Intelligent Chemical Space Exploration: the old and the new

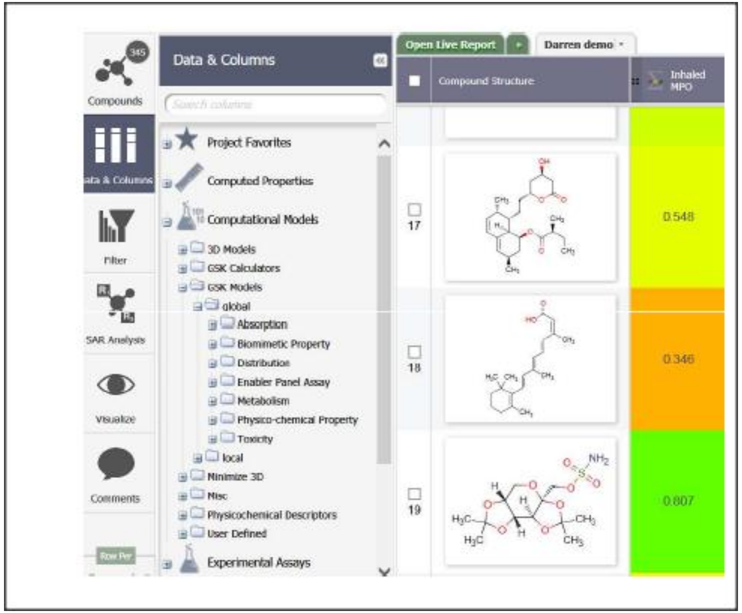
Chris Luscombe
GlaxoSmithKline

Objectives of Lead Optimisation

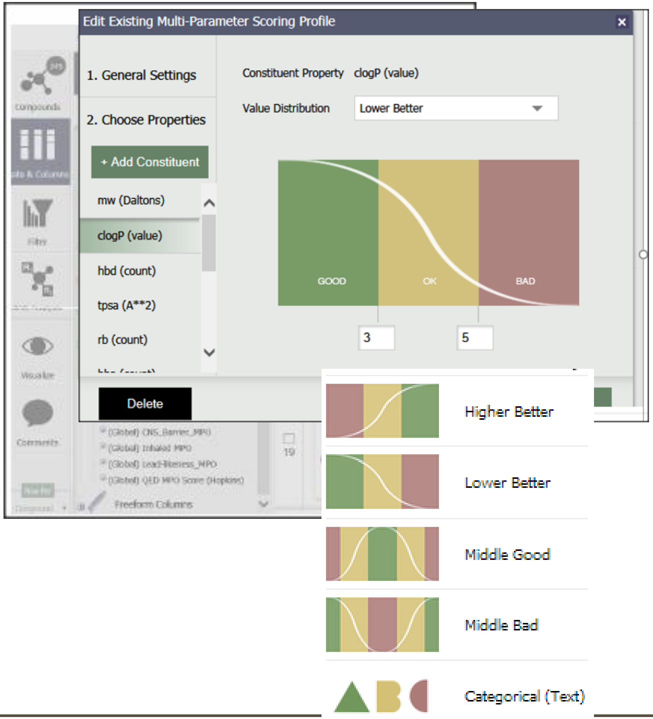
- Design Array experiments to answer SAR questions to enhance potency
- Improve physicochemical properties to enhance ADME
- Discover new monomer groups of interest.
  - Improve Selectivity
  - Establish IP

Multi parameter optimisation

60+ “machine learnt” predictive models published to end users



	Compound Structure	Inhaled MPO
17		0.548
18		0.346
19		0.807



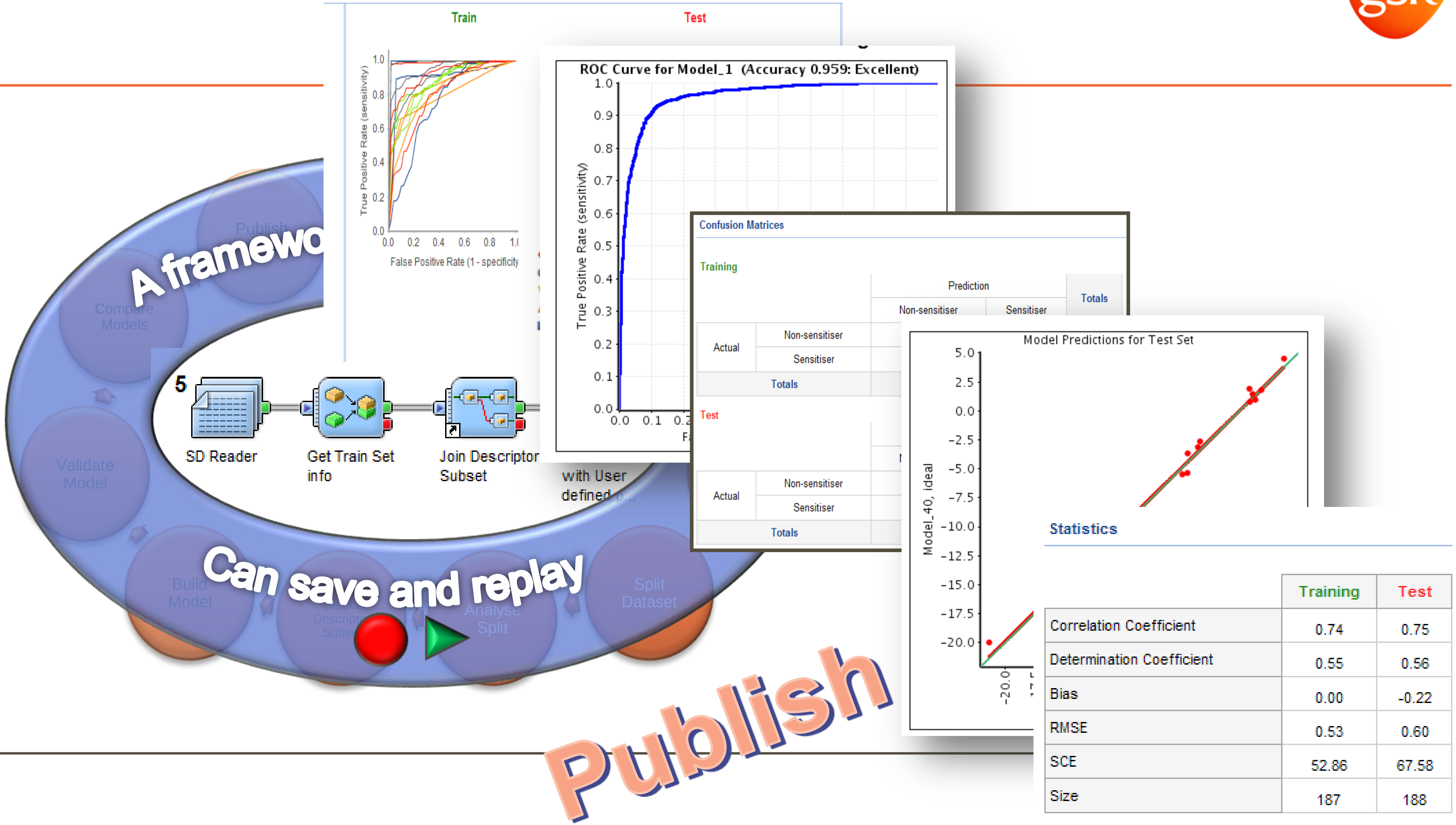
Constituent Property clogP (value)

Value Distribution Lower Better

3

5

QSAR WorkBench: Automating the Expert



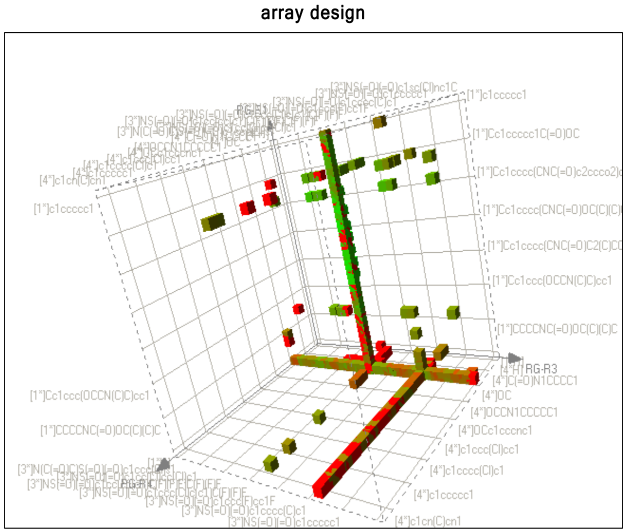
Statistics

	Training	Test
Correlation Coefficient	0.74	0.75
Determination Coefficient	0.55	0.56
Bias	0.00	-0.22
RMSE	0.53	0.60
SCE	52.86	67.58
Size	187	188

Exploring Chemical Space : Arrays based approaches

- Traditional Med Chem
  - Linear arrays (1 x n, 1 x m)
  - Cherry pick and make best combination
  - Assumes Free-Wilson compliance
- Combinatorial chemistry – make all combinations (m x n)
  - No assumptions on Free-Wilson
  - Resource intensive (synthesis and testing)
- SPARSE arrays
  - Make a defined subset of the full combinatorial array
  - Selection using ‘Design of Experiments’ (DOE)

Traditionally SAR determination



- Optimisation at a single position allows
  - Easy synthesis planning
  - Detailed understanding of SAR
- Assumes FW type additivity.

- This approach is widely used and reasonably successful but…

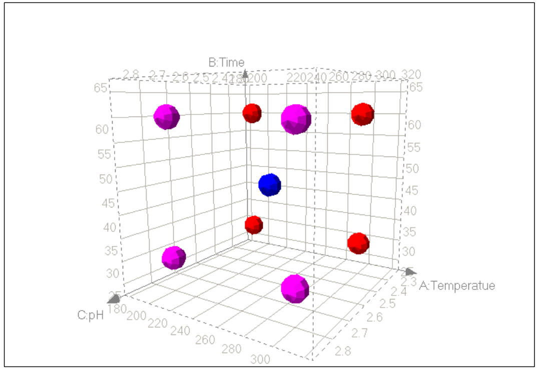
DOE in Medicinal Chemistry?

- Carrying out experiments in continuous property space is easy in domains where the levels are easily chosen such as in a chemical synthesis

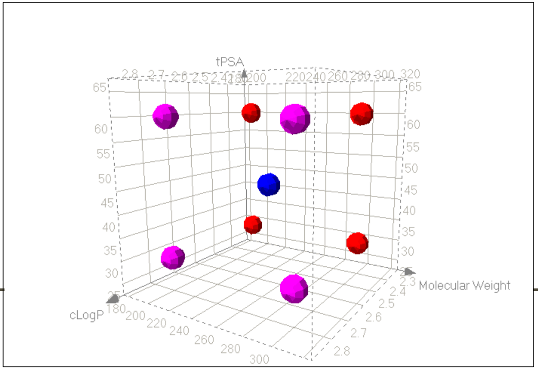
- Creating compounds with particular combinations of physico-chemical properties by modifying monomers around a template is not so easy

- So how do we use DOE to design compounds?

Chemical Reaction



Physcico-Chemical Space



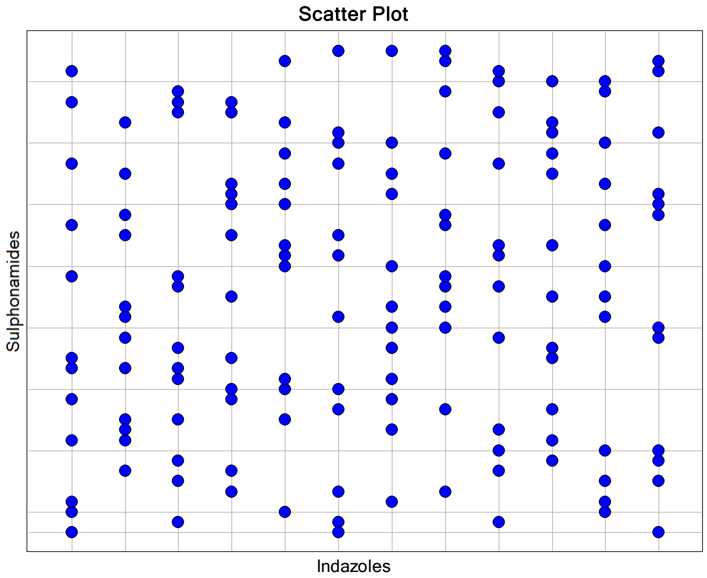
DOE in Medicinal Chemistry?

- We propose that Design of Experiments (DOE) based approaches can be applied to array scenarios where the full (e.g. M x N) array cannot be synthesized for practical reasons.
- By treating each monomer in the array as a categorical factor of the design, a balanced fractional (“Sparse”) array design can be generated.
- This novel approach can be successfully used to understand and exploit the SAR of a late stage optimization programme

Sparse array to evaluate defined N x M combinatorial space with a fractional subset

- **Design**
  - 12 Indazoles (R1)
    - Identified using classical SAR approaches
  - 48 sulphonyl chlorides monomers (R2)
    - selected from library using a variety of criteria
      - Lead-likeness score

R2



R1

- •12 monomers per R1
- •3 monomers per R2

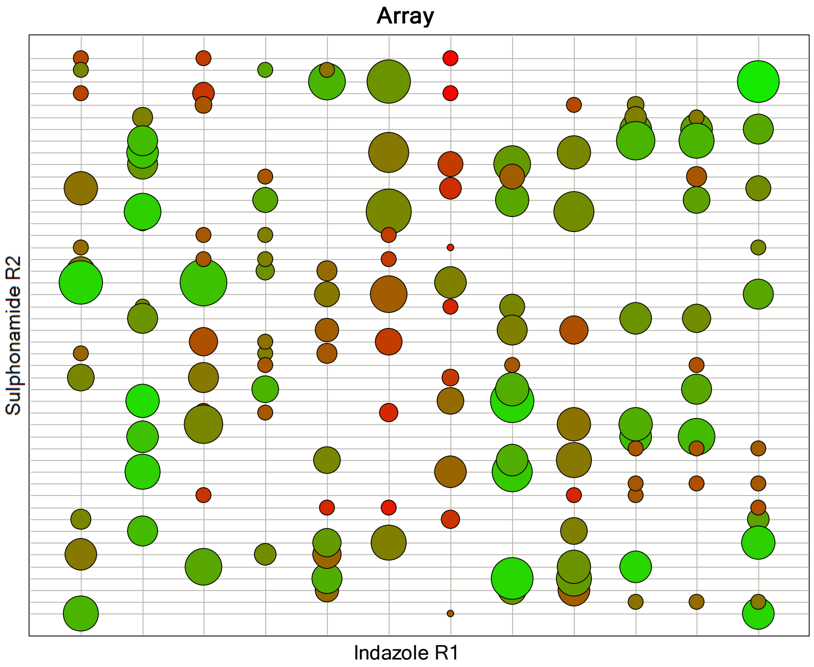
Questions

- Is the fraction selected sufficient to explore the chemistry space?
- Can we adequately assess monomer potential?
- Can we predict the ‘missing’ compounds?
- Is it a practical way to direct chemistry synthesis?
- Is it an efficient process?

- Does it work?

Measured Potency for the Sparse array

- 142 of 144 compounds from patchwork array were synthesised and tested
- Coloured for potency, sized by ligand efficiency
- Clear that some Indazoles are more promising than others



Predicted most potent compounds that haven’t already been synthesized

- All compounds subsequently synthesized had measured potencies within +/- 0.2 pIC50 of the predicted value
- Validated the Additivity assumption
- Identified promising alternatives which were sent for further PK analysis – potential back up to the current pre-candidate

C1
Predicted GTPgS = 7.6
BEI = 16.0
Measured = 7.6

C2
Predicted GTPgS = 7.5
BEI = 13.5
Measured = 7.6

C3
Predicted GTPgS = 7.5
BEI = 14.8
Measured = 7.3

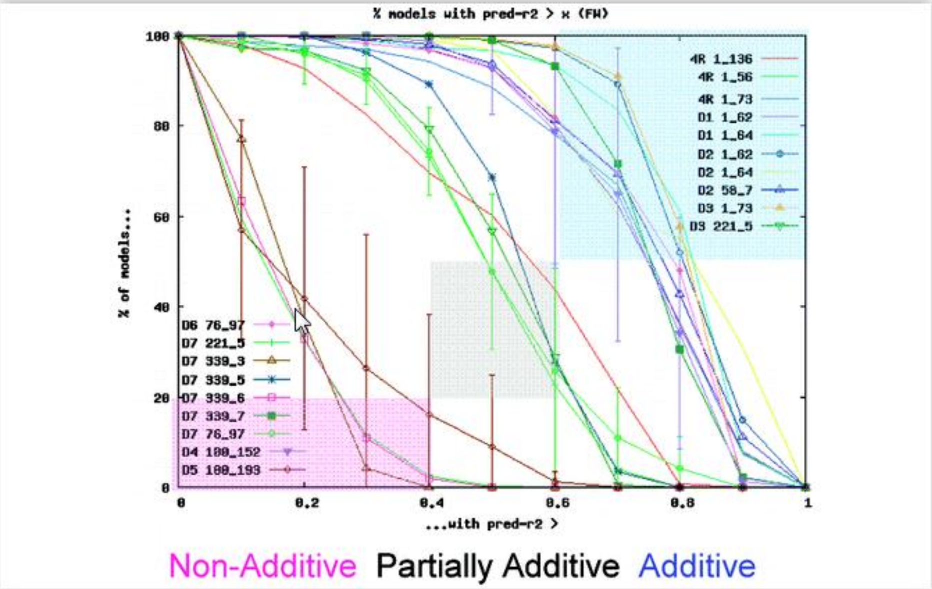
C4
Predicted GTPgS = 7.5
BEI = 14.2
Measured = 7.4

C5
Predicted GTPgS = 7.6
BEI = 15.6
Measured = 7.5

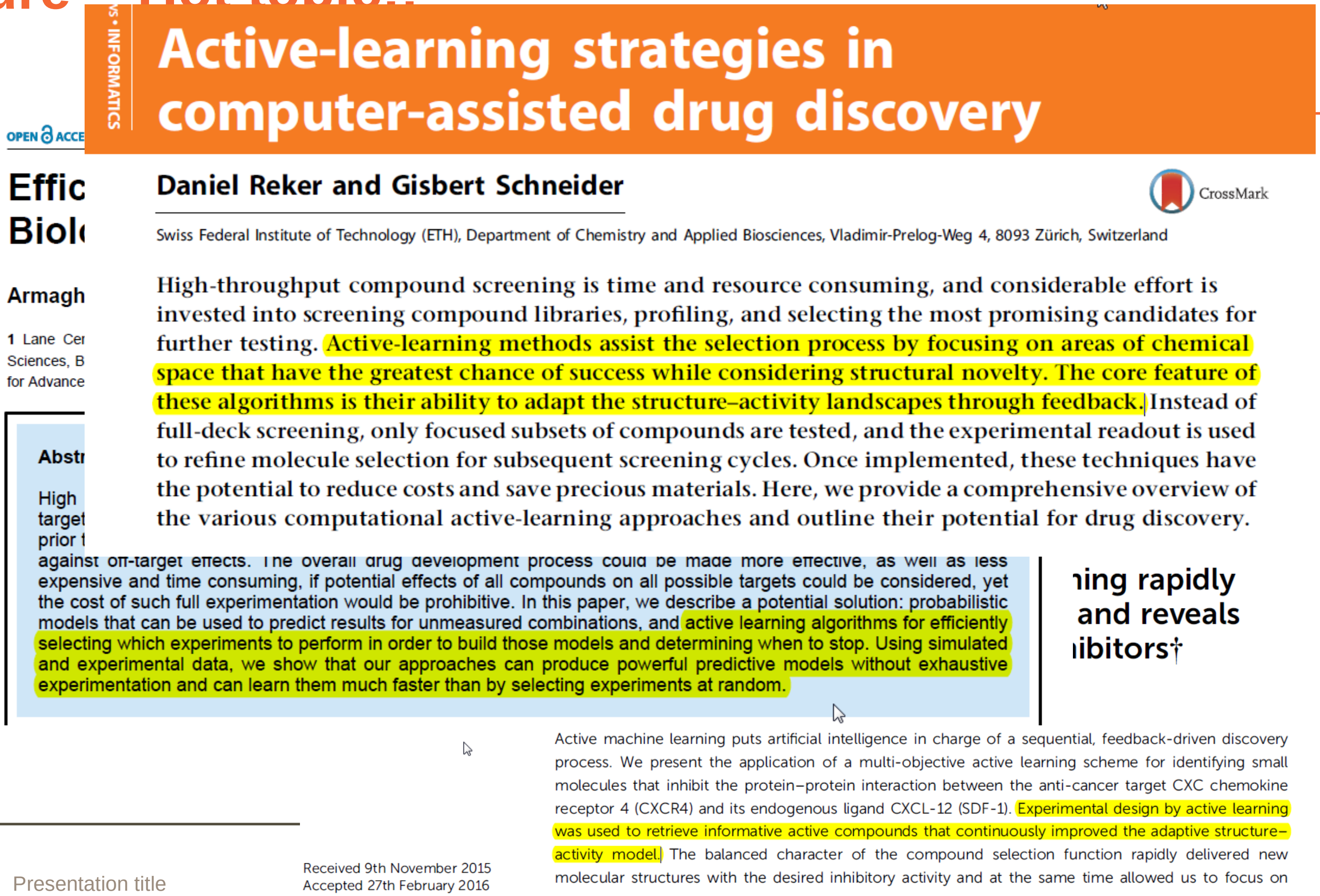
Assessment of Additive/Nonadditive Effects in Structure−Activity Relationships: Implications for Iterative Drug Design

J. Med. Chem. 51, 23, 7552-7562 *Yogendra Patel, Valerie J. Gillet, Trevor Howe, Joaquin Pastor, Julen Oyarzabal and Peter Willett*

- Free-Wilson (FW) analysis is based on the assumption that the contributions to activity made by substituents at different substitution positions are additive.
- We analyze eight near complete combinatorial libraries assayed on several different biological response(s) (GPCR, ion channel, kinase and P450 targets)
- only half-exhibit clear additive behavior, which leads us to question the concept of additivity that is widely taken for granted in drug discovery



Literature - Hot topic!!



VS • INFORMATICS

Active-learning strategies in computer-assisted drug discovery

Daniel Reker and Gisbert Schneider

CrossMark

Swiss Federal Institute of Technology (ETH), Department of Chemistry and Applied Biosciences, Vladimir-Prelog-Weg 4, 8093 Zürich, Switzerland

High-throughput compound screening is time and resource consuming, and considerable effort is invested into screening compound libraries, profiling, and selecting the most promising candidates for further testing. Active-learning methods assist the selection process by focusing on areas of chemical space that have the greatest chance of success while considering structural novelty. The core feature of these algorithms is their ability to adapt the structure–activity landscapes through feedback. Instead of full-deck screening, only focused subsets of compounds are tested, and the experimental readout is used to refine molecule selection for subsequent screening cycles. Once implemented, these techniques have the potential to reduce costs and save precious materials. Here, we provide a comprehensive overview of the various computational active-learning approaches and outline their potential for drug discovery.

OPEN ACCE

Effic

Biol

Armagh

1 Lane Cer

Sciences, B

for Advance

Abstr

High

target

prior t

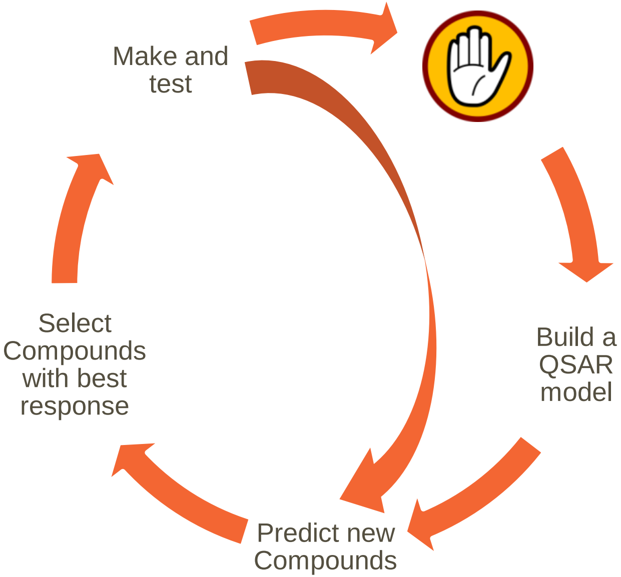
against off-target effects. The overall drug development process could be made more effective, as well as less expensive and time consuming, if potential effects of all compounds on all possible targets could be considered, yet the cost of such full experimentation would be prohibitive. In this paper, we describe a potential solution: probabilistic models that can be used to predict results for unmeasured combinations, and active learning algorithms for efficiently selecting which experiments to perform in order to build those models and determining when to stop. Using simulated and experimental data, we show that our approaches can produce powerful predictive models without exhaustive experimentation and can learn them much faster than by selecting experiments at random.

ning rapidly
and reveals
ibitors†

Received 9th November 2015
Accepted 27th February 2016

Active machine learning puts artificial intelligence in charge of a sequential, feedback-driven discovery process. We present the application of a multi-objective active learning scheme for identifying small molecules that inhibit the protein–protein interaction between the anti-cancer target CXC chemokine receptor 4 (CXCR4) and its endogenous ligand CXCL-12 (SDF-1). Experimental design by active learning was used to retrieve informative active compounds that continuously improved the adaptive structure–activity model. The balanced character of the compound selection function rapidly delivered new molecular structures with the desired inhibitory activity and at the same time allowed us to focus on

Classic use of inSilico to Guide Decisions – Passive Learning

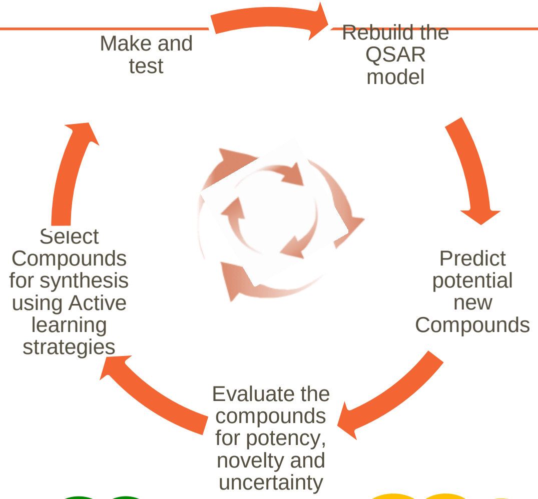


- The model is built and validated on available data.
- The model will be predictive for new compounds it 'knows' about – ie the Known Knowns
- The model doesn't 'learn' anything new.
- The rebuild cycle only rarely gets triggered

Typically the QSAR build is only done once

Iterate utilising the model(s)

Active Learning



Make and test

Rebuild the QSAR model

Predict potential new Compounds

Evaluate the compounds for potency, novelty and uncertainty

Select Compounds for synthesis using Active learning strategies

The model is built and validated on available data.

The model is updated <u>every</u> cycle

The choice of what to make next is guided by the <u>needs of the model</u> to improve as well as the programme objectives

Top N

This looks potent

Exploit

Novelty!

That's New!

The current model does not have any data in this space

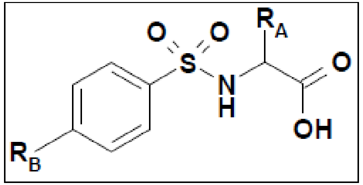
Uncertainty!
Potency maybe much higher than this

The current model predicts a low confidence in the prediction

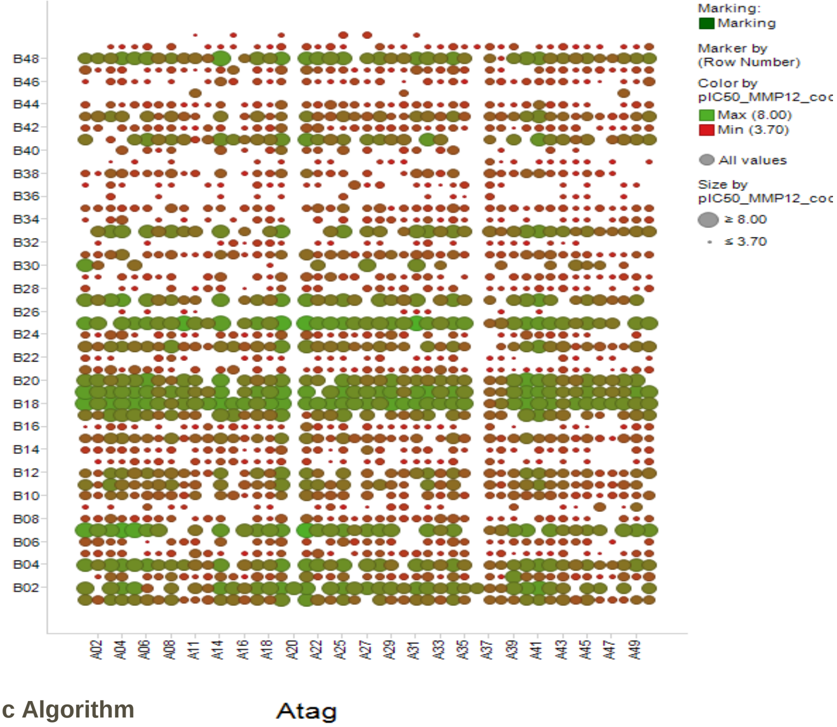
MMP12 50 x 50 monomer array

This is a fixed pool to test Active learning strategies

- Range of pIC50 (3.7 – 8.0)
- MMP-12 data set (1704 compounds)

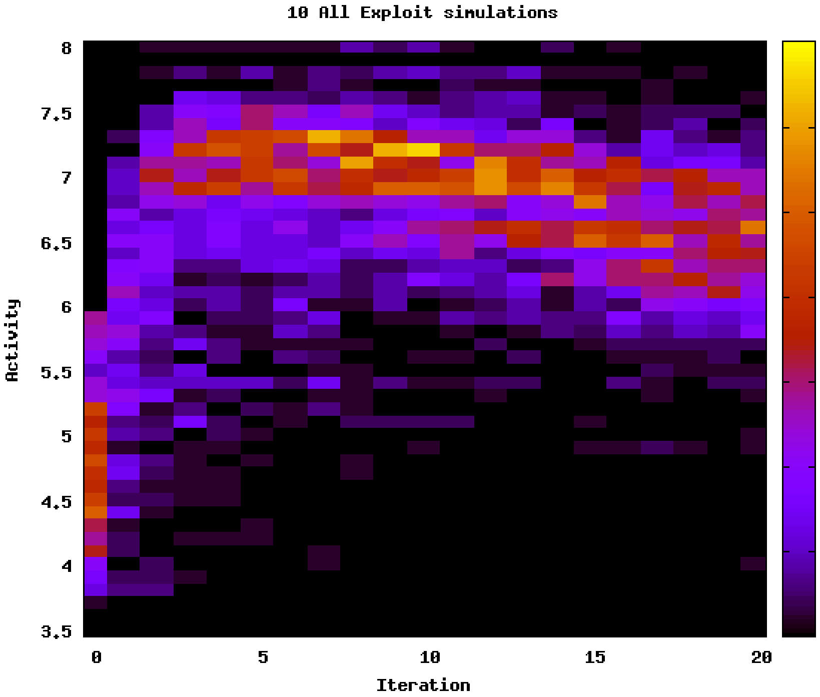


- Initialize by randomly taking ≈ 3% of the compounds with activity < 6 (about 37 compounds)
- Take 20 compounds per iteration and run for 20 iterations
- Questions to test :
  - Does Explore add value over just Exploit?
  - When should I Explore?

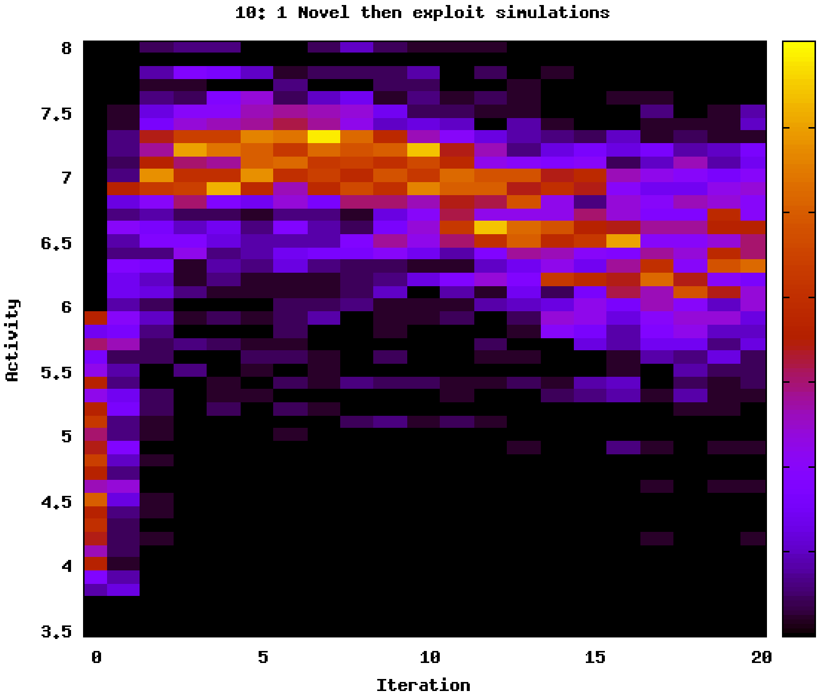


Automated Lead Optimization of MMP-12 Inhibitors Using a Genetic Algorithm
Stephen D. Pickett, Darren V. S. Green, David L. Hunt, David A. Pardoe, and Ian Hughes
ACS Medicinal Chemistry Letters **2011** *2* (1), 28-33

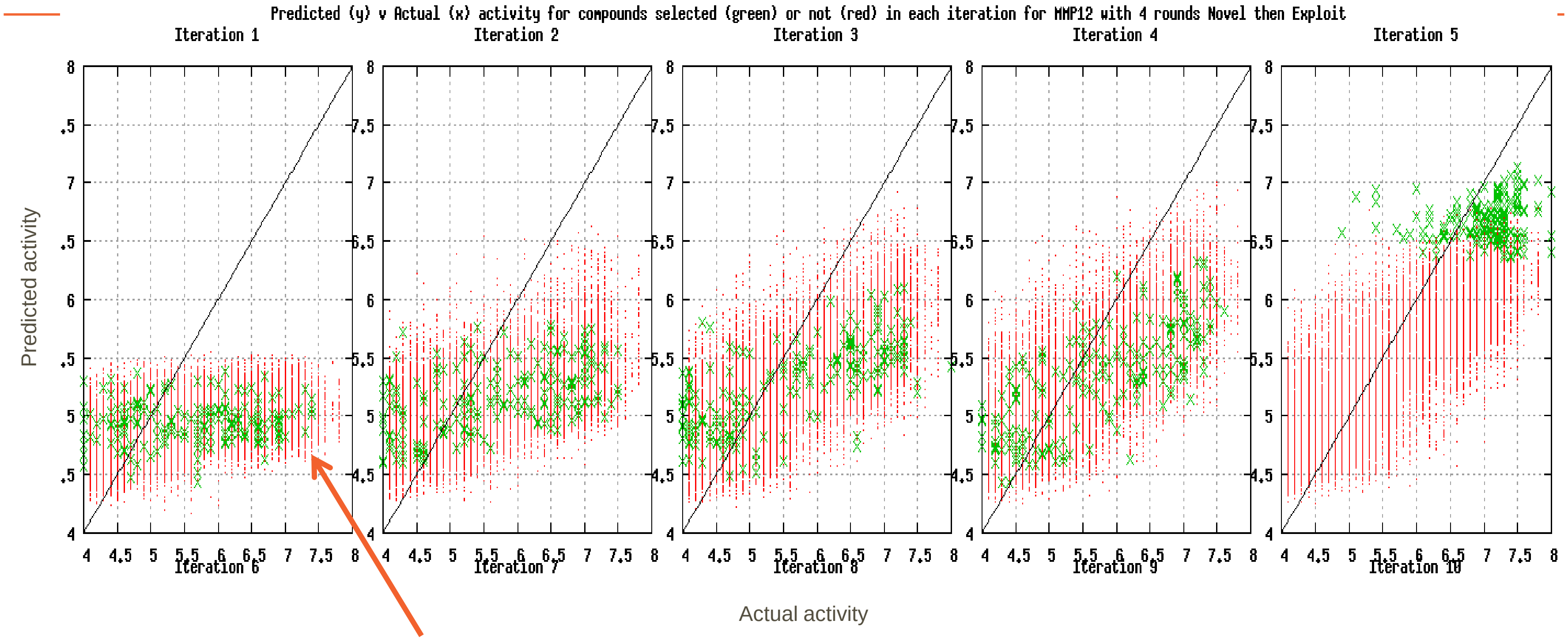
Exploit – ie just picking the predicted Top 20 each iteration (building a model after each round)



One round of Novel selection followed by Exploit



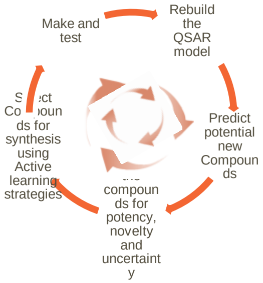
MMP 12, 4 novel then exploit

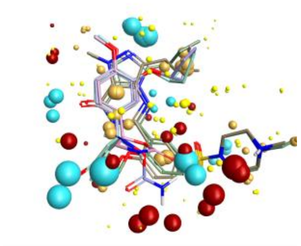


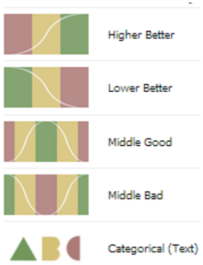
Initial iteration - all predictions are in range of activities found in training data
As RF model is used. Compounds with a whole range of activities are selected

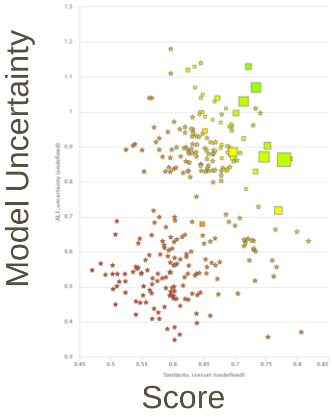
Combining Active Learning with MPO Selection

"Live" project example


Make and test
Rebuild the QSAR model
Predict potential new Compounds
the compounds for potency, novelty and uncertainty
Select Compounds for synthesis using Active learning strategies

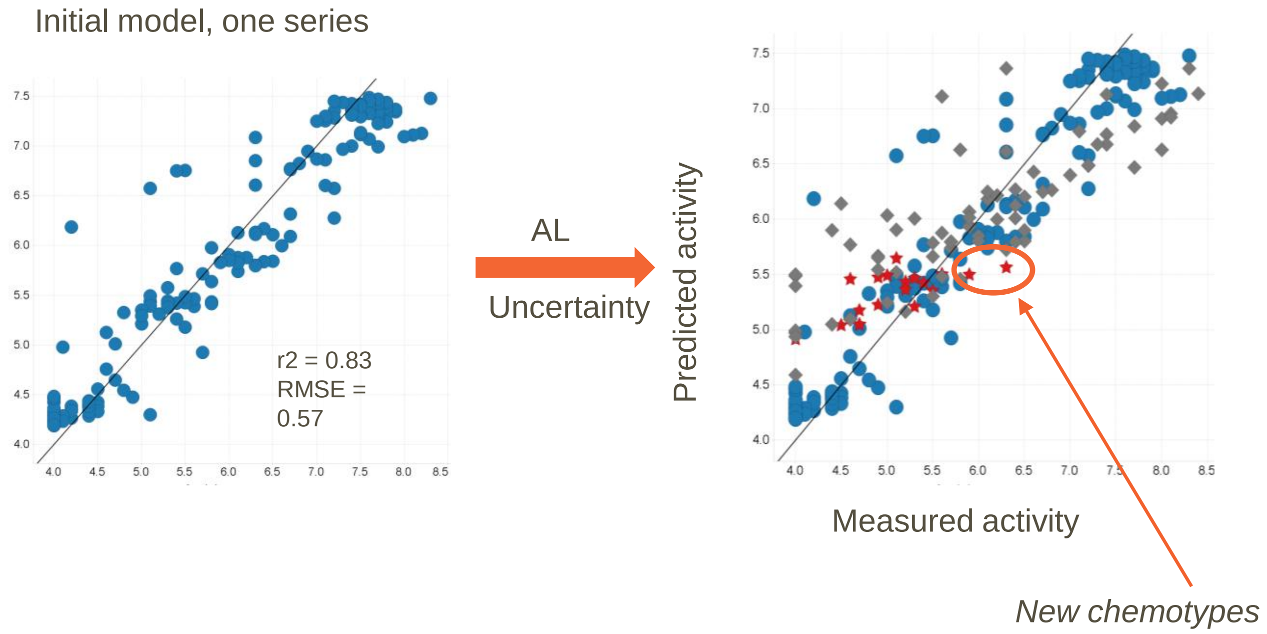






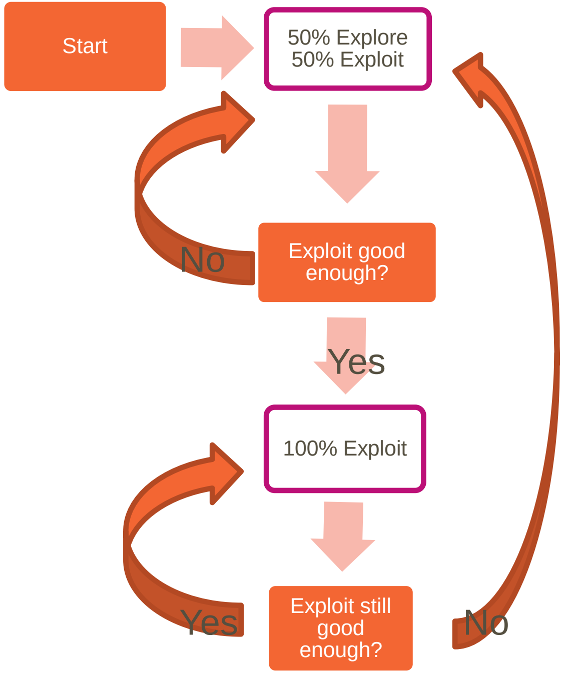
Active Learning – Example 2

Generating new series



19 compounds synthesized from AL model based on uncertainty –looking for positive surprises!

Adaptive strategy - when to explore ?

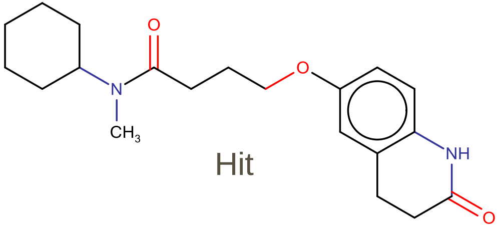


- “Good enough” depends on:
  - Resource remaining
  - Required/expected level of activity
- Uses Exploit compounds as a way of seeing how well active learning is doing
- Explore could be “Novel” in early iterations, “Uncertain” in later
- Not aware of adaptive strategies in AL-LO literature

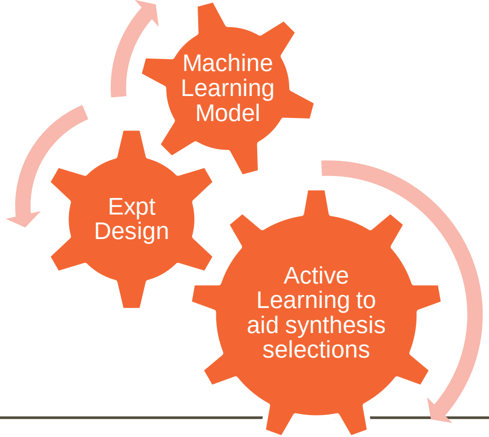
Driving Medicinal Chemistry using Active Learning

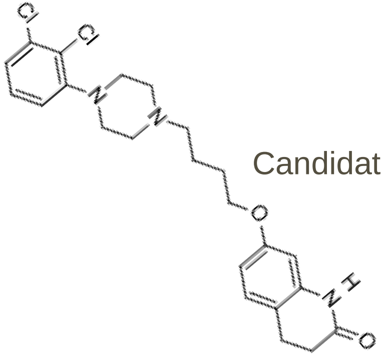
Experiments selected to improve models as well as drive programme goals

- Use Experimental design to efficiently scope SAR
  - Sparse Arrays
- Build insilico models to predict key properties
  - Choose experiments to enhance model building
- Embed Active Learning strategies to aid synthesis decisions
  - Explore and Exploit the model



Hit





Candidate

Acknowledgements

Computational

- Tessella
  - Nik Burkoff
- Molecular Design UK
  - Darren Green
  - Stephen Pickett
  - Martin Saunders
  - Sandeep Pal
  - Nick Barton
  - Alvaro Cortes

– Chemistry

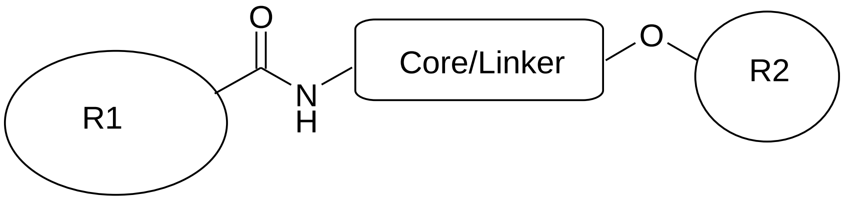
- Medicinal Chemistry
  - Hannah Davies
  - Heather Hobbs
  - Zoe Henley

Back Up slides

Free Wilson theory R1-Core-R2

- First mathematical technique for quantitative SAR

- Response = effect of Core + effect R1 substituent + effect of R2 substituent

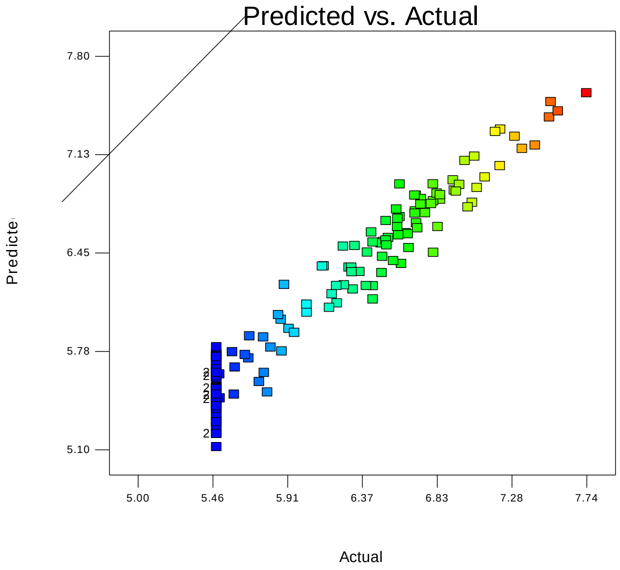


- Assumptions
  - No interactions between core and substituent
  - No interaction between substituents (R-groups)

- Can only explore chemical space defined by R-group combinations in the training set

FW analysis of monomer contribution

- A Free –Wilson analysis is a regression based approach to establish monomer contributions to a predictive model
- A high degree of fit suggests that the potency profile could be additive in nature.
  - The presence of outliers may imply non-additive behaviour
  - Assess potential interaction terms between monomers if the output appears to be non-additive



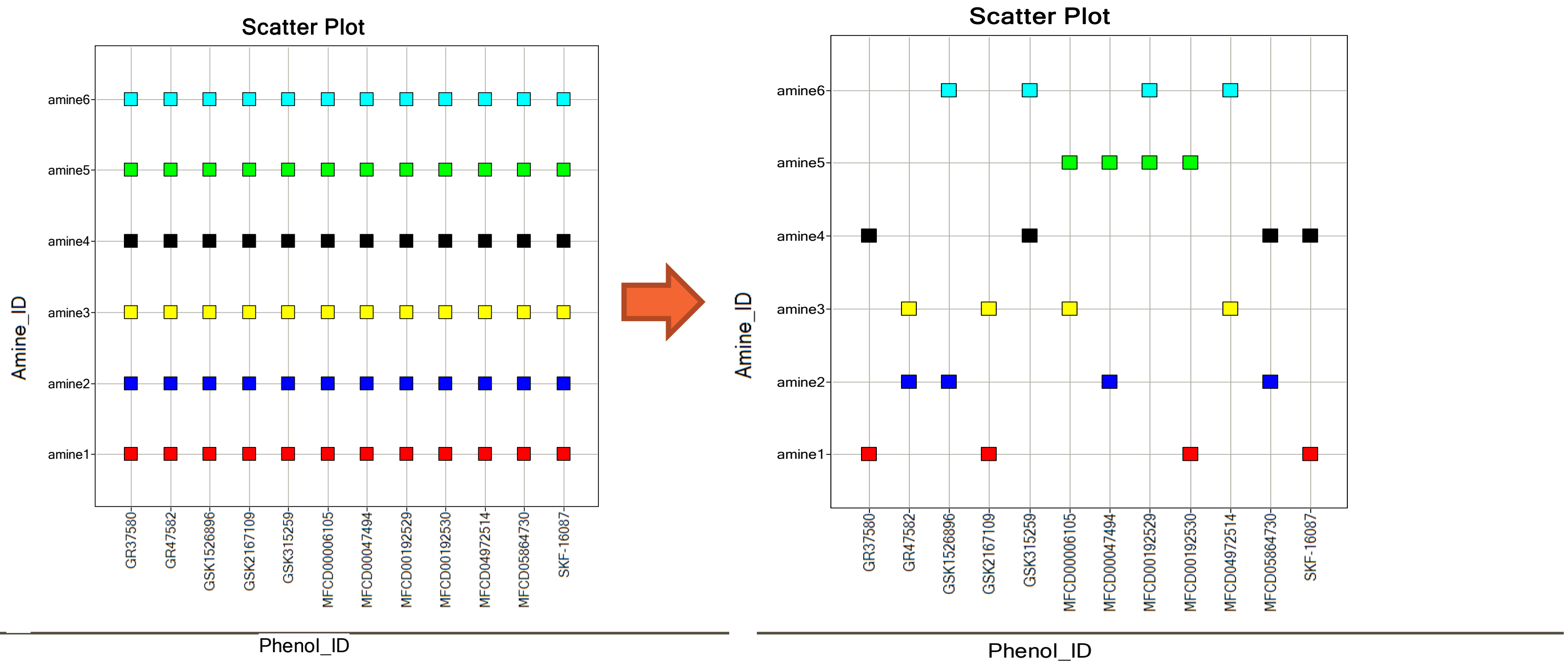
Design of Experiments (DOE)

- Experimental Design approaches are well established for the optimization of multi-factor experiments, such as reaction conditions.

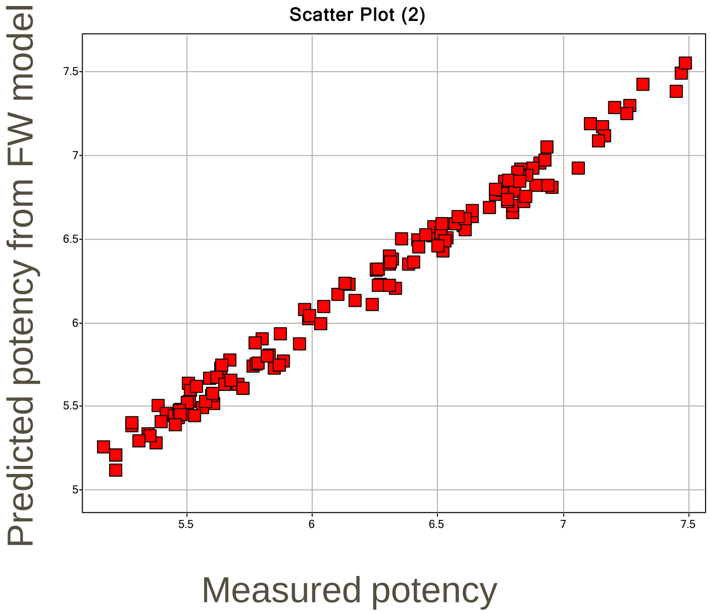
- Typically these domains utilize ‘continuous’ variables such as temperature, addition rate, time etc

- Can these same techniques be use where each variable is categorical?

Example of a Sparse Array
1/3rd fraction from an 6 x 12 array

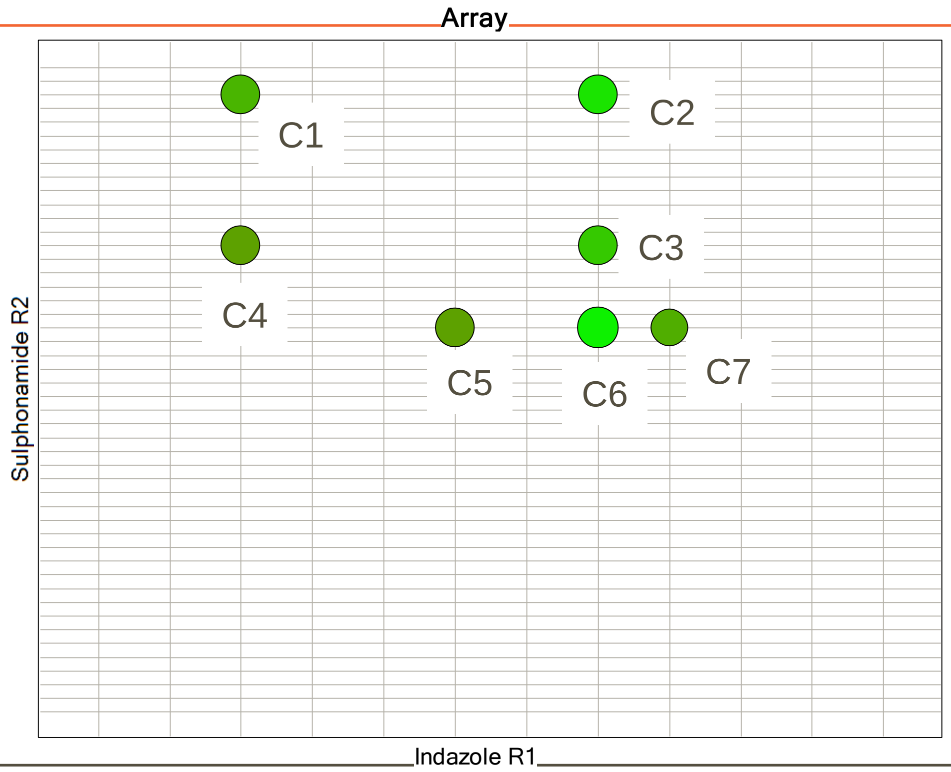


Sparse Array Data Analysis



- – Statistical analysis was done to evaluate ‘additivity’
- – Free Wilson model: Predicted potencies were plotted against measured potencies
- – The FW model show potential excellent additivity with no outliers.

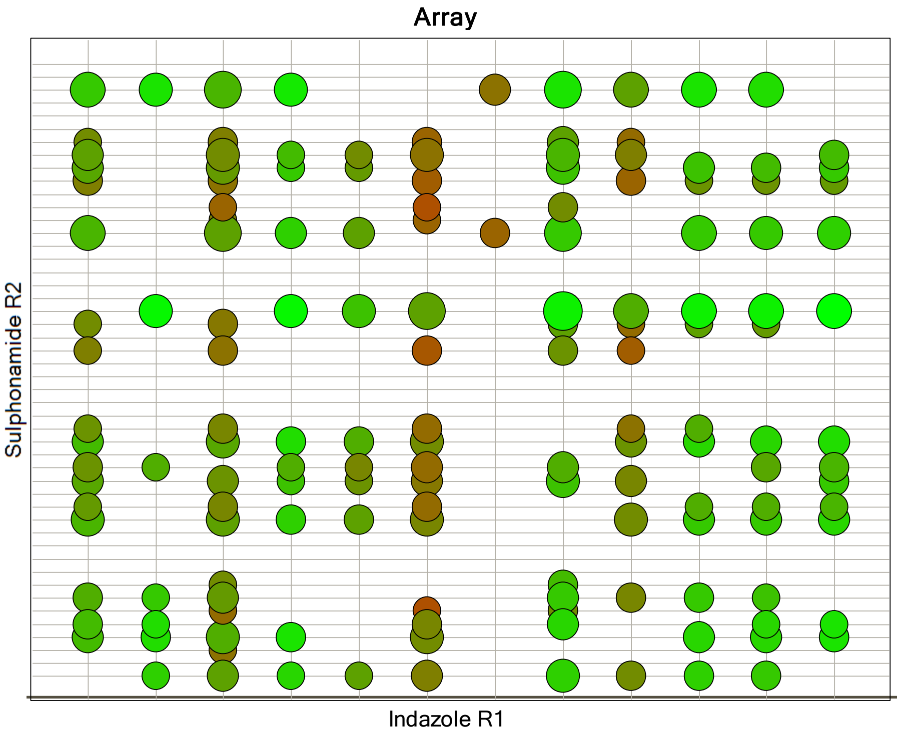
Find the predicted most potent compounds that haven't already been synthesized



RG-R2
(48 variants)

RG-R1 (12 variants)

Predicted Potency for the complete array of 576 compounds (Fit and Predict), only Actives (pIC50>6.5 shown)



RG-R2
(48 Variants)

RG-R1 (12 Variants)

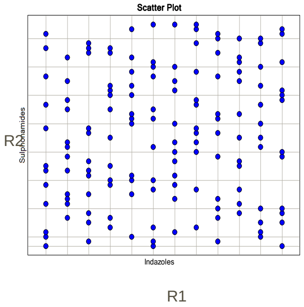
Start with an intent to model

Experimental Design - Sparse Arrays

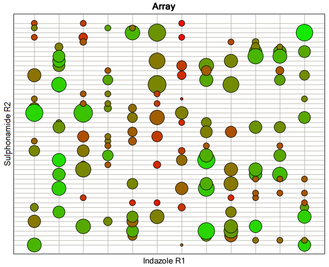
Evaluate defined N x M combinatorial space with a fractional subset

12 x 48 (576) sampled in 144 compounds

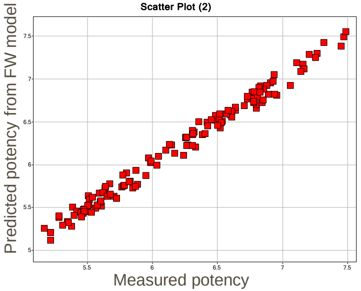
Design



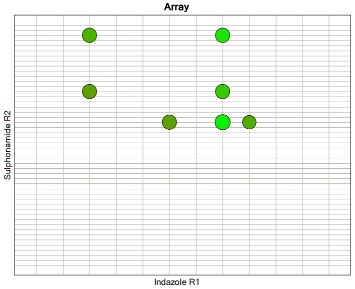
Make & Test



Model



Predict



- •12 monomers per R1
- •3 monomers per R2

Learnings from experience

- Ideally 3 examples minimum for each monomer within the design, although 2 will work for a robust assay and chemistry
- Need to have confidence in getting some active compounds
  - If all the compounds are inactive its difficult to fit a model!
- Confidence in ability to synthesize compounds
  - Some loss of particular compounds can be tolerated but if whole reactions fail then the array design will be compromised

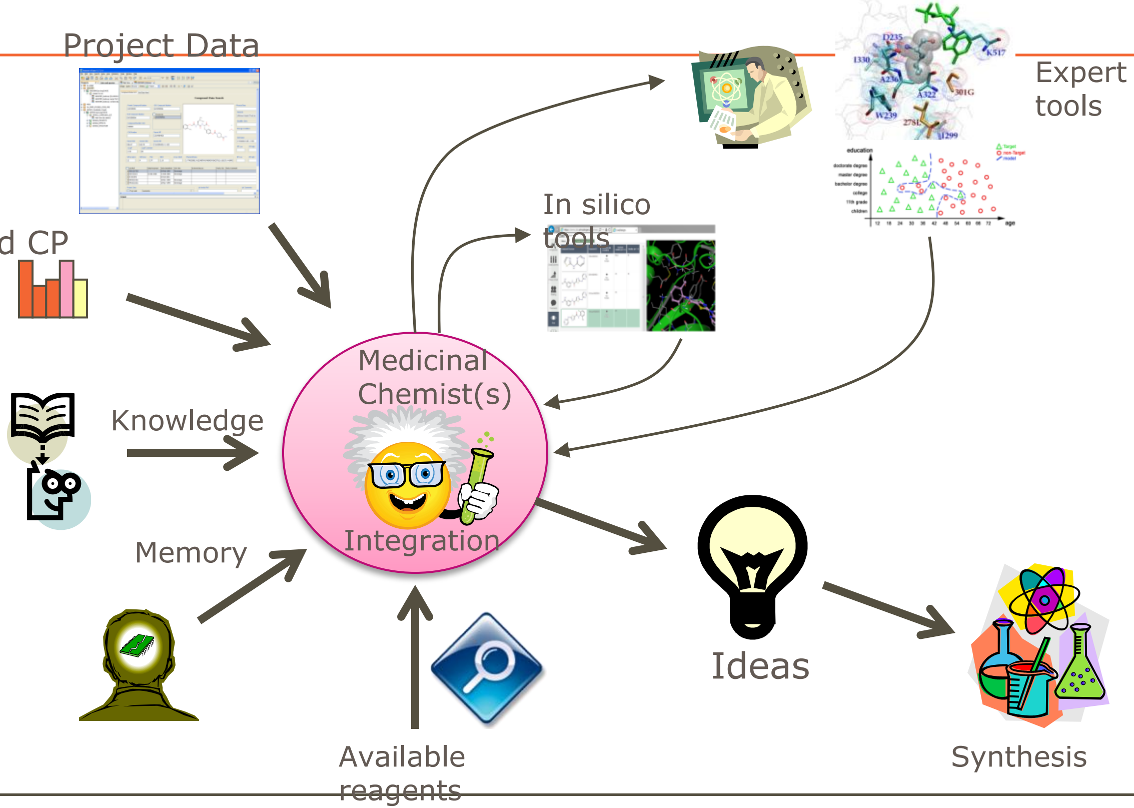
Summary

- Experimental Design may provide an alternative /complementary strategy
  - Initial exploration of new monomer space
  - Identification of back up compounds
  - Establish Addivity in the series
- Efficient Lead Optimisation by exploring more than one point of change at the same time on the molecular template
- Can unearth some surprises which may never have been found by traditional processes

- The data set generated is perfect for building QSARs

The Chemist Centric Design Process

Everything goes through the chemist(s), decisions are anchored by knowledge and intuition



Project Data

Desired CP

Knowledge

Memory

Medicinal Chemist(s)

Integration

Available reagents

In silico tools

Expert tools

Ideas

Synthesis

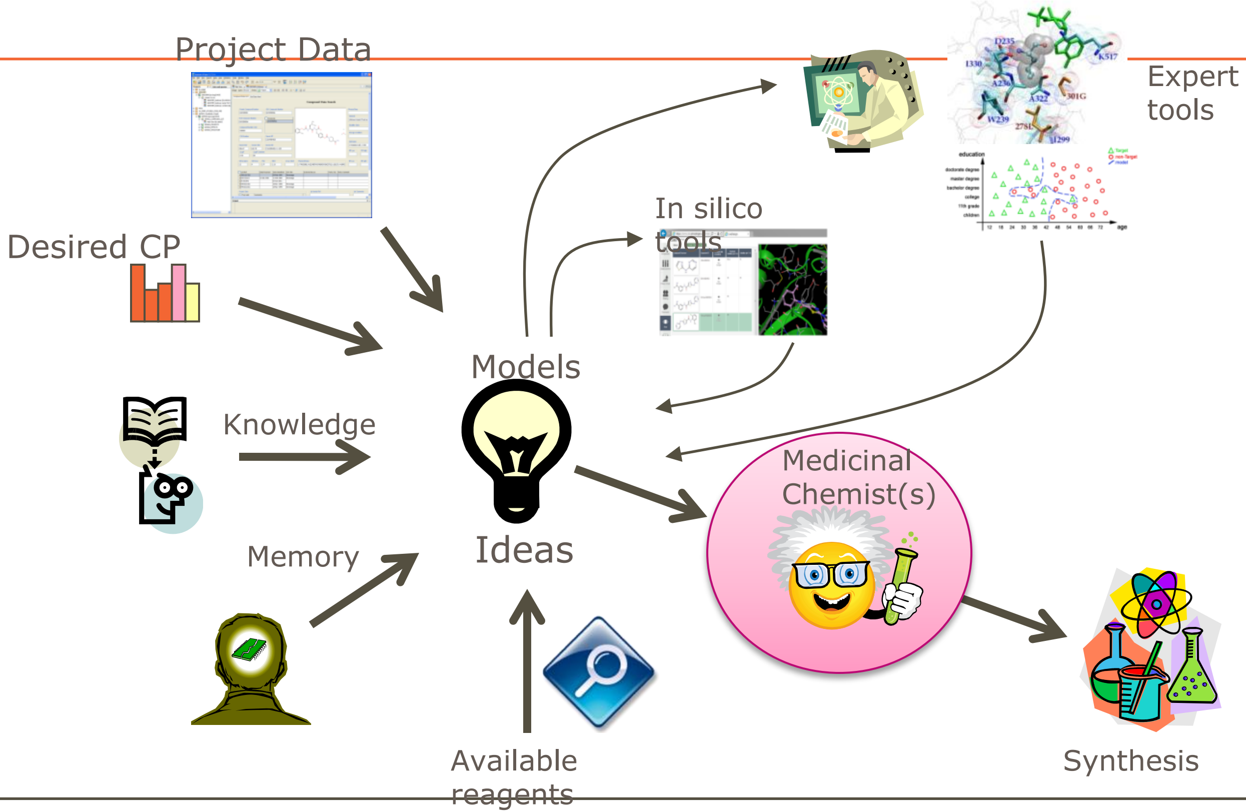
Where are we heading?

- **Quantification** is key to improving our processes
- **Chemist intuition** probably does not hold up to statistical analysis
- **Simple models** can add value to the design process, and better ones can spectacularly improve it

- **Molecule Design is experiencing a revolution**
  - Data, algorithms, computers
  - Requires Business Process Reengineering for the larger companies
  - In the near future, who and what constitutes a “Medicinal Chemist” will be very different

What if …

We put **systematic ideation and modelling** at the centre of the process?



Project Data

Desired CP

Knowledge

Memory

Models

Ideas

Available reagents

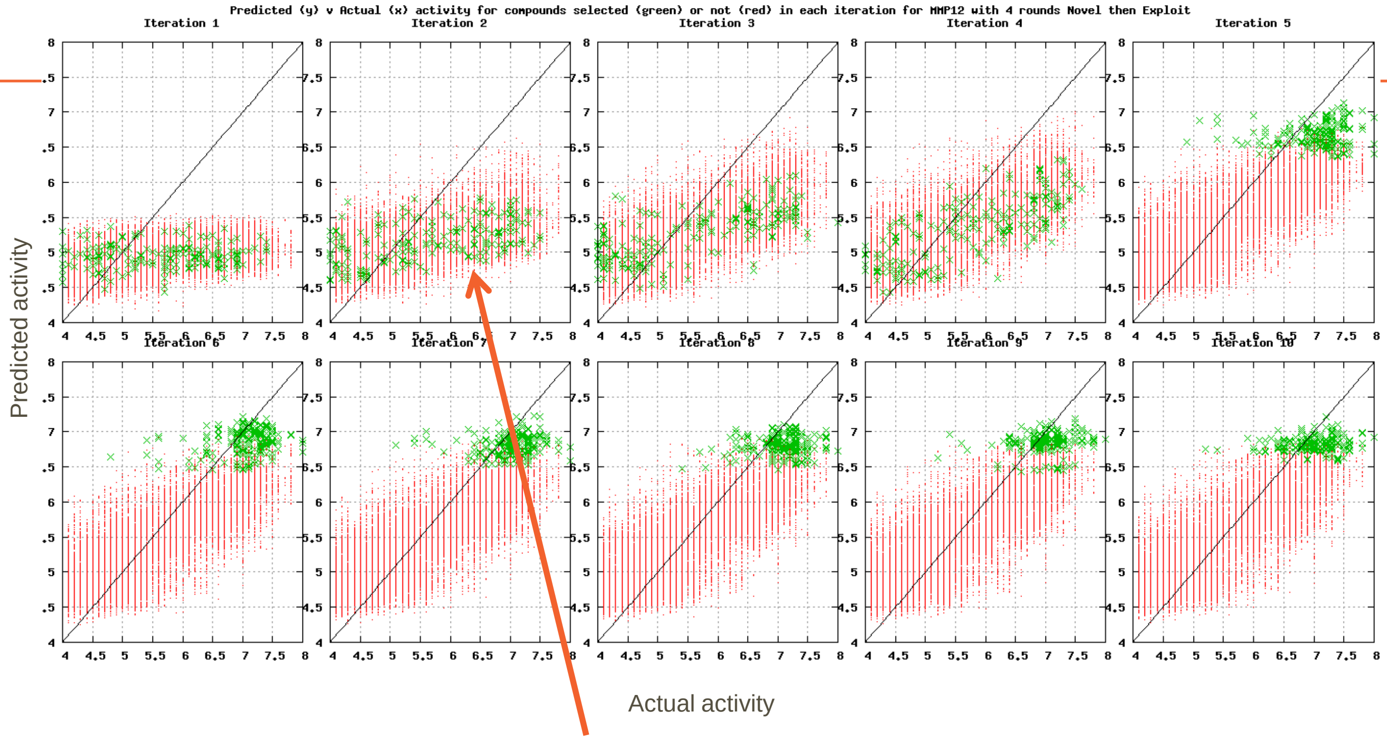
In silico tools

Expert tools

Medicinal Chemist(s)

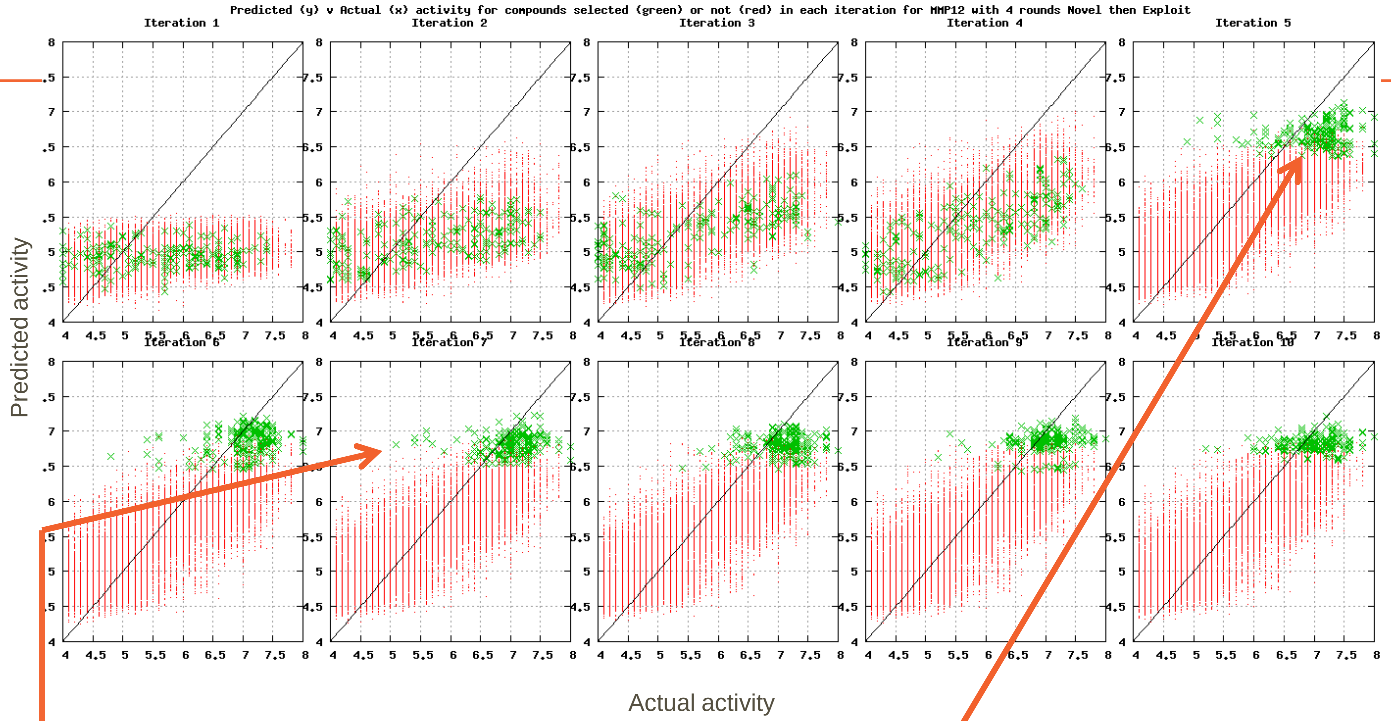
Synthesis

MMP 12, 4 novel then exploit



Model clearly improves after first iteration, but does not seem to get much better by selecting more novel compounds

MMP 12, 4 novel then exploit



Start exploiting after iteration 4, compounds are highly active (are found at right side of graph). Although some lower activity compounds are selected good compounds are selected for all 6 exploit iterations