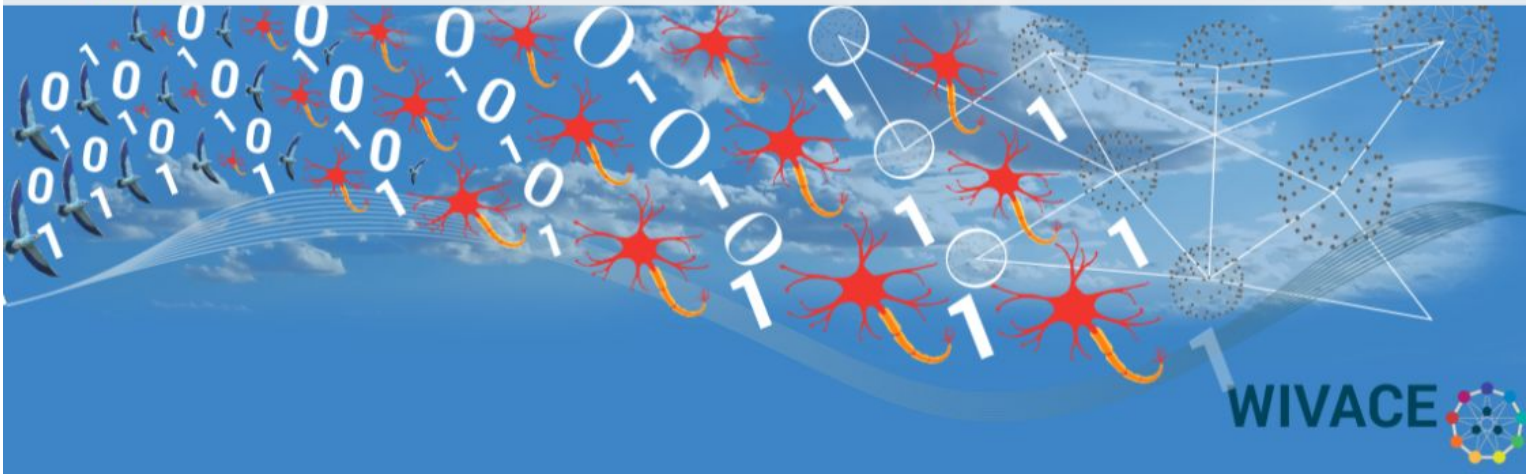
European Centre for Living Technology



WIVACE 2017

Book of Abstracts

Venice, 19-21 September 2017

For info please visit: http://wivace.org/2017

Università Ca'Foscari Venezia
Dipartimento di Scienze Ambientali, Informatica e Statistica

CCIS

Springer

Ca'Foscari Zattere
Cultural Flow Zone

WIVACE 2017 - Book of Abstracts
Andrea Roli, Debora Slanzi, Marco Villani eds.

Università degli Studi di Modena e Reggio Emilia
Dip. di Scienze Sociali, Cognitive e Quantitative
Viale Allegri 9, 42121 Reggio Emilia (RE)

ISBN 978-88-903581-3-5 (online)

Index

Evolving multi-objective optimization in high dimensional systems

Debora Slanzi[1,2], Valentina Mameli[1], Marina Khoroshiltseva[1], and Irene Poli[1,2]

[1] European Centre for Living Technology, S. Marco 2940, 30124 Venice, Italy
[2] Department of Environmental Sciences, Informatics and Statistics, Ca' Foscari University of Venice, Cannaregio 873, 30121 Venice, Italy
{debora.slanzi, valentina.mameli, marina.khoroshiltseva, irenpoli}@unive.it

Introduction

Discovering optimal values in high dimensional systems can be a challenging problem, in particular when the number of experimental tests (or observations) is small. Moreover the optimal values can involve different properties of the systems, introducing multiple (and possible conflicting) objective functions to be optimized simultaneously. This framing of the problem can make the search of the optimal values difficult.

In general, a multi-objective optimization problem can be described in the following way:
consider a vector valued objective function $f : C \to \mathbb{R}^k$ from a set $C \subseteq \mathbb{R}^d$ to real numbers $\mathbb{R}^k$, with $f(c) = (f_1(c), \ldots, f_k(c))^T$, where d is the dimension of each element of C; search the element $c_0 \in C$ such that $f(c_0) \leq f(c)$ for all $c \in C$ (minimization) or such that $f(c_0) \geq f(c)$ for all $c \in C$ (maximization).
Frequently, in multi-objective optimization, there does not exist a feasible solution, c_0, which minimizes (or maximizes) all objective functions simultaneously. Therefore, the goal is to achieve Pareto optimal solutions, that is, solutions that cannot be improved in any of the objectives without degrading at least one of the other objectives. In this research we will introduce a methodological approach to address multi-objective optimization in the context above described and related to a molecular system of interest for drug discovery.

1 Evolutionary inference for discovering the system optimal values

In order to develop an efficient approach able to achieve the optimal values of a system with a very small set of experimental tests, we developed a methodological approach based on evolutionary statistical inference for high dimensional experimental spaces and big data analysis. This approach, which we call m-EDO (multi-objective Evolutionary data Design for Optimization), drives the evolution towards the target by estimating and combining predictions from different stochastic models, such as Lasso Regression, Stepwise Regression, Boosting,

Neural Networks; see for example [2] and references therein. m-EDO is based on experimental data and is designed to discover the best solution through testing only an extremely small number of candidate solutions, making very efficient and effective the discovery process.

2 Lead optimization in a molecular system

A key problem that the drug discovery research field confronts is to identify small molecules, modulators of protein function, which are likely to be therapeutically useful. Common practices rely on the screening of vast libraries of small molecules (often 1-2 million molecules) in order to identify a molecule, known as a lead molecule, which specifically inhibits or activates the protein function. Such a molecule interacts with the required target, but generally lacks other essential attributes required for a drug candidate. Discovering the optimal lead molecule can then be framed as a multi-objective optimization problem. In this research we address the lead optimization of MMP-12 Inhibitors, using the combinatorial library and biological data made available (public domain) by [1]. This library consists of 2500 molecules characterized by their composition and by their experimental responses. The composition here considered is represented by a set of 22272 fragments, that we describe as binary variables (presence/absence). The high number of fragments give rise to the high dimensionality of the molecular system. For this system the experimental response variables here considered are: *Activity* at the target protein; *Solubility*; *Safety*; *ClogP*; *Molecular Weight*. The aim of this study is to develop a multi-objective optimization procedure based on experimental data (no simulation), and involving a very small number of experimental tests, to avoid waste of research time and resources.
We built m-EDO using the molecular library provided by Pickett et al. (2011) as a source of response variables for selected compositions. We assume that the compositions to test in the lab should be less than 140 (out of the 2500). Knowing the whole experimental space (complete library) allowed us to evaluate the performance of the approach in searching the best response values. These values of the response variables represent the target of our study, and are reported in the following:

- *Activity*, Y_1: the maximum value of Y_1 is **8**, which corresponds to the optimal value. The **99-th percentile** of the response variable distribution is **7.5** (**maximization of** Y_1).
- *Solubility*, Y_2: the maximum value of Y_2 is **-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: the maximum value of Y_3 is **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: the minimum value of Y_4 is **-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: the minimum value of Y_5 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).

The goal of the multi-objective optimization is to discover the three molecules that satisfy the constraints of the problem and reach their best response values. These molecules are represented (in red) in the following Pareto front representation of the molecule *Solubility* and *Safety* after having selected the molecules with an *Activity* greater than 6, a *ClogP* less than 3 and a *Molecular Weight* less than 450.

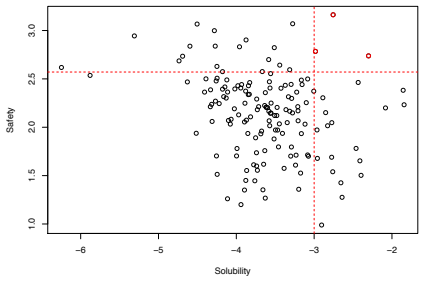


Fig. 1: Pareto front representation of the molecule *Solubility* and *Safety*, respecting the defined constraints for *Activity*, *ClogP* and *Molecular Weight*.

3 The best molecules

We built the EDO approach to optimize the five response variables for the lead optimization process, under the hypothesis to conduct a number of experimental tests less than 140 (on the total of 2500 candidate compositions). At first, to evaluate the performance of the approach, we developed the procedure for each single response variable for a single objective optimization. The evolution in EDO has been driven by the information achieved with the Lasso model, Stepwise regression, the Boosting model, and finally with a mixture of these models. Moreover, in order to evaluate the robustness of EDO we also repeated the procedure 1000 runs.
The results achieved for the single optimization process are represented in the following table:

Notice that EDO procedure is able to achieve the best response values in a very high proportion of 1000 runs, showing also a better performance of the Mixture of models with respect to the single model optimization. Concerning the response values in the region of optimality (1% best values of the distribution)

Objective		Lasso	Stepwise	Boosting	Mixture of Models	NN
Activity	Optimum	844	782	665	916	660
	Region of opt.	1000	995	998	1000	990
Solubility	Optimum	875	745	872	912	556
	Region of opt.	995	998	1000	1000	996
Safety	Optimum	387	358	278	467	228
	Region of opt.	1000	1000	1000	1000	999
ClogP	Optimum	848	821	917	918	760
	Region of opt.	950	946	981	1000	945
Molecular Weight	Optimum	738	822	751	887	346
	Region of opt.	905	966	956	1000	780

Table 1: Single objective optimization: number of runs (out of 1000 runs) in which EDO uncovers the optimum value and values in the region of optimality.

we observe that the Mixture of Models is able to achieve these values in all the 1000 runs and for all the variables.

We then developed the multi-objective optimization by using different approaches for combining the achieved response values and, in comparing the results, we noticed that the simple linear combination of the best values has a very good performance. In the following table we present the results achieved with the Lasso model, the Neural Networks model (hereafter NN) and the Mixture of Models. The three ways to optimize give similar results in discovering the best values, and the difference may lie in discovering just one, or at least one, or all three molecules. Mixture of Models again outperforms the alternatives in discovering at least one molecule of the three in more than 90% of 1000 runs.

Number of best molecules	Lasso	NN	Mixture of Models
0	130	161	92
1	43	59	51
2	320	288	384
3	506	491	472
At least one	869	838	907

Table 2: Multi-objective optimization: number of runs (out of 1000 runs) in which m-EDO uncovers the best molecules.

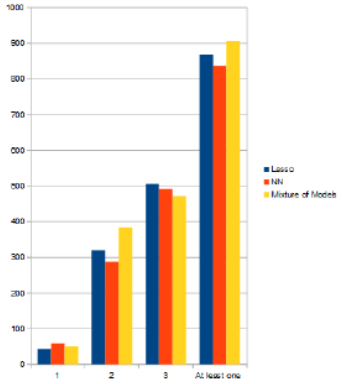


Fig. 2: Multi-objective optimization: best molecules found in 1000 runs.

From these results one can also see the value of the evolution principle in the search process: from the first generation there is a clear tendency for the procedure to converge towards the optimal values.

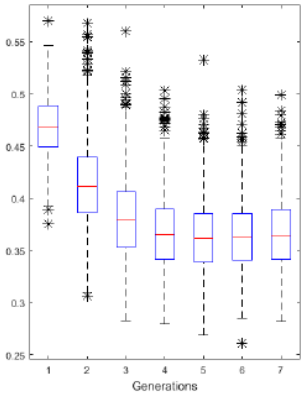


Fig. 3: Evolution through generations: box-plot of the molecule values achieved in 1000 runs at each generation with the Mixture of Models.

Acknowledgements

The authors would like to acknowledge the GlaxoSmithKline Medicines Research Centre (UK) for the very fruitful collaboration in developing this research.

References

1. Pickett, S.D., Green, D.V.S., Hunt, D.L., Pardoe, D.A., Hughes, I.: Automated lead optimization of MMP-12 inhibitors using a genetic algorithm. *A*CS Med. Chem. Lett. **2(1)**, 28–33 (2011).
2. Mameli, Valentina, Lunardon, Nicola, Khoroshiltseva, Marina, Slanzi, Debora, Poli, Irene: Reducing Dimensionality in Molecular Systems: A Bayesian Non-parametric Approach in Federico Rossi; Stefano Piotto; Simona Concilio, Advances in Artificial Life, Evolutionary Computation, and Systems Chemistry, Springer International Publishing, **708**, 114–125 (2017).