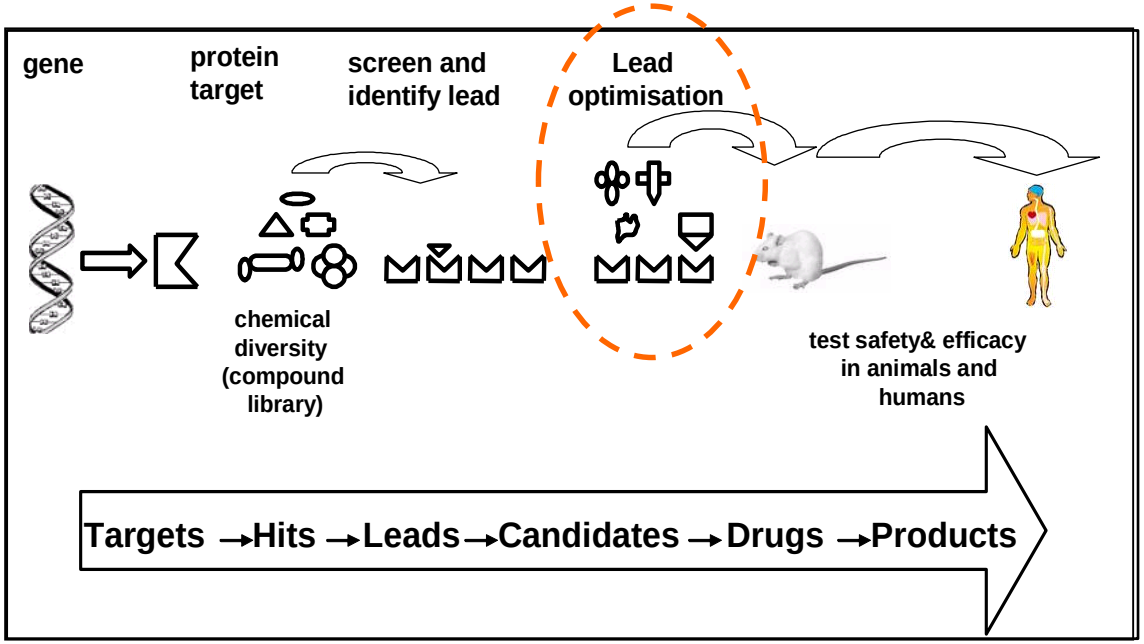
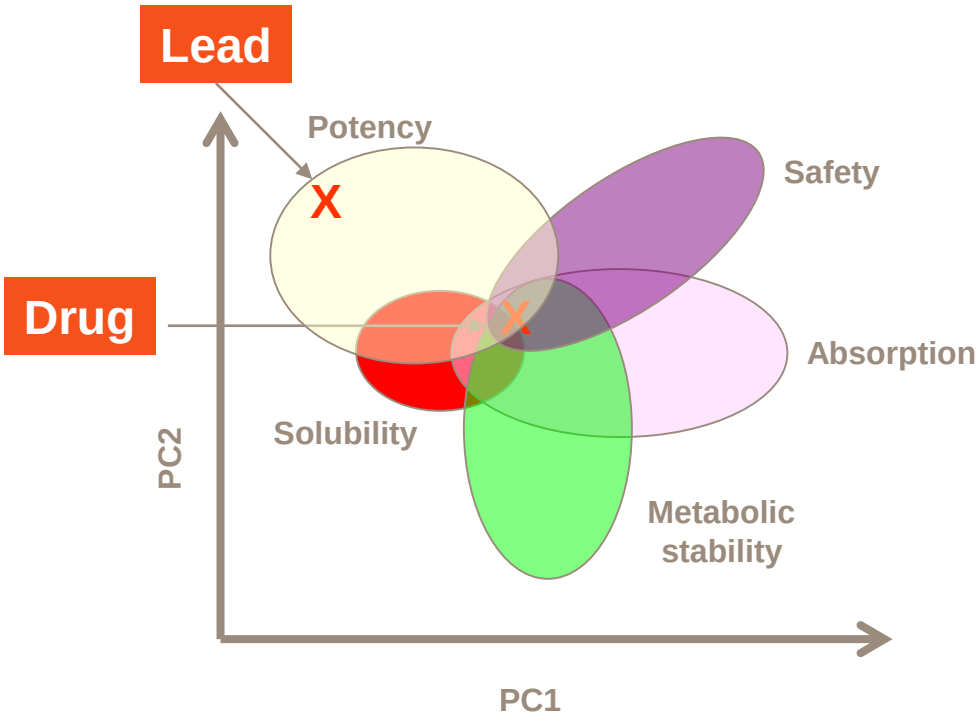
The Quantification of Lead Optimisation

Darren Green
Computational and Modelling Sciences
GSK

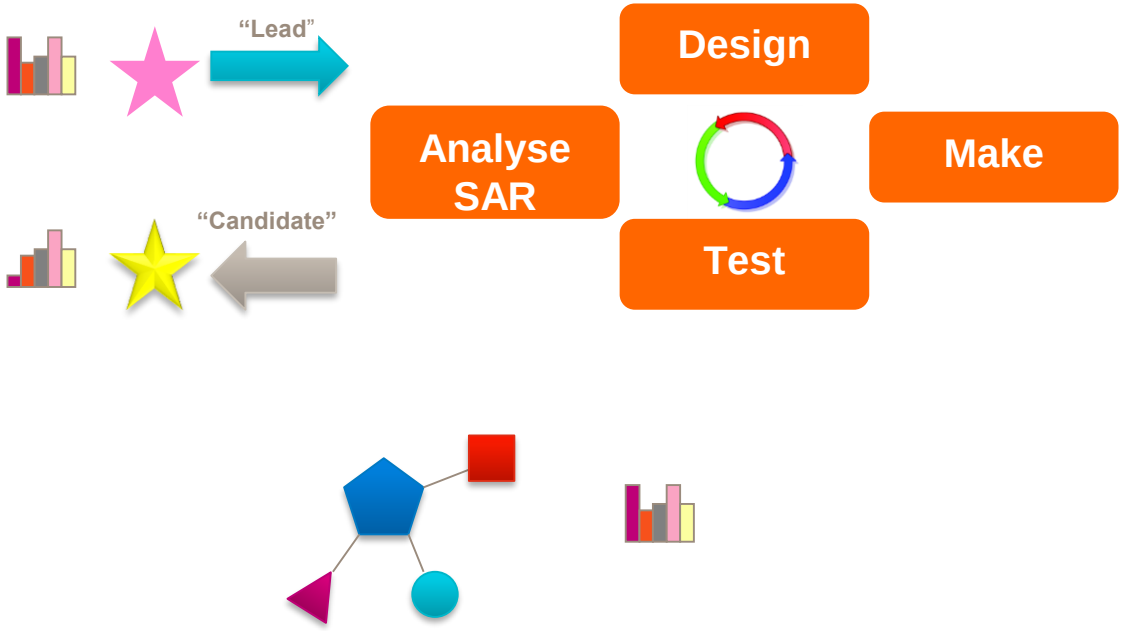
Lead Optimisation within Drug Discovery



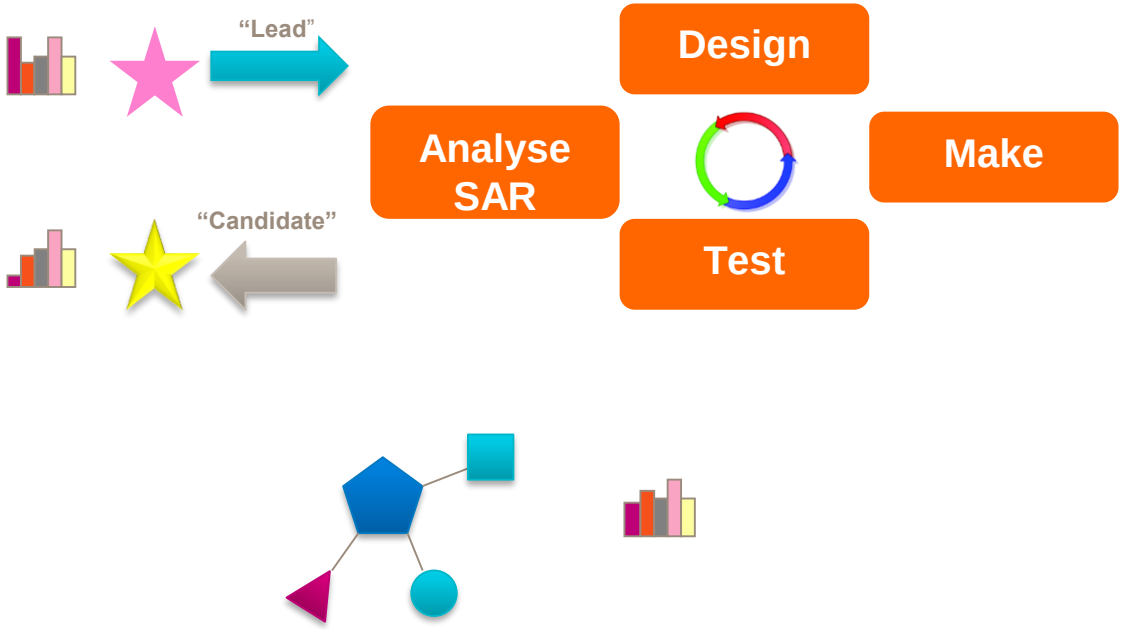
A multi-objective optimisation



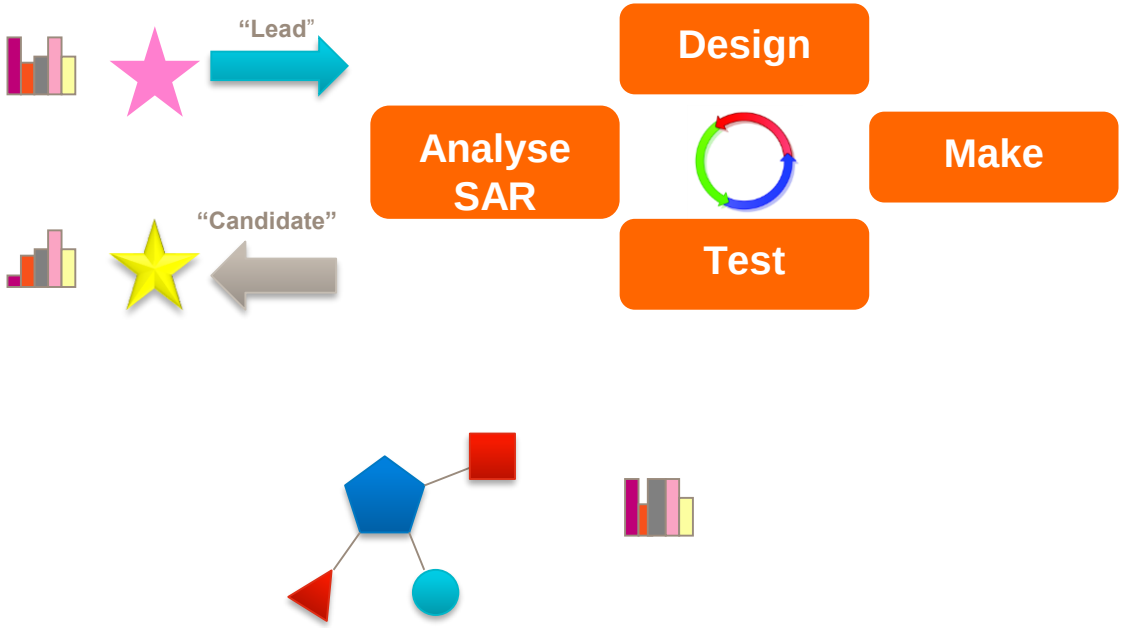
The Lead Optimisation cycle



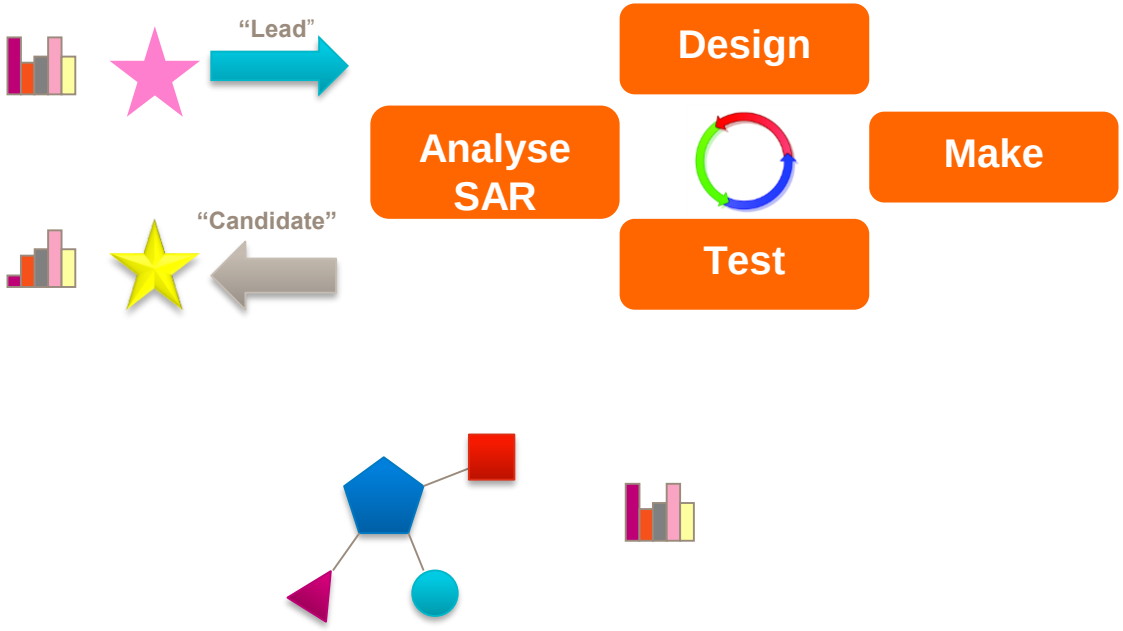
The Lead Optimisation cycle



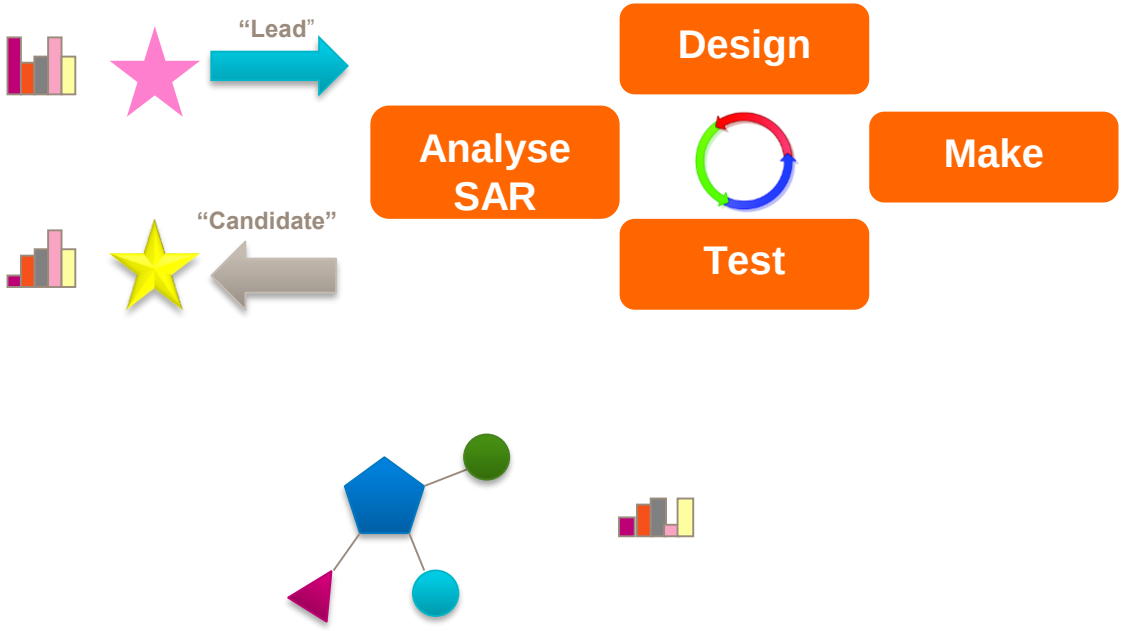
The Lead Optimisation cycle



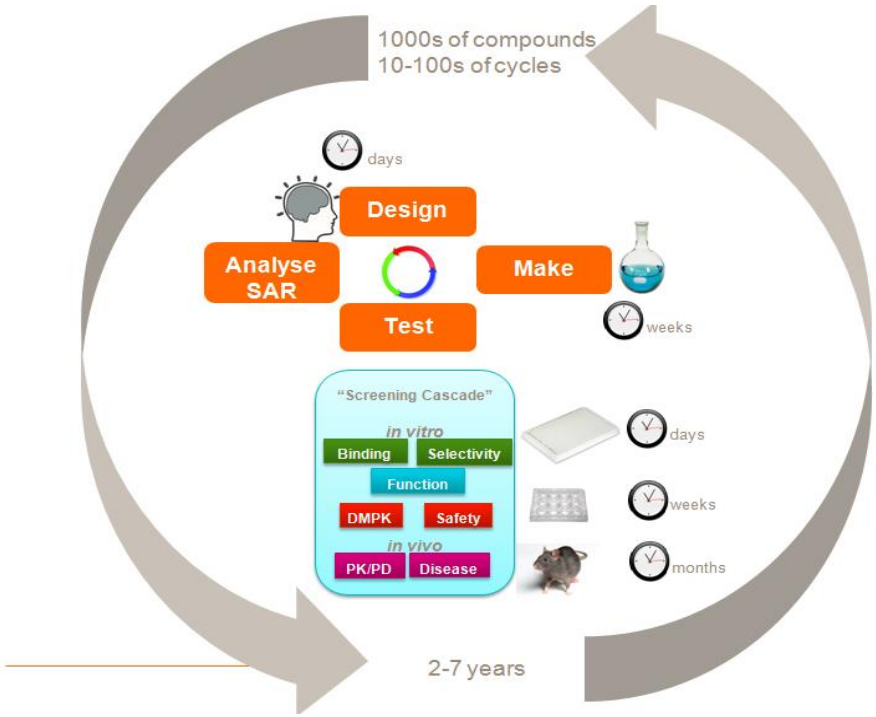
The Lead Optimisation cycle



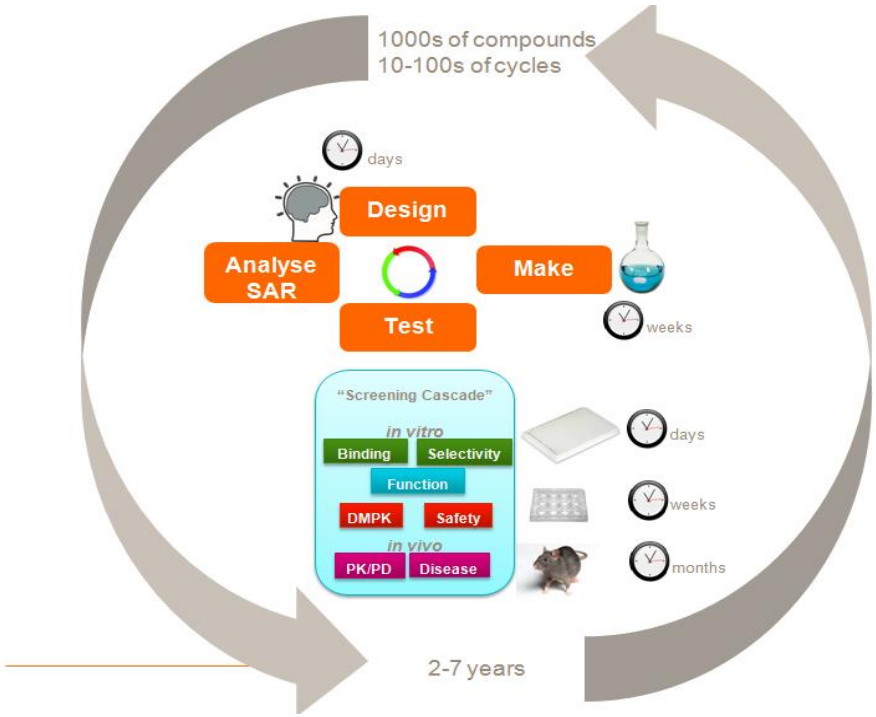
The Lead Optimisation cycle



The Discovery Cycle

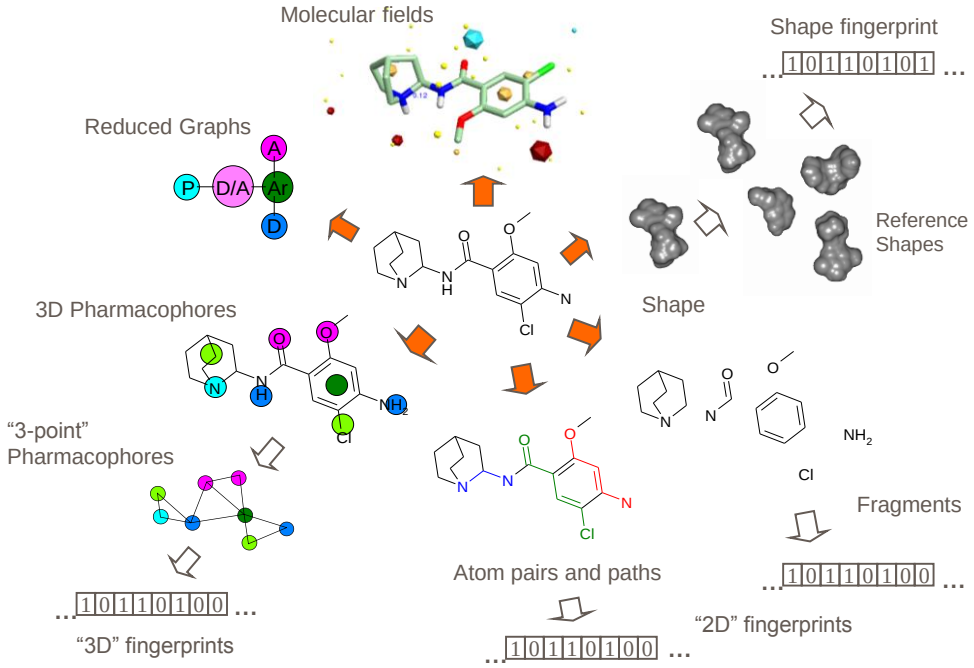


An ideal opportunity for Machine Learning?



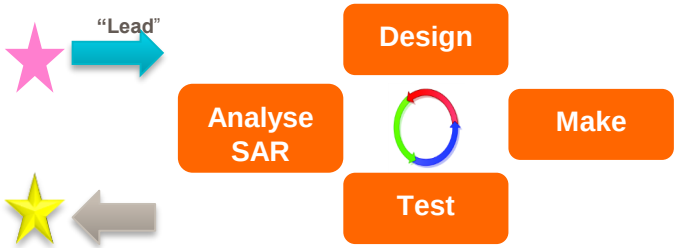
Quantitative Structure Activity Relationships: (QSAR)

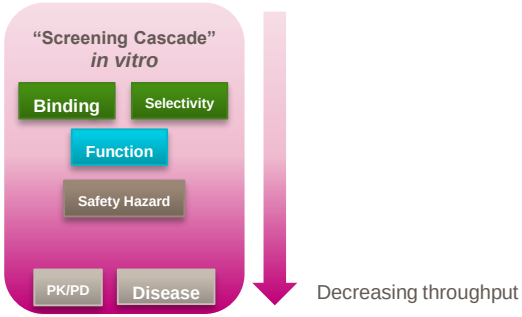
– A QSAR is a mathematical relationship between a measured parameter (e.g. biological activity) and the chemical characteristics of the molecule that was measured
  – since 1899 (and arguably1868)!



- MLR
- kNN
- PLS
- RF
- Bayes
- SVM
- NN
- deepNN
- etc etc

Improving the optimisation process





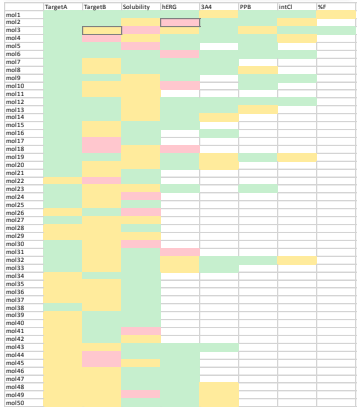
In order to have the most efficient optimisation
- Which compounds to make next?
- How to make them quickly?
- Which assays to run on which compound?

The nature of lead optimisation project data

Small
Start with 1, maybe 10 data points

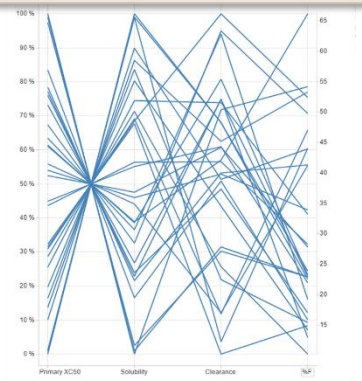
Slow & Expensive to grow

Sparse & Imbalanced

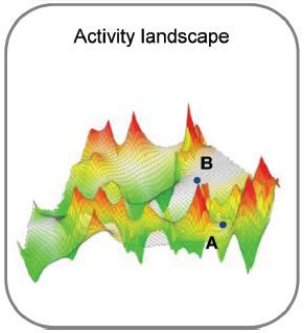


Large Search space

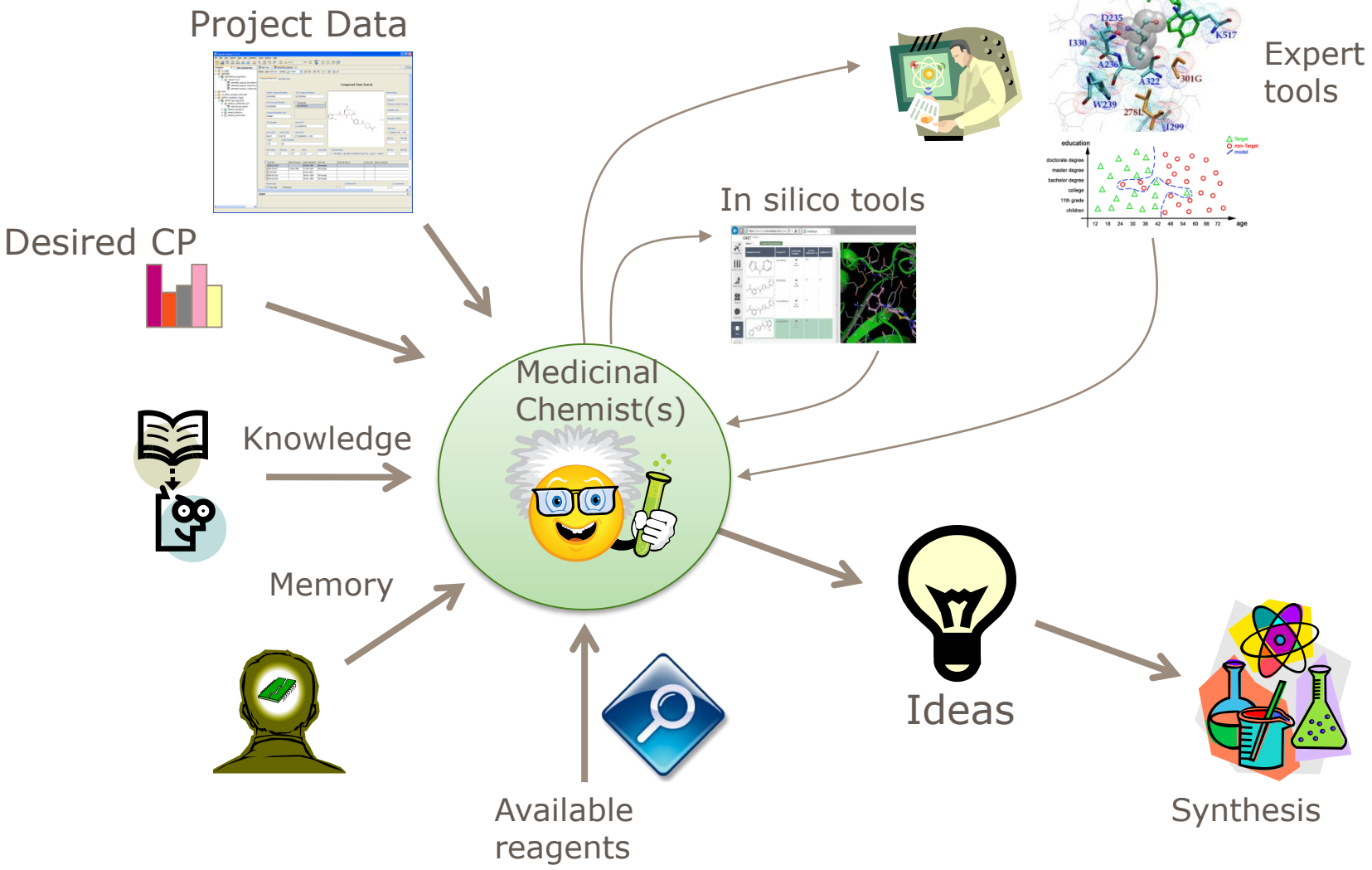
Competing & Multi-Objective



Discontinuous

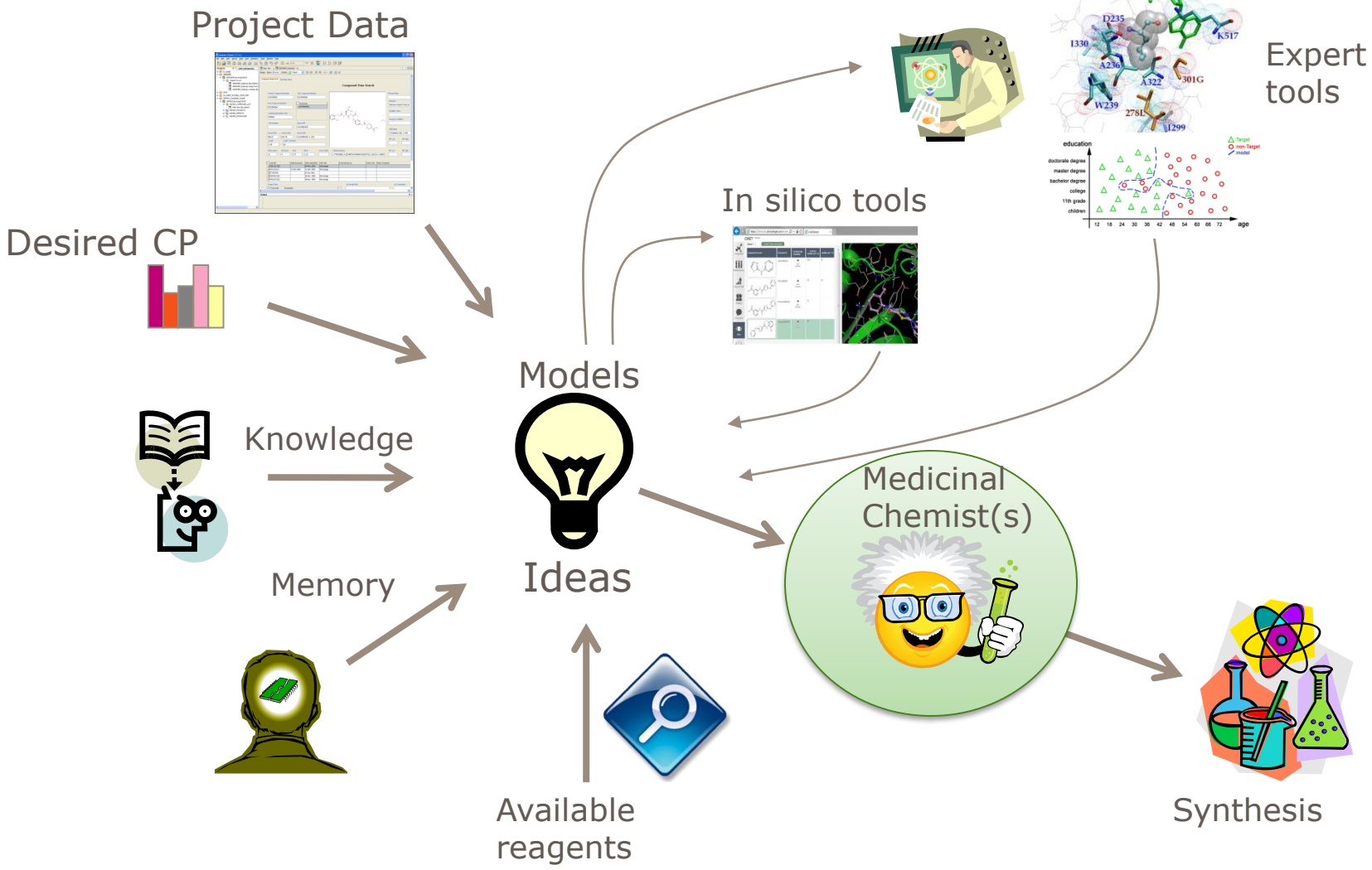


Current Model

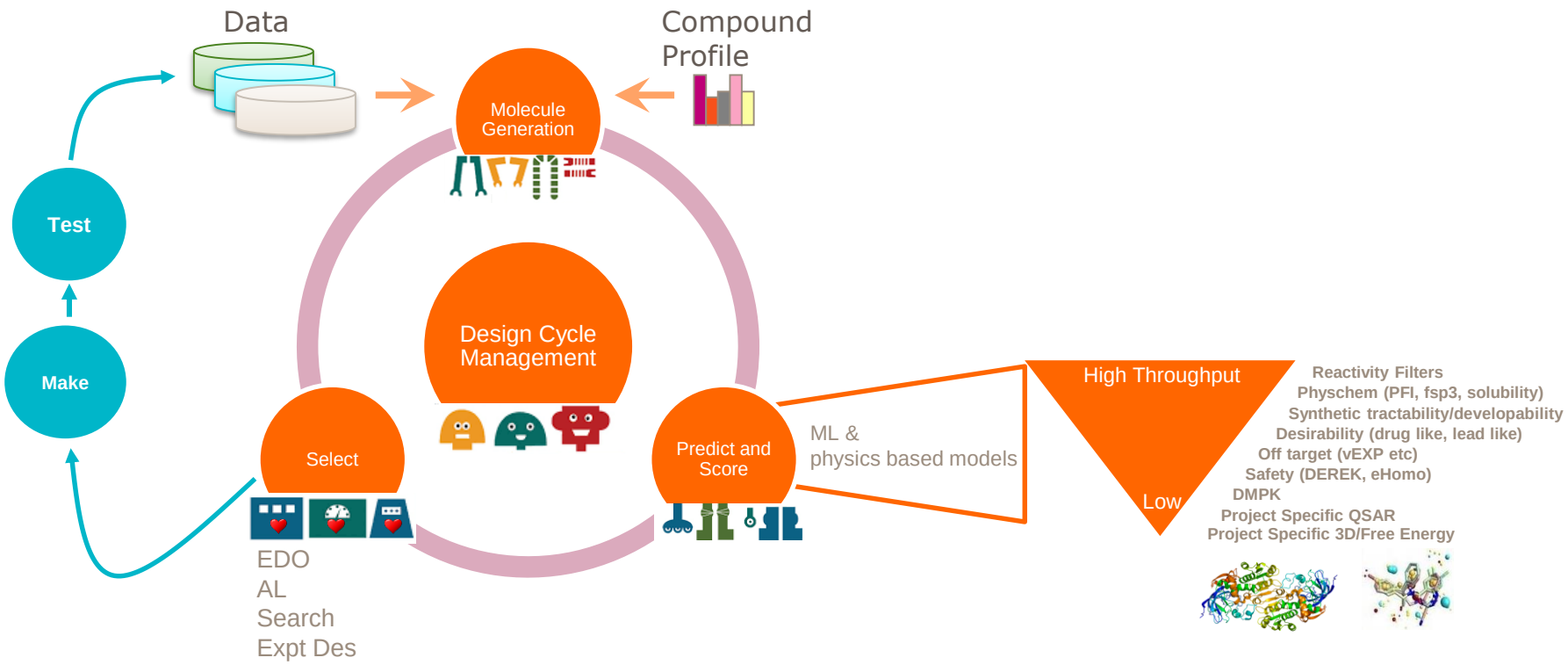


What if?

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



Building BRADSHAW: GSK's automated molecular design platform



George Box

1919-2013

All models are wrong; some models are useful

George Box

1919-2013

All models are wrong; some models are useful

For such a model there is no need to ask the question "Is the model true?". If "truth" is to be the "whole truth" the answer must be "No".

The only question of interest is
"Is the model illuminating and *useful*?“

George Box

1919-2013

All models are wrong; some models are useful

For such a model there is no need to ask the question "Is the model true?". If "truth" is to be the "whole truth" the answer must be "No".

The only question of interest is
"Is the model illuminating and *useful*?“

For our purposes, the question might be rephrased:

The only question of interest is:
"Is the model better than the current one?“

George Box

1919-2013

All models are wrong; some models are useful

For such a model there is no need to ask the question "Is the model true?". If "truth" is to be the "whole truth" the answer must be "No".

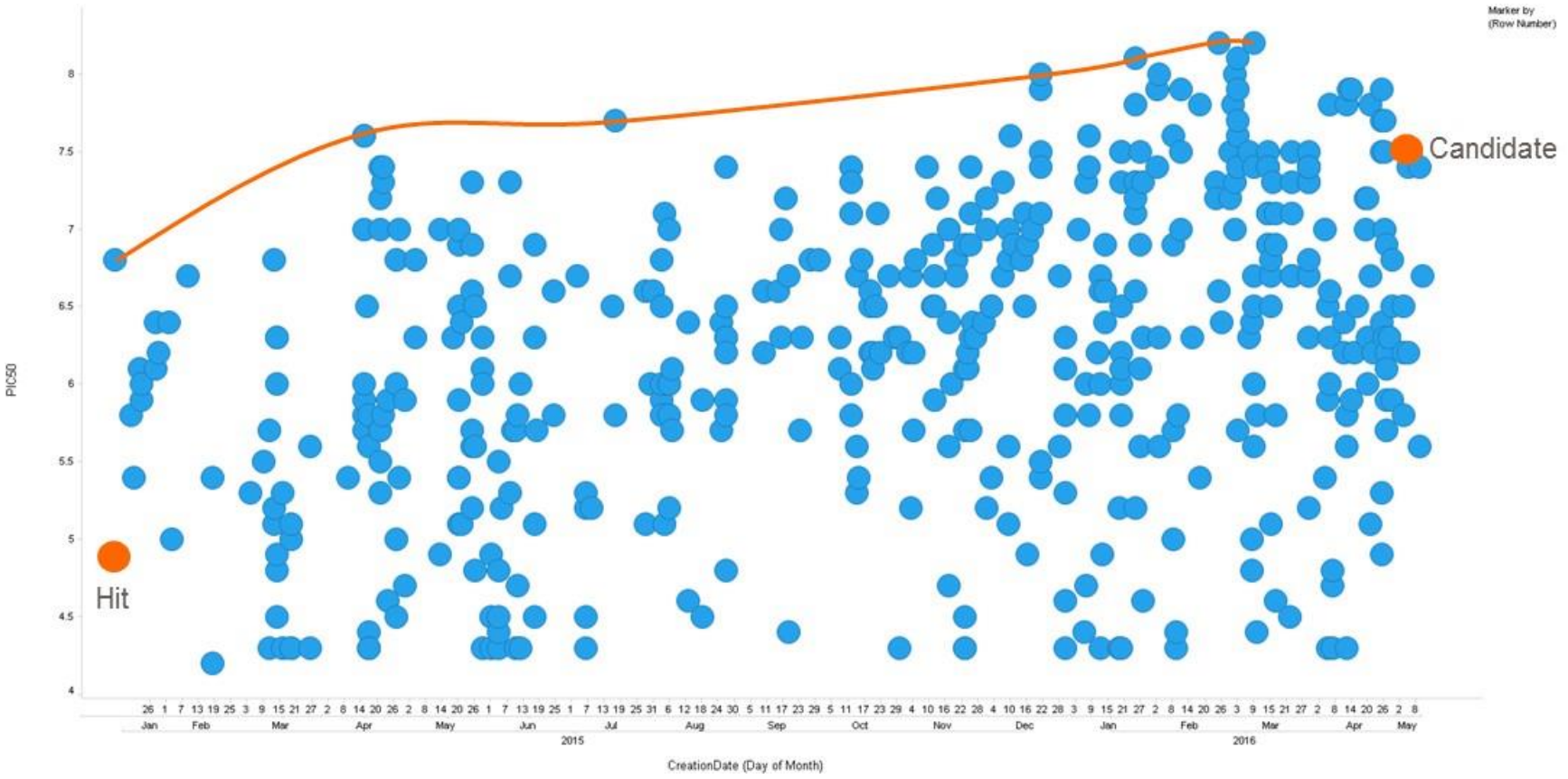
The only question of interest is
"Is the model illuminating and *useful*?“

For our purposes, the question might be rephrased:

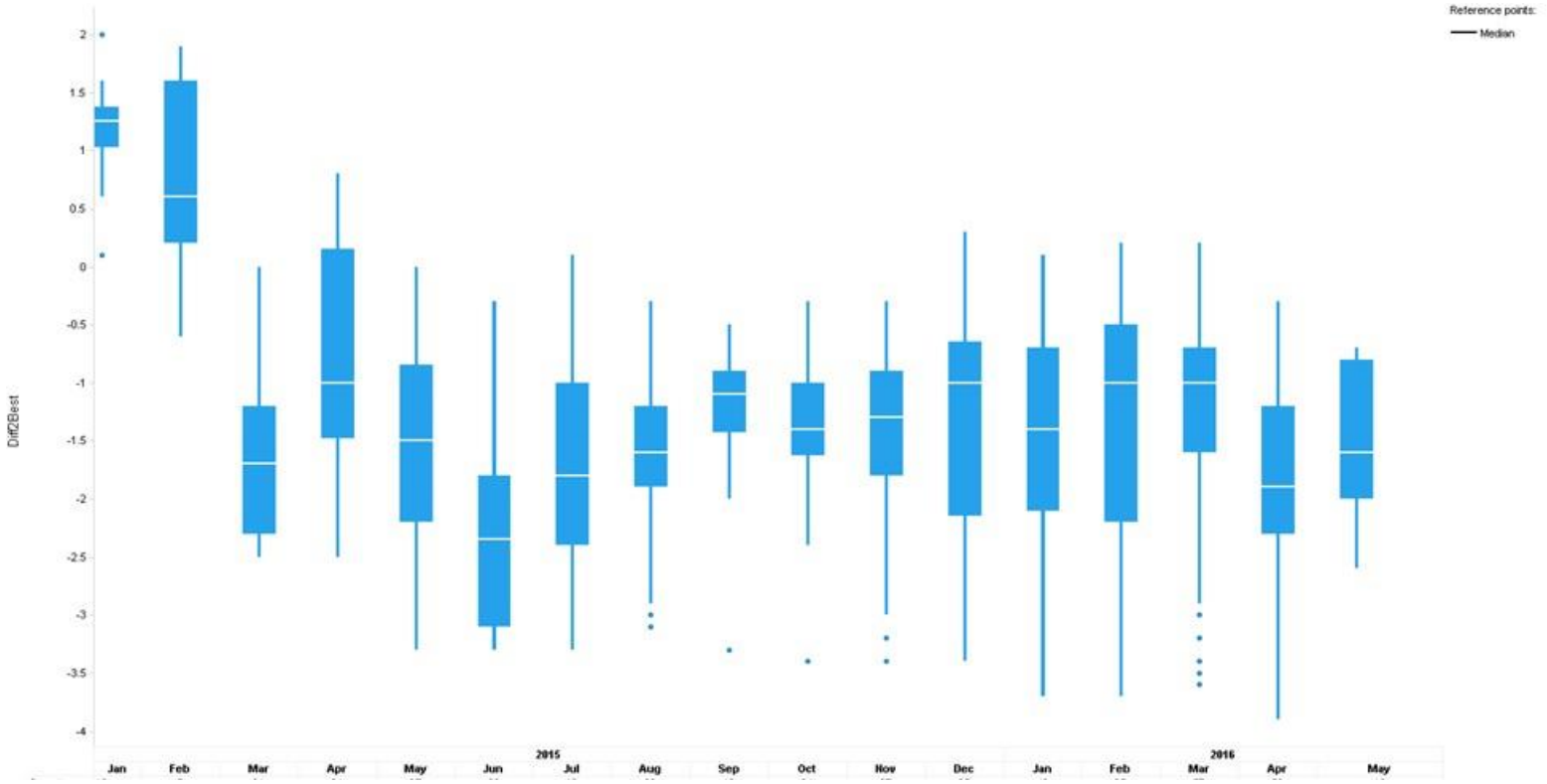
The only question of interest is:
"Is the model better than the current one?“

But how good is the current one? We shall have to quantify chemists’ intuition…

Quantitation case study



Method 1:
Δ(best compound known – activity achieved)



Method 2:
Kendall TAU

Kendall Tau Rank Correlation Coefficient

Discuss | Add to List | Share

Info | Videos | Discuss

In statistics, the **Kendall rank correlation coefficient**, commonly referred to as **Kendall's tau coefficient** (after the Greek letter ?), is a statistic used to measure the ordinal association between two measured quantities. A **tau test** is a non-parametric hypothesis test for statistical dependence based on the tau coefficient.

It is a measure of rank correlation: the similarity of the orderings of the data when ranked by each of the quantities. It is named after Maurice Kendall, who developed it in 1938,[1] though Gustav Fechner had proposed a similar measure in the context of time series in 1897.[2]

Intuitively, the Kendall correlation between two variables will be high when observations have a similar (or identical for a correlation of 1) rank (i.e. relative position label of the observations within the variable: 1st, 2nd, 3rd, etc.) between the two variables, and low when observations have a dissimilar (or fully different for a correlation of -1) rank between the two variables.

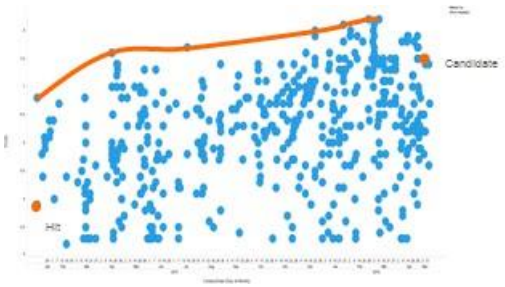
Both Kendall's τ and Spearman's ρ can be formulated as special cases of a more general correlation coefficient.

Definition

Let (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) be a set of observations of the joint random variables *X* and *Y* respectively, such that all the values of (x_i) and (y_i) are unique. Any pair of observations (x_i, y_i) and (x_j, y_j), where $i \neq j$, are said to be *concordant* if the ranks for both elements (more precisely, the sort order by *x* and by *y*) agree: that is, if both $x_i > x_j$ and $y_i > y_j$; or if both $x_i < x_j$ and $y_i < y_j$. They are said to be *discordant*, if $x_i > x_j$ and $y_i < y_j$; or if $x_i < x_j$ and $y_i > y_j$. If $x_i = x_j$ or $y_i = y_j$, the pair is neither concordant nor discordant.

The Kendall ? coefficient is defined as:

$$\tau = \frac{(\text{number of concordant pairs}) - (\text{number of discordant pairs})}{n(n-1)/2}.$$ [3]



τ = **0.094** for our dataset
i.e. no correlation/optimisation

http://www.like2do.com/learn?s=Kendall_tau_rank_correlation_coefficient

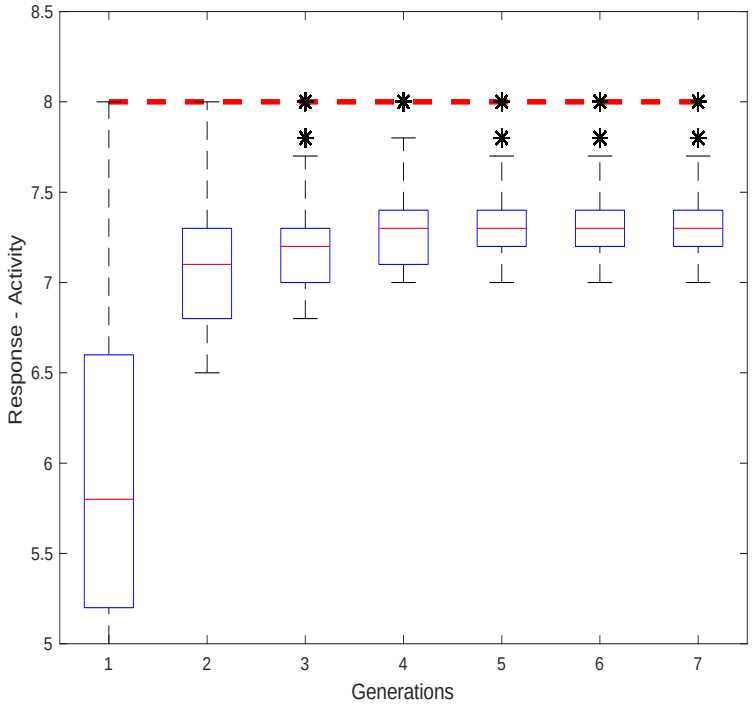
Simple model guided optimisation:
- build *one* QSAR model after 100 compounds, use it

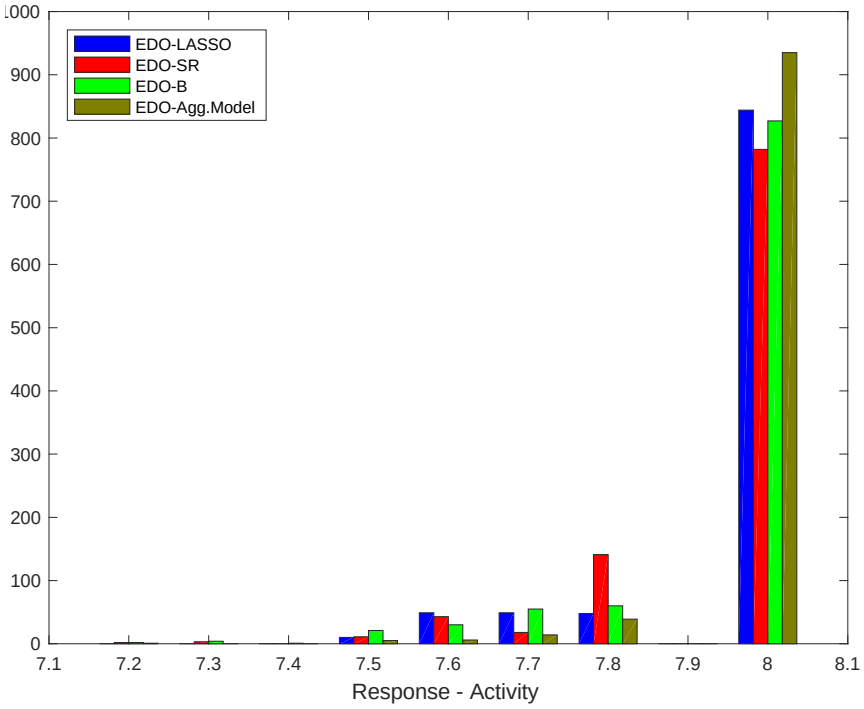
Saving =
~25% time and resource (!)

This is what *real* optimisation looks like

MMP-12 optimisation using an evolutionary optimisation method (EDO)

20 random start points followed by iterations of 20 selected compounds





Università Ca'Foscari Venezia

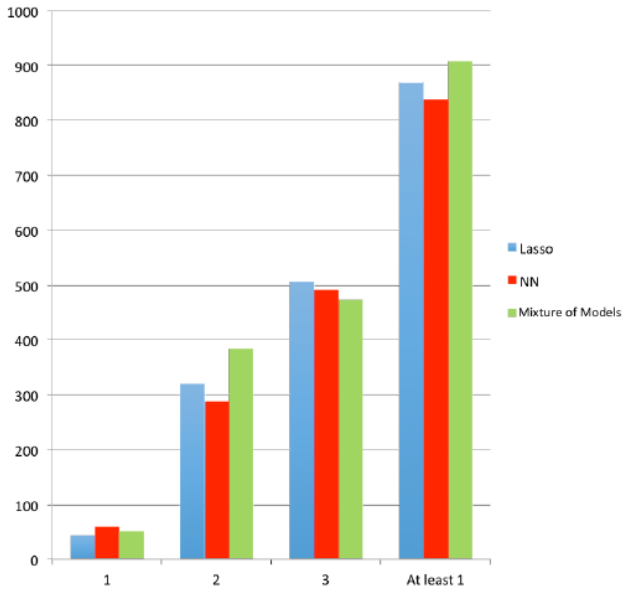
European Centre for Living Technology

Pickett, Green, Hunt, Pardoe, Hughes, Automated Lead Optimization of MMP-12, *ACS Med. Chem. Lett.* 2011,2.28-33

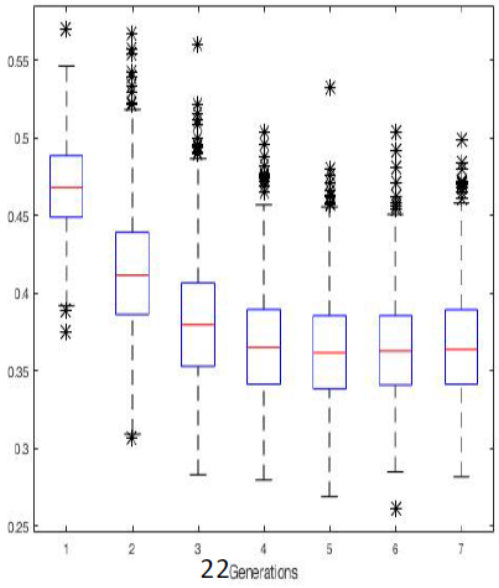
Multi-objectives: potent, soluble, safe

multi-objective optimization (m-EDO)

Best molecules found in 1000 runs

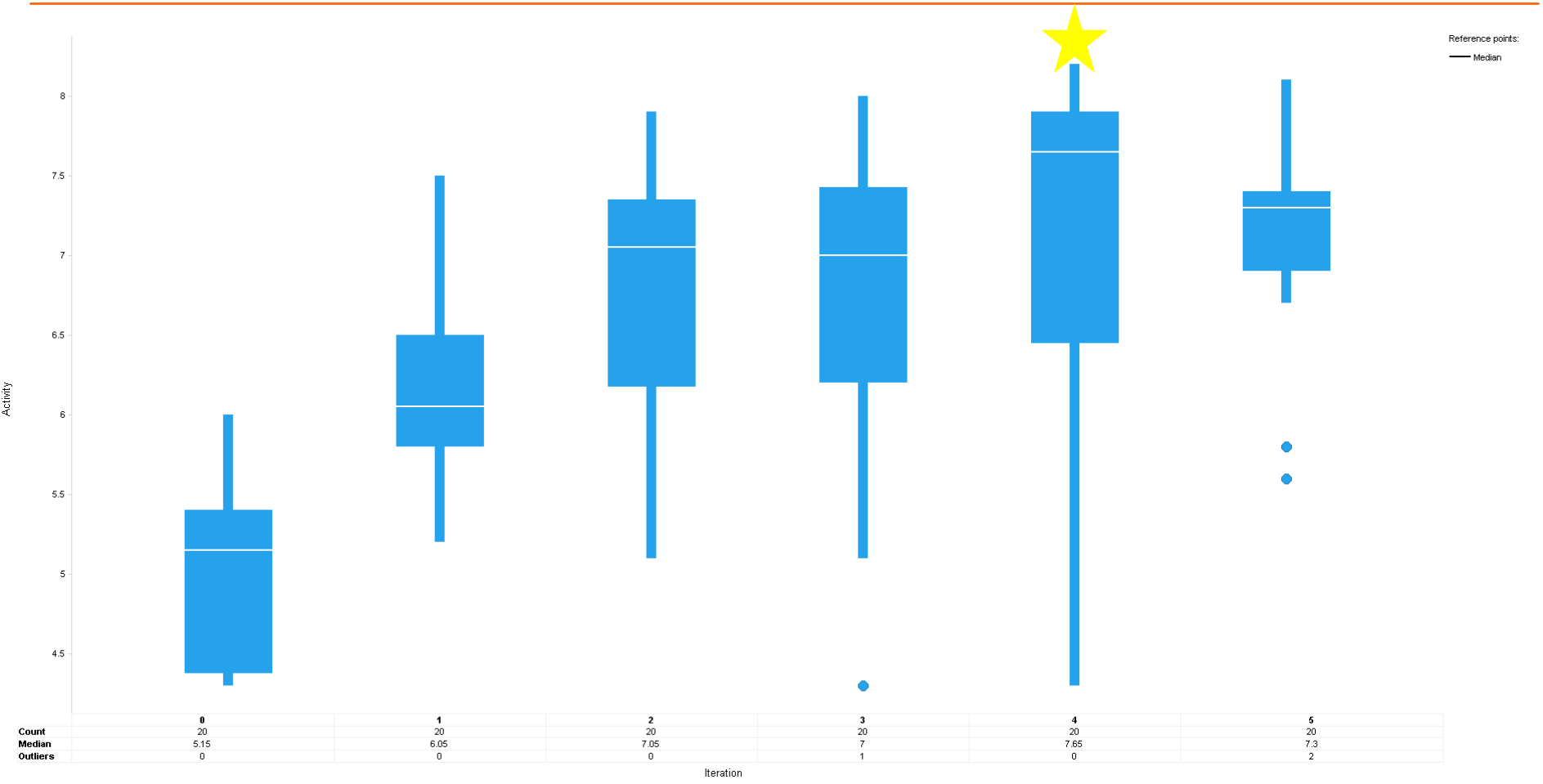


The boxplot of the molecule values achieved in 1000 runs at each generation with Mixture of Models



EDO on our Case study

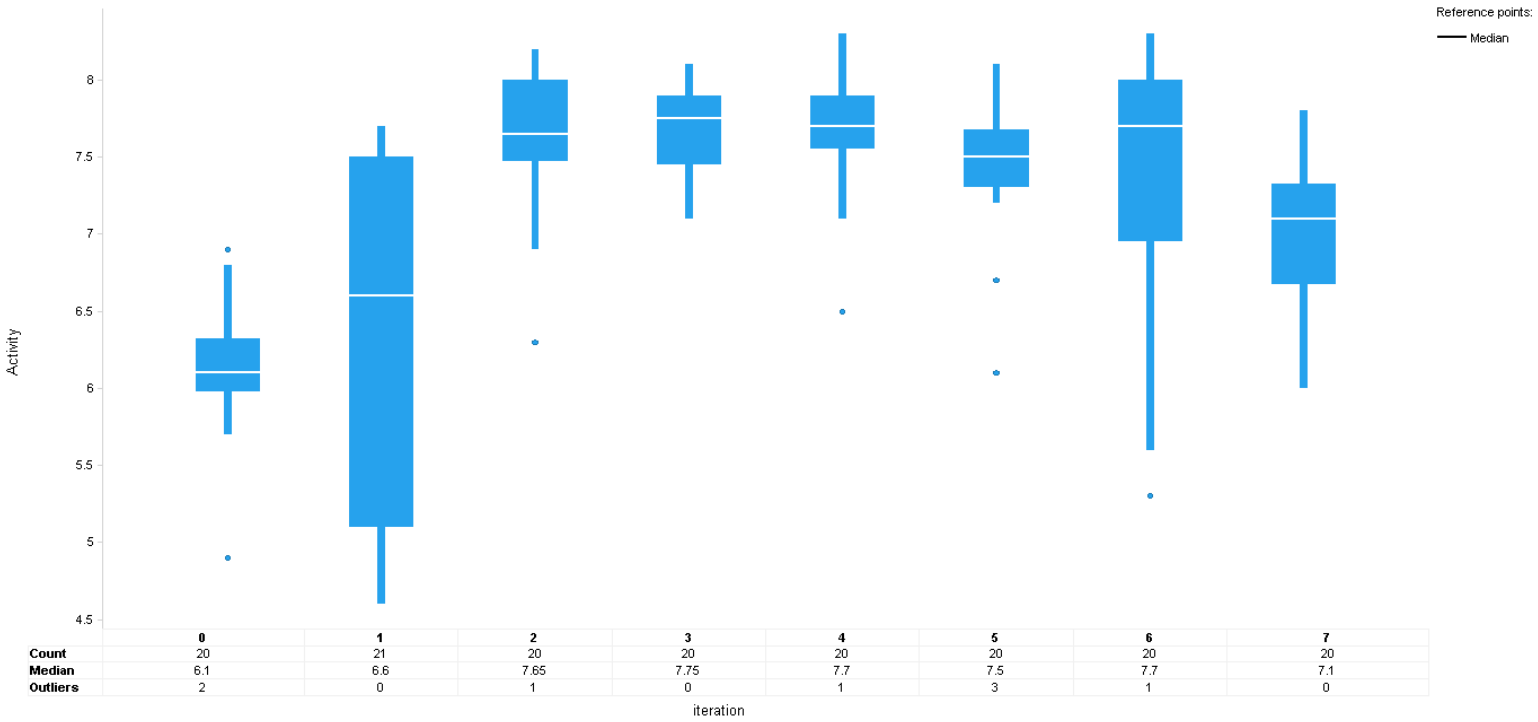
Best compound found in 100 compounds!



	0	1	2	3	4	5
Count	20	20	20	20	20	20
Median	5.15	6.05	7.05	7	7.65	7.3
Outliers	0	0	0	1	0	2

An example where EDO does not find the best molecule

Best molecule has Activity = 9.8

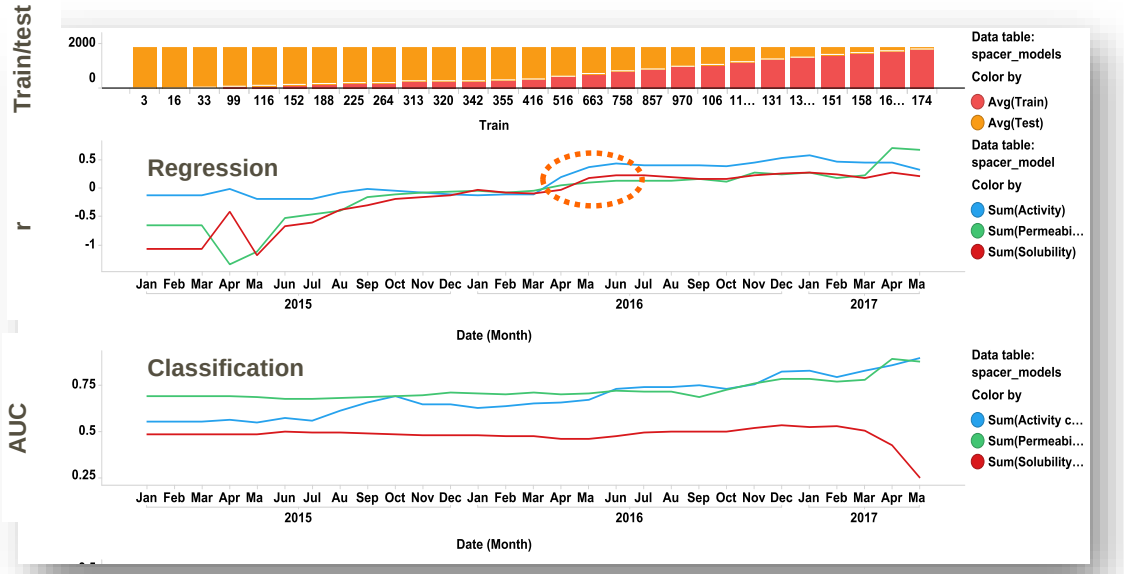


	0	1	2	3	4	5	6	7
Count	20	21	20	20	20	20	20	20
Median	6.1	6.6	7.65	7.75	7.7	7.5	7.7	7.1
Outliers	2	0	1	0	1	3	1	0

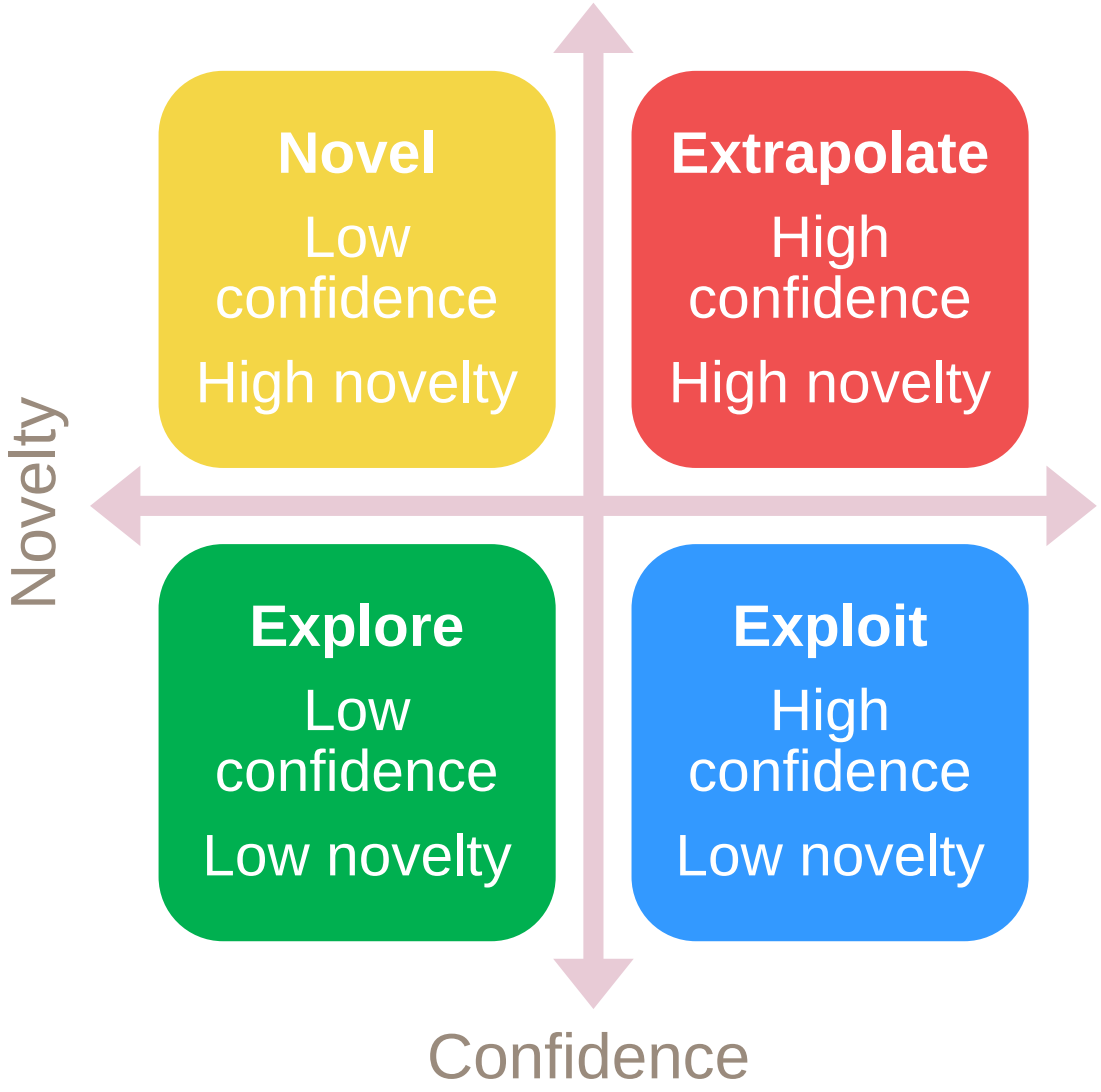
Project Telemetry

When should we intervene and start using models?

- Continual model building
  - Which endpoints are we now able to model?
  - When can we model with confidence?



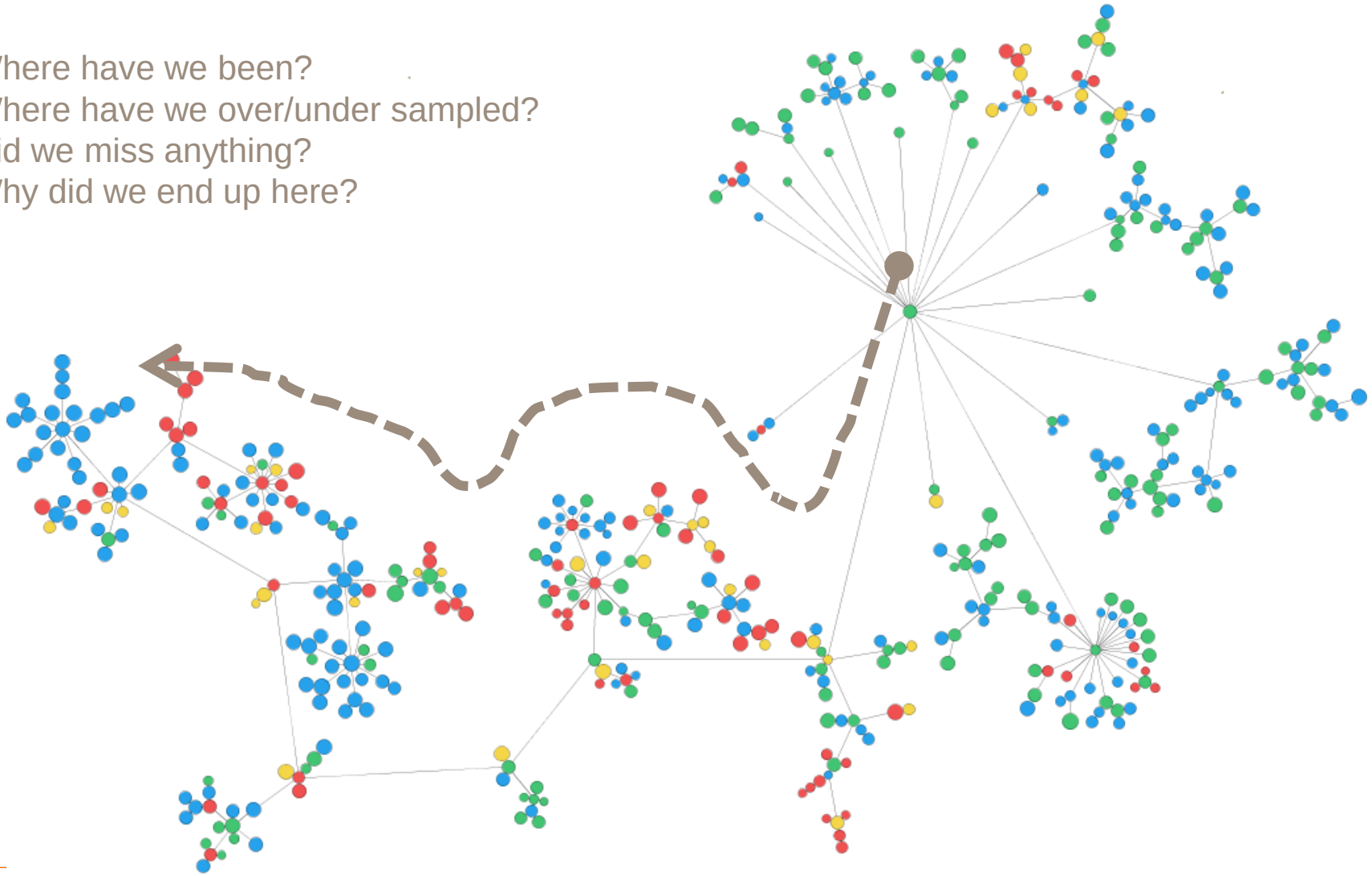
Using a language to describe optimisation steps



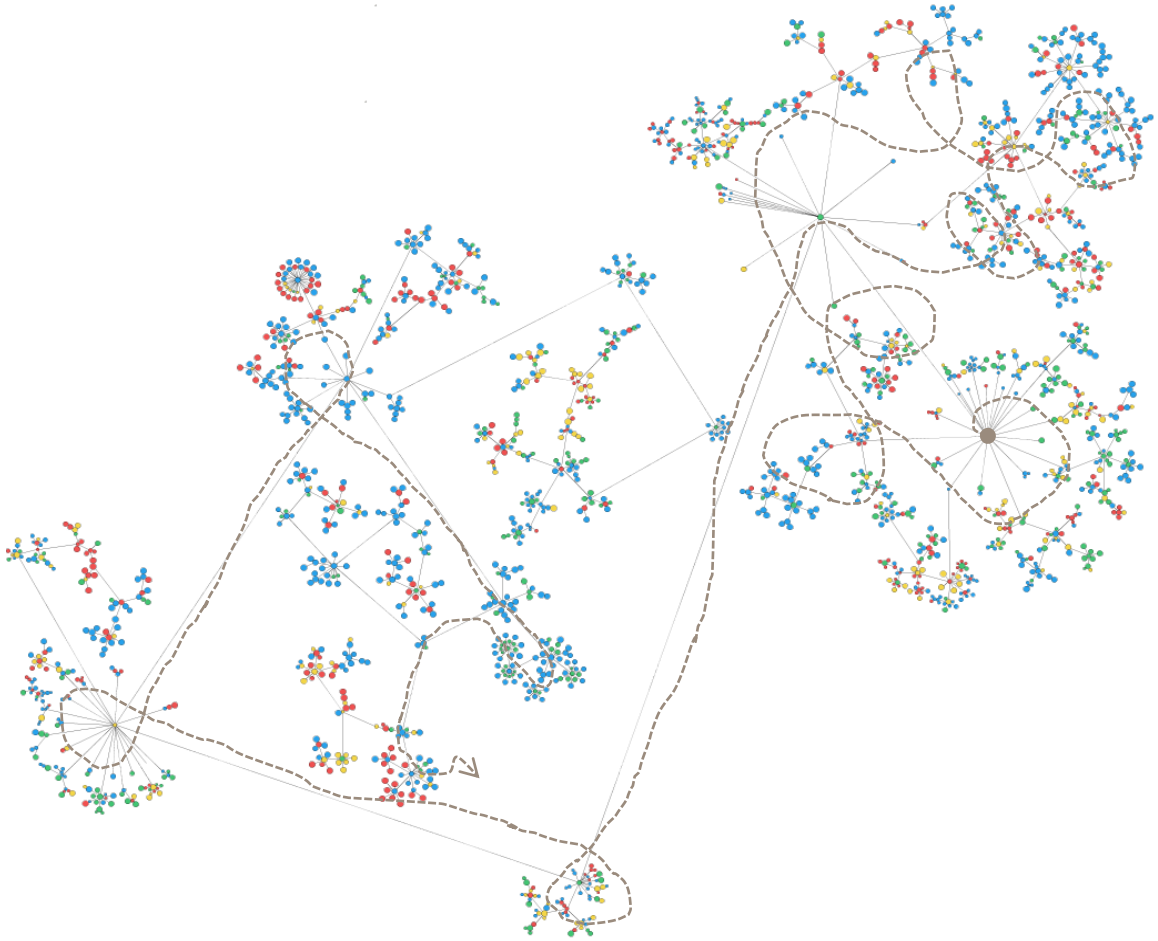
Visualisation of chemical space sampling

Reflect on what has been done or illustrate what we propose to do next

Where have we been?
Where have we over/under sampled?
Did we miss anything?
Why did we end up here?



Another real life example (!)



Lessons learnt from collaboration with ECLT

- Iterative model building (EDO) can be an effective optimisation strategy
- Our public test set (single parameter MMP12) maybe too simple- most methods work
- Encoding chemical structure for optimal model performance is not a solved problem
- For “real life” projects, we may need to generate data for initial model building (cf Chris Luscombe) before EDO-style optimisation
- Visualisation tools help people to understand what we are trying to achieve and reflect on previous performance
- Collaborating with people from different backgrounds and approaches is a productive and inspiring way of solving problems!

Acknowledgements

Stephen Pickett
Chris Luscombe
David Marcus
Paul Bamborough
Ian Wall
Denise Vlachou

Professor Irene Poli, Debora Slanzi, Alessandro Giovannelli, Marina Khoroshiltseva, Valentina Mameli (Università Ca'Foscari Venezia)
Professor Philip Brown (University of Kent)