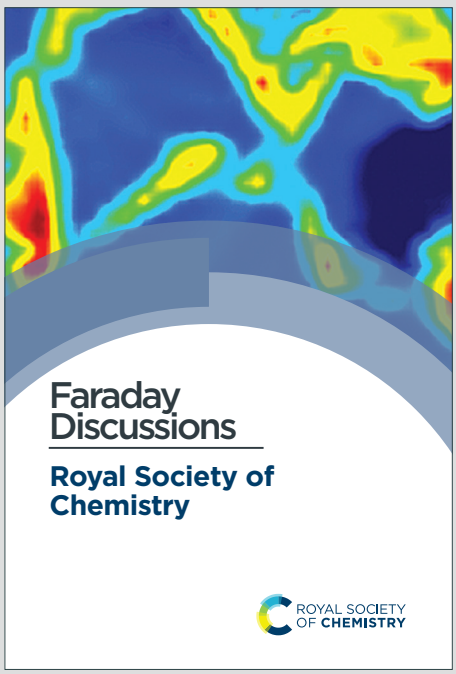
View Article Online
View Journal

Faraday Discussions

Accepted Manuscript



This is an Accepted Manuscript, which has been through the Royal Society of Chemistry peer review process and has been accepted for publication.

Accepted Manuscripts are published online shortly after acceptance, before technical editing, formatting and proof reading. Using this free service, authors can make their results available to the community, in citable form, before we publish the edited article. We will replace this Accepted Manuscript with the edited and formatted Advance Article as soon as it is available.

You can find more information about Accepted Manuscripts in the Information for Authors.

Please note that technical editing may introduce minor changes to the text and/or graphics, which may alter content. The journal's standard Terms & Conditions and the Ethical guidelines still apply. In no event shall the Royal Society of Chemistry be held responsible for any errors or omissions in this Accepted Manuscript or any consequences arising from the use of any information it contains.

This article can be cited before page numbers have been issued, to do this please use: A. S. Nair, L. Foppa and M. Scheffler, *Faraday Discuss.*, 2026, DOI: 10.1039/D5FD00159E.

ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/faraday_d

View Article Online
DOI: 10.1039/D5FD00159E

Journal Name

ARTICLE TYPE

Cite this: DOI: 00.0000/xxxxxxxxxx

Interpretable Bayesian Optimization for Catalyst Discovery†

Akhil S. Nair,*[a,b] Lucas Foppa,[a] and Matthias Scheffler[a]

Received Date
Accepted Date

DOI: 00.0000/xxxxxxxxxx

Bayesian Optimization (BO) efficiently explores vast design spaces using probabilistic surrogate models, enabling the guided discovery of materials with desired properties. However, most BO frameworks rely on the knowledge of a few key physical parameters (features) correlated with the materials property of interest. This is a challenge in heterogeneous catalysis, where the material properties are governed by an intricate interplay of multiple physical processes, and the mentioned parameters are typically unknown. Here, we introduce the Sparse Adaptive Representation–based Bayesian Optimization (SARBO) framework that utilizes the sure independence screening and sparsifying operator (SISSO) symbolic-regression method for on-the-fly selection of key physical parameters correlated with materials properties during BO. Crucially, SISSO takes into account nonlinear relationships and interactions between multiple parameters when selecting key features. We demonstrate that SARBO enables efficient navigation of the materials spaces and outperforms widely used feature-selection approaches for the simulated discovery of single- and dual-atom alloy surface sites capable of activating CO_2, a critical step in the CO_2 reduction reaction.

1 Introduction

The discovery of new materials is essential for progress in many technologies, including catalysis and energy storage[1,2]. However, the population of possible materials or *materials space*, is practically infinite. Hence, the *direct* screening of materials via high-throughput experimentation or computational simulations is inefficient. This has motivated the development of autonomous materials discovery platforms, where closed-loop systems selectively acquire the next materials to be investigated and incorporate the feedback from experiments or simulations to improve the efficiency of materials discovery[3,4]. Bayesian Optimization (BO) has been an essential component of such platforms, as it is well-suited for optimizing objective functions that link physical parameters describing the material properties, termed *features*, to a certain target materials property[5,6]. In the context of materials discovery, the optimization could refer to identifying materials with optimal properties, e.g., with high or low values. BO has two main components: a surrogate model and an acquisition function[7]. The surrogate (machine-learning) model approximates the objective function, mapping the features to the material properties based on the observed data. Gaussian Process (GP) and Random Forests (RF) are the most common surrogate models for BO due to their probabilistic formulation[8] and robustness to noise[9], respectively. GP models the objective function non-parametrically as a distribution over smooth functions, enabling both accurate predictions and principled uncertainty estimates. RF uses an ensemble of decision trees, providing strong predictive performance while offering practical uncertainty estimates through ensemble variance. The acquisition function uses predictions and uncertainty estimates from the surrogate model to suggest the next materials from the *design space*, i.e. the collection of candidate materials represented through a set of features, available for evaluation. The selected materials are then evaluated, added to the training data, and the surrogate model is updated, completing one BO iteration. Thus, BO balances the selection of materials with predicted desired properties, i.e., the model *exploitation*, and the *exploration* of portions of the design space that may not be well represented in the training set. By focusing on such carefully selected samples, BO significantly reduces the effort, e.g., the cost for evaluation of the materials property, necessary to identify materials with desired properties[10,11].

Despite its successes in materials discovery applications[12–14], the efficiency of BO decreases as the number of features increases. This is because the search domain increases exponentially with the number of features, making it difficult to identify optimal regions within an affordable number of property evaluations[15]. This is a limitation in fields such as heterogeneous catalysis, where the catalytic performance of a material is governed by an intricate interplay of multiple physical processes, and the key features correlated with the catalyst performance are often unknown. In such cases, BO might utilize significant resources to

[a] The NOMAD Laboratory at the Fritz Haber Institute of the Max Planck Society, Faradayweg 4-6, D-14195 Berlin, Germany
[b] Institut für Chemie und Biochemie, Freie Universität Berlin, Arnimallee 22, 14195 Berlin, Germany

Open Access Article. Published on 27 January 2026. Downloaded on 4/29/2026 3:49:16 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

Faraday Discussions Accepted Manuscript

View Article Online
DOI: 10.1039/D5FD00159E

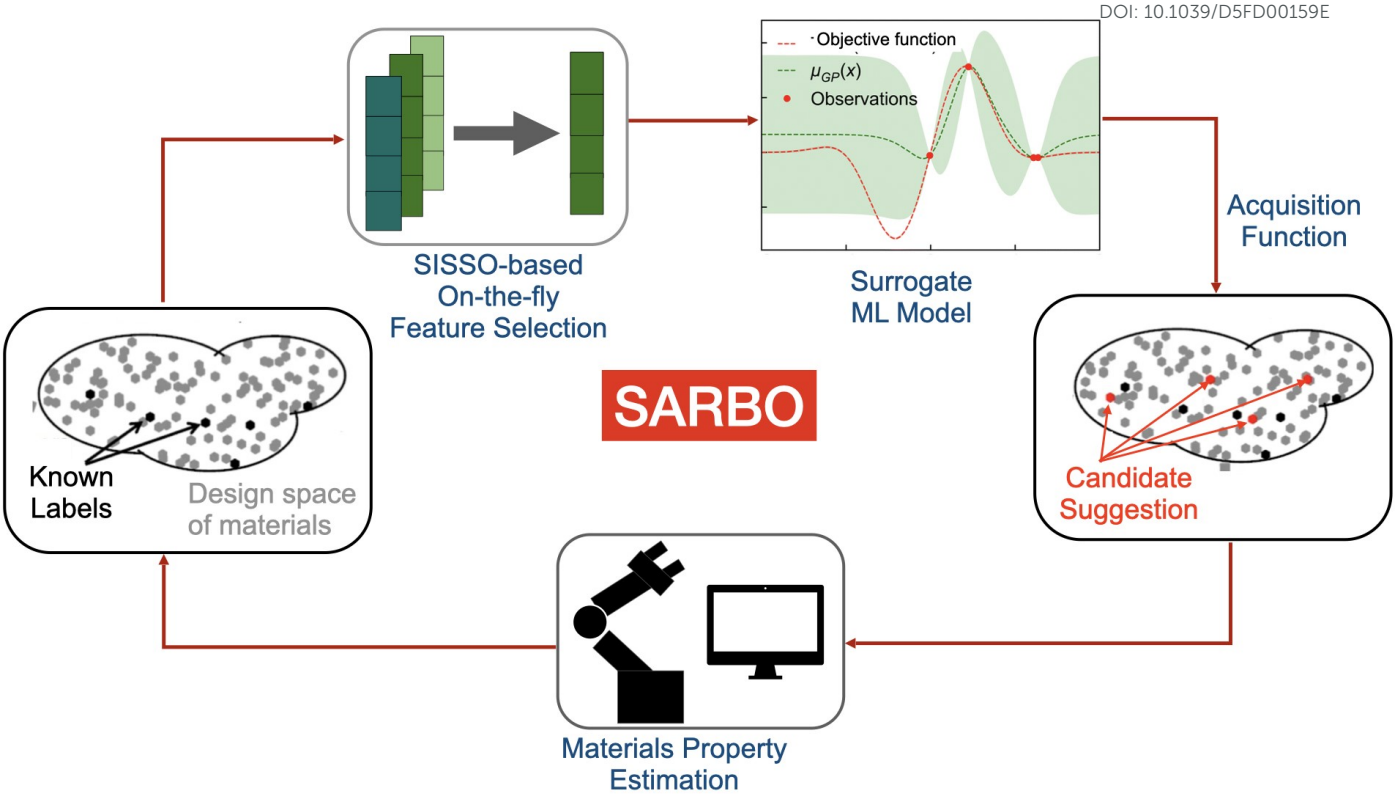


Fig. 1 Schematic of the SARBO framework. The iterative cycle begins with known labelled data, proceeds through SISSO-based on-the-fly feature selection to adaptively represent materials, training a surrogate ML model, candidate suggestion via acquisition function and property estimation until exhausting a pre-defined budget or achieving a desired optimization goal.

explore irrelevant portions of the design space. In order to mitigate this "curse of dimensionality" challenge in BO, several methods have been proposed where the surrogate model parameters can be tuned to focus on the features that matter most, ignoring irrelevant ones[16,17]. However, such methods (e.g., automatic relevance determination[16]) are typically tied to a specific surrogate model (e.g. GP) and therefore are not easily transferable to other model classes. *Feature-selection* (FS) methods, in turn, automatically identify a few relevant features, out of many offered, potentially relevant features. Then, the surrogate model uses only these most relevant features during BO[18]. However, most commonly used FS methods rely on linear correlation metrics and consequently fail to capture complex, non-linear feature interactions. Therefore, they may neglect features that appear redundant or weakly correlated with the target property when considered individually, but that might become relevant when they are combined with other features. In addition, the FS is typically performed based on the initial data set and this fixed set of features is utilized throughout the BO iterations. This might be an issue in materials science applications, since the relevance of features can change as new materials are acquired and included in the training data set. Indeed, the initial training set might not be representative of the entire design space. Thus, BO with a fixed feature set may overlook relevant materials. To address this issue, Mahyar et al. introduced adaptive FS within each BO iteration, demonstrating improved performance over traditional BO approaches[19]. However, this framework did not account for the non-linear interactions between features. As a result, even with adaptive representations, features which are potentially relevant to describe the target property might be excluded, causing FS methods to function as bad advisors, limiting the effectiveness of BO.

In this work, we present the Sparse Adaptive Representation–based Bayesian Optimization (SARBO) framework to address the curse of dimensionality faced by standard BO approaches (Fig.1). This framework utilizes the Sure-Independence Screening and Sparsifying Operator (SISSO) symbolic-regression method for on-the-fly selection of features, taking into account nonlinear target-feature relationships and interaction between features[20]. SISSO identifies, at each iteration of the BO, a small set of physico-chemical features correlated with the target property, also termed as *materials genes*[21]. Then, a surrogate model, e.g. GP, RF or the SISSO model itself, is trained only with the features selected by SISSO. This strategy refines the design space description on-the-fly. In addition, SISSO is an inherently interpretable method in the sense that it provides models as analytical expressions or *descriptors*. These expressions make the most relevant correlations between the features and the target property explicitly. In the prior work[22], we have demonstrated how SISSO can be directly used as a surrogate model for BO, combining both its feature selection, predictive ability and uncertainty estimates by considering ensembles of SISSO models. In the current study, we repurpose SISSO as a general FS method that can be utilized with any surrogate model. In particular, SISSO-based FS can be

Open Access Article. Published on 27 January 2026. Downloaded on 4/29/2026 3:49:16 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

Faraday Discussions Accepted Manuscript

View Article Online
DOI: 10.1039/D5FD00159F

Open Access Article. Published on 27 January 2026. Downloaded on 4/29/2026 3:49:16 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

combined with GP, taking advantage of the principled uncertainty estimates provided by this approach. We critically evaluate the performance of the SARBO approach against standard BO integrating alternative FS strategies, demonstrating its superior performance in identifying the surface sites of single- and dual-atom alloys for CO_2 activation, a critical step in CO_2 reduction reaction.

2 Methods

2.1 Dataset

The dataset considered in this study consists of 218 surface sites of single-atom alloys (SAAs) and 16 surface sites of dual-atom alloys (DAAs)[23]. In these materials, one (two) isolated atom(s) of a metal, termed *dopant(s)*, is (are) dispersed and stabilized on a different metal, termed *host*[24–30] (Fig. 2). SAAs and DAAs display unique electronic properties compared to monometallic systems or stoichiometric alloys[31–34] that can be used to design new catalysts. The SAAs and DAAs considered in this work were constructed using surface terminations (111), (100), (110), (211) and (0001) of Cu, Pd and Zn hosts with dopants corresponding to elements of periodic table groups 8–13. In addition, 6 monometallic sites of Pd(211) surface is also included in the dataset. The CO_2 chemisorption was evaluated for each of the surface sites using density functional theory (DFT) calculations with the mBEEF[35] meta-generalized gradient approximation for electron exchange and correlation as implemented in the FHI-aims code[36,37]. mBEEF has been shown to provide an accurate description of CO_2 binding on SAA surfaces by previous benchmarking studies[38,39]. The maximum C–O bond elongation in chemisorbed CO_2, denoted as $\Delta d_{\max}^{C-O}$, is taken as the target property. It is estimated as;

$$\Delta d_{\max}^{C-O} = d_{chem}^{C-O} - d_{eq}^{C-O} \quad (1)$$

where d_{chem}^{C-O} is the largest distance between C and O atoms in the chemisorbed CO_2 molecule, and d_{eq}^{C-O} is the C-O bond distance in the gas-phase CO_2 molecule. This bond elongation measures the ability of the surface site to activate CO_2, a prerequisite for subsequent steps in CO_2 reduction. The larger $\Delta d_{\max}^{C-O}$, weaker is the C-O bond and the more likely it is that the molecule can be converted[40,41]. Thus, the BO approach in this work aims at finding surface sites of SAAs and DAAs that maximize $\Delta d_{\max}^{C-O}$.

The 26 offered features can be classified in five different types: elemental properties of the (i) host and (ii) dopant such as *d*-orbital radius and electronegativity, (iii) properties related to the CO_2 adsorption such as its adsorption energy, (iv) elemental and geometrical properties of the surface sites such as averaged *d*-orbital radius and coordination number, and (v) properties of the surface sites plus nearest neighbors such as generalized coordination number[42] (see Table S1 for the full list of features). It should be noted that adsorption-related properties are included as features to model $\Delta d_{\max}^{C-O}$ because they are required to achieve accurate SISSO models. For example, including these features increased the training performance from an R^2 of 0.4 to 0.74. This approach, however, presupposes knowledge of adsorption properties that are only available after performing DFT calculations. This reliance could, in principle, be alleviated by increasing the model complexity or by adopting the hi-SISSO strategy[43], although we find that the latter only slightly improve the performance of the model in this case (see Fig. S1). Because our focus is on evaluating optimization performance using a dataset in which these properties are available for all surface sites rather than predicting $\Delta d_{\max}^{C-O}$ for previously unseen surface sites, we did not explore further alternatives to circumvent the use of the adsorption-related properties. For the case of DAAs, we have only considered bridge sites between the two dopant atoms for CO_2 adsorption and hence the elemental properties of the dopant are taken as the average of the properties of the two atoms. This dataset is particularly suitable for our study as $\Delta d_{\max}^{C-O}$ could be governed by multiple factors of various origins, such as dopant and host chemistry, surface structure and local coordination. Thus, the different SAAs and DAAs may require descriptions based on different features, making this a useful test case for adaptive feature selection.

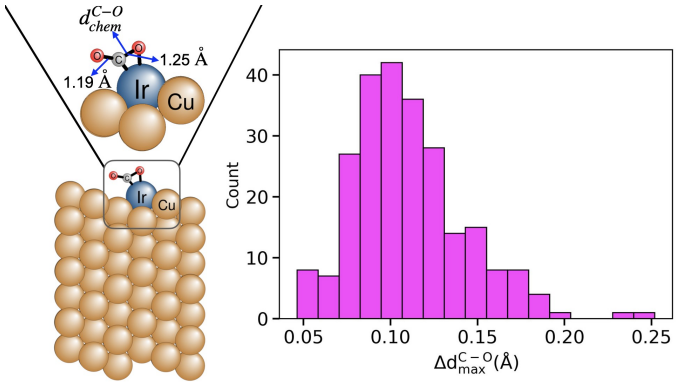


Fig. 2 (left) Representative atomic structure of the Ir@Cu(211) SAA surface. The local coordination environment of the single-atom is shown in an enlarged view. The target property $\Delta d_{\max}^{C-O}$ is evaluated as the difference between the largest distance between C and O atoms in chemisorbed CO_2 (indicated by d_{chem}^{C-O}) and the C-O bond length in the gas phase CO_2 molecule, (right) Distribution of $\Delta d_{\max}^{C-O}$ values across the dataset of 240 surface sites in the considered dataset.

2.2 Sparse Adaptive Representation-based Bayesian Optimization (SARBO)

The SARBO framework is designed for efficient and interpretable materials discovery in high-dimensional feature spaces. BO operates on two different pools of materials: (i) an initial labeled pool of (few) materials $\mathscr{D}_{\text{init}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{M}$ for which the target property is known and (ii) a large pool of unlabeled candidate materials $\mathscr{D}_{\text{design}} = \{\mathbf{x}_j\}_{j=1}^{P}$, for which the target property is not known. Each material $\mathbf{x}$ is represented by a set of initially offered features ϕ_{total}.

The closed-loop SARBO proceeds as follows:

1. **Descriptor Learning with SISSO:** The core component of SARBO utilizes SISSO for adaptive feature selection. During each BO iteration, SISSO processes the current $\mathscr{D}_{\text{init}}$ to construct an extensive space of candidate features Φ by applying a set of mathematical operations to ϕ_{total}. These operations are unary (log, exponential, square) and binary (addition, subtraction, muliplication). Thus, they capture

Faraday Discussions Accepted Manuscript

View Article Online
DOI: 10.1039/D5FD00159F

Table 1 Performance comparison of feature selection (FS) methods using Gaussian Process (GP), Random Forest (RF), and SISSO. Considered FS methods include recursive feature elimination (RFE), mutual information (MI), minimum redundancy maximum relevance (MRMR), and SISSO-based feature selection. Metrics reported are the coefficient of determination (R^2), mean absolute error (MAE), and root mean squared error (RMSE). For each method, absolute values are shown, and values in parentheses indicate ratios relative to the baseline (NoFS) using the complete set of features without any selection. For SISSO, the labels "SM" and "FS" indicate the scenarios when it is used directly as the surrogate model and only for feature selection, respectively.

FS Method	R^2			MAE (Å)			RMSE (Å)		
	GP	RF	SISSO-SM	GP	RF	SISSO-SM	GP	RF	SISSO-SM
NoFS	0.763	0.702	0.753	0.012	0.013	0.013	0.017	0.019	0.017
SISSO-FS	0.844 (1.106)	0.719 (1.024)	0.753 (1.0)	0.010 (0.835)	0.013 (0.970)	0.013 (1.0)	0.014 (0.812)	0.019 (0.971)	0.017 (1.0)
MI	0.644 (0.844)	0.646 (0.920)	0.695 (0.923)	0.017 (1.411)	0.015 (1.111)	0.015 (1.141)	0.021 (1.225)	0.021 (1.091)	0.019 (1.111)
RFE	0.650 (0.852)	0.655 (0.933)	0.648 (0.861)	0.016 (1.395)	0.015 (1.087)	0.017 (1.321)	0.021 (1.215)	0.021 (1.077)	0.021 (1.193)
MRMR	0.671 (0.880)	0.597 (0.850)	0.648 (0.861)	0.015 (1.269)	0.016 (1.157)	0.017 (1.321)	0.020 (1.177)	0.022 (1.164)	0.021 (1.193)

nonlinearities and interactions between features. Then, the sure-independence screening procedure followed by ℓ_0-regularized regression is applied to select a model of the form:

$$P(\mathbf{x}) = c_0 + \sum_{k=1}^{K} c_k d_k(\phi_{\text{relevant}})$$

where $\{d_k\}$ represent the selected expressions. These expressions are components of a descriptor vector, and they depend on few, relevant features ϕ_{relevant}, from all initially offered features. K is the descriptor dimensionality and c_k are fitted coefficients. We note that even though the model above is linear, the nonlinearities are included in the expressions generated by SISSO. Also, the term "features" in this work refers to what is called "primary features" in the SISSO approach[44]. The immense space of "SISSO-created features" (billions), i.e. candidate descriptors, only plays a hidden role in the current work.

2. **Adaptive Representation:** The SISSO model is used to redefine the design space representation for subsequent BO iterations. The materials are represented using only the set of features $\phi_{\text{relevant}} \subset \phi_{\text{total}}$ that appear in the SISSO model. This creates a minimal, relevant subset of the original set of features.

3. **Surrogate Modeling:** A surrogate model $f(\mathbf{x})$ is trained on $\mathscr{D}_{\text{init}}$ represented only by ϕ_{relevant} to approximate the objective function $y(\mathbf{x})$. The model provides a mean prediction $\mu(\mathbf{x})$ and an uncertainty estimate $\sigma(\mathbf{x})$ for all $\mathbf{x} \in \mathscr{D}_{\text{design}}$. This work demonstrates SARBO using GP and RF as surrogates where $\sigma(\mathbf{x})$ is estimated as the standard deviation of the posterior distribution and predictions across ensemble trees, respectively.

4. **Acquisition and Evaluation:** We employ Expected improvement (EI)[45] and upper confidence bound (UCB)[46] as acquisition functions as they balance exploitation and exploration for BO. For both acquisition functions, the next candidate(s) are selected by maximizing:

$$\mathbf{x}_{\text{next}} = \arg\max_{\mathbf{x} \in \mathscr{D}_{\text{design}}} \alpha(\mathbf{x}; \mu, \sigma),$$

where $\alpha(\mathbf{x})$ represents the either EI or UCB scores. The target property y_{next} for $\mathbf{x}_{\text{next}}$ is evaluated , and the dataset is updated: $\mathscr{D}_{\text{init}} \leftarrow \mathscr{D}_{\text{init}} \cup \{(\mathbf{x}_{\text{next}}, y_{\text{next}})\}$.

3 Results and Discussions

3.1 Comparison of Feature Selection Performance

We first compared the performance of SISSO-based feature selection (FS) with three other FS methods: recursive feature elimination (RFE)[47], mutual information (MI)[48], and minimum redundancy maximum relevance (MRMR)[49]. RFE iteratively prunes features based on their importance scores obtained from a RF model. MI retains features with the highest mutual information with the target property. Finally, MRMR selects features that maximize relevance to the target property while minimizing redundancy among them. RFE and MI are performed with `scikit-learn`[50]. MRMR is performed with the `feature-engine` package[51]. We trained GP and RF models for the target $\Delta d_{\text{max}}^{\text{C-O}}$ by splitting the dataset into 80% train and 20% test data. The test-set performance of the models obtained with the features selected by the methods SISSO, RFE, MI, and MRMR is shown in Table 1 along with a baseline (labelled NoFS) corresponding to using the entire set of features.

Let us first examine the performance of the GP models across the different FS methods. GP trained on SISSO selected features (labelled SISSO-FS) show the best performance among all methods, achieving the highest R^2 (0.844) and the lowest MAE (0.01 Å) and RMSE (0.014 Å) values. This represents an improvement over the full-feature baseline. In contrast, the MI, RFE, and MRMR approaches all lead to reduced GP performance, with consistently higher errors. A similar trend is observed for the RF models: SISSO-based FS again provides the highest performance (R^2 = 0.719, MAE = 0.013 Å, RMSE = 0.019 Å). However, the relative gains are more modest than for GP, reflecting the robustness of RF to irrelevant or redundant features via ensemble averaging. The remaining FS methods once again underperform relative to the full-feature baseline, exhibiting lower R^2 and higher MAE, RMSE values. When SISSO is used as a surrogate model (labelled SISSO-SM) with its default FS, it is observed to have an intermediate performance between GP and RF. When combined with other FS methods, SISSO-SM generally performs better than RF for MI but is slightly worse than GP, and for RFE and MRMR, it tends to underperform compared to GP while being comparable to or slightly worse than RF. Table S2 shows the features selected by each FS method. While an overlap is observed for

Open Access Article. Published on 27 January 2026. Downloaded on 4/29/2026 3:49:16 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

Faraday Discussions Accepted Manuscript

View Article Online
DOI: 10.1039/D5FD00159E

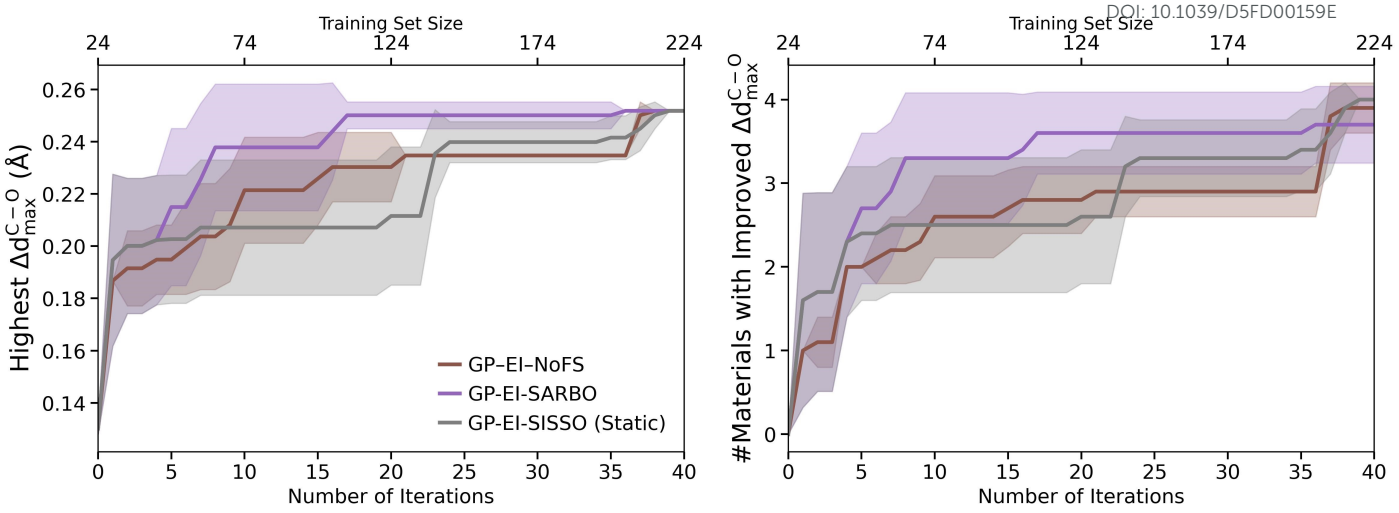


Fig. 3 Performance comparison of SARBO against baseline Bayesian Optimization (BO) methods: BO with no feature selection (NoFS) and BO with a static, one-time SISSO feature selection (SISSO-Static). (left) Evolution of the maximum Δd_{max}^{C-O} value over 40 iterations and (right) cumulative count of materials improving the target property. Training set sizes at selected iterations are shown on the top x-axis of both panels. Shaded regions indicate standard deviation across 10 independent trials.

the selected features, only SISSO-FS uniquely captures effects associated with the host surface, single-atom dopant, surface site and the coordination environment, which are likely critical for describing CO_2 activation. In contrast, the other feature selection methods systematically overlook at least one of these key factors. The enhanced performance observed with SISSO-FS compared to using the full set of features can be attributed to the reduction of redundancy and irrelevant information present in the full feature set. By relying on the most informative features, surrogate models can more effectively capture the underlying relationships, leading to improved prediction accuracy and generalization.

Overall, these results highlight SISSO as the most effective FS strategy among the considered methods, and GP as the surrogate model achieving the highest predictive performance, particularly when combined with SISSO-selected features. This suggests that even when using non-SISSO surrogate models such as GP or RF, employing SISSO solely for feature selection can provide substantial gains, as it identifies the most informative features while leveraging the flexibility of other surrogate models to learn complex feature–target relationships. Hence, in the following discussion, we focus on GP and RF surrogate models.

3.2 Performance of SARBO Framework

We evaluate SARBO's performance through *after-the-fact* simulated BO campaigns. After-the-fact means that the target property is not evaluated for a selected candidate surface site via DFT calculations, but it is taken from the complete dataset. Indeed, a small fraction (f_{train}) of the total dataset is used as the initial data ($\mathscr{D}_{\text{init}}$) for training, while the remaining data constitute a hidden candidate pool ($\mathscr{D}_{\text{design}}$). This approach mimics real-world materials discovery scenarios where the optimal candidates are unknown *a priori* and has been adapted for comparing and benchmarking various BO methodologies in previous studies[52–54]. The optimization is performed for N_{iter} iterations. At each iteration, the surrogate models ($\mathscr{M} \in \{\text{GP}, \text{RF}\}$) are trained. They provide predictions for the targets Δd_{max}^{C-O}, along with uncertainty estimates $\sigma(\mathbf{x})$. The acquisition functions ($\alpha \in \{\text{EI}, \text{UCB}\}$) then select a batch ($b$) of candidates $\mathbf{x}_{\text{next}} = \arg\max \alpha(\mathbf{x})$, whose values of the target Δd_{max}^{C-O} are added to the training dataset. Finally, the FS method selects the relevant features for the next iteration. Since the optimization objective targets maximization of Δd_{max}^{C-O}, we reserve a fraction (f_{top}) of top candidates (highest Δd_{max}^{C-O} value) in $\mathscr{D}_{\text{design}}$, excluding them from $\mathscr{D}_{\text{init}}$. We monitor the performance of the BO using two metrics: the highest observed Δd_{max}^{C-O} value at each iteration and the cumulative count of surface sites improving the target property, i.e., with Δd_{max}^{C-O} values larger than the best known value in the previous iteration. Each BO campaign is performed N_{trial} times with different randomly selected initial training sets. Thus, confidence intervals are obtained for the BO campaigns. For all the results presented in the manuscript, we consider f_{train}=0.1, f_{top}=0.2, b=5, N_{iter}=40 and N_{trial}=10, unless stated otherwise.

We begin by comparing the performance of SARBO against standard BO under two distinct scenarios. The first is a baseline where BO operates on the full feature set without any selection (*NoFS*). The second, labelled *Static*, employs a SISSO-derived feature set determined once at the outset and held fixed throughout all BO iterations. This static condition represents a data-rich ideal, where sufficient data exists *a priori* for feature selection, a scenario uncommon in typical BO experiments but plausible when leveraging pre-existing data from low-fidelity simulations or a closely related property.

Fig. 3 shows the results for the GP surrogate model and the EI acquisition function. The evolution of the highest Δd_{max}^{C-O} values identified by BO with the number of iterations is shown in the left panel. The highest Δd_{max}^{C-O} value around 0.25 Å is identified at N_{iter}=17 by SARBO. Conversely, N_{iter}=37 is needed for the standard BO approach utilizing the full set of features. While

Open Access Article. Published on 27 January 2026. Downloaded on 4/29/2026 3:49:16 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

Faraday Discussions Accepted Manuscript

View Article Online
DOI: 10.1039/D5FD00159E

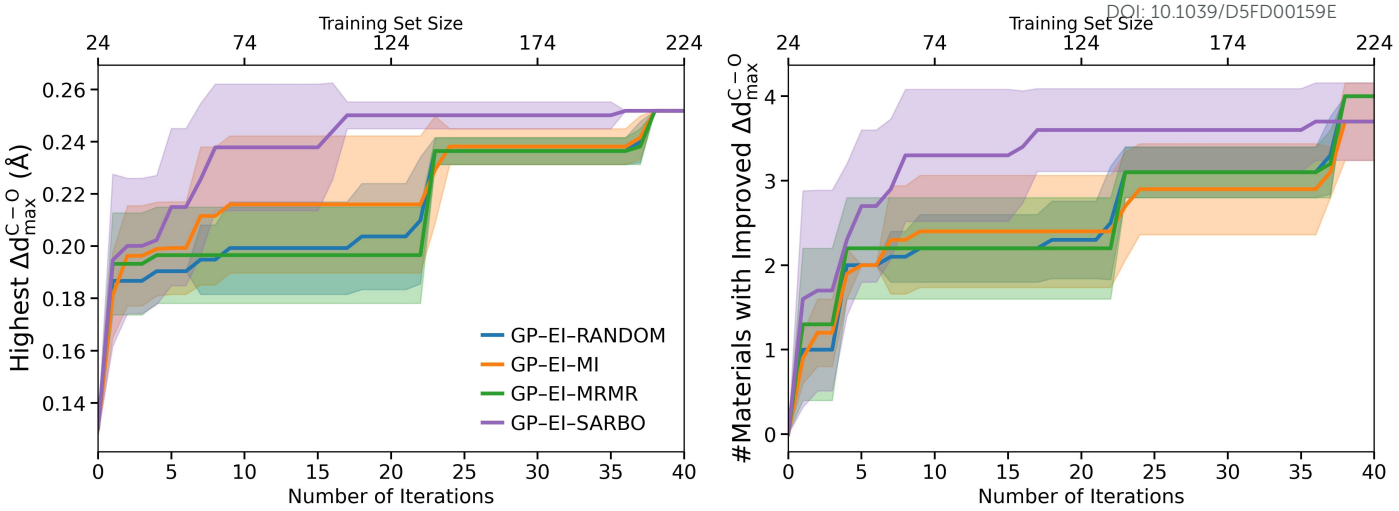


Fig. 4 Performance comparison between BO with different feature selection strategies : (left) maximum $\Delta d_{\max}^{\mathrm{C-O}}$ evolution, (right) cumulative improvements in target property. Shaded regions indicate standard deviation across 10 independent trials.

the static feature selection method shows a marginal performance improvement over the standard BO in the initial and final stages, it still required N_{iter}=38 to achieve the highest target value. We also analyzed the evolution of the number of surface sites with improved $\Delta d_{\max}^{\mathrm{C-O}}$ identified by BO with the number of iterations (Fig. 3, right panel). SARBO exhibits a small number of early, high-impact improvements and quickly reaches the vicinity of the global optimum, after which few further improvements are possible. In contrast, both standard and static feature set-based BO accumulate many incremental improvements over a larger number of iterations, reflecting slower progress through lower-quality regions of the design space. Overall, these results demonstrate that SARBO achieves substantially faster convergence toward the optimal region of the design space, indicating that the adaptively selected features from SISSO models provide a more informative and efficient representation of the design space for the surrogate model. Although an initial, one-time feature selection can improve performance, it is ultimately outperformed by SARBO. This static approach also carries the prerequisite of a substantial initial dataset for the feature selection itself. These results collectively underscore the critical advantage of SARBO's adaptive feature adaptation.

We now compare the performance of various FS methods when integrated into an adaptive BO loop like the one used in SARBO, for the case of the GP surrogate model with the EI acquisition function. Fig. 4 compares SARBO with BO approaches that employ MI and MRMR. Additionally, we also consider a random-selection baseline (RANDOM) where features are selected randomly without any measurement of feature importance. For consistency, all methods use the same number of features selected by SARBO. It can be seen that SARBO outperforms the other FS strategies in identifying surface sites with the highest $\Delta d_{\max}^{\mathrm{C-O}}$. Indeed, SARBO reaches the maximum mean value around 0.25 Å by iteration $N_{\mathrm{iter}} = 17$, whereas the other methods require $N_{\mathrm{iter}} = 38$ to achieve the same performance. This advantage is already evident at early iterations: at $N_{\mathrm{iter}} = 10$, the highest $\Delta d_{\max}^{\mathrm{C-O}}$ values are 0.20 (RANDOM), 0.21 (MI), 0.19 (MRMR), and 0.24 Å(SARBO). A similar trend appears in the cumulative number of improved candidates. SARBO identifies improved candidates earlier and with larger mean values relative to the other methods. Although MI and MRMR show slight improvements over SARBO at later iterations, these gains occur only after substantially more evaluations, indicating that they require greater sampling effort for marginal benefit. These results are consistent with the ability of SISSO to explore large spaces of possibly nonlinear expressions, thus incorporating important nonlinear interactions when selecting relevant features.

To assess the generality of the performance trends observed with the GP-EI setup, we have also performed after-the-fact BO runs using RF as the surrogate model and UCB as the acquisition function (Fig. S2). Compared to GP-EI, it is observed that the RF-UCB-based BO campaigns show slightly faster convergence in early iterations. The high $\Delta d_{\max}^{\mathrm{C-O}}$ values are identified at $N_{\mathrm{iter}} = 7$ instead of 17 for SARBO. Nevertheless, the performance of BO with different FS methods remains the same: SARBO > MI > MRMR > RANDOM. These results suggest that SARBO's adaptive feature selection provides benefits for BO experiments involving alternate choices surrogate models and acquisition functions.

To understand the origin of the difference in BO performance of different FS methods, we investigated which types of features were selected by each FS method. The five types of features contained in the dataset are host (8 features), dopant (7 features), CO_2 adsorption (2 features), surface site (4 features), and local coordination environment (5 features) - see Methods section for details. We computed the total frequency of selection for each feature category per trial, summing over all 40 iterations shown in Fig. 4. The box plots in Fig. 5 show the distribution of these aggregated counts from the 10 trials. Distinct selection patterns emerge across different FS Methods. The random feature selection exhibits the expected uniform distribution across categories, serving as an unbiased baseline. SARBO demonstrates a strong preference towards host- and adsorption-related features,

Open Access Article. Published on 27 January 2026. Downloaded on 4/29/2026 3:49:16 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

Faraday Discussions Accepted Manuscript

View Article Online
DOI: 10.1039/D5FD00159E

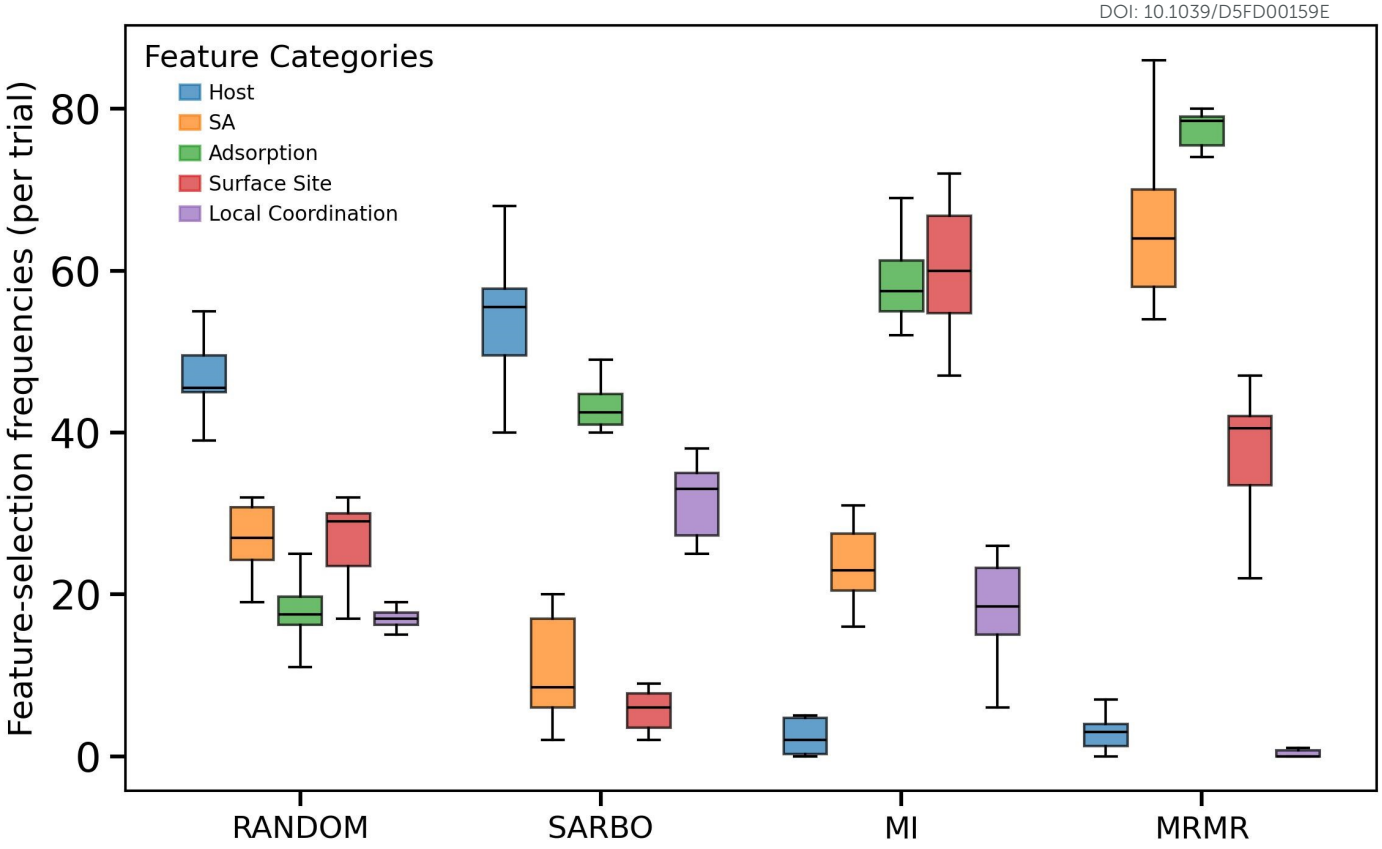


Fig. 5 Distribution of feature category selection frequencies across 10 BO trials using GP-EI. For each category, boxplots display the distribution of cumulative selection counts per trial, with the central line indicating the median, box boundaries representing the interquartile range and whiskers showing the full data range.

followed by local-coordination related ones. MI primarily selects surface sites relateded features, whereas MRMR preferes SA-related features. All methods select adsorption-related features to some extent, consistent with their direct correlation to the target property $\Delta d_{\max}^{\mathrm{C-O}}$. The distinctive feature selection pattern of SARBO suggests it identifies alternative physical relationships that may capture different aspects of the catalytic behavior. This unique feature selection pattern enables SARBO to effectively guide the optimization toward promising regions of the design space, as evidenced by its superior performance in identifying CO_2-activating sites.

Conclusions

This work introduces the Sparse Adaptive Representation-based Bayesian Optimization (SARBO) framework, which applies the SISSO method for efficient feature selection. Crucially, the feature selection is performed at each iteration of the optimization. SARBO addresses the challenge of high dimensional feature spaces in BO, and it thus improves the efficiency of BO when the few, key features required to describe the materials property or function are not known beforehand. We demonstrate the superior optimization performance of SARBO compared to standard BO and popular feature selection methods for the identification of surface sites of single and dual-atom alloys capable of activating the CO_2 molecule. We ascribe its superior performance to the ability to capture non-linear relationships and interaction between features that are often critical in governing target material properties but overlooked by widely used feature selection methods. Indeed, the features selected by SISSO take into account the different factors that contribute to the CO_2 activation, such as the chemistry and surface structure of host, adsorption energetics and the effect of local coordination of the single-atom on the host surface. Furthermore, SARBO introduces a vital layer of interpretability to the typically black-box BO process. We anticipate that this work will enhance closed-loop frameworks by enabling more efficient exploration of vast design spaces, accelerating materials discovery efforts both for computational studies and for autonomous experimentation.

Author contributions

Conceptualization, Methodology, Software: A.S.N; Writing: A.S.N., L.F.; Supervision: M.S.

Conflicts of interest

The authors declare no conflicts of interest.

Data availability

The code base of SARBO, along with all the datasets and supporting scripts, is available at the GitLab repository `https://gitlab.com/akhilsnair/SARBO`.

Open Access Article. Published on 27 January 2026. Downloaded on 4/29/2026 3:49:16 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

Faraday Discussions Accepted Manuscript

View Article Online
DOI: 10.1039/D5FD00159F

Open Access Article. Published on 27 January 2026. Downloaded on 4/29/2026 3:49:16 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

Faraday Discussions Accepted Manuscript

Acknowledgements

This work was funded by the ERC Advanced Grant TEC1p (European Research Council, Grant Agreement No. 740233). A.S.N. acknowledges Herzain I. Rivera-Arrieta for providing the SAA dataset. A.S.N. thanks Beate Paulus and the German Research Foundation (DFG) for their support through the Walter-Benjamin Fellowship Program (project No. 540316537).

References

1 C. Liu, F. Li, L.-P. Ma and H.-M. Cheng, *Advanced materials*, 2010, **22**, E28–E62.

2 J. Kibsgaard and I. Chorkendorff, *Nature Energy*, 2019, **4**, 430–433.

3 N. J. Szymanski, B. Rendy, Y. Fei, R. E. Kumar, T. He, D. Milsted, M. J. McDermott, M. Gallant, E. D. Cubuk, A. Merchant *et al.*, *Nature*, 2023, **624**, 86–91.

4 F. Strieth-Kalthoff, H. Hao, V. Rathore, J. Derasp, T. Gaudin, N. H. Angello, M. Seifrid, E. Trushina, M. Guy, J. Liu *et al.*, *Science*, 2024, **384**, eadk9227.

5 K. Terayama, M. Sumita, R. Tamura and K. Tsuda, *Accounts of Chemical Research*, 2021, **54**, 1334–1346.

6 P. Honarmandi, V. Attari and R. Arroyave, *Computational Materials Science*, 2022, **210**, 111417.

7 P. I. Frazier, *Recent advances in optimization and modeling of contemporary problems*, Informs, 2018, pp. 255–278.

8 C. Williams and C. Rasmussen, *Advances in neural information processing systems*, 1995, **8**, N/A.

9 M.-H. Roy and D. Larocque, *Journal of Nonparametric Statistics*, 2012, **24**, 993–1006.

10 A. G. Kusne, H. Yu, C. Wu, H. Zhang, J. Hattrick-Simpers, B. DeCost, S. Sarker, C. Oses, C. Toher, S. Curtarolo *et al.*, *Nature communications*, 2020, **11**, 5966.

11 A. Deshwal, C. M. Simon and J. R. Doppa, *Molecular Systems Design & Engineering*, 2021, **6**, 1066–1086.

12 J. Peng, J. K. Damewood, J. Karaguesian, R. Gómez-Bombarelli and Y. Shao-Horn, *Joule*, 2021, **5**, 3069–3071.

13 J. K. Pedersen, C. M. Clausen, O. A. Krysiak, B. Xiao, T. A. Batchelor, T. Löffler, V. A. Mints, L. Banko, M. Arenz, A. Savan *et al.*, *Angewandte Chemie*, 2021, **133**, 24346–24354.

14 K. J. Jenewein, L. Torresi, N. Haghmoradi, A. Kormányos, P. Friederich and S. Cherevko, *Journal of Materials Chemistry A*, 2024, **12**, 3072–3083.

15 B. Shahriari, K. Swersky, Z. Wang, R. P. Adams and N. De Freitas, *Proceedings of the IEEE*, 2015, **104**, 148–175.

16 M. Mørup and L. K. Hansen, *Journal of Chemometrics: A Journal of the Chemometrics Society*, 2009, **23**, 352–363.

17 Q. Liang, A. E. Gongora, Z. Ren, A. Tiihonen, Z. Liu, S. Sun, J. R. Deneault, D. Bash, F. Mekki-Berrada, S. A. Khan *et al.*, *npj Computational Materials*, 2021, **7**, 188.

18 L. Long, C. Cartis and P. F. Shustin, *arXiv preprint arXiv:2412.09183*, 2024.

19 M. Rajabi-Kochi, N. Mahboubi, A. P. S. Gill and S. M. Moosavi, *Chemical Science*, 2025, **16**, 5464–5474.

20 R. Ouyang, S. Curtarolo, E. Ahmetcik, M. Scheffler and L. M. Ghiringhelli, *Physical Review Materials*, 2018, **2**, 083802.

21 L. Foppa, L. M. Ghiringhelli, F. Girgsdies, M. Hashagen, P. Kube, M. Hävecker, S. J. Carey, A. Tarasov, P. Kraus, F. Rosowski *et al.*, *MRS bulletin*, 2021, **46**, 1016–1026.

22 A. S. Nair, L. Foppa and M. Scheffler, *npj Computational Materials*, 2025, **11**, 1–7.

23 H. I. Rivera-Arrieta and L. Foppa, *ACS catalysis*, 2025, **15**, 2916–2926.

24 G. Di Liberto and G. Pacchioni, *Advanced Materials*, 2023, **35**, 2307150.

25 G. Di Liberto, L. Giordano and G. Pacchioni, *ACS Catalysis*, 2024, **14**, 45–55.

26 R. T. Hannagan, G. Giannakakis, M. Flytzani-Stephanopoulos and E. C. H. Sykes, *Chemical Reviews*, 2020, **120**, 12044–12088.

27 R. T. Hannagan, G. Giannakakis, R. Réocreux, J. Schumann, J. Finzel, Y. Wang, A. Michaelides, P. Deshlahra, P. Christopher, M. Flytzani-Stephanopoulos, M. Stamatakis and E. C. H. Sykes, *Science*, 2021, **372**, 1444–1447.

28 J. Schumann, Y. Bao, R. T. Hannagan, E. C. H. Sykes, M. Stamatakis and A. Michaelides, *The Journal of Physical Chemistry Letters*, 2021, **12**, 10060–10067.

29 Z.-K. Han, D. Sarker, R. Ouyang, A. Mazheika, Y. Gao and S. V. Levchenko, *Nature Communications*, 2021, **12**, 1833.

30 M. Andersen, S. V. Levchenko, M. Scheffler and K. Reuter, *ACS Catalysis*, 2019, **9**, 2752–2759.

31 M. T. Darby, R. Réocreux, E. C. H. Sykes, A. Michaelides and M. Stamatakis, *ACS Catalysis*, 2018, **8**, 5038–5050.

32 R. Réocreux and M. Stamatakis, *Accounts of Chemical Research*, 2021, **55**, 87–97.

33 W. Li, S. E. Madan, R. Réocreux and M. Stamatakis, *ACS Catalysis*, 2023, **13**, 15851–15868.

34 Z. Chen and P. Zhang, *ACS omega*, 2022, **7**, 1585–1594.

35 J. Wellendorff, K. T. Lundgaard, K. W. Jacobsen and T. Bligaard, *The Journal of chemical physics*, 2014, **140**, N/A.

36 V. Blum, R. Gehrke, F. Hanke, P. Havu, V. Havu, X. Ren, K. Reuter and M. Scheffler, *Computer Physics Communications*, 2009, **180**, 2175–2196.

37 J. W. Abbott, C. M. Acosta, A. Akkoush, A. Ambrosetti, V. Atalla, A. Bagrets, J. Behler, D. Berger, B. Bieniek, J. Björk *et al.*, *arXiv preprint arXiv:2505.00125*, 2025.

38 L. Kabalan, I. Kowalec, C. R. A. Catlow and A. J. Logsdail, *Physical Chemistry Chemical Physics*, 2021, **23**, 14649–14661.

39 P. S. Schmidt and K. S. Thygesen, *The Journal of Physical Chemistry C*, 2018, **122**, 4381–4390.

40 H.-J. Freund and M. W. Roberts, *Surface Science Reports*, 1996, **25**, 225–273.

41 A. Mazheika, Y.-G. Wang, R. Valero, F. Viñes, F. Illas, L. M. Ghiringhelli, S. V. Levchenko and M. Scheffler, *Nature Communications*, 2022, **13**, 419.

42 F. Calle-Vallejo, *Advanced Science*, 2023, **10**, 2207644.

43 L. Foppa, T. A. Purcell, S. V. Levchenko, M. Scheffler and L. M. Ghiringhelli, *Physical Review Letters*, 2022, **129**, 055301.

44 T. A. Young, T. Johnston-Wood, V. L. Deringer and F. Duarte,

View Article Online
DOI: 10.1039/D5FD00159E

Chemical science, 2021, **12**, 10944–10955.

45 J. Mockus, *Towards global optimization*, 1998, **2**, 117.

46 A. I. Forrester and A. J. Keane, *Progress in aerospace sciences*, 2009, **45**, 50–79.

47 X.-w. Chen and J. C. Jeong, Sixth international conference on machine learning and applications (ICMLA 2007), 2007, pp. 429–435.

48 A. Kraskov, H. Stögbauer and P. Grassberger, *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 2004, **69**, 066138.

49 H. Peng, F. Long and C. Ding, *IEEE Transactions on pattern analysis and machine intelligence*, 2005, **27**, 1226–1238.

50 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, *the Journal of machine Learning research*, 2011, **12**, 2825–2830.

51 S. Galli, *Journal of Open Source Software*, 2021, **6**, 3642.

52 J. H. Montoya, K. T. Winther, R. A. Flores, T. Bligaard, J. S. Hummelshøj and M. Aykol, *Chemical Science*, 2020, **11**, 8517–8532.

53 C. K. Borg, E. S. Muckley, C. Nyby, J. E. Saal, L. Ward, A. Mehta and B. Meredig, *Digital Discovery*, 2023, **2**, 327–338.

54 L. Kavalsky, V. I. Hegde, E. Muckley, M. S. Johnson, B. Meredig and V. Viswanathan, *Digital Discovery*, 2023, **2**, 1112–1125.

Open Access Article. Published on 27 January 2026. Downloaded on 4/29/2026 3:49:16 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

Faraday Discussions Accepted Manuscript

View Article Online
DOI: 10.1039/D5FD00159E

Data Availability Statement

The code base of SARBO, along with all the datasets and supporting scripts, is available at Gitlab repository https://gitlab.com/akhilsnair/SARBO.

Open Access Article. Published on 27 January 2026. Downloaded on 4/29/2026 3:49:16 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

Faraday Discussions Accepted Manuscript