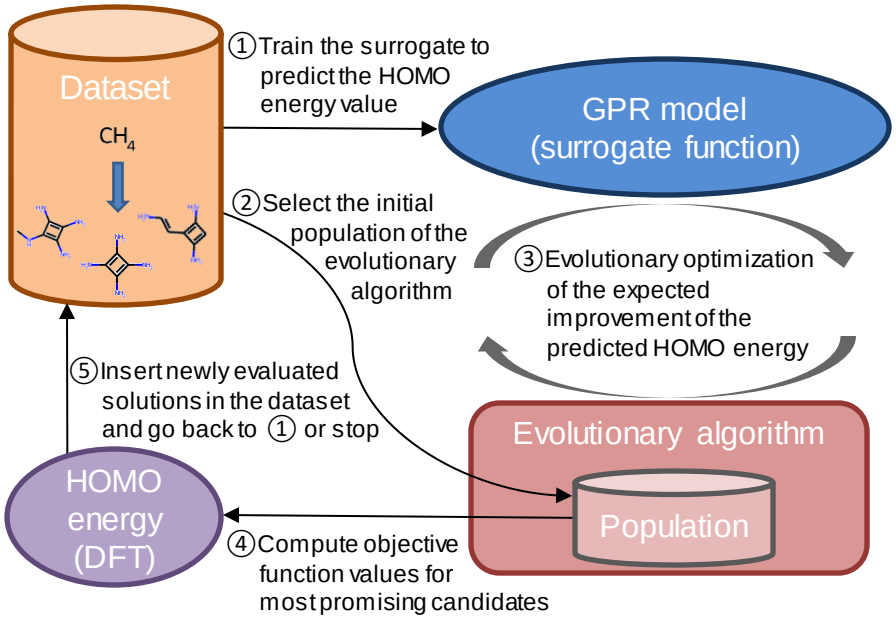
P16

# Surrogate-based black-box framework to optimize electronic properties for *de novo* organic molecular materials

Jules Leguy[1], Thomas Cauchy[2], Béatrice Duval[1], Benoit Da Mota[1]
[1]Univ Angers, LERIA, SFR MATHSTIC, F-49000 Angers, France
[2]Univ Angers, CNRS, MOLTECH-ANJOU, SFR MATRIX, F-49000 Angers, France

*De novo* molecular generation of molecular organic materials is hindered by the **cost of quantum mechanics** calculations required to assess the properties. Quick machine-learning estimators of these properties may suffer from generalization issues, in particular in the chemical space of materials [1].
We propose here to tackle the problem with a **surrogate-based black box optimization** approach. It consists in optimizing the values predicted by a machine learning model to provide candidates that are evaluated by DFT calculations. The model is thus used as a **surrogate of the property** to be optimized, and is re-trained regurlarly using all discovered data points, so that it is suited to the chemical space of the actual optimization problem. Our method is evaluated by maximizing the HOMO energy.

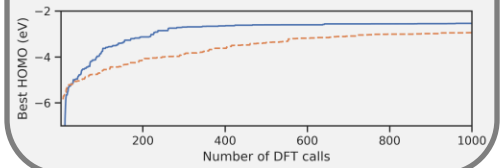
At the start of the experiment, the dataset of solutions contains **only the methane** and its associated HOMO value (-10.58 eV). Knowledge to predict accurately the objective value of the candidates must yet be acquired.
At the end of the experiment, the best solutions found are derivatives of polyamino-cyclobutadiene, that have **high HOMO energy** values (above –2.70 eV).




The surrogate function is defined as a **Gaussian process regression** (GPR) model, that provides an **uncertainty** estimation of its predictions. It is used with a RBF kernel, applied on the **MBTR** descriptor [2].

We use an **evolutionary algorithm** (EvoMol [3]) as an internal optimizer providing candidate solutions. In our setting, the function that is optimized is neither the DFT nor the surrogate function directly, but the **expected improvement of the surrogate**, that takes the uncertainty into account.
At each step of the main loop, the population of the evolutionary algorithm is initialized with a subset of the dataset, drawn randomly based on the HOMO value. After evolutionary optimization, **the best candidates are submitted to DFT evaluation.**

Our approach (blue straight line) is assessed against a direct evolutionary optimization of the DFT-estimated HOMO, thus without using a surrogate model (orange dashed line). The knowledge learnt by the surrogate allows to find **better scoring solutions with less calls to the costly DFT evaluation**.




[1] Marta Glavatskikh et al., "Dataset's Chemical Diversity Limits the Generalizability of Machine Learning Predictions," *Journal of Cheminformatics* 11, no. 1 (December 2019)
[2] Haoyan Huo and Matthias Rupp, "Unified Representation of Molecules and Crystals for Machine Learning," *ArXiv:1704.06439 [Cond-Mat, Physics:Physics]*, January 2, 2018
[3] Jules Leguy et al., "EvoMol: A Flexible and Interpretable Evolutionary Algorithm for Unbiased de Novo Molecular Generation," *Journal of Cheminformatics* 12, no. 1 (September 16, 2020): 55