Bayesian Optimization with Partial Evaluations for Materials Design

Poompol Buathong[1] Jiayue Wan[2] Peter I. Frazier[2]
[1]Center for Applied Mathematics [2]Operations Research & Information Engineering
Cornell University

Introduction

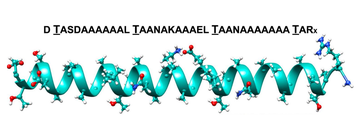
- **Materials design and drug discovery** is a challenging problem because (1) the design space is huge, and (2) measuring a molecule's functionality in a physical experiment is time-consuming and expensive.
- **Bayesian optimization** (BO) [1], a sequential experimental design strategy for optimizing expensive-to-evaluate functions, enables efficient discovery of molecules with desirable chemical properties.
- In chemical design problems, domain experts often use **computational chemistry models** to study the molecular mechanisms underlying a chemical property of interest. Such models provide important additional information not available from physical experiments but are ignored by standard BO algorithms.
- We show that **BO of function networks** (BOFN) [2] can combine information from computational chemistry models and physical experiments to improve molecular discovery. However, BOFN is inefficient for molecular discovery when used directly because it requires that physical experiments be run on all molecules evaluated via computational chemistry.
- Therefore, we propose BOFN with **partial evaluations** (pBOFN), which supports efficient molecular discovery by allowing physical experiments to be skipped for molecules whose computational chemistry calculations lack promise, reducing the number of expensive physical experiments.

Applications

Antifreeze Protein (AFP) Discovery

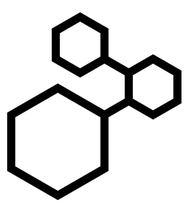
- Our primary application, part of a Multi-University Research Initiative supported by the Air Force and led by UCSD, is discovery of peptides that inhibit ice growth.
- Proteins that inhibit ice growth (recrystallization) appear in nature, e.g., in fish that live in the Arctic and Antarctic.
- We wish to improve on these proteins for use in aerospace (to prevent ice formation on aircraft) and for cryopreservation of biological tissue.
- There are also applications more broadly in molecular discovery, inverse problems / model calibration, and reinforcement learning.

Winter flounder



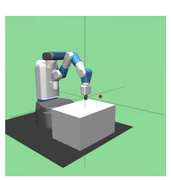
Helical AFP1

Other Applications



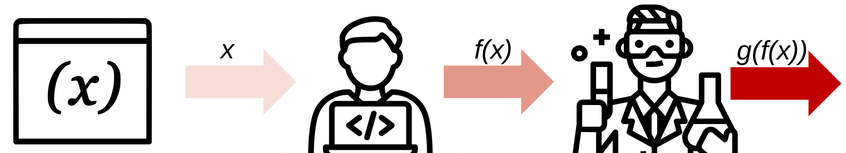
Molecule property optimization

Model calibration



Robot movement

Simple Function Network in Materials Design

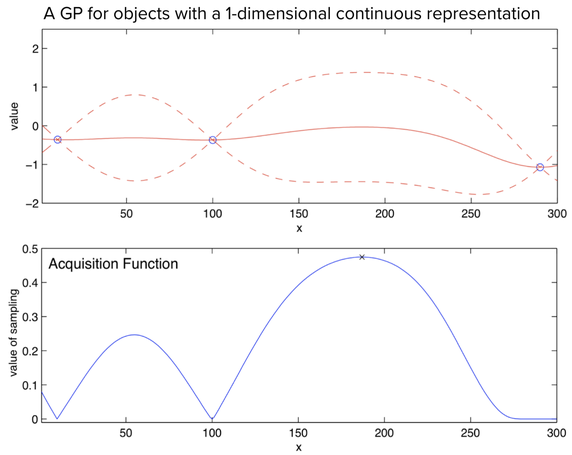


Input x
e.g. peptide/molecule

Computational model f
e.g. computational chemistry simulation, approximate mathematical model

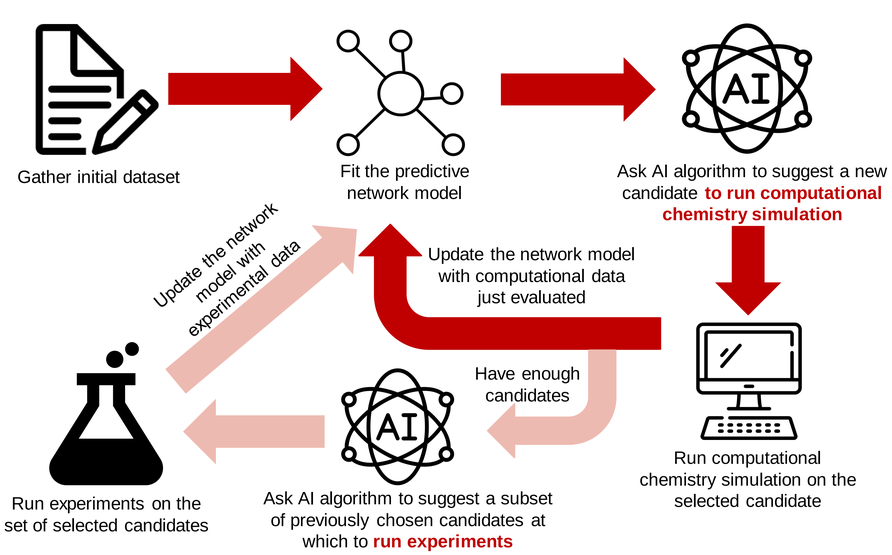
Physical experiment g
e.g. chemical experiment

Bayesian Optimization



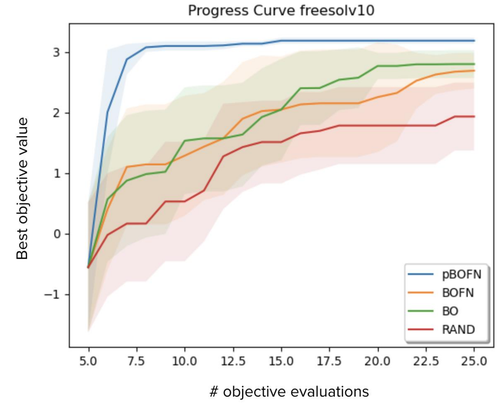
- **Bayesian optimization** is an approach to optimizing objective functions that take a long time to evaluate.
- It often uses a **Gaussian process** (GP) statistical model. In BOFN, we use the composition of a network of GPs.
- BO computes the value of information (i.e., **acquisition function**), which quantifies the value of evaluating the objective at a given point.
- BO recommends running the next experiment at the point with the **highest** value of information.

Workflow for pBOFN



Preliminary Results

- We test the performance of the proposed pBOFN framework on the **FreeSolv dataset** [3]: a curated database containing both calculated and experimental free energies of 642 small molecules.
- **Experiment setup**
  - Goal: maximizing experimental free energy
  - We use a variational autoencoder (VAE) to represent molecules in a continuous space
  - Baseline methods: random sampling (RAND), BO, BOFN and pBOFN
  - We use 5 randomly selected initial observations and 20 iterations
  - In each iteration, pBOFN selects 20 molecules for which to calculate computational free energies and selects only one most-promising candidate at which to evaluate the experimental free energy.
  - 10 replications
- **Results**: pBOFN outperformed all baseline algorithms in terms of solution quality and number of physical experiments required. This is because it leverages computational values that are highly predictive of the experimental free energies.



Conclusion and Future Work

- We proposed a novel Bayesian Optimization framework for molecular design that uses computational chemistry models and physical experiments together to discover promising molecules faster than using physical experiments alone.
- We tested the algorithm on a chemistry dataset (FreeSolv).
- The results show that the proposed algorithm provides a better solution using fewer physical experiments than existing methods.
- Future work includes incorporating heterogeneous costs for different information sources and applying the framework to a broader range of applications.

Acknowledgements and References

This work was supported by the Air Force Office of Scientific Research, FA9550-0-1-0351.

[1] Frazier, P. I. (2018). Bayesian optimization. In *Recent advances in optimization and modeling of contemporary problems* (pp. 255-278). Informs.

[2] Astudillo, R., & Frazier, P. (2021). Bayesian optimization of function networks. *Advances in Neural Information Processing Systems*, *34*, 14463-14475.

[3] Mobley, D. L., & Guthrie, J. P. (2014). FreeSolv: a database of experimental and calculated hydration free energies, with input files. *Journal of computer-aided molecular design*, *28*(7), 711-720.