Supplementary materials for Adaptive Exploration and Optimization of Materials Crystal Structures

Arvind Krishna
Department of Statistics and Data Science,
Northwestern University, Evanston, IL 60208

Huan Tran
School of Material Science and Engineering,
Georgia Institute of Technology, Atlanta, GA 30318

Chaofan Huang
H. Milton Stewart School of Industrial & Systems Engineering,
Georgia Institute of Technology, Atlanta, GA 30332

Rampi Ramprasad
School of Material Science and Engineering,
Georgia Institute of Technology, Atlanta, GA 30318

V. Roshan Joseph
H. Milton Stewart School of Industrial & Systems Engineering,
Georgia Institute of Technology, Atlanta, GA 30332

1 Additional Figures

Figure S1 illustrates the definition of "boundary" and "interior" in a two-dimensional example.

Figure S3 (left) shows the expected improvement as a percentage of the current minimum estimate of potential energy from the energy-data. The expected improvement has a decreasing trend with the number of iterations. As the algorithm learns the potential energy surface and exploits promising locations for global minimum, a lesser improvement in the current global minimum estimate is expected in further iterations. Figure S3 (right) shows the potential energy of the first $n = 320$ observations of the known-data (black circles), followed by that of the 53 fingerprints (red circles) iteratively added to the known-data during the adaptive expansion procedure, which are in-turn followed by the 95 fingerprints (blue) added to the known data during the Bayesian optimization procedure. This figure makes it clear that the non-adaptive expansion algorithm expands the domain space without considering

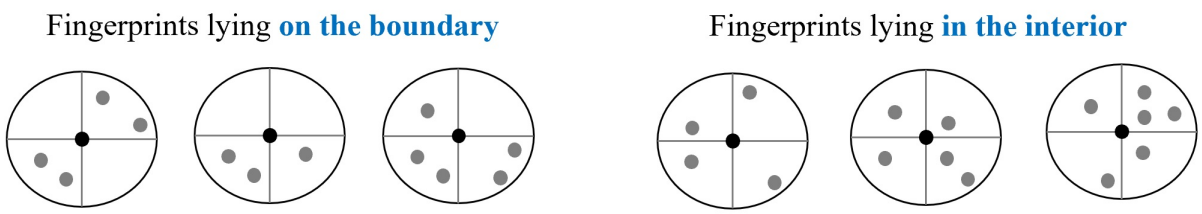

Figure S1: Examples of two-dimensional fingerprints that lie (left): on the boundary of the domain space spanned by the candidate set; (right): in the interior of the domain space spanned by the candidate set. Note that the radius of the circle is r.

the potential energy, the adaptive expansion algorithm drives the expansion towards lower-energy regions, and the Bayesian optimization procedure explores the candidate set and exploits the low-energy regions for the global minimum of potential energy.

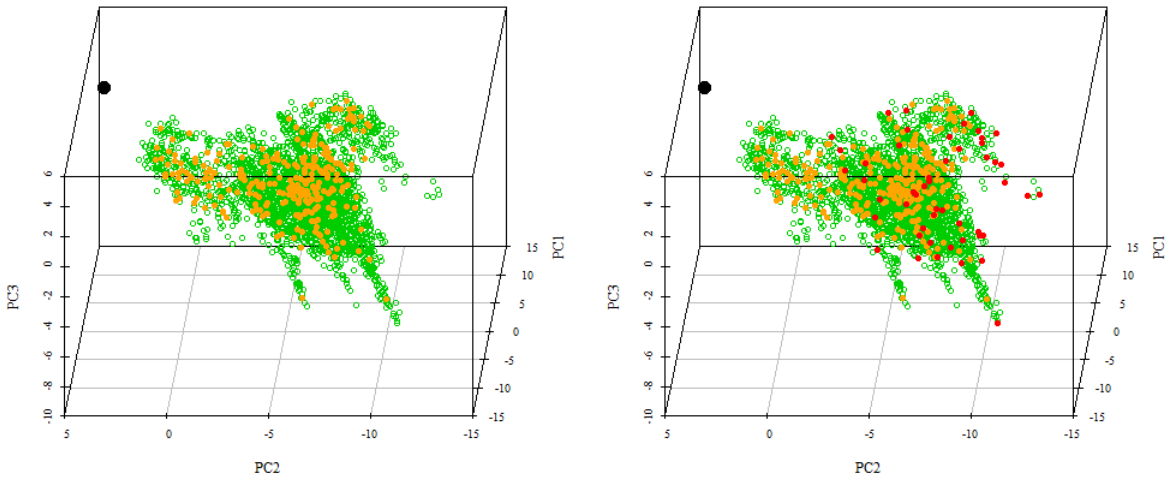

Figure S2: (left): Initial set of fingerprints (orange) over the candidate set of fingerprints (green); (right): Initial set of fingerprints augmented by the space-filling MaxPro design (red).

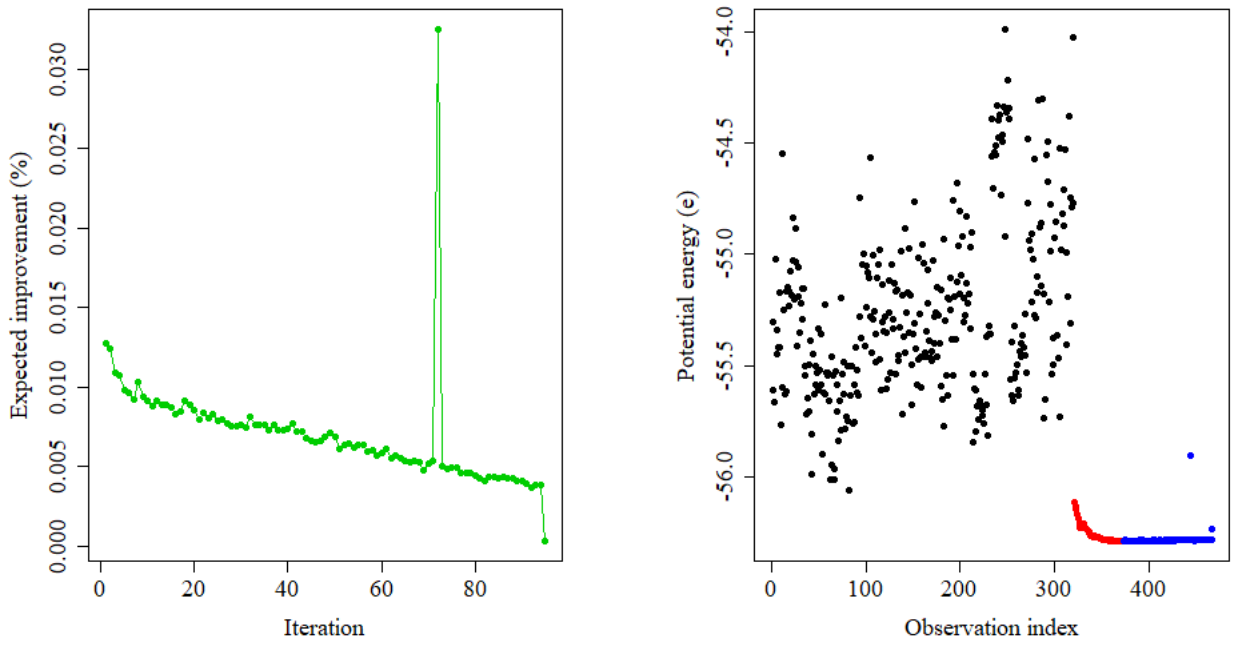

Figure S3: (left): Percentage of expected improvement with respect to the current energy minimum estimate; (right): Potential energy of all the fingerprints in the known-data - for the initial known fingerprints (in black), for the fingerprints added by adaptive expansion (in red), for the fingerprints added during the Bayesian optimization procedure (in blue).

2 Computational complexity

Here are the details of the computational complexity of our methodology. The non-adaptive expansion step has a complexity of $O((n_0+n_1)^2p^2)$ as it requires computation of $N_1 = n_0+n_1$, p-dimensional fingerprints' Euclidean distance to their nearest neighbor in each of their $2p$ neighborhoods.

The computational complexity of fitting a GP over k points is $O(k^3)$. In the adaptive expansion step, the first GP is based on n_0 fingerprints (assuming $n_0 = 10p$), and is updated after the addition of every 10 fingerprints to the candidate set. Thus the last GP is based on $(n_0 + n_2/10)$ fingerprints. This leads to a complexity of $O((n_0 + n_2/10)^3 n_2)$ for the adaptive expansion procedure. As in non-adaptive expansion, the adaptive expansion algorithm has a complexity of $O(((n_0 + n_1)n_2 + n_2^2/2)p^2)$ associated with Euclidean distance computation to the nearest neighbor for each fingerprint added to the candidate set. However, since this cost is negligible compared to the cost of fitting the GP model, it is ignored.

In case of Bayesian optimization, a GP is fit repeatedly until the expected improvement becomes negligible. The first GP is based on $(n_0 + n_2/10)$ fingerprints, while the last one is based on $(n_0 + n_2/10 + n_3)$ fingerprints. This leads to a computational complexity of $O((n_0 + n_2/10 + n_3)^3 n_3)$ for the Bayesian optimization procedure.