Accelerating Black-Box Molecular Property Optimization by Adaptively Learning Sparse Subspaces

Farshud Sorourifar SOROURIFAR.1@OSU.EDU
Thomas Banker THOMAS_BANKER@BERKELEY.EDU
Joel A. Paulson PAULSON.82@OSU.EDU
Department of Chemical and Biomolecular Engineering,
The Ohio State University, Columbus, OH 43210, USA

Abstract

Molecular property optimization (MPO) problems are inherently challenging since they are formulated over discrete, unstructured spaces and the labeling process involves expensive simulations or experiments, which fundamentally limits the amount of available data. Bayesian optimization (BO), which is a powerful and popular framework for efficient optimization of noisy, black-box objective functions (e.g., measured property values), thus is a potentially attractive framework for MPO. To apply BO to MPO problems, one must select a structured molecular representation that enables construction of a probabilistic surrogate model. Many molecular representations have been developed, however, they are all high-dimensional, which introduces important challenges in the BO process – mainly because the curse of dimensionality makes it difficult to define and perform inference over a suitable class of surrogate models. This challenge has been recently addressed by learning a lower-dimensional encoding of a SMILE or graph representation of a molecule in an unsupervised manner and then performing BO in the encoded space. In this work, we show that such methods have a tendency to "get stuck," which we hypothesize occurs since the mapping from the encoded space to property values is not necessarily well-modeled by a Gaussian process. We argue for an alternative approach that combines numerical molecular descriptors with a sparse axis-aligned Gaussian process model, which is capable of rapidly identifying sparse subspaces that are most relevant to modeling the unknown property function. We demonstrate that our proposed method substantially outperforms existing MPO methods on a variety of benchmark and real-world problems. Specifically, we show that our method can routinely find near-optimal molecules out of a set of more than $>$ 100k alternatives within 100 or fewer expensive queries.

1. Introduction

Molecular property optimization (MPO) is the process of systematically improving the structural and/or functional properties of molecules to meet specific objectives. MPO is a critical step in a variety of scientific and engineering applications including chemistry, drug discovery, and material science. One can generally formulate MPO problems as follows

$$m^\star \in \underset{m \in \mathcal{M}}{\operatorname{argmax}} \; F(m), \tag{1}$$

where m is a molecule from the discrete set $\mathcal{M}$ and $F : \mathcal{M} \to \mathbb{R}$ is an unknown objective function that maps a molecule to a performance value. The goal is thus to find the "best"

molecule $m^\star$ that has the best performance $F(m^\star) \geq F(m)$ for all $m \in \mathcal{M}$ where $\mathcal{M}$ is a set with large but finite cardinality $|\mathcal{M}| < \infty$. This problem is easily solved if we could perfectly measure the property for every molecule; however, in reality, (i) we often only get noisy observations $y = F(m) + \varepsilon$ and (ii) the number of candidate molecules is very large (millions or more) such that we can only observe a relatively small number of options. In recent years, there has been a substantial amount of work on the use of machine learning as an effective tool to address these challenges (Bartók et al., 2017; von Lilienfeld and Burke, 2020). Most work on machine learning for MPO can be divided into two categories: *guided search* and *translation* (Hoffman et al., 2022). In guided search, the idea is to construct predictive models over some type of molecular representation to sequentially select promising molecules for testing. Translation, on the other hand, treats the molecule generation problem as a sequence-to-sequence translation problem, which requires additional information that is not always available. Thus, in this work, we focus on developing a guided search approach.

To develop a guided search strategy, one must first select an effective numerical representation of the molecule. We can think of this representation as a function $R : \mathcal{M} \to \mathcal{X}$ that maps from molecule space $\mathcal{M}$ to a numerical feature space $\mathcal{X}$. As long as this mapping is invertible, we can equivalently express (1) as finding $m^\star = R^{-1}(x^\star)$ where $x^\star \in \text{argmax}_{x \in \mathcal{X}} f(x)$ and $f(x) = F \circ R^{-1}(x)$ is the objective as a function of the structured numerical representation vector x. The space $\mathcal{X}$ could be continuous or discrete and many different representations have been proposed in the literature including SMILES strings (Anderson et al., 1987), molecular graphs (Wieder et al., 2020), molecular fingerprints (Cereto-Massagué et al., 2015), and molecular descriptors (Moriwaki et al., 2018). An important challenge with all of these representations is that they are naturally high-dimensional, which complicates the optimization problem. A recent line of work has looked to address this challenge by learning a lower-dimensional continuous latent representation in which one can more easily execute efficient search strategies such as Bayesian optimization (BO) (Frazier, 2018). One of the most common examples is the combined use of a variational autoencoder and BO (Gómez-Bombarelli et al., 2018; Griffiths and Hernández-Lobato, 2020). Let $z = E(x)$ denote an encoded latent representation of x. These methods proceed by using BO to maximize $g(z) = f(D(z))$ over z (where a decoder transforms back to $x = D(z)$). If E is learned using little-to-no property data, then there is no driving force for g to have smoothness or continuity properties that would allow one to efficiently search over the latent space $\mathcal{Z}$ even if it has a smaller dimension than $\mathcal{X}$. Furthermore, it is possible that sparsity in the behavior of $f(x)$ is lost when transforming to $g(z)$.

In this paper, we propose the **Mol**ecular **D**escriptors and **A**ctively **I**dentified **S**ubspaces (MolDAIS) framework, which is a new strategy to approach MPO problems that work directly in a numerical molecular feature space. In particular, we argue that molecular descriptors (i.e., outcomes of mathematical procedures applied to the symbolic representation of a molecule) (Todeschini and Consonni, 2000) combined with Gaussian process surrogate models defined on sparse axis-aligned subspaces (SAAS) provide an effective framework for MPO in the low-data regime. The motivation for our approach can be traced back to ideas in interpretable machine learning, which often posit the existence of a relatively small number of well-selected (understandable) features that can be used to accurately predict the desired property of interest. Since these key descriptors may not be known *a priori*, one can attempt to learn them with sparse regression methods such as SISSO (Ouyang

et al., 2018). These sparse regression methods, however, require that the target property depend linearly on the important descriptors in the feature set. They also do not directly capture uncertainty, which is crucial for navigating the exploration-exploitation tradeoff in BO. It turns out that we can address these both of challenges by taking advantage of the SAAS prior developed in (Eriksson and Jankowiak, 2021). Not only can SAAS sparsely pick out features from x, the space is systematically updated as new information is collected, enabling adaptive learning of sparse and interpretable subspaces. The latter point significantly simplifies the inference task, which reduces the amount of data needed to make useful property predictions. We demonstrate the advantages of MolDAIS on three unique problems by comparing to existing MPO approaches. First, we consider a benchmark logP problem for which we can consistently find the best molecule out of 250k candidates in only 100 iterations (substantially outperforming state-of-the-art alternatives). Second, we consider two real-world problems related to optimization over a class of organic molecules whose properties are computed from an expensive density function theory simulation.

2. The MolDAIS Framework

2.1 Molecular Descriptors

Molecular descriptors are defined as the *"final result of a logical and mathematical procedure, which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardized experiment"* (Todeschini and Consonni, 2000). Many types of molecular descriptors have been developed. Here, we focus on a set of more than 1800 descriptors that can be efficiently computed from the open-source Mordred software program (Moriwaki et al., 2018). Examples of the descriptors include outcomes of rotationally and translationally invariant operations on a molecular graph as well as quantities relevant in chemistry (such as atomic weights). Using these descriptors, we can cast (1) as the following equivalent optimization problem

$$x^\star \in \underset{x \in \mathcal{X}}{\operatorname{argmax}} \; f(x), \tag{2}$$

where $f(x) = F(R^{-1}(x))$ denotes the performance value as a function of the Mordred representation $x = R(m)$ and $\mathcal{X} = [0,1]^D$ is the normalized numerical Mordred space defined over a D-dimensional hypercube ($D \sim 2000$ in this work).

2.2 Bayesian Optimization with Sparse Axis-aligned Subspaces

The challenge with solving (2) directly is that $\mathcal{X}$ is a high-dimensional space so it is difficult to apply efficient optimization strategies, such as Bayesian optimization (BO), due to the curse of dimensionality that manifests when attempting to build a surrogate model for $f(x)$ from initial data. In particular, Gaussian processes (GPs) are non-parametric function priors that are commonly used in BO due to their flexibility and natural uncertainty quantification abilities (Williams and Rasmussen, 2006). A GP over an input space $\mathcal{X}$ is fully specified by a prior mean function $\mu_0 : \mathcal{X} \to \mathbb{R}$ and prior covariance (or kernel) function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. As commonly done, we will assume $\mu_0(x) = 0$ for all $x \in \mathcal{X}$, which can be achieved in practice by normalizing the input data. The kernel function encodes information about the smoothness and rate of change of the unknown function, with a popular

choice being the squared exponential (SE) kernel given by

$$k^{\psi}(x, x') = \sigma_k^2 \exp\left\{-\tfrac{1}{2}\sum_{i=1}^{D} \rho_i (x_i - x_i')^2\right\}, \tag{3}$$

where $\psi = \{\rho_1, \ldots, \rho_D, \sigma_k^2\}$ are the hyperparameters of the kernel that consist of inverse lengthscales $\{\rho_i\}_{i=1}^D$ for each dimension and the output scale σ_k^2. For fixed ψ, the posterior distribution $p(f(x_\star)|\mathcal{D}_n) \sim \mathcal{N}(\mu_n(x_\star), \sigma_n^2(x_\star))$ at a test point $x_\star$ given past observations $\mathcal{D}_n = \{(x_i, y_i)\}_{i=1}^n$ can be analytically computed. Given this predictive distribution, standard BO proceeds by selecting the next evaluation point $x_{n+1} \in \operatorname{argmax}_{x\in\mathcal{X}} \alpha_n(x)$ that maximizes an acquisition function $\alpha_n : \mathcal{X} \to \mathbb{R}$, which should be chosen to provide a good measure of the potential benefit of querying f at every $x \in \mathcal{X}$ in the future. By defining our utility function as the best observed sample $u(\mathcal{D}_n) = \max_{(x,y)\in\mathcal{D}_n} y_n$, we can derive the *expected improvement* (EI) acquisition function as the expected increase in utility $\alpha_n(x) = \mathrm{EI}_n(x|\eta, \psi) = \mathbb{E}_n\{u(\mathcal{D}_{n+1}) - u(\mathcal{D}_n)\}$ where $\eta = \max_n y_n$ is the best observed value so far. For a standard GP model, EI has a simple closed-form solution (Jones et al., 1998).

The main challenge with the standard BO strategy for (2) is that, when D is large, the space of possible functions mapping $\mathcal{X}$ to $\mathbb{R}$ is too large to learn even assuming some degree of smoothness imparted by the SE kernel (3). Without additional prior information, a natural way to deal with this challenge is to assume a hierarchy of relevance in the dimensions such that we can focus on the subset of features in $x = \{x_1, x_2, \ldots, x_D\}$ that are important for our property of interest. We choose to exploit the sparse axis-aligned subspace (SAAS) prior to accomplish this task. In short, SAAS is a prior over the kernel hyperparameters that induces sparse structure in the inverse squared lengthscales ρ_i. Small values of ρ_i imply dimension i is *unimportant* in the prediction, so a half-Cauchy prior is used to concentrate their values near zero. Only once observations provide enough evidence that dimension i is important can ρ_i escape zero. As more data is accumulated, the number of dimensions allowed to escape increases, allowing more ρ_i to be "activated", leading to a richer class of functions. Interested readers are referred to Eriksson and Jankowiak (2021) for details on the SAAS prior as well as the training procedure that exploits an established Hamiltonian Monte Carlo method to generate L approximate posterior samples for the kernel hyperparameters $\{\psi_l\}_{l=1}^L$. Therefore, given n previous property evaluations for different molecules, the next molecule we want to sample is selected according to $m_{n+1} = R^{-1}(x_{n+1})$

$$x_{n+1} \in \operatorname*{argmax}_{x\in\mathcal{X}} \ \tfrac{1}{L}\sum_{l=1}^{L} \mathrm{EI}_n(x|\eta, \psi_l). \tag{4}$$

These steps have been implemented in the BoTorch package (Balandat et al., 2020) for which (4) can be tackled using efficient gradient-based optimization methods.

3. Experiments

3.1 Baseline Methods

We compare MolDAIS to several baselines that are a mixture of different choices of the starting molecular feature space, the learned latent representation, and type of BO method. All methods are provided with 10 randomly chosen initial samples and a budget of 90 additional samples and results are shown for 5 independent replications of each algorithm.

LADDER is a recently proposed MPO method that combines molecular fingerprints (FP) with a junction-tree autoencoder (JTAE) to construct a latent representation of the FP space (Deshwal and Doppa, 2021). A standard BO method is then used to actively select new samples in the latent space. We use the default settings from the original paper

SAAS-FP-JTAE is a variant of LADDER that uses the same latent representation (learned by applying a JTAE to molecular FPs) but now uses the SAASBO strategy described in Section 2.2. This is meant to study the impact of the starting molecular descriptor space, as we expect the SAAS prior to be less effective in the encoded latent space.

SBO refers to the standard BO method that optimizes directly over the high-dimensional molecular descriptor $\mathcal{X}$ space using the EI acquisition function. This is meant to study the importance of the SAAS prior for accelerating convergence.

SBO-PCA is a slight modification to SBO that replaces the high-dimensional $\mathcal{X}$ space with a new lower-dimensional latent space constructed by applying principal component analysis (PCA) to the unlabeled molecular descriptor data. The size of the latent space is chosen such that 99% of the variance in the $\mathcal{X}$ data is captured.

SLR refers to sequential linear regression, which is a very naive version of MolDAIS that uses sparse linear regression to identify a model with a small number of non-zero coefficients. This linear surrogate model is then maximized directly to select the next sample.

Random refers to a standard random search strategy wherein the next sample is chosen uniformly at random in $\mathcal{X}$. This method, which does not exploit any past data, is meant to provide a lower bound on performance.

3.2 Maximizing LogP over the Zinc Molecule Dataset

We first consider the problem of finding molecules with the best drug-like properties (Kusner et al., 2017). In particular, the goal is to maximize the water-octanol partition coefficient (logP) over the space of molecules. As done in previous work, we consider the Zinc molecule data set that consists of $250,000$ commercially-available molecules.

The results are shown in Figure 1 (Left), which shows that MolDAIS clearly outperforms all other methods, as it is able to find the highest logP value by sampling less than 0.04% of the candidates in all replicates. This represents a more than 30% improvement over the best candidates found with all other methods under the same conditions. Even a reduced space variant of MolDAIS, which removes any feature suspected to highly correlate to logP, still outperforms all other tested methods. Interestingly, LADDER performs relatively similar to the other methods on average and actually results in a larger distribution of outcomes.

3.3 Maximizing Solvation Free Energy for Quinone Molecule Class

Next, we consider the problem of finding the best molecule from the quinone class (Tabor et al., 2019) that has the largest Gibbs free energy of solvation ΔG_{solv}, which is an important property for battery materials and pharmaceuticals. Although one can compute ΔG_{solv} using density functional theory (DFT), it requires two separate simulations in different phases, which takes around 24 CPU hours per molecule on a supercomputing cluster. We consider a space of $> 100,000$ quinone molecules and look to maximize ΔG_{solv}. The results are shown in Figure 1 (Middle), which again outperform all considered alternatives. MolDAIS finds the global solution within around 10 iterations. Since the latent space built

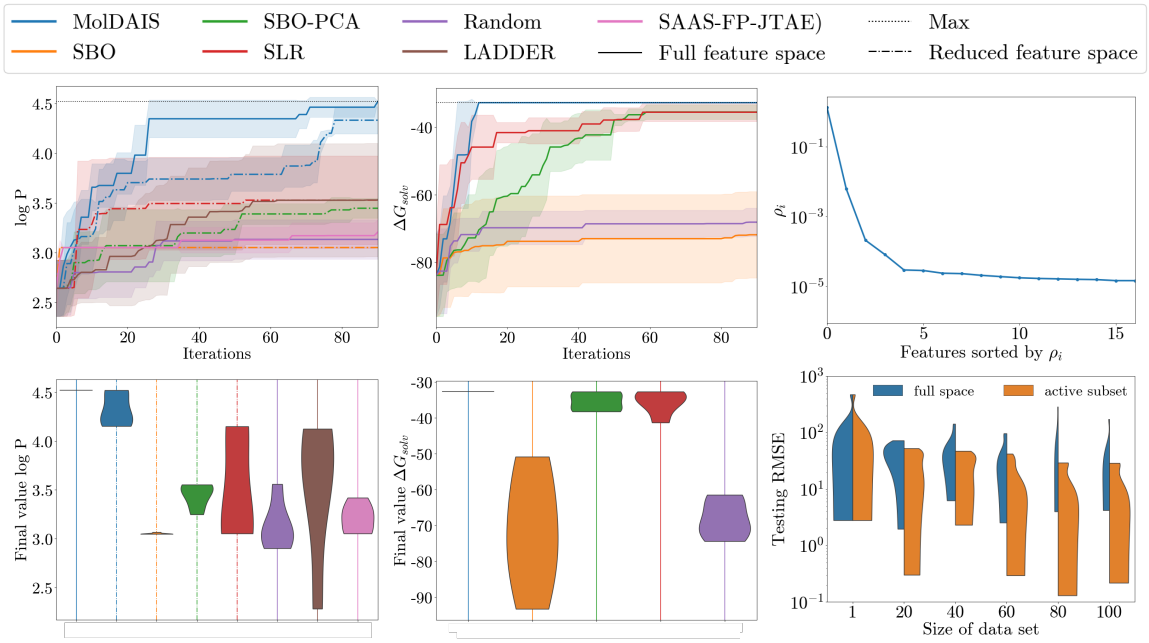

Figure 1: **Left:** logP maximization. **Middle:** Solvation free energy maximization. **Right:** Feature importance and test accuracy for solvation free energy GP model.

with LADDER may not be compatible with the quinone class, we do not directly compare against LADDER or SAAS-FP-JTAE in this example. To gain more insight into why MolDAIS can achieve such strong results, we perform additional analysis on the SAAS-GP model (Figure 1; Right). The top plot shows the ρ_i values sorted in decreasing order; we see a clear separation in terms of the top 4 values and the remaining values, indicating a small number of Mordred descriptors are needed to predict ΔG_{solv}. Furthermore, the bottom plot shows the root mean squared error (RMSE) on a held-out set of 100 test molecules based on randomly sampled training sets of different sizes. We see that the RMSE values are significantly smaller even when only 20 data points are known, indicating the SAAS prior does enable learning a model structure with improved prediction accuracy. See Appendix A for similar results achieved on a reduction potential maximization problem.

4. Conclusions

In this work, we propose a new molecular property optimization (MPO) method, MolDAIS, that can efficiently identify high-performance molecules in the low-data regime. MolDAIS combines molecular descriptors with a sparse axis-aligned subspace (SAAS) prior to adaptively learn sparse and interpretable subsets of the high-dimensional molecular feature space that can be used within a Bayesian optimization (BO) framework to directly balance exploration and exploitation of the search space. We empirically show that MolDAIS outperforms existing MPO methods on benchmark and real-world problems. Furthermore, in some cases, MolDAIS can find near globally optimal molecules with 100 or less queries out of more than 100k candidates without any prior information (i.e., in a fully black-box manner).

References

Eric Anderson, Gilman D Veith, and David Weininger. *SMILES, a line notation and computerized interpreter for chemical structures*. US Environmental Protection Agency, Environmental Research Laboratory, 1987.

Maximilian Balandat, Brian Karrer, Daniel Jiang, Samuel Daulton, Ben Letham, Andrew G Wilson, and Eytan Bakshy. BoTorch: A framework for efficient Monte-Carlo Bayesian optimization. *Advances in Neural Information Processing Systems*, 33:21524–21538, 2020.

Albert P Bartók, Sandip De, Carl Poelking, Noam Bernstein, James R Kermode, Gábor Csányi, and Michele Ceriotti. Machine learning unifies the modeling of materials and molecules. *Science Advances*, 3(12):e1701816, 2017.

Adrià Cereto-Massagué, María José Ojeda, Cristina Valls, Miquel Mulero, Santiago Garcia-Vallvé, and Gerard Pujadas. Molecular fingerprint similarity search in virtual screening. *Methods*, 71:58–63, 2015.

Aryan Deshwal and Jana Doppa. Combining latent space and structured kernels for bayesian optimization over combinatorial spaces. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 8185–8200. Curran Associates, Inc., 2021. URL `https://proceedings.neurips.cc/paper_files/paper/2021/file/44e76e99b5e194377e955b13fb12f630-Paper.pdf`.

David Eriksson and Martin Jankowiak. High-dimensional bayesian optimization with sparse axis-aligned subspaces. In *Uncertainty in Artificial Intelligence*, pages 493–503. PMLR, 2021.

Peter I. Frazier. A tutorial on bayesian optimization, 2018.

Rafael Gómez-Bombarelli, Jennifer N Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D Hirzel, Ryan P Adams, and Alán Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Central Science*, 4(2):268–276, 2018.

Ryan-Rhys Griffiths and José Miguel Hernández-Lobato. Constrained Bayesian optimization for automatic chemical design using variational autoencoders. *Chemical Science*, 11 (2):577–586, 2020.

Samuel C Hoffman, Vijil Chenthamarakshan, Kahini Wadhawan, Pin-Yu Chen, and Payel Das. Optimizing molecules using efficient queries from property evaluations. *Nature Machine Intelligence*, 4(1):21–31, 2022.

Donald R Jones, Matthias Schonlau, and William J Welch. Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13:455–492, 1998.

Matt J Kusner, Brooks Paige, and José Miguel Hernández-Lobato. Grammar variational autoencoder. In *International Conference on Machine Learning*, pages 1945–1954. PMLR, 2017.

Hirotomo Moriwaki, Yu-Shi Tian, Norihito Kawashita, and Tatsuya Takagi. Mordred: a molecular descriptor calculator. *Journal of Cheminformatics*, 10(1):4, Feb 2018. ISSN 1758-2946. doi: 10.1186/s13321-018-0258-y. URL `https://doi.org/10.1186/s13321-018-0258-y`.

Runhai Ouyang, Stefano Curtarolo, Emre Ahmetcik, Matthias Scheffler, and Luca M. Ghiringhelli. Sisso: A compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates. *Phys. Rev. Mater.*, 2:083802, Aug 2018. doi: 10.1103/PhysRevMaterials.2.083802. URL `https://link.aps.org/doi/10.1103/PhysRevMaterials.2.083802`.

Daniel P. Tabor, Rafael Gómez-Bombarelli, Liuchuan Tong, Roy G. Gordon, Michael J. Aziz, and Alán Aspuru-Guzik. Mapping the frontiers of quinone stability in aqueous media: implications for organic aqueous redox flow batteries. *J. Mater. Chem. A*, 7: 12833–12841, 2019. doi: 10.1039/C9TA03219C. URL `http://dx.doi.org/10.1039/C9TA03219C`.

R. Todeschini and V Consonni. *Frontmatter*, pages i–xxi. John Wiley & Sons, Ltd, 2000. ISBN 9783527613106. doi: https://doi.org/10.1002/9783527613106.fmatter. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/9783527613106.fmatter`.

O Anatole von Lilienfeld and Kieron Burke. Retrospective on a decade of machine learning for chemical discovery. *Nature Communications*, 11(1):4895, 2020.

Oliver Wieder, Stefan Kohlbacher, Mélaine Kuenemann, Arthur Garon, Pierre Ducrot, Thomas Seidel, and Thierry Langer. A compact review of molecular property prediction with graph neural networks. *Drug Discovery Today: Technologies*, 37:1–12, 2020.

Christopher KI Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.

Appendix A. Maximizing Reduction Potential

To further illustrate the flexibility of the proposed MolDAIS method, we consider another variant of an MPO problem defined over the quinone molecule class considered in Section 3.3. Specifically, we now consider the reduction (or redox) potential that are a measure of energy required to reduce or oxide a molecule relative to the standard hydrogen electrode, which can again be predicted using DFT. The results for maximizing redox potential E^0 are shown in Figure 2. Even though this is a completely different property than ΔG_{solv}, MolDAIS still outperforms all considered alternative methods, though the SBO-PCA method does perform fairly close towards the final iterations. We again see a clear separation in terms of the number of relevant dimensions (those with relatively large ρ_i values), though the overall values are higher than in the ΔG_{solv} that is likely the source of slightly worse performance. Similarly, from a prediction quality point-of-view, the SAAS-GP still outperforms the traditional GP, though there is less of a gap than in the previous cases.

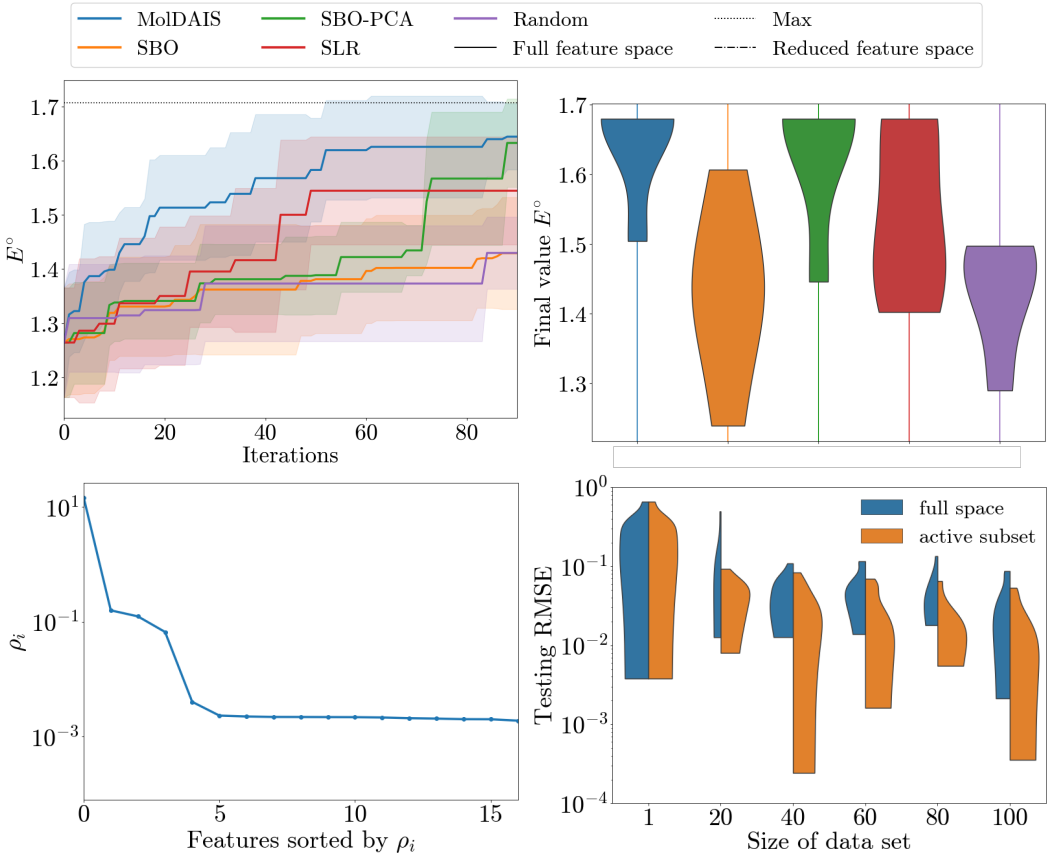

Figure 2: **Top Left:** Convergence results for maximization of redox potential versus number of iterations. **Top Right:** The distribution in the best maximum found at the final iteration over the replicates. **Bottom Left:** Inverse squared lengthscale values sorted in descending order for the SAAS-GP model at the final iteration. **Bottom Right:** Test RMSE values for the standard GP and SAAS-GP models given random training sets of different sizes.