed and quality of hit to lead design programs by using quantitative methods to; **predict**, **design**, and **select** molecules

- Applicability domains are useful warnings
- We frequently want to make predictions outside of these
- Some algorithms are better suited for extrapolation
- Can we say which is better in the drug discovery context



Discovering Highly Potent Molecules from an Initial Set of Inactives Using Iterative Screening

Isidro Cortés-Ciriano*†, Nicholas C. Firth‡§, Andreas Bender†, and Oliver Watson§

† Centre for Molecular Informatics, Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, United Kingdom

‡ Centre for Medical Image Computing, Department of Computer Science, UCL, London WC1E 6BT, United Kingdom

§ Evariste Technologies Ltd, Goring on Thames RG8 9AL, United Kingdom

J. Chem. Inf. Model., Article ASAP

DOI: 10.1021/acs.jcim.8b00376

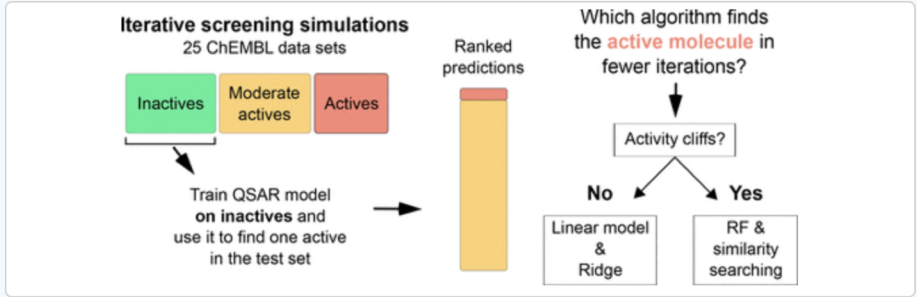
Publication Date (Web): August 21, 2018

*isidrolauscher@gmail.com.

Cite this: *J. Chem. Inf. Model.* XXXX, XXX, XXX-XXX

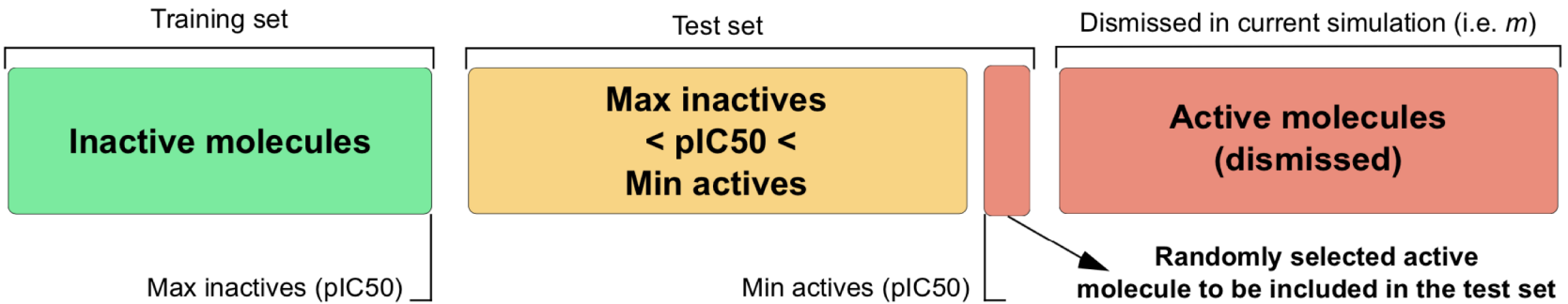
RIS Citation GO

Abstract



PREDICTION

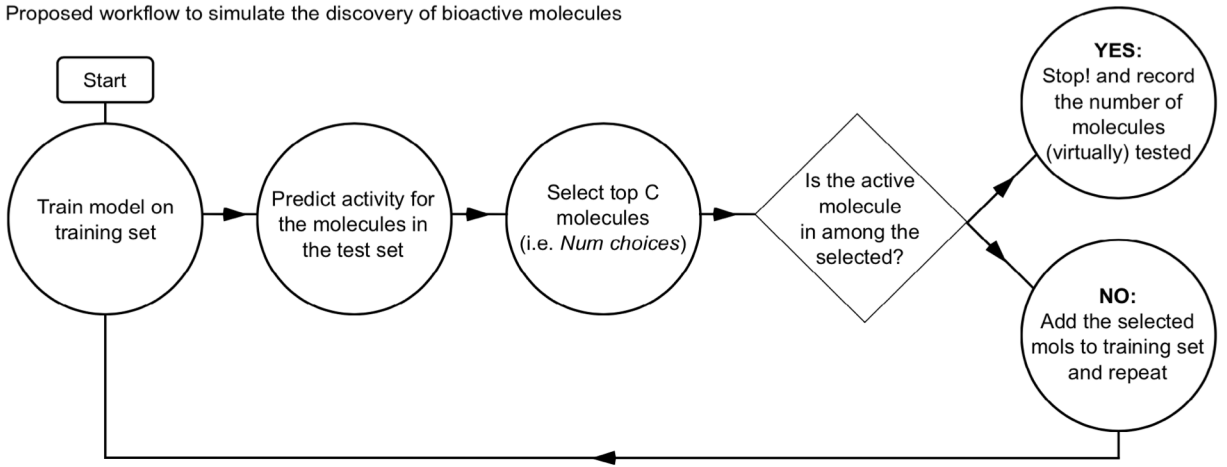
EVARISTE TECHNOLOGIES



PREDICTION

EVARISTE TECHNOLOGIES

- 25 ChEMBL targets
- 9 algorithms
- ‘ECFP4’ Fingerprints
- 250 repetitions
- In each repeat hyperparameter optimization was done



ChEMBL

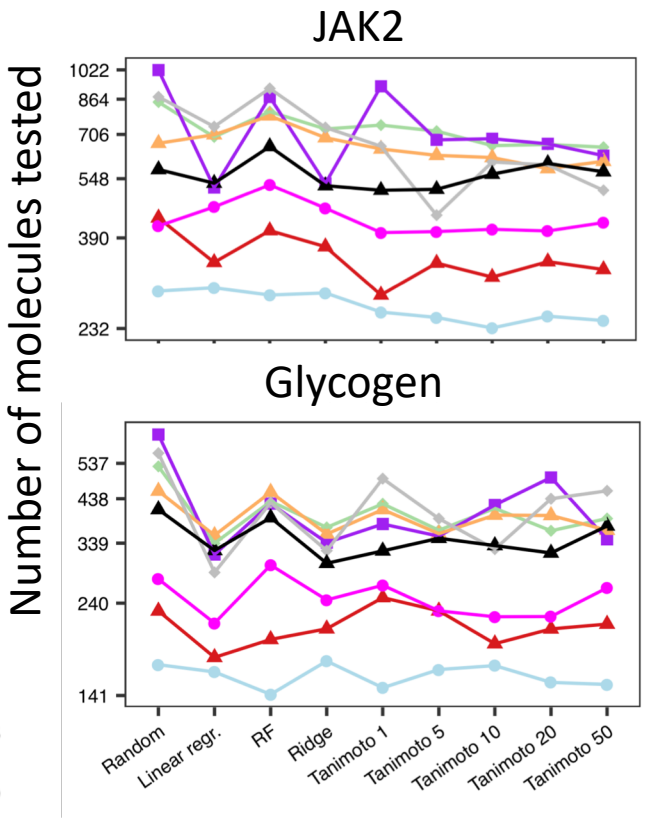
RDKit

Open-Source Cheminformatics and Machine Learning

scikit learn

EVARISTE TECHNOLOGIES

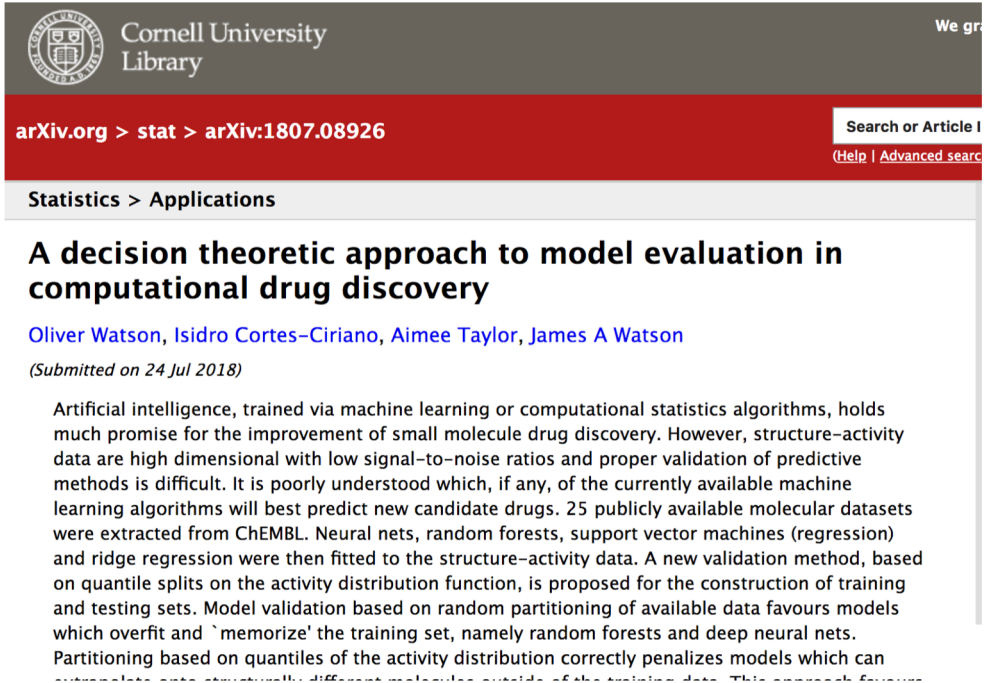
- Random Forest and similarity searching not much better than random
- Ridge Regression often much faster than Random Forest
- Throughout the study we noticed that MAE rarely agreed with our results



PREDICTION

EVARISTE TECHNOLOGIES

- We noticed that different cutoffs could affect the performance of an algorithm
- When should we be using each algorithm?
- We wanted to do this in a more formal way



Cornell University Library

We gr

arXiv.org > stat > arXiv:1807.08926

Search or Article I

(Help | Advanced searc

Statistics > Applications

A decision theoretic approach to model evaluation in computational drug discovery

Oliver Watson, Isidro Cortes-Ciriano, Aimee Taylor, James A Watson

(Submitted on 24 Jul 2018)

Artificial intelligence, trained via machine learning or computational statistics algorithms, holds much promise for the improvement of small molecule drug discovery. However, structure-activity data are high dimensional with low signal-to-noise ratios and proper validation of predictive methods is difficult. It is poorly understood which, if any, of the currently available machine learning algorithms will best predict new candidate drugs. 25 publicly available molecular datasets were extracted from ChEMBL. Neural nets, random forests, support vector machines (regression) and ridge regression were then fitted to the structure-activity data. A new validation method, based on quantile splits on the activity distribution function, is proposed for the construction of training and testing sets. Model validation based on random partitioning of available data favours models which overfit and `memorize' the training set, namely random forests and deep neural nets. Partitioning based on quantiles of the activity distribution correctly penalizes models which can

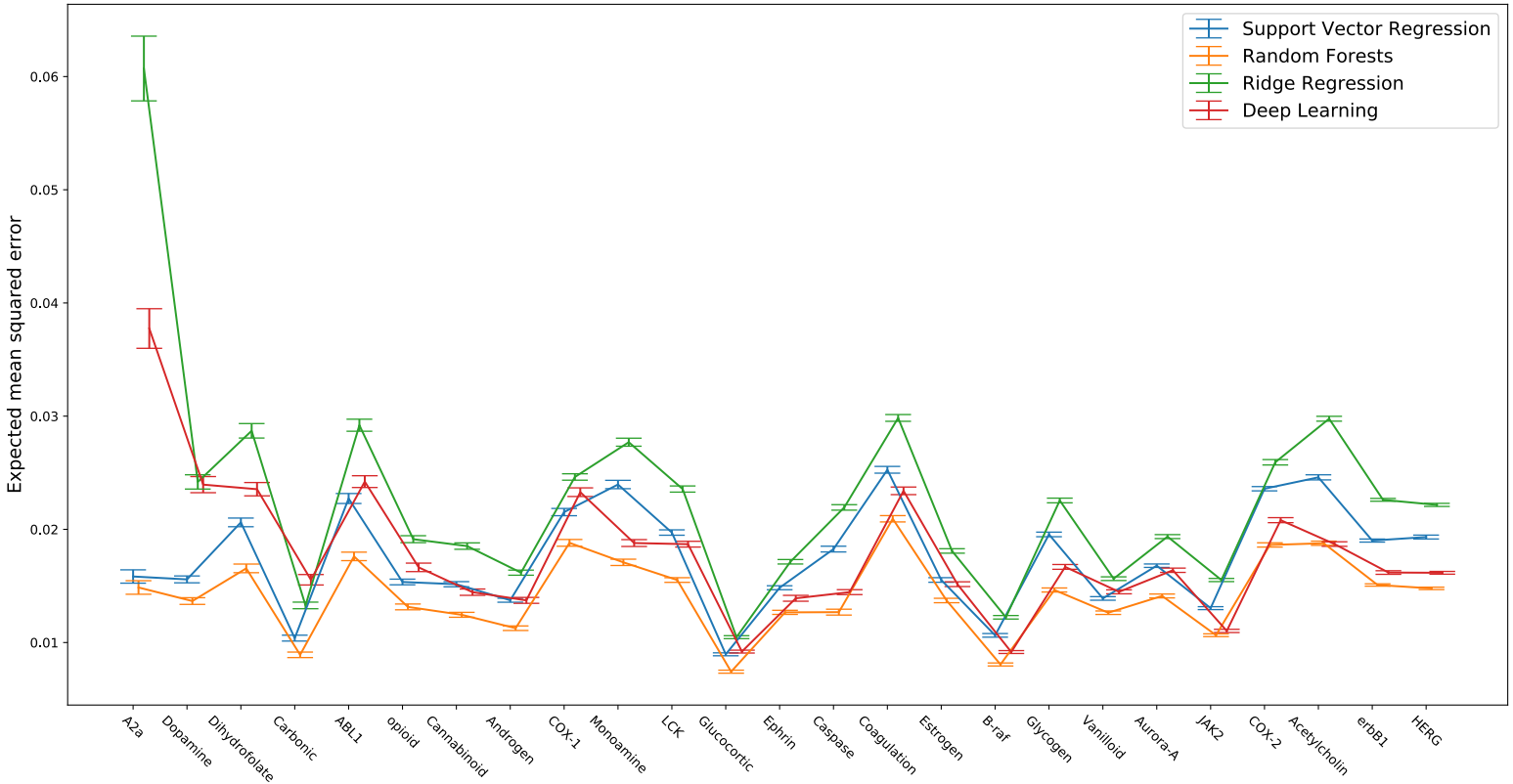
- Similar to the previous study we were interested in comparing this with ranks
- Formalized the loss function and the training set splits
- Used the same data and descriptors

$$L^{\gamma}_{\min} = \frac{1}{N_{test} - N_{\gamma}} \min_{j=N_{\gamma}..N_{test}} \mathrm{Rank}_{M_t}(\mathbf{x_j})$$

$$L^{\gamma}_{\mathrm{sum}} = \frac{1}{N_{\gamma}(N_{test} - N_{\gamma})} \left(\sum_{j=N_{\gamma}}^{N_{test}} \mathrm{Rank}_{M_t}(\mathbf{x_j}) - N_{\gamma}(N_{\gamma} - 1)/2 \right)$$

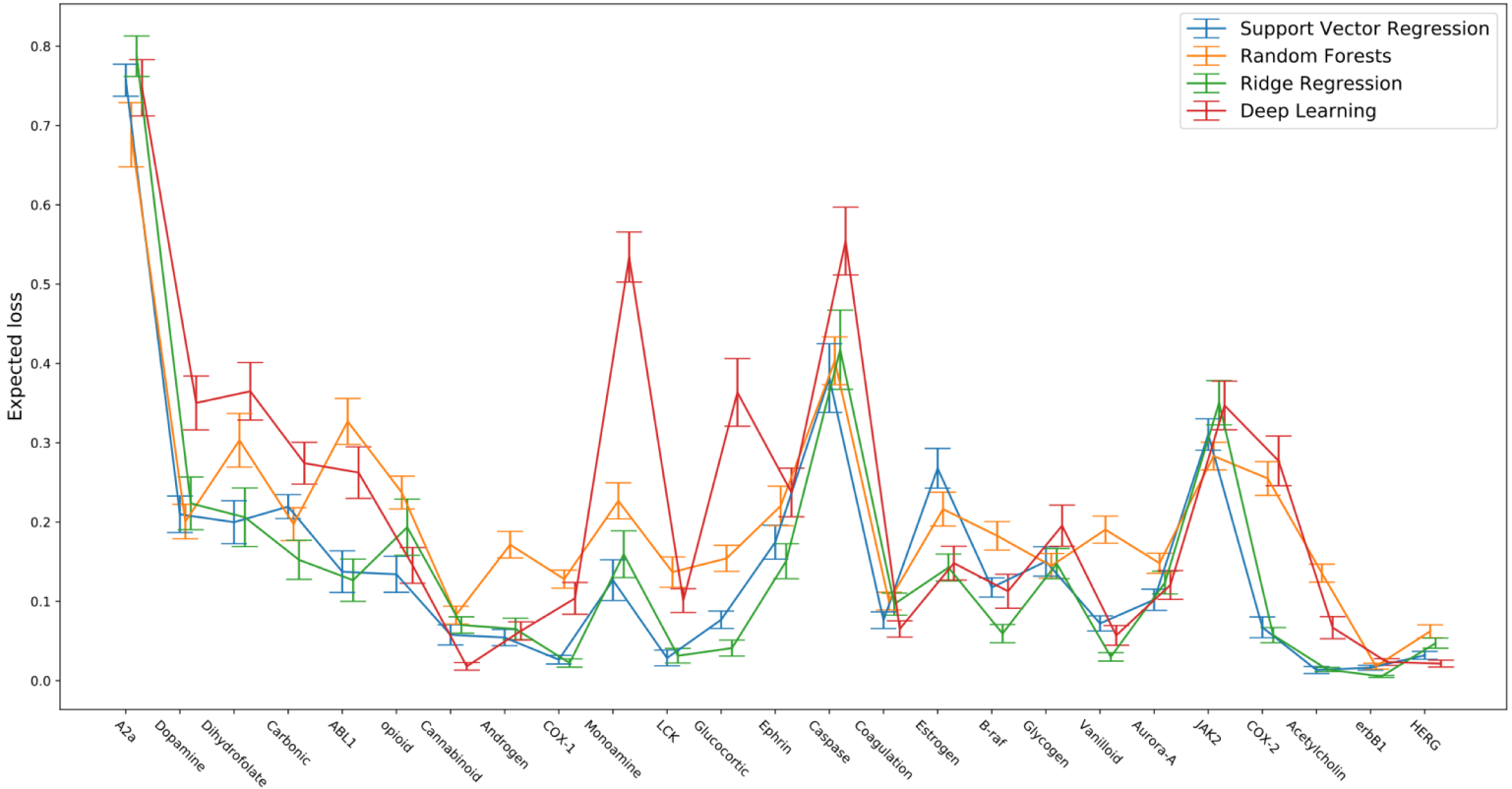
PREDICTION

EVARISTE TECHNOLOGIES



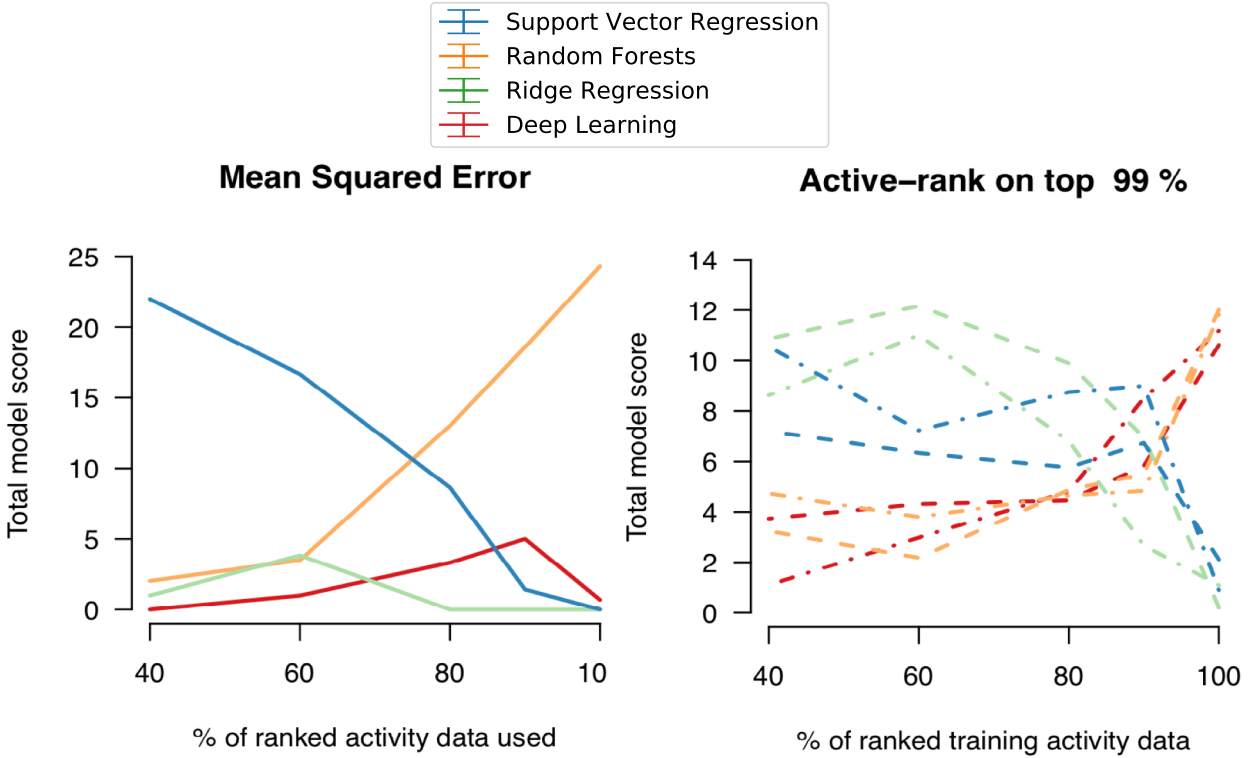
PREDICTION

EVARISTE TECHNOLOGIES



EVARISTE TECHNOLOGIES

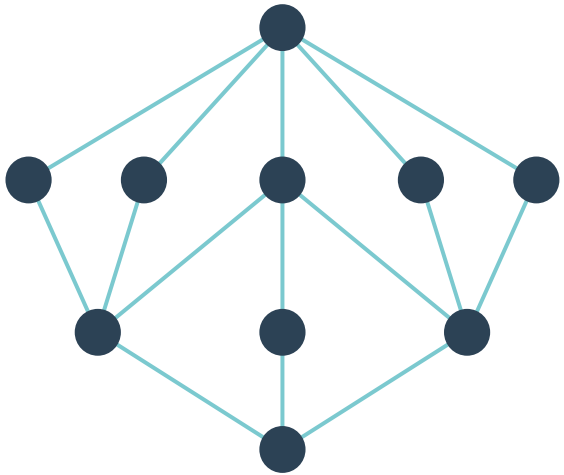
- If you don't have very active data, use a less complex model
- The more active the molecules you have the more complex the model you should use

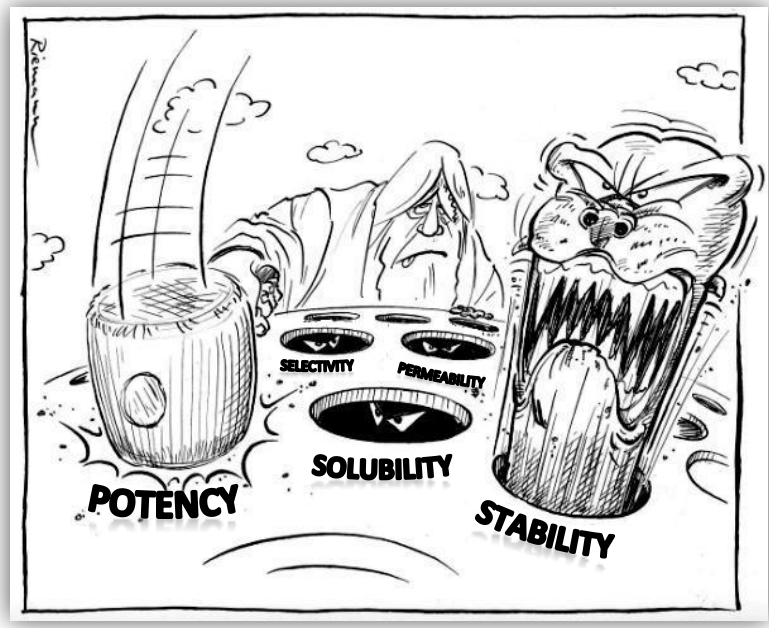


OVERVIEW

EVARISTE
TECHNOLOGIES

- Evariste
- Prediction
- Design
- Selection





- Optimizing in multiple objectives simultaneously often leads into synthetically challenging chemical space that teams are more reluctant to explore
- DND is an efficient method of searching chemistry space to satisfy multiple criteria
- We have chosen to use a fragment based approach to balance synthetic tractability and exploration of chemical space

JOURNAL OF
CHEMICAL INFORMATION
AND MODELING

Article

pubs.acs.org/jcim

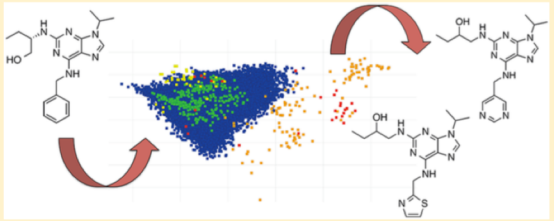
MOARF, an Integrated Workflow for Multiobjective Optimization: Implementation, Synthesis, and Biological Evaluation

Nicholas C. Firth, Butrus Atrash, Nathan Brown,* and Julian Blagg*

Cancer Research UK Cancer Therapeutics Unit, Division of Cancer Therapeutics, The Institute of Cancer Research, London, SM2 5NG, U.K.

Supporting Information

ABSTRACT: We describe the development and application of an integrated, multiobjective optimization workflow (MOARF) for directed medicinal chemistry design. This workflow couples a rule-based molecular fragmentation scheme (SynDiR) with a pharmacophore fingerprint-based fragment replacement algorithm (RATS) to broaden the scope of reconnection options considered in the generation of potential solution structures. Solutions are ranked by a multiobjective scoring algorithm comprising ligand-based (shape similarity) biochemical activity predictions as well as physicochemical property calculations. Application of this iterative workflow to optimization of the CDK2 inhibitor Seliciclib (CYC202, R-roscovitine) generated solution molecules in desired physicochemical property space. Synthesis and experimental evaluation of optimal solution molecules demonstrates CDK2 biochemical activity and improved human metabolic stability.



DESIGN

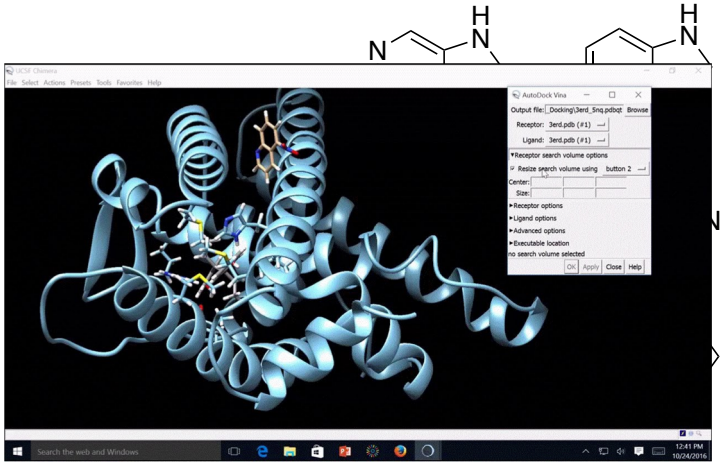
EVARISTE TECHNOLOGIES

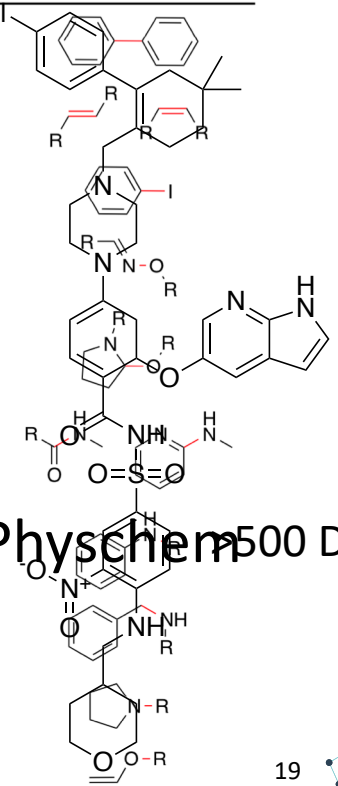
Input Molecule(s) → Fragment

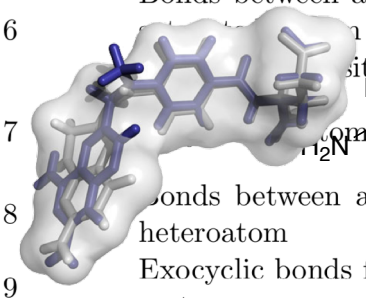
Replace a fragment

Filter

Score



	Rule	Example
1	l-heteroaryl or heteroaryl- ond	
2	E- or Z-Alkene	
3	Carbon-iodine b sys- tem	
4	een any two heteroatoms in an m	
5	inkages	
6	Bonds between a heteroatom and an un- n containing a heteroatom	
7	sition om bonds	
8	onds between a benzylic carbon and a heteroatom	
9	Exocyclic bonds from nitrogen in a cyclic system	
10	Enolic bonds	



QSAR + Physchem

< 500 Da

DESIGN

EVARISTE TECHNOLOGIES

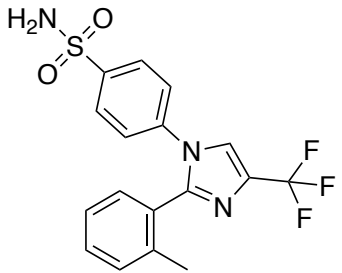
- Had to completely start over due to IP
- Pick a starting point
- Collect all COX-1 and COX-2 ChEMBL data
- Remove all other published molecules from our dataset
- Build Random Forest models for both COX1 and COX2

COX-1

Sum z-score

COX-2

J. Med. Chem. (2002) 45:4847-4857

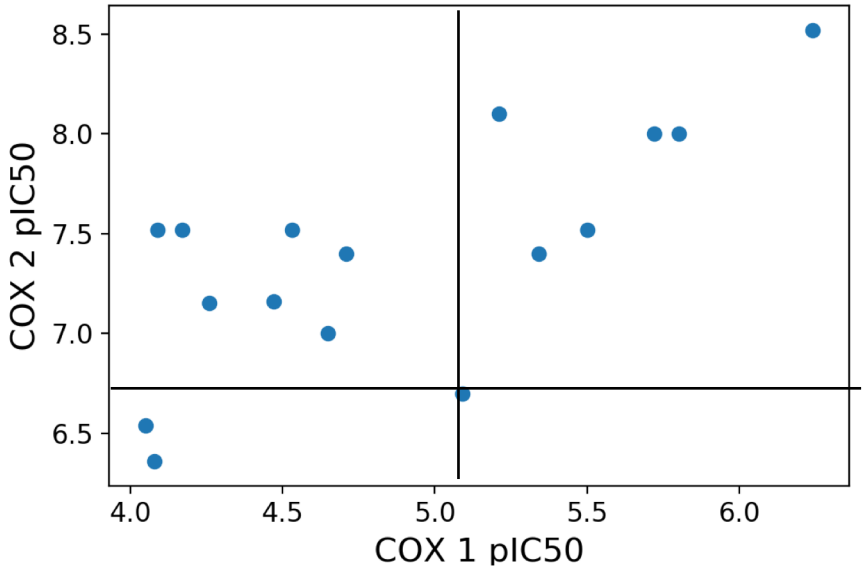


COX-1 pIC50 = 5.1
COX-2 pIC50 = 6.7

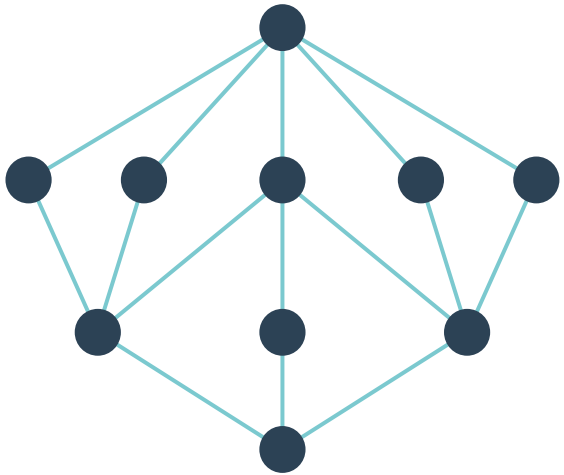
- Ten iterations
- Keep top 100 molecules
- Make 1000 fragment replacements

DESIGN

- No prospective testing available **yet**
- A lot of the molecules designed began to look like another, more effective series
- In the first two iterations 16 molecules were designed from the same series
- Increase in COX-2 and decrease in COX-1
- In process of purchasing some of the others, using compound selection method

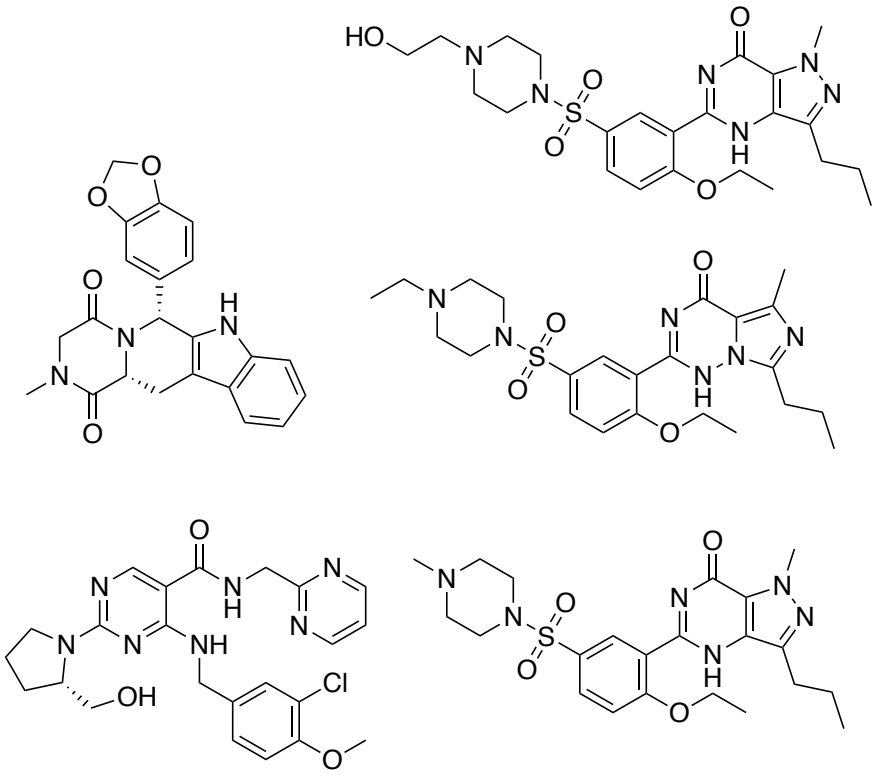


- Evariste
- Prediction
- Design
- Selection



SELECTION

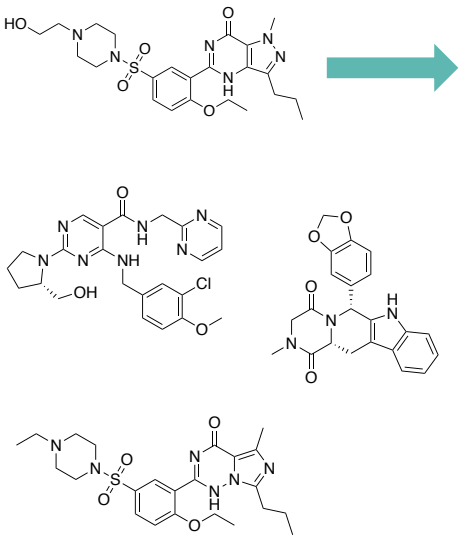
EVARISTE TECHNOLOGIES



- No model is perfect
- Only have finite resources to make and test molecule
- Constantly striving to maximise the probability of experimental success
- We want to chose the best **combination** of molecules to optimize our chances of success

SELECTION

EVARISTE TECHNOLOGIES



Model 1

Model 2

⋮

Model *n*

$y_1 \pm \varepsilon_1$

$y_2 \pm \varepsilon_2$

⋮

$y_n \pm \varepsilon_n$

1.0	0.8	0.4	0.6
0.8	1.0	0.5	0.7
0.4	0.5	1.0	0.5
0.6	0.7	0.5	1.0

Molecular Similarity

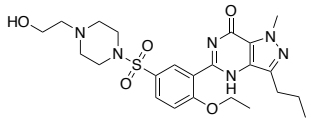
P(matches candidate profile)

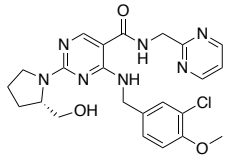
SELECTION

EVARISTE TECHNOLOGIES

Maximise the probability that our set contains a molecule that fits our profile, given the models

P(matches candidate profile)

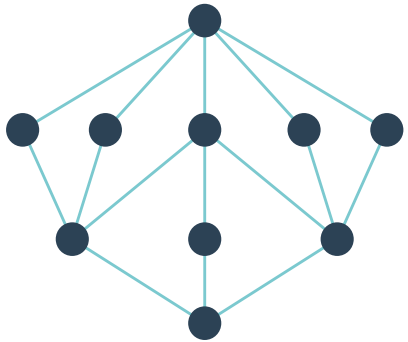




ACKNOWLEDMENTS

EVARISTE TECHNOLOGIES

- Evariste Colleagues
- Andreas Bender x 2
- Greg Landrum
- RDKit community
- scikit-learn community



Questions?