Bayesian Optimization for the Discovery of Redox Active Quinones

Giacomo De Gobbi [1] **Robert Peharz** [1] **Stefan Spirk** [2 3] **Janine Maier** [2 3] **Reyhan Yagmur** [3]

Abstract

Traditional computational chemistry techniques are often a severe bottleneck in scaling up the discovery of materials or new useful molecules. Machine learning techniques have proven effective to overcome such limitations and often lead to unprecedented results. In this work we focus on finding candidate molecules that are benzoquinones derivatives to be used in organic flow batteries. We present a sampling algorithm equipped with chemistry-based constraints in order to generate a molecular library and we utilize a Bayesian optimization strategy to select the best suited molecules.

1. Introduction

The transition to sustainable energy is one of the most pressing problems in the current century, both on global and national level. On a global level, we face problems connected with climate change, pollution, and depletion of resources. On a national level, dependencies on energy supplies from politically instable regions, as for example in Europe, further fuels political and economic challenges. Consequently, many government have progressed towards substituting fossil energy, in particular electricity, by renewable sources.

Renewable sources, however, have their well-known set of challenges as they are subject to seasonal changes and depend on weather and daytime. The main challenge we face is to store the surplus energy from renewables in storage facilities such as mechanical, physical, thermal or chemical storages on a mid or even long term. Among these options, organic flow batteries are a promising recent approach with the potential to make a difference in energy storage technology (Huskinson et al., 2014), with a vast number of options for realizing redox active molecules based on organic chemistry.

[1]Institute of Theoretical Computer Science, TU Graz, Austria [2]Institute of Bioproducts and Paper Technology, TU Graz, Austria [3]Ecolyte GmbH, Austria. Correspondence to: Giacomo De Gobbi <giacomo.degobbi@tugraz.at>.

AI4Science Workshop of the 41 st International Conference on Machine Learning, Copyright 2024 by the author(s).

Unfortunately, several problems concerning the mass adoption of this technology remain (Luo et al., 2019). Firstly, even though certain types of organic molecules might be promising, such as anthraquinones, phazines, thiophenazines, etc., and reach voltages of even above $1\,V$ in single cell settings, their investigation in larger setups or integrated systems is hardly ever reported. Another, nearly equally important point, is the lack of suitable starting materials for organic flow batteries at large scale. Typically, the molecules reported for organic flow batteries in literature are complicated to synthesize using expensive pathways or the raw materials are not produced at sufficient scale.

This work is part of a project aiming at above mentioned challenges, based on AI-guided development of flow battery components. Particular emphasis is on the redox active species already produced at industrial scale from biobased sources, such as waste materials from the wood industry, such as quinones and their precursors such as vanillin and related aldehydes. In particular, Bayesian optimization and active learning strategies are employed to find derivatives with the lowest redox potential.

2. Methods

The design of functional molecules often relies on combinatorial, high-throughput screening strategies enabled by high-performance computing. Despite the successes of high throughput experimentation in chemistry, biology, and materials science, these approaches typically employ exhaustive searches that scale exponentially with the size of the search space. Data-driven strategies that can adaptively search parameter spaces without the need for exhaustive exploration are thus replacing traditional design of experiment approaches in many instances (Hickman et al., 2022).

These strategies use machine-learnt surrogate models trained on all data generated through the experimental campaign, and are updated each time new data is collected. One such approach is Bayesian optimization which, based on the surrogate model, defines a utility function that prioritize experiments based on their expected informativeness and performance (Mockus, 2012). In our case we employed Bayesian optimization to discover promising molecules with a minimal number of computationally expensive quantum chemistry calculations.

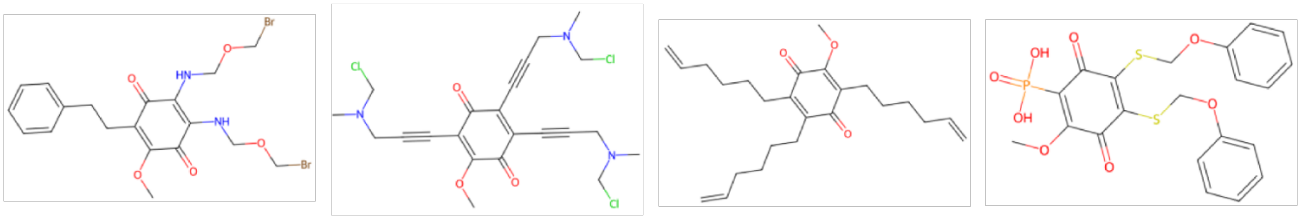

Figure 1. Four random examples from our molecular library.

Table 1. Candidate functional groups for molecules generation.

SMILES	
CX	N(C)(C)COX
CCX	NCOCX
CS(=O)(=O)O[NA]	N[O-][O-]
COX	NS(=O)(=O)O[NA]
CCL	OX
C2=CC=CC=C2	OCX
C=CX	OCS(=O)(=O)O[NA]
C=CCX	OC2=CC=CC=C2
C=CCCX	OC2=C(F)C(F)=C(F)C(F)=C2(F)
C#CX	OCL
C#CCX	OS(=O)(=O)O[NA]
C#COX	SX
C#CNX	SCX
C=NX	SCCX
C=NCX	SCOX
C=NCCX	SCS(=O)(=O)O[NA]
C=NOX	S(=O)(=O)O[NA]
CONX	C2N=CCS2
CONCX	C2N=CC=N2
CONCCX	[N+]2=CC=CC=C2.[CL-]
C#N	[N+](C)(C)(C).[CL-]
[F]	[SI](C)(C)(C)
[CL]	S(=O)CX
[BR]	[SI](O)(C)(C)
NX	S(=O)(=O)CX
NCX	P(=O)(O)(O)
N(C)CX	SSCX
N(C)COX	H

2.1. Molecular library generation

In order to produce a dataset for our BO strategy we follow the procedure of (Jain et al., 2023). We start with a core molecule, 2-Methoxy-1,4-benzoquinone, in SMILES representation. Then, we utilize a sampling algorithm to add in the position 2, 3 and 5 of the benzene ring, different atoms and functional groups from Table 1.

The symbol *X* it is used to indicate another call of the sampling algorithm to continue the chain, if not present means that the procedure needs to stop. In the first call of the sampler all elements have equal probability to be attach to the core molecule, while in the following iterations the probability of elements that do not have the symbol *X* increase.

When a molecules in SMILES representation is generated, in order to be added to the library needs to respect the following constraints:

- Oxygen cannot bind directly with another oxygen or with an halogen.
- In the molecule should be present just one halogen element.
- No halogen is allowed next to a triple bond.
- No nitrogen-nitrogen single bond.
- Sodium can only bind to oxygen.
- Halogens cannot bind with nitrogen or sulfur.
- Maximum 1 triple bond for each growth point in the core molecule.
- The number of hetero-atoms should be under 40% of the total atoms in the molecule.
- Complexity of the molecule must be ≥ 20 and ≤ 80.
- Synthesizability score ≥ 0.5.
- Log-Solubility ≥ -2.

The first conditions are heuristics based on basic chemical knowledge and are necessary to avoid infeasible molecules. The complexity constraint in this work is calculated as the number of non hydrogen atoms in the whole molecule excluding the atoms from the core. To impose synthesizability constraint we used *RAScore* (Thakkar et al., 2021), that is based on the computer-aided synthesis planner *AiZythFinder*. *RAscore* predicts the probability of *AiZythFinder* (Genheden et al., 2020) being able to identify a synthehtic route for target organic molecule. For computing the Log-Solubility of the molecules we use *AqSolPred*, a ML model trained on *AqSolDB* (Sorkun et al., 2019) that is the largest publicly available aqueous solubility dataset. Some examples of our library are shown in Figure 1.

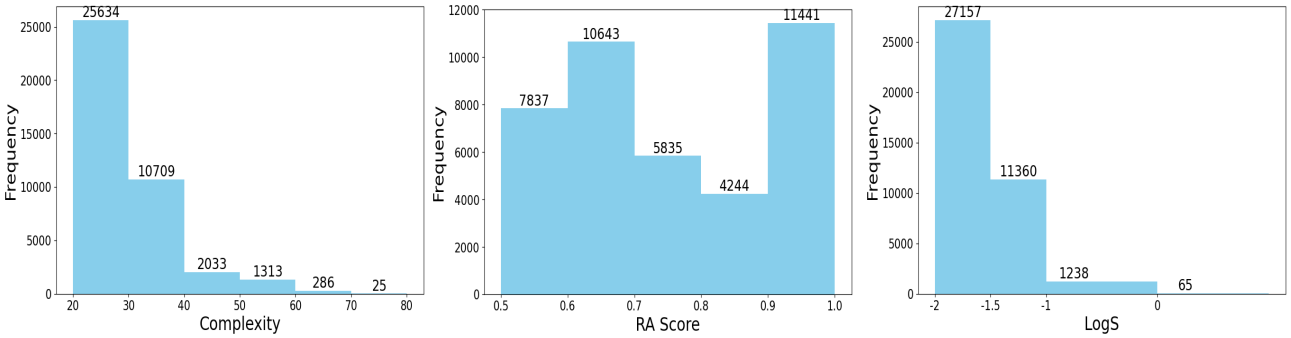

Figure 2. Distribution of constrained variable in the molecualr library

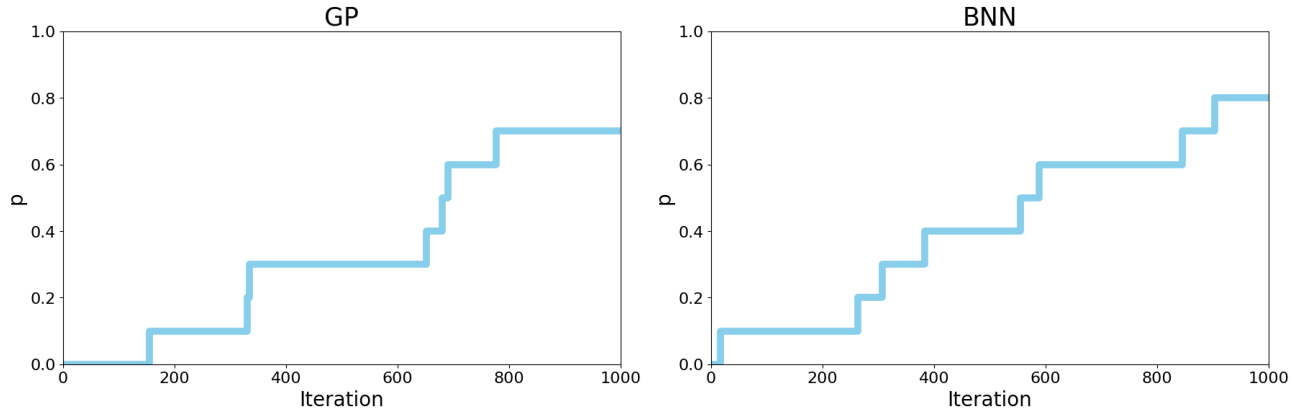

Figure 3. Results of the two surrogate models on our pre-liminary library.

2.2. Feature and Label Calculation

After the molecular library generation, we used the *RDkit* library to generate 205 molecular features based on the SMILES representation and then applied PCA in order to reduce the features dimension to 30. For the calculation of the label (redox potential), was calculated as:

$$E_{\text{red}} = -\frac{\Delta G \cdot 2625.50}{2 \cdot 96.485} + 0.699$$

where the Gibbs free energy is calculated as:

$$\Delta G = G_{OX} + \Delta G_{MHQ} - G_{RED} - G_{MQ}$$

All the above parameter were calculated with density functional theory (DFT) using *PySCF* library (Sun et al., 2018), at the B3LYP/6-31++G level of theory. We have compared the Gibbs free energy with the ground state energy of a subset of generated molecules and we noticed a not significant difference. For this reason, we directly use the ground state energy in place of Gibbs free energy in the above formula for computational efficiency and faster screening of molecules.

2.3. Bayesian Optimization Strategy

To start the strategy we select at random n molecules from the molecule library and we use DFT to compute the redox potentials for these n molecules. Using E_{red} values as dependent variables and the 30 features as independent variables, we develop two different surrogate models. We use a Gaussian Process Regression (GPR) model with the Matern kernel and a Bayesian Neural Network (BNN) with architecture described in (Häse et al., 2018). These models predict the mean μ and standard deviation σ of E_{red} values for the other molecules in the library. These two quantities are used in the expected improvement (EI) acquisition function to guide the selection of the next molecules for E_{red} evaluation, aiming to minimize E_{red} in the dataset labeled with DFT-calculated values. We compute the EI for each molecule in the library and select the one with the highest EI for the next DFT calculation. This newly evaluated molecule is then added to the labeled dataset, marking the completion of one Bayesian Optimization (BO) iteration. Each iteration involves retraining the surrogate model with the updated labeled dataset to predict the EI of all molecules,

thereby selecting another unlabeled molecule. This iterative process generally enhances the surrogate model predictive accuracy, progressively identifying more optimal molecules for DFT calculations. The BO runs until it either finds optimal candidate or reaches a maximum number of iterations allowed.

3. Results

We generated $40,000$ molecules as a preliminary library whose distribution over complexity, RAscore and solubility is shown in Figure 2. We then split the library in 10 subsets each containing $15,000$ random molecules to test our BO strategy. In this case we pre-compute all redox potentials with DFT for testing the surrogate models reducing the computational time. We start selecting $n = 500$ and if the minimum was selected in this initial training set we removed it and sample another point from the subset. We fixed the maximum number of the Bayesian Optimization steps at 1000.

In Figure 2 and 3 we can see the results of our experiments. On the x-axis we put the iteration step when the surrogate model find the minimum in the subset while in the y-axis we put a CDF proxy, specifically $p = \sum_{n=1}^{N} \mathbb{1}_{\{it_{(n)} \leq it_{max}\}}/N$, that indicate if the surrogate model was able to find the minimum in each subsets. As we can see BNN outperform GP given that was able to find the minimum in more subsets and in less iterations.

Importantly, these results demonstrate BO's efficacy in scaling up screening for promising molecules. In particular, BO based on BNNs was able to find the highest scoring candidate among a library of $40k$ in less than $1k$ iterations in a majority of runs.

4. Conclusion and Future Works

We have demonstrated that given a molecules library designed with chemistry domain knowledge, our BO strategy was able to identify the best suited molecules in few iterations. In future weeks, further constraints and a bigger library molecules (at least $500k$) will be used in order to fully exploit the power of the procedure in a real case scenario.

Ultimately, we will test the top K molecules with the lowest redox potential, synthesizing and characterizing them using state-of -the-art techniques (e.g., multidimensional NMR spectroscopy, XRD, in case crystals can be obtained, IR spectroscopy). The standard reduction potentials and their dependency on the pH value will be investigated by cyclovoltammetry. The activation energies will be determined using the Randles-Sevcik as well as the Levich approach. Solubility and stability will be also determined at different pH values and conditions (ambient vs N2 atmosphere, UV, temperature). In cases where molecules provide sufficient stability (1 week without decomposition when stored in solution) and solubility (at least 0.5 mol/L), they will be tested in single cell flow batteries.

Acknowledgements

This project has received funding from the European Union's EIC Pathfinder Challenges 2022 programme under grant agreement No 101115293 (VanillaFlow). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Innovation Council. Neither the European Union nor the European Innovation Council can be held responsible for them.

Co-funded by
the European Union

References

Genheden, S., Thakkar, A., Chadimova, V., Reymond, J.-L., Engkvist, O., and Bjerrum, E. A fast robust and flexible open-source software for retrosynthetic planning. *ChemRxiv*, 2020.

Hickman, R. J., Aldeghi, M., Häse, F., and Aspuru-Guzik, A. Bayesian optimization with known experimental and design constraints for chemistry applications. *Digital Discovery*, 1(5):732–744, 2022.

Huskinson, B., Marshak, M., and Suh, C. e. a. A metal-free organic–inorganic aqueous flow battery. *Nature*, 505: 195–198, 2014.

Häse, F., Roch, L. M., Kreisbeck, C., and Aspuru-Guzik, A. Phoenics: A bayesian optimizer for chemistry. *ACS central science*, 4:1134–1145, 2018.

Jain, A., Shkrob, I. A., Doan, H. A., Robertson, L. A., Zhang, L., and Assary, R. S. In silico discovery of a new class of anolyte redoxmers for non-aqueous redox flow batteries. *Digital Discovery*, 2(4):1197–1208, 2023.

Luo, J., Hu, B., Hu, M., Zhao, Y., and Liu, T. L. Status and prospects of organic redox flow batteries toward sustainable energy storage. *ACS Energy Letters*, 4(9):2220–2240, 2019.

Mockus, J. Bayesian approach to global optimization: theory and applications. *Springer Science & Business Media*, 37, 2012.

Sorkun, M., Khetan, A., and Er, S. Aqsoldb, a curated reference set of aqueous solubility and 2d descriptors for a diverse set of compounds. *Sci Dat*, 6, 2019.

Sun, Q., Berkelbach, T. C., Blunt, N. S., Booth, G. H., Guo, S., Li, Z., Li, J., McClain, J., Sharma, S., Wouters, S., and Chan, G. K.-L. Pyscf: the python-based simulations of chemistry framework. *WIREs Comput. Mol. Sci.*, 8(1): 1197–1208, 2018.

Thakkar, A., Chadimová, V., Bjerrum, E. J., Engkvist, O., and Reymond, J.-L. Retrosynthetic accessibility score (rascore) – rapid machine learned synthesizability classification from ai driven retrosynthetic planning. *Chem. Sci.*, 12:3339–3349, 2021.