Multi-objective Bayesian Optimization with Heuristic Objectives for Biomedical and Molecular Data Analysis Workflows

Alina Selega[1,3] **Kieran R. Campbell**[1,2,3,4]

[1]Lunenfeld-Tanenbaum Research Institute
[2]University of Toronto
[3]Vector Institute
[4]Ontario Institute for Cancer Research

Abstract Many practical applications require optimization of multiple, computationally expensive, and possibly competing objectives that are well-suited for multi-objective Bayesian optimization (MOBO). However, for many types of biomedical data, measures of data analysis workflow success are often heuristic and therefore it is not known a priori which objectives are useful. Thus, MOBO methods that return the full Pareto front may be suboptimal in these cases. Here we propose a novel MOBO method that adaptively updates the scalarization function using properties of the posterior of a multi-output Gaussian process surrogate function. This approach selects useful objectives based on a flexible set of desirable criteria, allowing the functional form of each objective to guide optimization. We demonstrate the qualitative behaviour of our method on toy data and perform proof-of-concept analyses of single-cell RNA sequencing and highly multiplexed imaging datasets for univariate input optimization.

1 Broader Impact Statement

As our general approach is applicable to any multi-objective optimization scenario (while tailored to biomedical analyses), we acknowledge that it could be used in highly diverse applications. While unlikely, these could include ethically dubious bioinformatics analyses e.g. those pertaining to genetic testing of embryos, or biomedical analysis systems that are systematically biased against certain populations. We strongly caution against any such use of our method without a thorough ethical review process and urge biomedical practitioners to carefully consider how their methods will affect all members of society. From the perspective of environmental impact, optimizing parameters of bioinformatics workflows with our approach may be more efficient with respect to computational resources compared to exhaustively evaluating all parameter values, which could lead to a reduction in carbon emissions.

2 Submission Checklist

1. For all authors...

   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] Our abstract and introduction outline the scope of our work (multi-objective Bayesian optimization for bioinformatics workflows in the context of many heuristic objectives) and clearly enumerate our contributions (introduction and evaluation of a novel method which uses desirable characteristics of the objective posterior functional form to optimize useful objectives).

   (b) Did you describe the limitations of your work? [Yes] We describe the limitations of our method in the discussion, namely we highlight the fact that we make no optimality claims

on the ability to explore the entire Pareto front; we point out that our method does not remove subjectivity from the analysis as many important quantities in our method are set by the user; and we discuss scenarios where the examples of desirable behaviours we propose may be limiting. We also clearly state in the abstract, introduction, and discussion that the current work evaluates our method on univariate inputs only.

(c) Did you discuss any potential negative societal impacts of your work? [Yes] We highlight in the discussion that there are controversial analyses in bioinformatics such as genetic testing of embryos, and that, as with any bioinformatics method, there is nothing stopping practitioners using our work in such scenarios. We do not cite any relevant work to avoid drawing attention to that research domain and emphasize the need for thorough ethical review in such applications.

(d) Have you read the ethics author's and review guidelines and ensured that your paper conforms to them? `https://automl.cc/ethics-accessibility/` [Yes] We read the ethics and accessibility guidelines and ensured our paper conforms to them. For example, we created high quality figures, ensured that our references are up-to-date in terms of publication information and authors' names by using Rebiber, and used inclusive language throughout the manuscript.

2. If you are including theoretical results...

(a) Did you state the full set of assumptions of all theoretical results? [Yes] We state all assumptions in Section 3 which outlines the details of our method.

(b) Did you include complete proofs of all theoretical results? [Yes] Proofs are provided in the Appendix E.

3. If you ran experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results, including all requirements (e.g., `requirements.txt` with explicit version), an instructive `README` with installation, and execution commands (either in the supplemental material or as a URL)? [Yes] All code required to re-run the experiments has been submitted as supplementary material. We also include a `README` file with installation instructions, code execution commands, and a dedicated `Pipfile` file containing all details about the dependencies.

(b) Did you include the raw results of running the given instructions on the given code and data? [No] Due to the large number of outputs of our method (acquisitions, behaviours, inclusion probabilities, objective and meta-objective values, and acquisition function values at each acquisition step), we did not include all raw outputs for all repetitions of the experiments that we ran but we provided instructions on how to re-run these analyses. We also included all processed data required to generate the figures and tables in our paper.

(c) Did you include scripts and commands that can be used to generate the figures and tables in your paper based on the raw results of the code, data, and instructions given? [Yes] We provide all notebooks and required processed data used to generate the figures and tables from the raw results in the supplementary materials. We also provide the original data we used for one of our experiments that the code can be executed on. We were not able to provide the data for the other experiment due to the size limit on the supplementary materials zip file (50 MB).

(d) Did you ensure sufficient code quality such that your code can be safely executed and the code is properly documented? [Yes] Our code can be safely executed in an environment with

required dependencies and we made efforts to follow best practises by following tutorials from GPyTorch and BoTorch, making our code modular, and increasing accessibility via documentation.

(e) Did you specify all the training details (e.g., data splits, pre-processing, search spaces, fixed hyperparameter settings, and how they were chosen)? [Yes] All these details are summarized in Appendices A1, A2, A3, A4, B2, B3 and Supplementary Tables 3 and 4.

(f) Did you ensure that you compared different methods (including your own) exactly on the same benchmarks, including the same datasets, search space, code for training and hyperparameters for that code? [Yes] We compared all methods on the same benchmarks, including random seeds, initial training sets, hyperparameter search space, hyperparameters of the training procedure, and code implementing the optimized objectives.

(g) Did you run ablation studies to assess the impact of different components of your approach? [Yes] We ran ablation studies to assess the impact of different desirable behaviours on our approach and found that no one behaviour drove the performance. The details of these experiments are described in Appendix A4 and Supplementary Tables 6, 7.

(h) Did you use the same evaluation protocol for the methods being compared? [Yes] All methods had the same batch size, number of acquisitions, and initial training set and we used the same meta-objectives measuring the concordance between each methods' results and expert labels to evaluate all methods via three regret metrics we constructed. We also compared all methods using a hypervolume measure (Supplementary Figure 12).

(i) Did you compare performance over time? [Yes] We compared performance of all methods over time by computing the regret metrics we constructed as functions of the acquisition step (Appendix A5 and Supplementary Figure 10).

(j) Did you perform multiple runs of your experiments and report random seeds? [Yes] We ran 100 repetitions of each experiment and reported random seeds in the associated `yaml` files provided in the supplementary materials.

(k) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] All results in the tables are reported with error bars representing standard deviation w.r.t. the random seed after running experiments multiple times and shaded regions in the plots denote the 95% confidence interval across multiple runs of the experiment.

(l) Did you use tabular or surrogate benchmarks for in-depth evaluations? [Yes] We evaluated all methods in the scenarios reflecting the scope of our work, namely the existence of many objectives some of which may not be useful for optimization. We used a surrogate benchmark simulating useful and not useful objectives (Section 4.1) and two real bioinformatics workflows applied to highly multiplexed imaging and molecular sequencing datasets (Sections 4.2 and 4.3).

(m) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No] For Bayesian optimization methods, the compute bottleneck lies with function evaluation rather than the Bayesian optimization procedure itself, so we do not include a discussion of compute time/resources.

(n) Did you report how you tuned hyperparameters, and what time and resources this required (if they were not automatically tuned by your AutoML method, e.g. in a NAS approach; and also hyperparameters of your own method)? [N/A] We did not tune any hyperparameters beyond those optimized by the considered methods. We fixed all other hyperparameters following best bioinformatics practises or existing tutorials for the compared methods.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

   (a) If your work uses existing assets, did you cite the creators? [Yes] We appropriately cited all datasets used in this work and all libraries we used to implement our method or performed comparisons with.

   (b) Did you mention the license of the assets? [No] All datasets we used are public domain and have no applicable license.

   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] Code required to run experiments has been submitted in the supplemental material.

   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [No] All datasets used are fully anonymized and public domain not requiring further consent/ethics approval.

   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No] No data used here has personally identifiable information.

5. If you used crowdsourcing or conducted research with human subjects...

   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A] We did not use crowdsourcing nor conducted research with human subjects.

   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A] We did not use crowdsourcing nor conducted research with human subjects.

   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A] We did not use crowdsourcing nor conducted research with human subjects.