Università
Ca'Foscari
Venezia

Evolutionary Design for Optimization (EDO)
with high dimensional small data

Irene Poli

European Centre for Living Technology,

Ca' Foscari University of Venice

Joint work with D. Slanzi, V. Mameli, F. Della Marra, A. Giovannelli, M.Khoroshiltseva, D. De March, M. Borrotti.
at *European Centre for Living Technology.*

With the collaboration of D. Green, C. Luscombe and Molecular Design Group, *GlaxoSmithKline (GSK), Research Medicines Research Centre, Stevenage, UK.*

BLOOM project (*www.ecltech.org)*

Venice, October 20, 2018

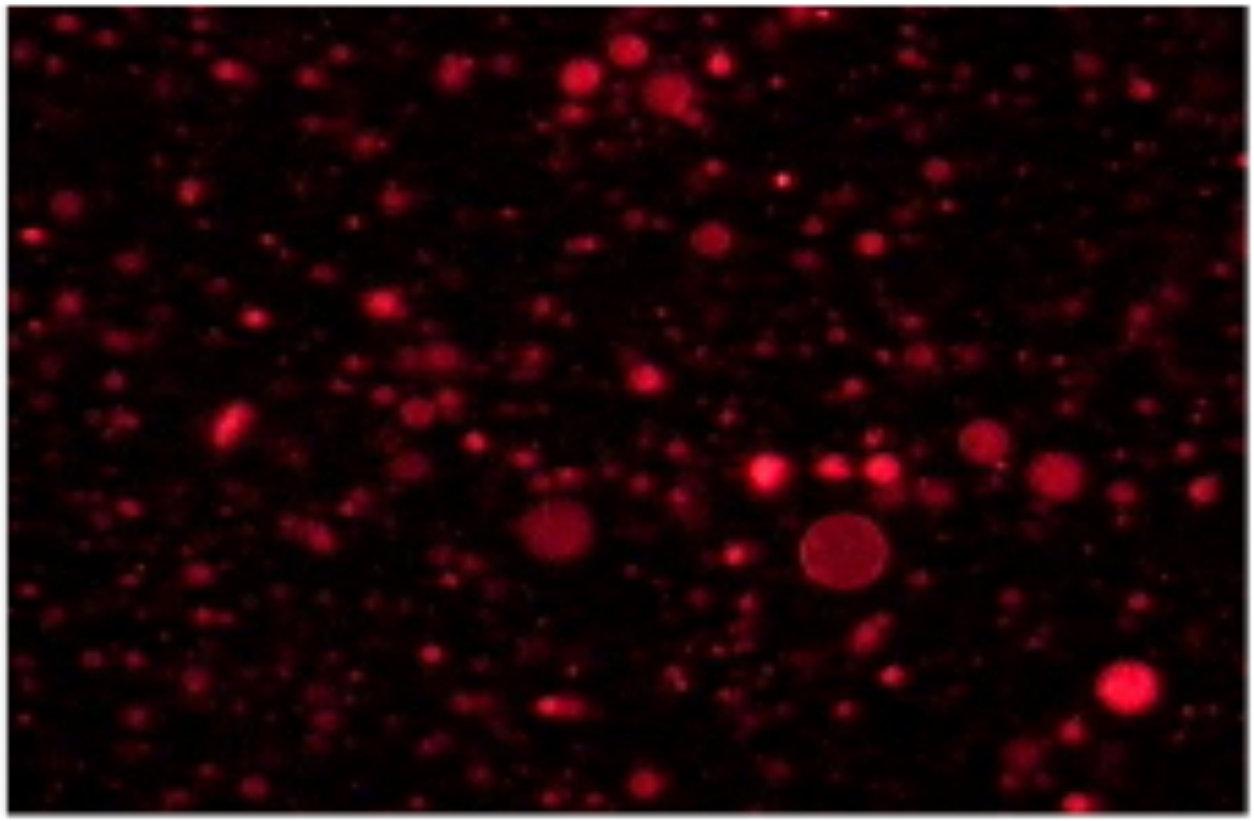
Outline

1) Motivating problems.

2) The Evolutionary Design for Optimization (EDO).

3) EDO for lead molecular optimization in drug discovery field.

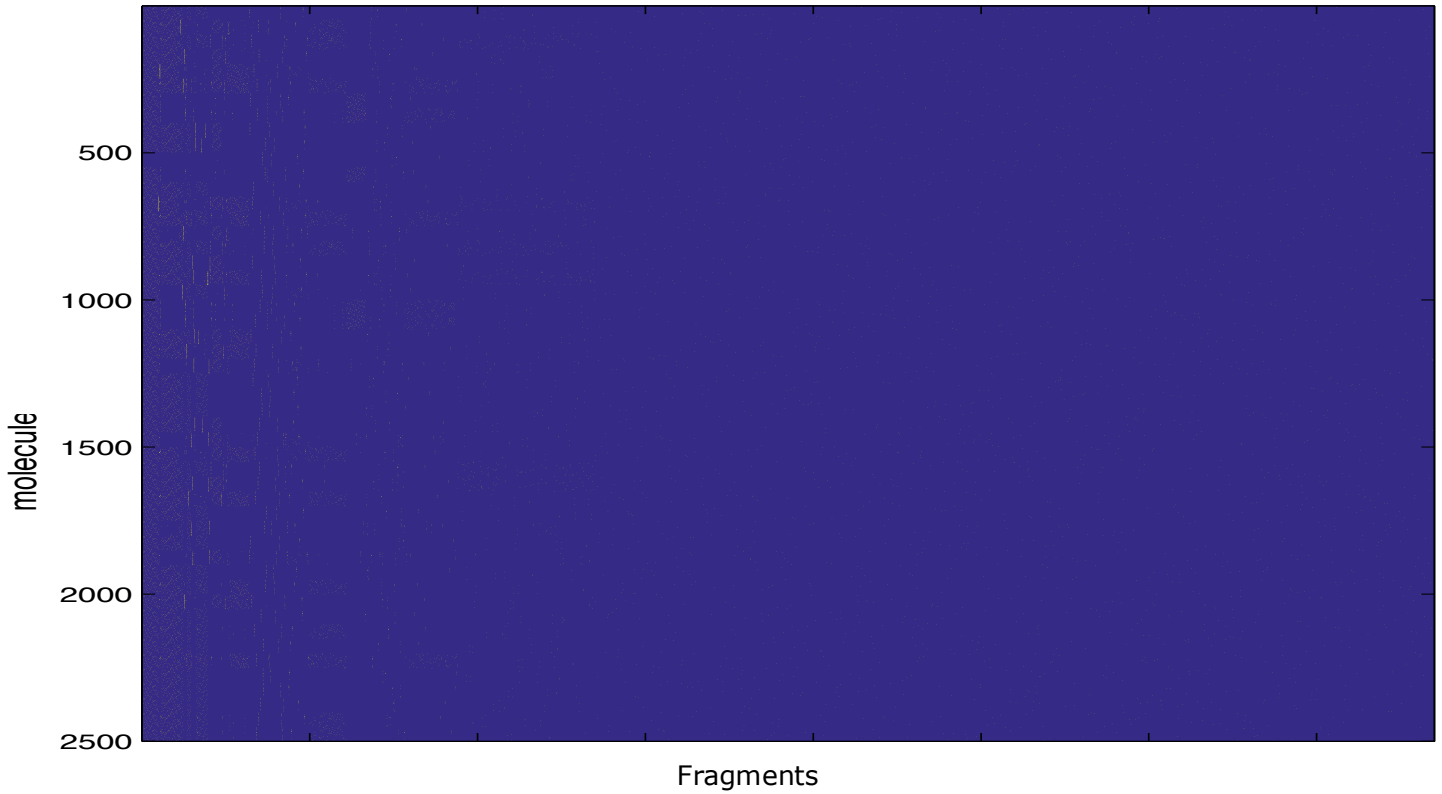
Building artificial cells



Generate the optimal molecular composition that give rise a cell vesicle (PACE, EU project)

Building molecules

with properties for drug development process



Generate the fragment composition that give rise to a candidate drug
(BLOOM project)

Design for Optimization

The systems described are *complex systems* in which a very large number of interacting entities give rise to *structures* whose properties should be *optimized* in order to accomplish a particular task.

These systems are characterized by *High Dimensionality*, and their study is developed with experimentation.

Experimentation is expensive, time consuming, frequently polluting and developed on living organisms.

Q: how to design experiments for optimization with the smallest number of design points?

Design for Optimization

Optimization generally consists of finding the **best values of some objective function** given a defined domain. Best values are typically the maximum or minimum values of the function, but other values can be considered.

Some notation:

we consider a finite space **X,** whose elements are the candidate sets of predictors (the compositions to be tested in laboratory), $\mathbf{X} = (\mathbf{x}_1,\ldots,\mathbf{x}_n)'$, with $\mathbf{x_i} = (x_{i1},\ldots,x_{ip})$, i=1,…,n, and

define an objective function f: **X** → R, from some set of **X** to real numbers.

We search the element $\mathbf{x}_o \in \mathbf{X}$ such that $f(\mathbf{x_o}) \geq f(\mathbf{x})$ for all $\mathbf{x} \in \mathbf{X}$ (maximization).

→ We do not know the function f, we must assume and estimate it by testing (or observing) the n elements of **X.**

→ We assume to test a very small set of **X**, namely n_m, with $n_m << p$.

Our **goal** is to find the element $\mathbf{x}_o \in \mathbf{X}$ considering just n_m elements of **X.**

Design for Optimization

Frequently more than one objective function is to be optimized so we address a ***multi-objective optimization problem,*** described by a vector-valued objective function

$$\boldsymbol{f} : \mathbf{X} \rightarrow R^k \ , \quad \boldsymbol{f}(\mathbf{x}) = (f_1(\mathbf{x}) \ \ldots \ f_k(\mathbf{x})).$$

Objectives can be *conflicting*.

In several problems is not possible to find the solution that can optimize simultaneously all of them.

Decision are then taken with some trade offs (combining solutions or deriving Pareto Optimal solutions, that is, solutions that cannot be improved in any of the objectives without degrading at least one of the other objectives).

Which strategy with $n_m << p$

The strategy requires:

- The selection of the **design** of experiments: *size* and *structure* of each single experimental point.

- The selection of the **model** to describe the relation between the structure and the experimental response: model selection, variables selection, dimension reductions, etc..

- The selection of the **Optimization criteria**: select the best composition with respect to a particular target.

Which strategy with $n_m << p$

Which **design**? Which **model?** Which optimization **criteria?**

Several and different approaches have been proposed in the literature
e.g. Factorial design; Response surface methodology; Optimal Design; V-Optimal design;

but the *high dimensionality* makes unfeasible or extremely hard to adopt classical design

the Evolutionary Design for Optimization (EDO)

We propose an **evolutionary strategy** (from the field of nature inspired computation) where the design is regarded as a small population of experimental points that *evolves in search* of the optimal solution.

The evolution is driven by *models on data* collected by the experimentation.

A simple and general description of the strategy is the following:

- A **first population** of experimental points $\mathbf{x_{j1}} = (x_{j11},\ldots,x_{jp1})$, $j=1,\ldots,n_{m1}$ is randomly chosen (and possibly prior information). The **experiment** is conducted, and an n_{m1}-dimensional response vector $\mathbf{y_1}=(y_{11},\ldots, y_{nm1,1})'$ is generated.
- A set of predictive **models** A_l, $l=1,\ldots,L$, are selected and estimated on this first collected data (model selection uncertainty, variable selection uncertainty).
- **Predictions** on the entire experimental space are derived, a prediction **combining method** is defined, and the best predicted values become the second population $\mathbf{x_{j2}} = (x_{j12},\ldots,x_{jp2})$, $j=1,\ldots,n_{m2}$ (second generation of the algorithm).

The procedure is iterated until a defined stopping rule is satisfied.

the Evolutionary Design for Optimization (EDO)

Design:

Is sequential and is selected according to the optimization criterion.

Models:

We draw inference from data using different statistical models for high dimensionality, built at any generation of the evolution:

Penalized Regression models (Lasso models), Clustering penalized regressions,
Stepwise Regression, Bayesian Regressions with different classes of priors,
Random Forests, Boosting,
Bayesian Networks, Neural Networks models

The information from these models is evaluate *in robustness, and prediction accuracy.*

Optimization:

At any generation we combine the best values in prediction from different models and select them as members of the new generation to be evaluated in laboratory.

Lead Optimization of MMP-12 Inhibitors

MMP -12 :

- **Matrix metalloproteinase -12**, is an **enzyme** that in humans is encoded by the *MMP12 gene*.
- This enzyme is involved in inflammation pathways and their precise biological *role* is still unknown.
- Various Research Centres are investigating the use of MMP-12 inhibitors in treatment of chronic obstructive pulmonary disease (COPD), asthma, rheumatoid arthritis and cancer.

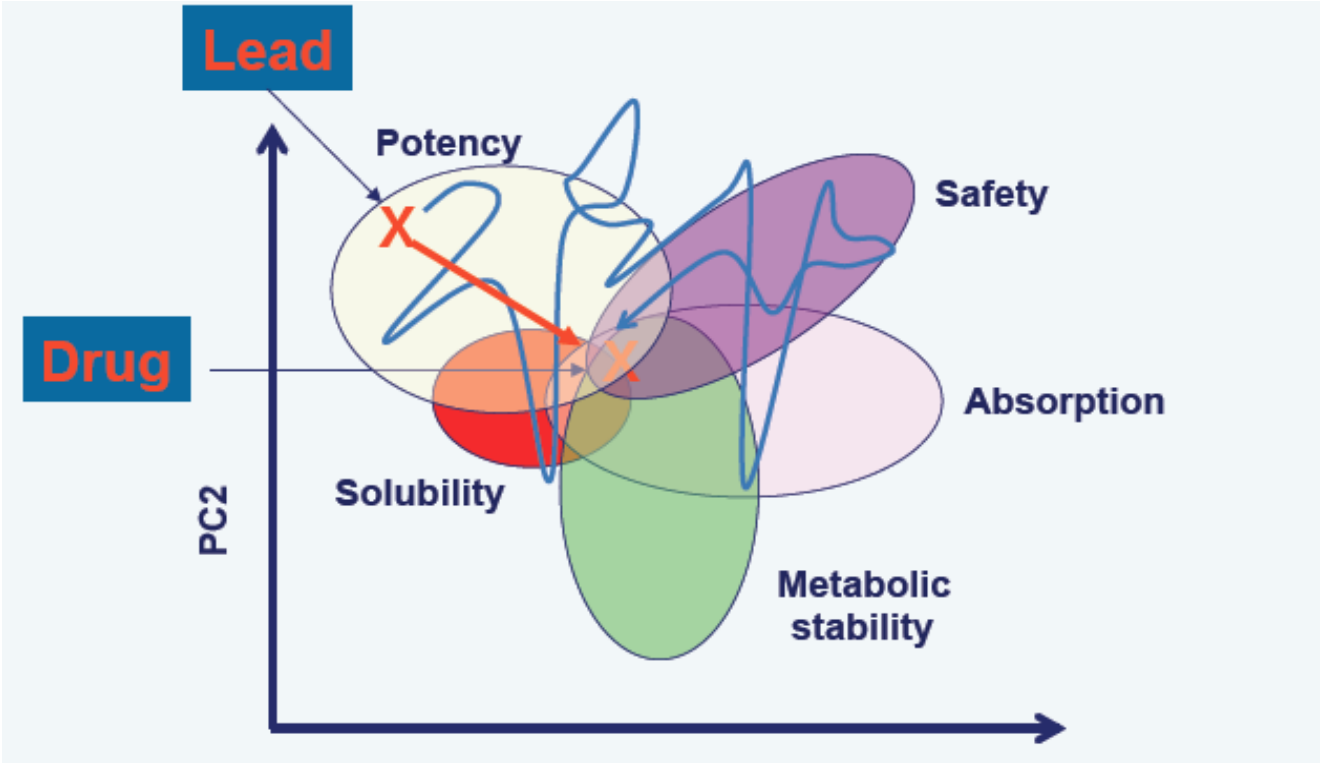
Lead Optimization of MMP-12 Inhibitors

A key phase of new drug discovery process concerns the **identification of small molecules modulators of protein function**, under the hypothesis that their activity can affect a particular disease state.

Current practice relies on the **screening of vast libraries of small molecules (often 1-2 million molecules)** in order to identify *a molecule* which specifically inhibits or activates the protein function, commonly known as **a *Lead Molecule, (LM)*.**

The lead molecule interacts with the required target, but generally lacks *other attributes* needed for a drug candidate, such as: absorption, distribution, metabolism and excretion (ADME).

Lead Optimization as a multi-objective optimisation problem



(GSK-group representation)

EDO for Lead optimization of MMP-12

The problem description:

A set of **2500 molecules** is considered, **n=2500**;

Each molecule is represented by a set of 22272 **Fragments**, **p=22272.**

Fragments are described by their presence or absence: the **X** matrix consists of **binary variables;**

X is a matrix (2500 x 22272).

Discover the best molecule in terms of a set of properties testing just 140 experimental points, $\mathbf{n_m = 140.}$

The **X** matrix is known and is made public by **Pickett and *al.*[1].**

The properties to be optimized are five ***unknown*** **response variables (Y_1,...,Y_5),** namely**:**

Y_1= Activity; Y_2=Solubility; Y_3=Safety; Y_4=cLogP, Y_5=Molecular Weight.

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

EDO for Lead optimization of MMP-12

We address then the problem of the simultaneous optimization of

Multiple and Competing Objectives:

- maximise the **Activity**; maximise the **Solubility;** maximise the **Safety;**
- minimise **cLogP**; minimise **MW**.

the experimental space: 22272 Fragments

sparse high dimensional space

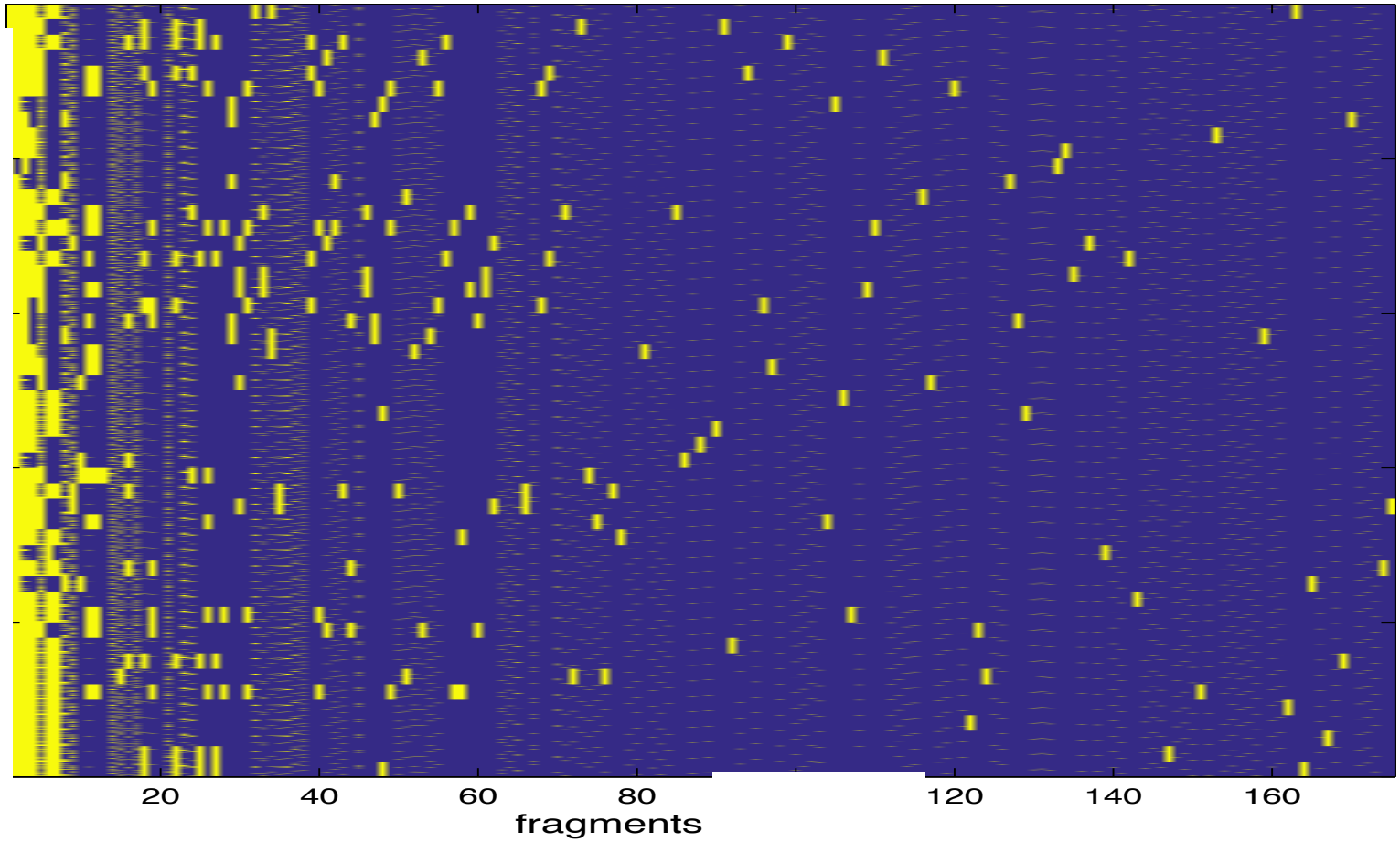
Dimensionality reduction

With the aim to reduce the dimensionality of the experimental space we proceed on different stages and **procedures NOT involving the response function** values, namely Y_i.

At the *first stage* we evaluated the *multicollinearity condition* and derived the smaller experimental space consisting of **4059 Fragments,** and on this space we developed the EDO approach pursuing the single objective optimization.

As a *second stage, based on a different representation of the **X** matrix (in terms of ATAG-BTAG)* we adopted the *Galois conceptual clustering (a machine learning paradigm for unsupervised classification*) and reduce the search space to **175 Fragments** on which we developed single objective and multi-objective optimization.

A fragment selection representation: 175 fragments



EDO for Lead optimization of MMP-12

We built a design consisting of **7** populations of **20** molecules each, with a total of **140 tests.**

- We selected the first population, a subset of **X**, say $\mathbf{X_1}$ (n_{m1} x p); selection has been at random and tested in the lab to achieve $\mathbf{y}_{11}$ (n_{m1} - vector).
- On this data set ($\mathbf{y_{11}}$,$\mathbf{X_1}$) we build and estimate a set of **statistical models** to **predict** the unknown response values of all the **x** not selected (2500-20), and we **combine** them.
- We chose and combined the best predicted values (highest values if we pursue maximization) for building the next generation $\mathbf{X_2}$.
- We iterate the algorithm until $\mathbf{X_7}$.

We evaluated the robustness of the procedure by repeating EDO for **1000 runs** (1000 repetitions with a different initial random set of the experimental units).

The full experimental space

To evaluate the performance of EDO the full library of tests on the 2500 molecules have been provided, making available the responses of all the 2500 molecules, that is $(Y_{i1},\ldots,Y_{i5})$, $\mathbf{X} = (\mathbf{x}_1,\ldots,\mathbf{x}_{2500})'$, with $\mathbf{x_i} = (x_{i1},\ldots,x_{ip})$, i=1,…,2500.

Activity (Y_1) presents a maximum value of **8**, and this corresponds to the optimal value. The 99-th percentile of the response variable distribution is **7.5** (**maximization** of Y_1).

Solubility (Y_2) presents a maximum value of **-1.766**, which corresponds to the optimal value. The 99-th percentile of the response variable distribution is **-2.415** (**maximization** of Y_2).

Safety (Y_3) presents a maximum value of **3.6262**, which corresponds to the optimal value. The 99-th percentile of the response variable distribution is **3.2309** (**maximization** of Y_3).

clogP (Y_4) presents a minimum value of **-2.505**, which corresponds to the optimal value. The 1-th percentile of the response variable distribution is **0.033** (**minimization** of Y_4).

Molecular Weight (Y_5) presents a minimum value of Y5 is **291.3**, which corresponds to the optimal value. The 1-th percentile of the response variable distribution is **339.3** (**minimization** of Y_5).

EDO performance based on 4059 Fragments and 140 tests.

Number of runs (out of 1000 runs) in which EDO finds the optimum value and values in the region of optimality (99-th percentile)

		Lasso	StepWise	Boosting	Comb. Predic.	Clu-Bayesian	Clu-H-single	Clu-H-average	CLU-SVR
Activity	Optimum	582	773	578	841	745	820	807	803
	Reg. of opt.	**995**	**996**	**998**	**1000**	**981**	**996**	**998**	**998**
Solubility	Optimum	852	660	777	895	693	792	783	756
	Reg. of opt.	**997**	**1000**	**1000**	**1000**	**997**	**997**	**996**	**996**
Safety	Optimum	450	325	276	425	392	307	328	345
	Reg. of opt.	**1000**	**1000**	**1000**	**1000**	**1000**	**999**	**1000**	**999**
cLogP	Optimum	915	799	839	899	661	859	804	855
	Reg. of opt.	**992**	**1000**	**1000**	**1000**	**912**	**956**	**955**	**958**
Molecular weigth	Optimum	569	663	519	651	926	641	695	653
	Reg. of opt.	**922**	**1000**	**1000**	**1000**	**991**	**887**	**907**	**898**

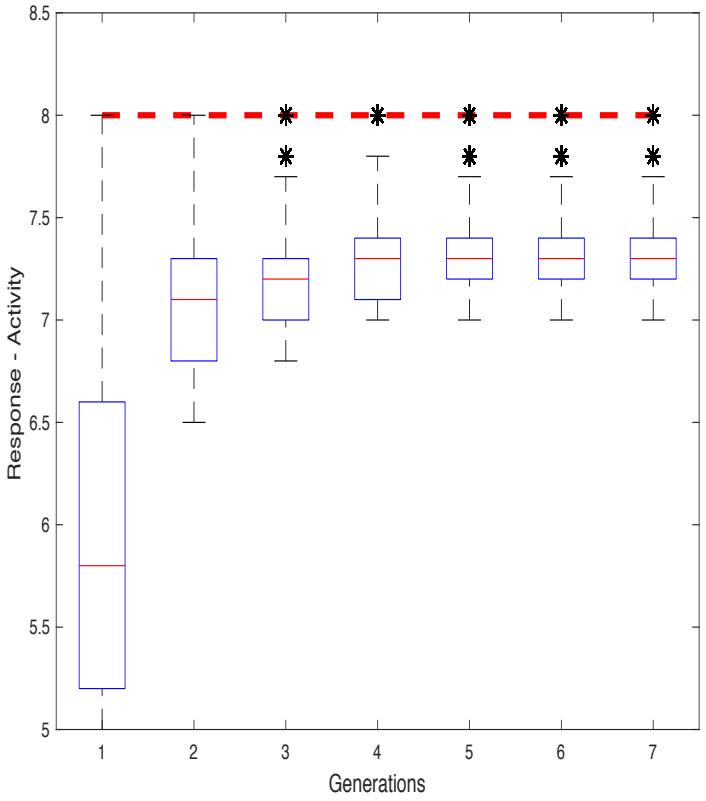
EDO performance based on 175 Fragments and 140 tests

Number of runs (out of 1000 runs) in which EDO discovers the optimum value and values in the region of optimality (99-th percentile)

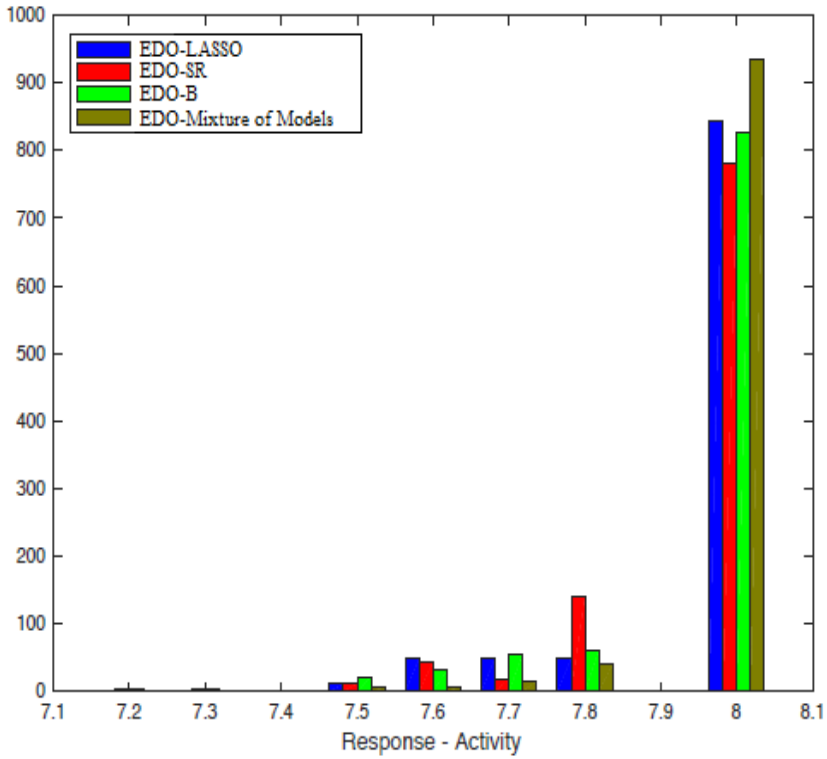
		Lasso	Stepwise	Boosting	**Comb. Predictions**	NN	Quantile R	SVR
Activity	Optimum	844	782	665	916	660	827	848
	Reg. of opt.	**1000**	**995**	**998**	**1000**	**990**	**997**	**999**
Solubility	Optimum	875	745	872	912	556		
	Reg. of opt.	**995**	**998**	**1000**	**1000**	**996**		
Safety	Optimum	857	858	911	920	628		
	Reg. of opt.	**1000**	**1000**	**1000**	**1000**	**999**		
ClopgP	Optimum	848	821	917	918	760		
	Reg. of opt.	**950**	**946**	**981**	**1000**	**945**		
MW	Optimum	738	822	751	887	746		
	Reg. of opt.	**905**	**966**	**956**	**1000**	**880**		

Activity optimization

Evolution through generations:
experimental values at each generation achieved in 1000 runs with Comb. Pred.

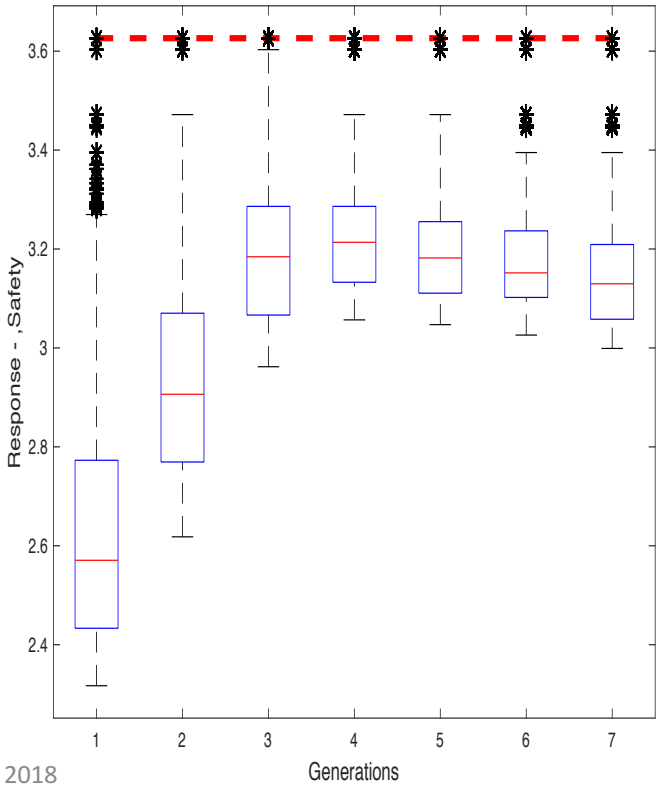


Experimental response values achieved in 1000 runs
in the region of optimality

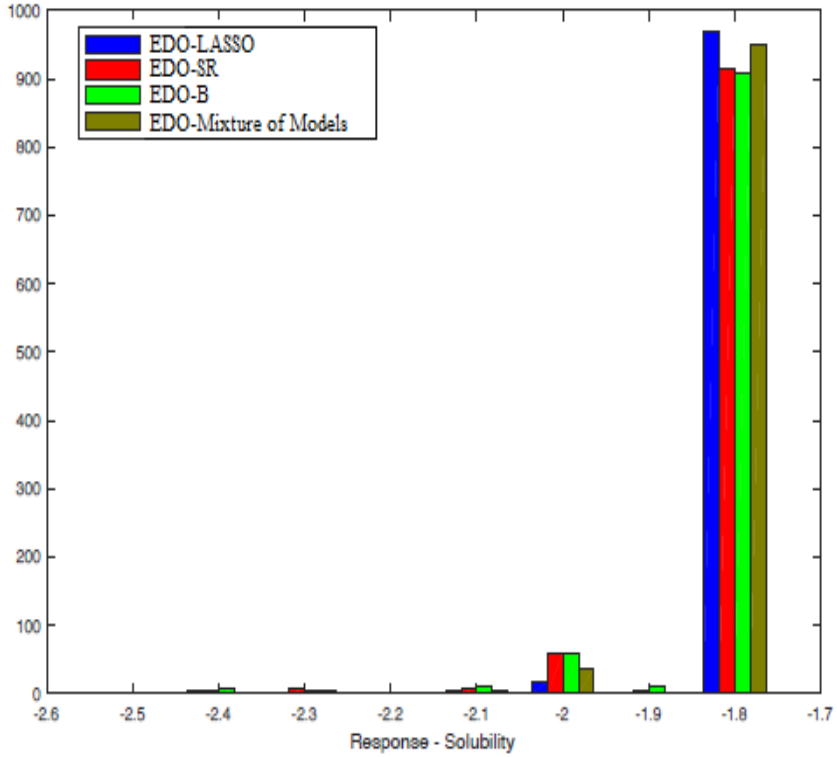


Solubility optimization

Evolution through generations:
experimental values at each generation achieved in 1000 runs with Comb. Pred.

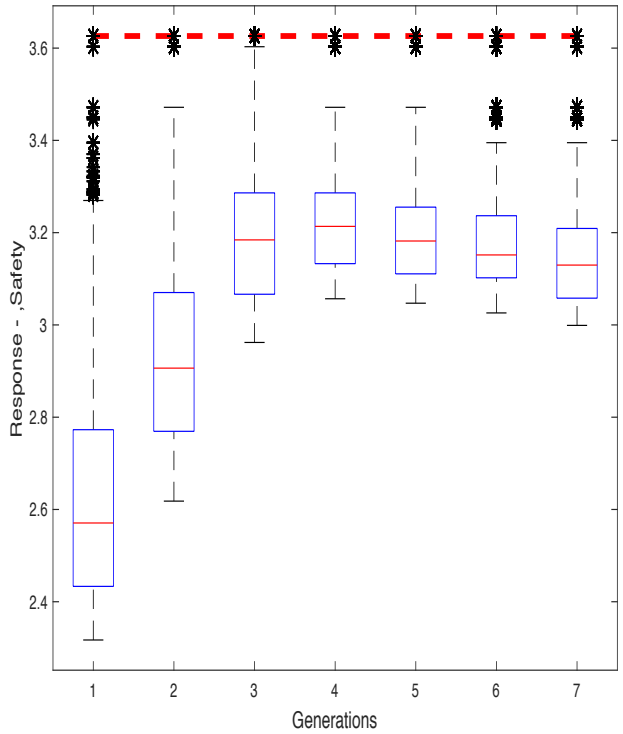


Experimental response values achieved in 1000 runs
In the region of optimality

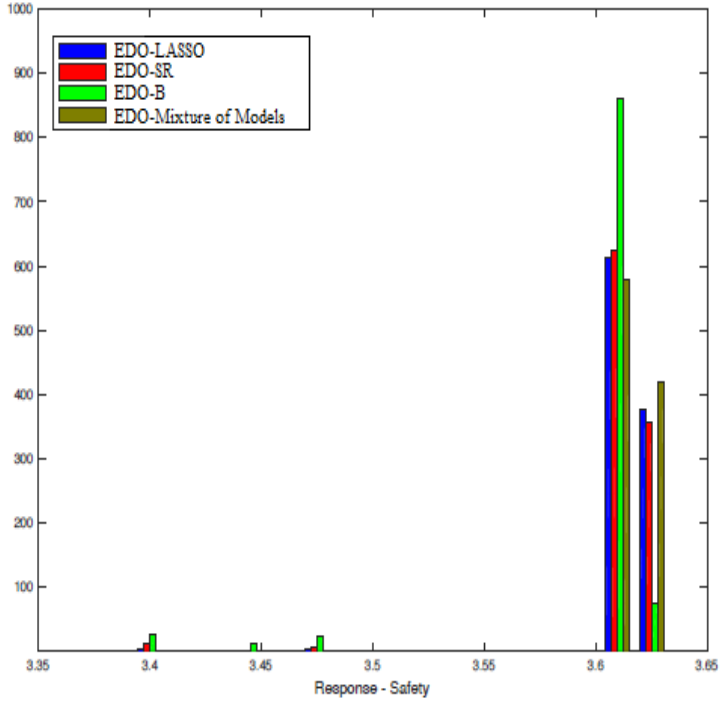


Safety optimization

Evolution through generations:
experimental values at each generation achieved in 1000 runs with Comb. Pred.

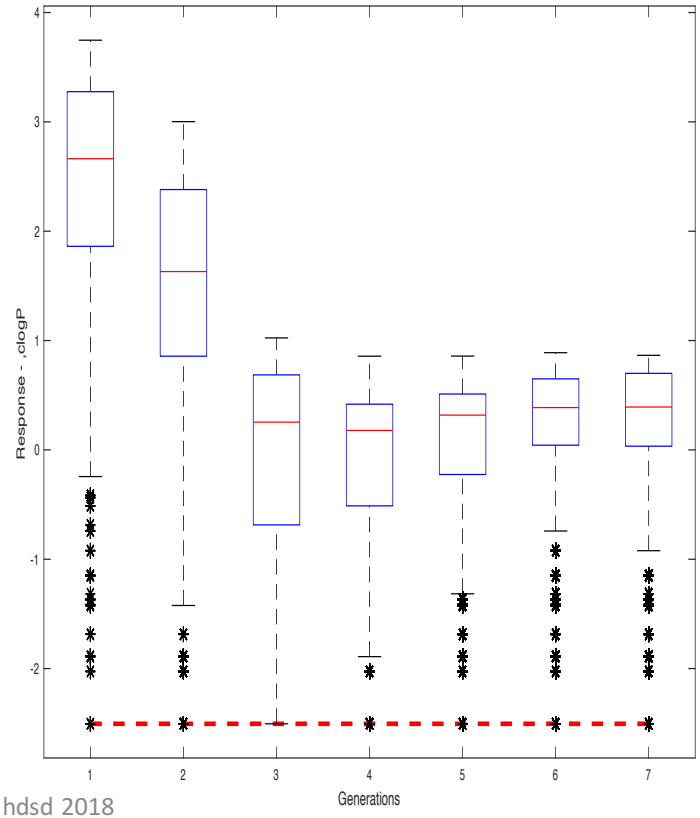


Experimental response values achieved in 1000 runs
In the region of optimality

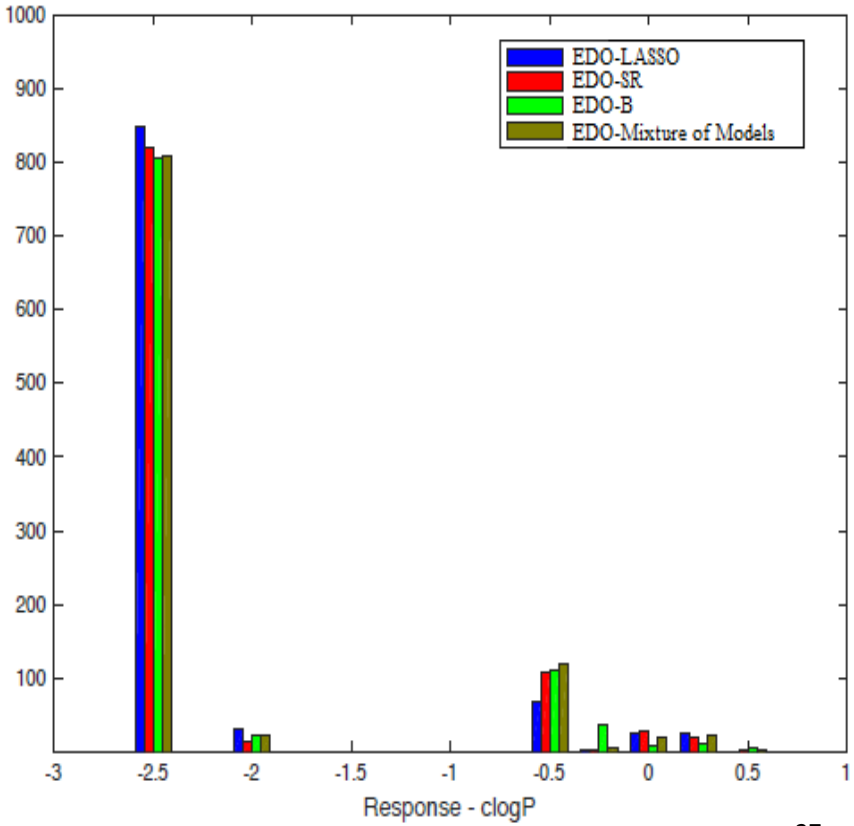


clogP optimization

Evolution through generations:
experimental values at each generation achieved in 1000 runs with Comb. Pred.



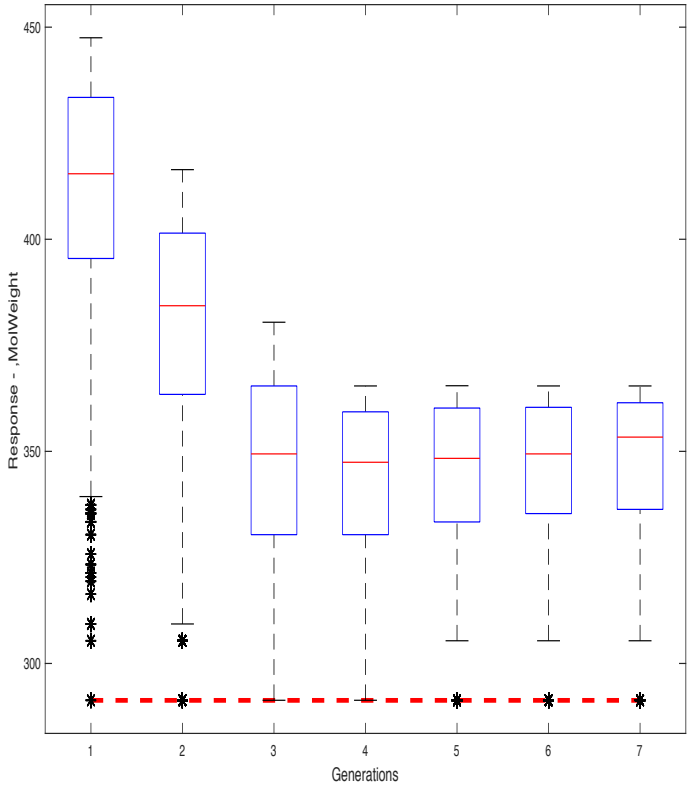
Experimental response values achieved in 1000 runs
in the region of optimality



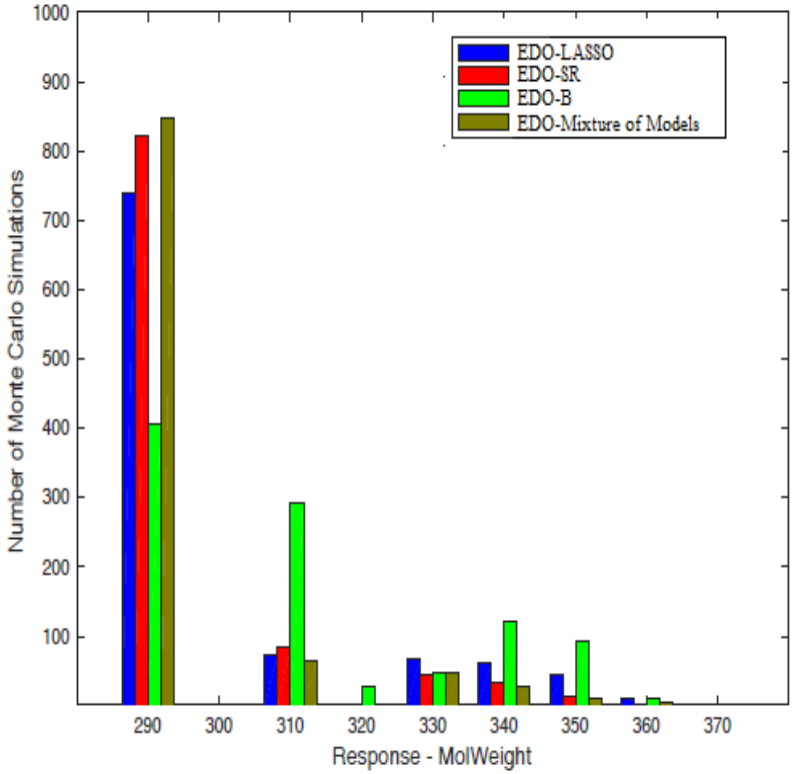
Molecular Weight optimization

Evolution through generations:

experimental values at each generation achieved in 1000 runs with Comb. Pred.



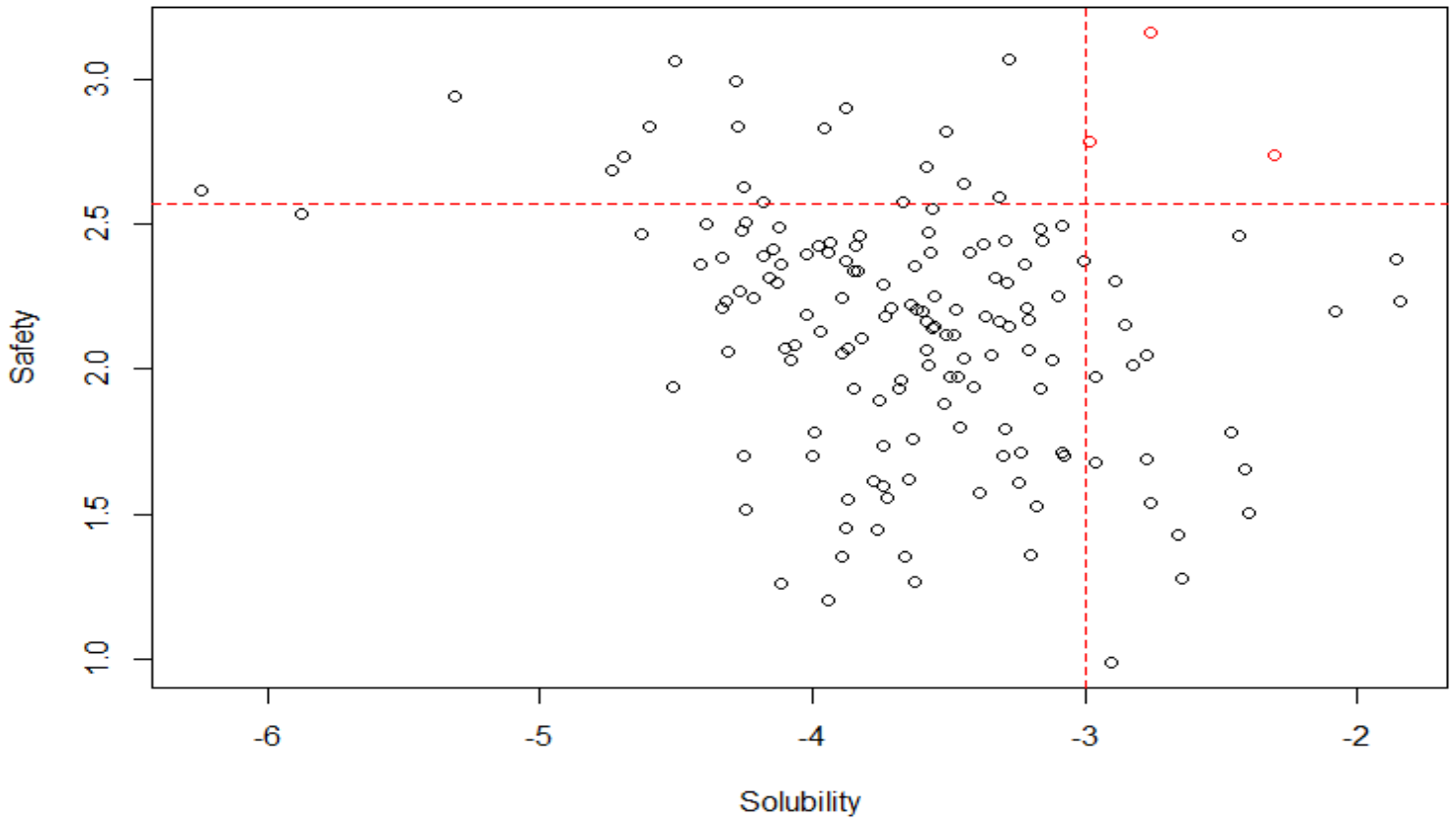
Experimental response values achieved in 1000 runs **in the region of optimality**



Multi-objective optimization

From a selected subset of the library we know our target: the 3 molecules (in red) that satisfy the proposed constraints (Pareto Front).

Activity: $y_1 > 6$.
Solubility: $y_2 -3$
Safety: $y_3 > 2.5$
ClogP: $y_4 \leq 4$

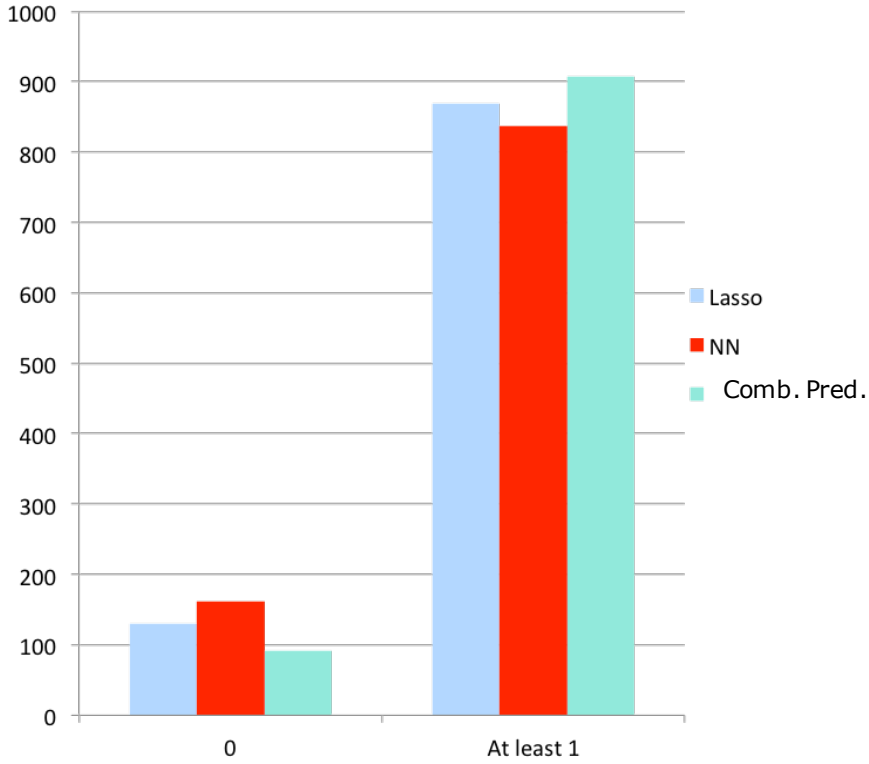


multi-objective optimization (m-EDO)

Number of runs (out of 1000 runs)
in which m-EDO discovers the best molecules

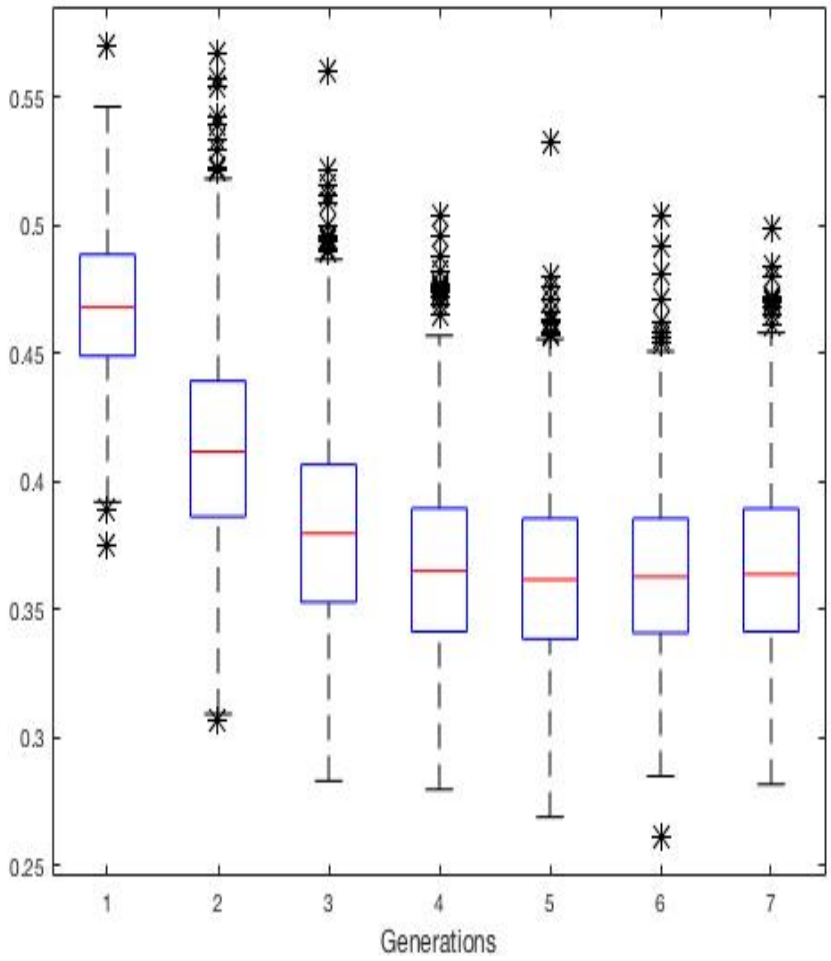
Number of best molecules	Lasso	NN	Comb. Pred.	Quantile R	SVR
0	130	161	92	363	215
1	43	59	51	82	35
2	320	288	384	166	136
3	506	491	472	389	614
At least 1	**869**	**838**	**907**	**637**	**785**

Best molecules obtained in 1000 runs



multi-objective optimization (m-EDO)

Best molecules found in 1000 runs at each generation with Comb. Pred.



Concluding:

EDO approach

is able to reach the target testing just a very small set of molecules:
(1-5%) of the total number of candidate tests.

is *fast*, (thus saves time, resources, and unnecessary experimentation)
is *effective,*
is *robust,*
is *easy* to use and interpret.

Current and future research

Clustering procedures for binary predictors in penalized regressions

Bayesian regression models for prediction

Model selection for optpimization

Combining Predictions for Optimization

Acknowledgements:

Thanks to **Darren Green, Chris Luscombe and Molecular Design group at *GlaxoSmithKline (GSK), Medicines Research Centre, Stevenage, UK*,** for their fundamental support in driving the research, giving suggestions and helpful advice.

Thanks to **Phil Brown** and the researchers of the **European Centre for Living Technology,** for helpful and constructive suggestions.

Thanks for the attention