Optimization Formulations for Self-Driving Inorganic Materials Design

Ruoxi Zhao[1,2], Stefan M Wild[2]
[1]University of California, Riverside, [2]Lawrence Berkeley National Laboratory

Abstract

Numerical optimization methods have the potential to develop efficient and fully autonomous discovery of new solid-state materials. Unlike traditional datasets where the inputs are continuous numbers, material synthesis consists of information such as precursors, product weight fraction, and products which are categorical data. This would be a challenge if we use black-box optimization methods such as Bayesian optimization and genetic algorithms. In this project, we examine how different optimization problem formulations can quantify the effects of different precursors synthesized under different temperatures using X-Ray Diffraction (XRD) data measured in an online fashion. Our goal is for these formulations to allow for more robust and efficient discovery of pathways to a target material.

Background

Conventional high-temperature synthesis based on solid-state reaction is often used for the preparation of inorganic materials, and there is no clear road map to optimize the solid-state synthesis of novel inorganic materials due to the difficulty to model solid-state reactions, which uses density functional theory (DFT) calculations based on thermochemical data.[1] Efforts such as LBNL's A-Lab[2] seek to leverage computer-aided optimization in a closed-loop solid-state synthesis setting to improve this situation.

We started by looking at the YBCO dataset in the ARROWS repository[3], as this is a comprehensive dataset that includes both positive and negative outcomes, which means the reactions do and do not yield sufficiently pure target material - YBCO. Such a dataset can be used as a benchmark on testing ARROWS and compare its performance to other optimization methods. We focused on the experimental dataset, which consists of different combinations of the 11 precursors, the temperature of each combination (at 600, 700, 800, 900, and 1000 degrees Celsius), whether the combination was experimentally verified, as well as the resulting products and measured XRD from the experiments.

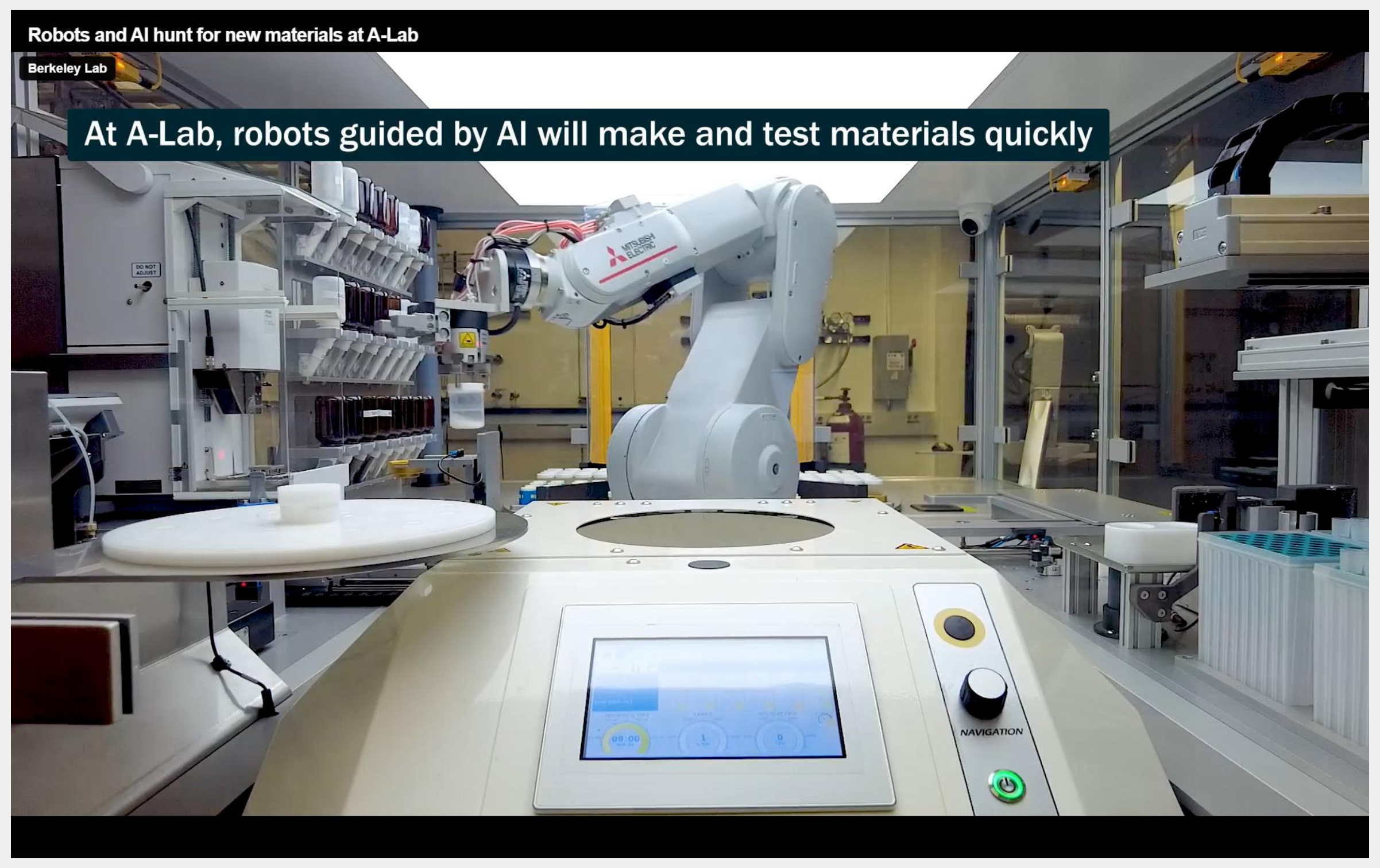

Figure 1: Robotic arm conducting materials synthesis experiments at LBNL's A-Lab.[2]

YBCO Experimental Dataset

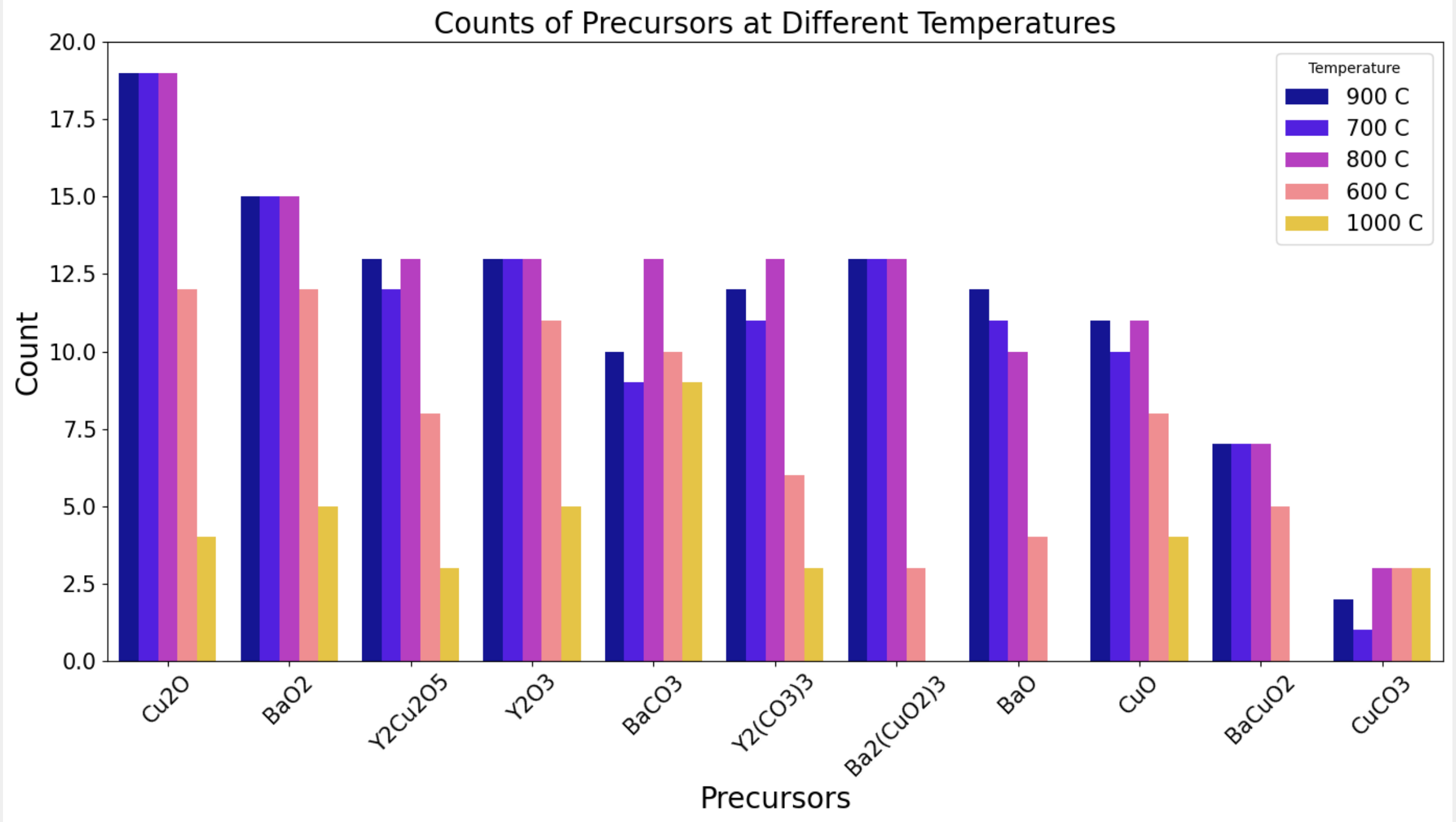

Figure 2: Experimentally verified combinations in the YBCO dataset: These involve 11 precursors at 5 temperatures.

Quantifying Objectives based on XRD Data

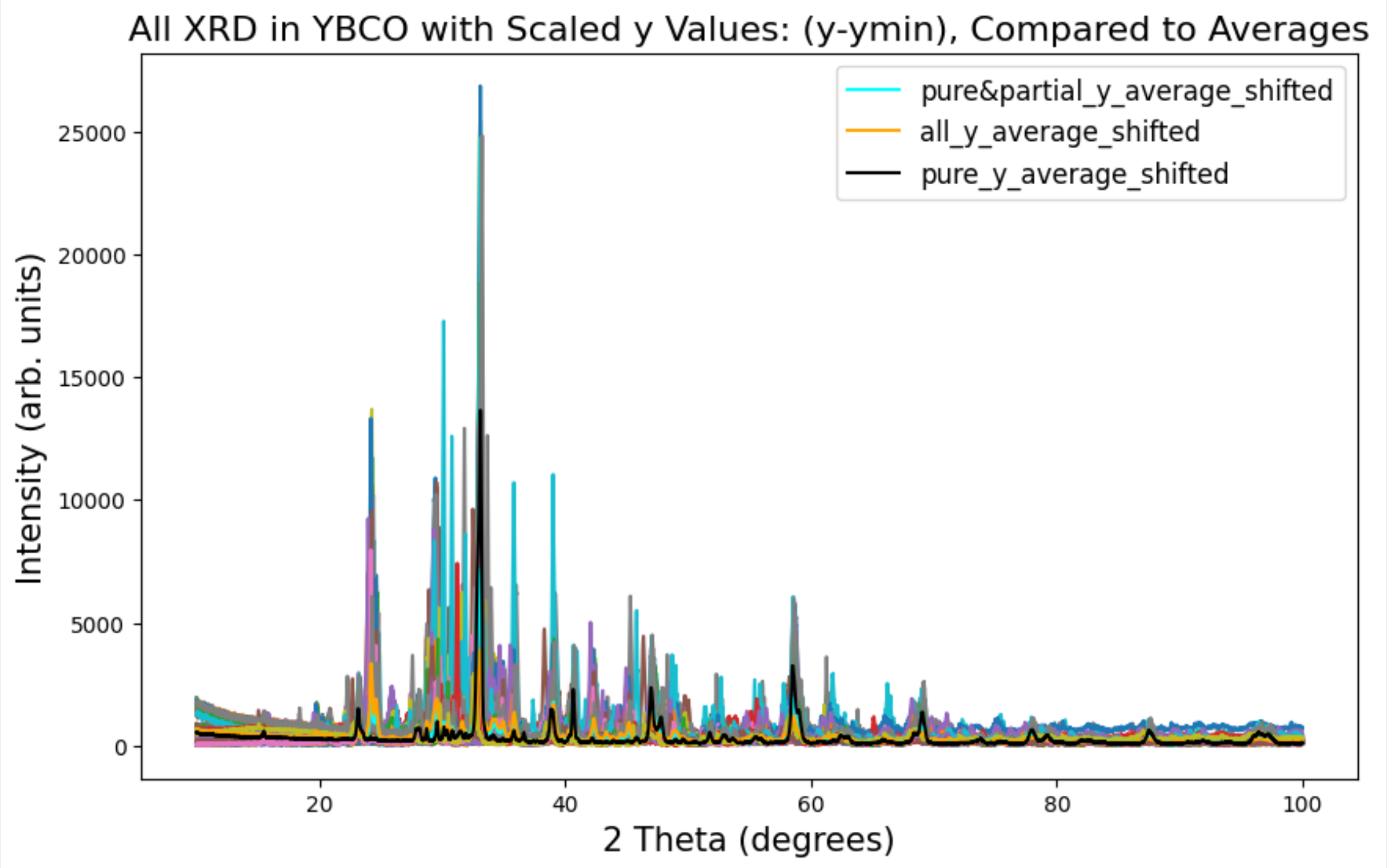

Figure 3: XRD traces in the dataset visualized with one way of approximating the YBCO target (black line).

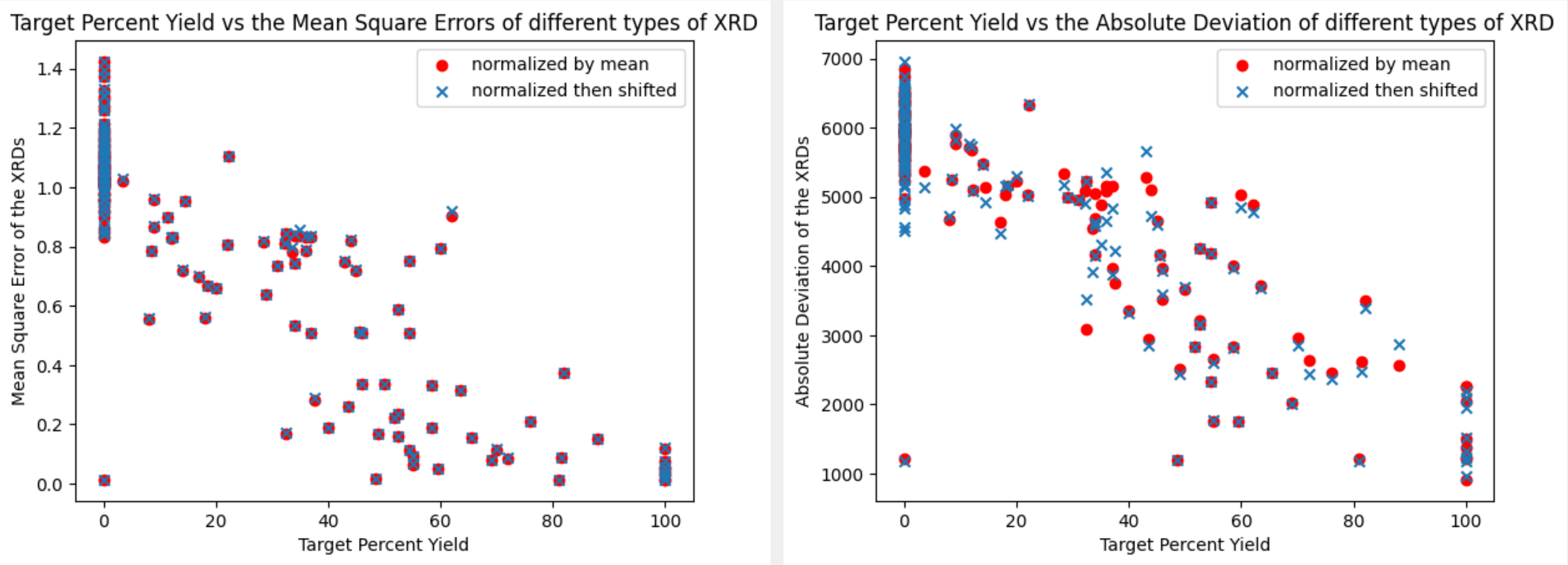

Figure 4: Negative correlation trends would indicate that an objective is a good predictor of YBCO yield:
(left) Target percent yield vs. the MSE of the 149 experimental XRD.
(right) Target percent yield vs. the absolute deviation of the 149 experimental XRD.

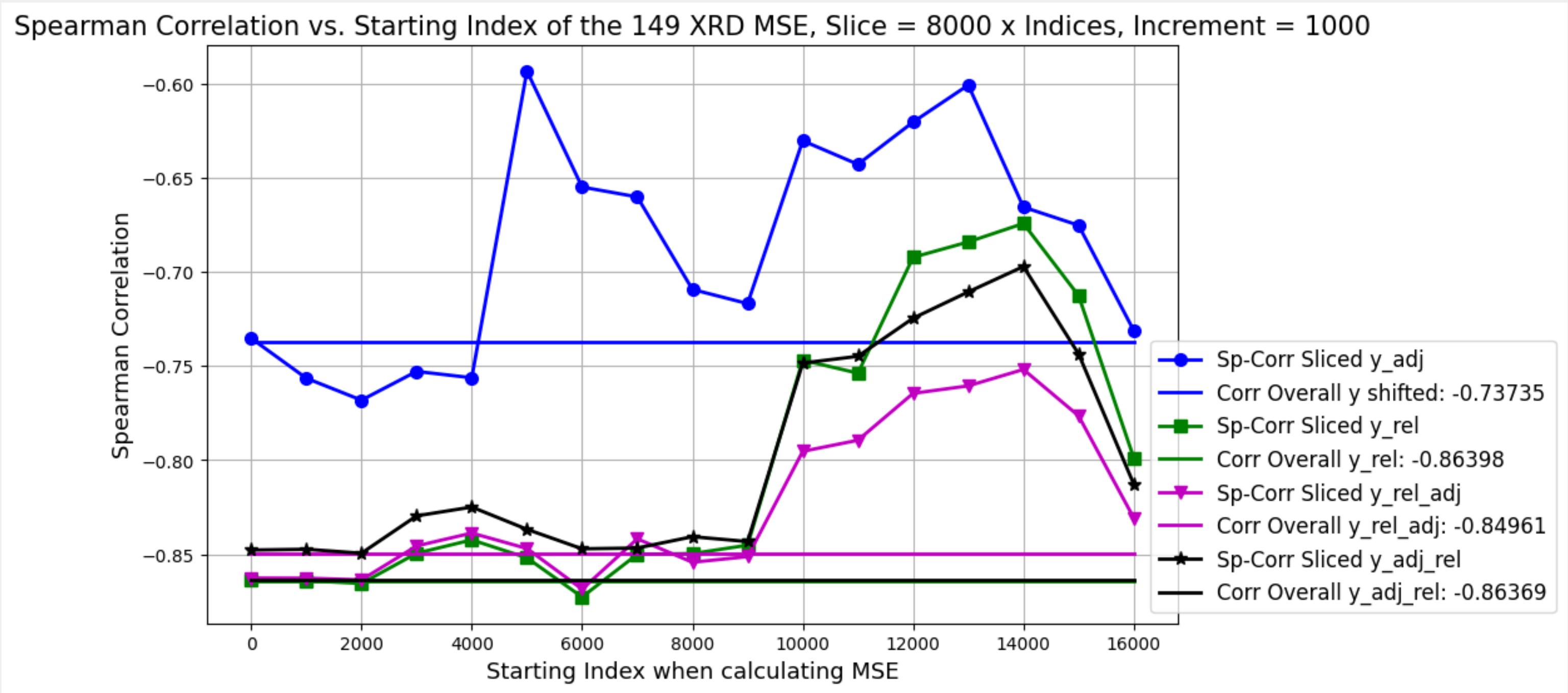

Figure 5: The effects of subsampling XRD indices on the Spearman correlation between subsampled MSE and YBCO yield.

Progress and Results

We first analyzed the dataset using Python and Pandas dataframes. We read the JSON format file as dataframes, extracted and stored useful information such as target percent yield and XRD of each experiment. To better operate with the data that had different sizes of x entries, we used linear interpolation to make sure each experiment has the same data size. Then we looked at XRD from the eight experiments that yielded pure YBCO, took the average and set that as our target function. Since some of the XRD are slightly shifted up, and the location of the peaks matters more, therefore we shifted all XRD by setting each minimum y value to zero. We later tried normalizing each y by its corresponding average. To gather more information, we sliced each XRD with the size of 8000 x indices and increment of 1000, then we analyzed the Spearman correlation between the corresponding MSE and target percent yield. Then we repeated the process to compare the absolute deviation. We observed that normalized y has the best Spearman correlation with target percent yield, roughly at -0.86398.

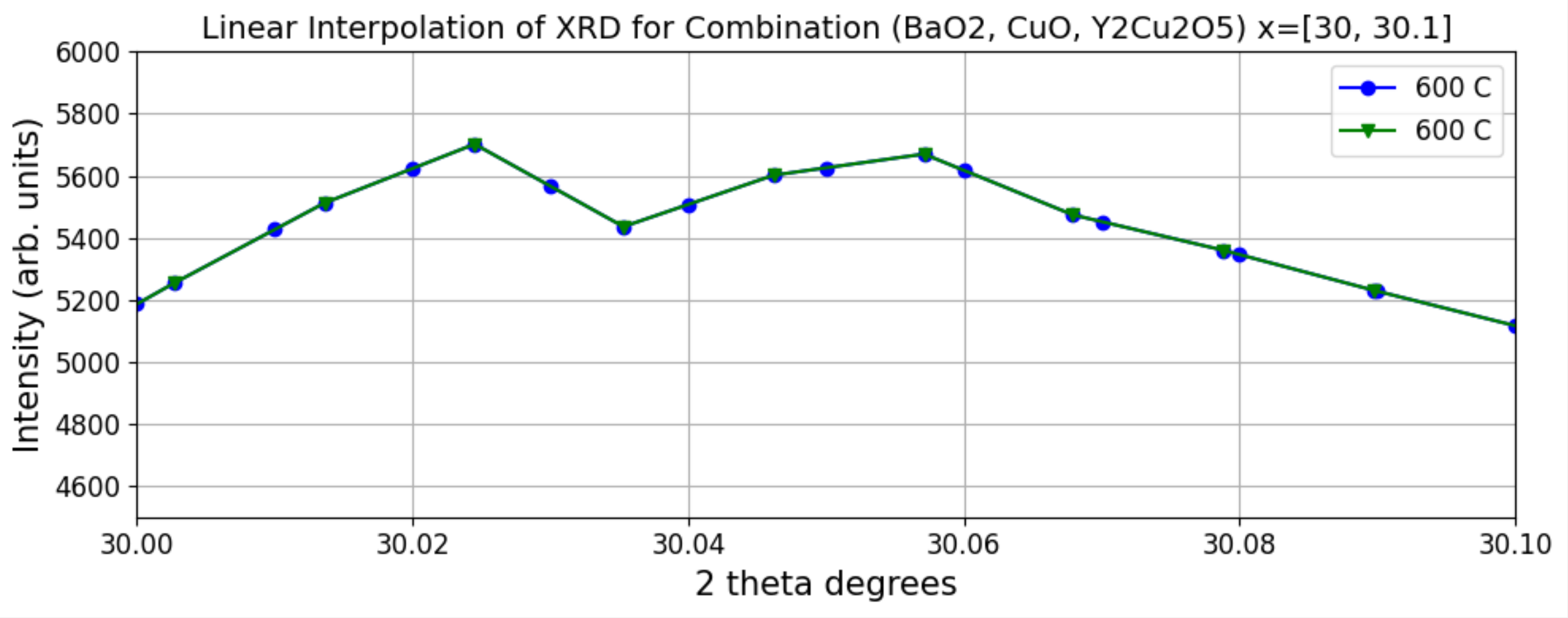

Figure 6: Linear interpolation of all XRD entries, blue is the interpolated dataset.

Future Work

- Optimize the objective function to minimize the experiments that did better than the target
- Determine a threshold as cut-off line to accurately identify and locate peaks in the target XRD patterns
- Compare different objective functions and correlations to improve our methods
- Develop a fully autonomous algorithm based on our findings

References

[1] Szymanski, N.J., Nevatia, P., Bartel, C.J. et al. Autonomous and dynamic precursor selection for solid-state materials synthesis. *Nat Commun 14*, 6956 (2023). https://doi.org/10.1038/s41467-023-42329-9
[2] Biron, Lauren. (2023, April 17). Meet the autonomous lab of the future. *Berkeley Lab News Center*. https://newscenter.lbl.gov/2023/04/17/meet-the-autonomous-lab-of-the-future/
[3] Szymanski, N.J., Nevatia, P., Bartel, C.J. et al. *GitHub* - njszym/ARROWS. GitHub. https://github.com/njszym/ARROWS

Acknowledgments

We are grateful to Lauren Walters (LBNL) for turning us onto this problem and dataset. Many thanks to Jeffery Donatelli and the Math for Experimental Data Analysis group for helpful discussions. Special thanks to Mary Ann Leung, the Sustainable Horizons Institute, and the Sustainable Research Pathways Program for connecting me with Dr. Wild and this amazing opportunity to work at the Berkeley Lab.

Contact information: Ruoxi Zhao: rzhao1129@gmail.com