Check for updates

Adaptive subspace Bayesian optimization over molecular descriptor libraries for data-efficient chemical design

Farshud Sorourifar, [ID] [a] Thomas Banker[ab] and Joel A. Paulson [ID] *[a]

The discovery of molecules with optimal functional properties is a central challenge across diverse fields such as energy storage, catalysis, and chemical sensing. However, molecular property optimization (MPO) remains difficult due to the combinatorial size of chemical space and the cost of acquiring property labels *via* simulations or wet-lab experiments. Bayesian optimization (BO) offers a principled framework for sample-efficient discovery in such settings, but its effectiveness depends critically on the quality of the molecular representation used to train the underlying probabilistic surrogate model. Existing approaches based on fingerprints, graphs, SMILES strings, or learned embeddings often struggle in low-data regimes due to high dimensionality or poorly structured latent spaces. Here, we introduce Molecular Descriptors with Actively Identified Subspaces (MolDAIS), a flexible molecular BO framework that adaptively identifies task-relevant subspaces within large descriptor libraries. Leveraging the sparse axis-aligned subspace (SAAS) prior introduced in recent BO literature, MolDAIS constructs parsimonious Gaussian process surrogate models that focus on task-relevant features as new data is acquired. In addition to validating this approach for descriptor-based MPO, we introduce two novel screening variants, which significantly reduce computational cost while preserving predictive accuracy and physical interpretability. We demonstrate that MolDAIS consistently outperforms state-of-the-art MPO methods across a suite of benchmark and real-world tasks, including single- and multi-objective optimization. Our results show that MolDAIS can identify near-optimal candidates from chemical libraries with over 100 000 molecules using fewer than 100 property evaluations, highlighting its promise as a practical tool for data-scarce molecular discovery.

1 Introduction

The discovery and design of molecules with tailored properties is essential not only to scientific inquiry but also to advancing engineering applications across sectors such as energy storage,[1–5] pharmaceuticals,[6–9] catalysis,[10–13] and soft electronics.[14–18] For instance, designing high-performance organic molecules has enabled more sustainable redox-active materials for aqueous batteries,[19] tunable ligands for selective catalysis,[20] and novel organic semiconductors for neuromorphic devices.[21] Realizing these types of breakthroughs hinges on either explicitly or implicitly solving a molecular property optimization (MPO) problem, which requires efficient identification of candidates with optimal properties from a vast, high-dimensional chemical design space.

Machine learning (ML) offers a promising strategy for accelerating MPO, especially in low-data regimes where

molecular simulations and/or wet-lab experiments are costly.[22,23] However, progress is often limited by three key challenges: (i) representing molecules in a form amenable to predictive modeling and optimization, (ii) building uncertainty-aware surrogate models from limited labeled data, and (iii) reasoning over high-dimensional representations where only a small subset of features may influence the target property.

A wide range of molecular representations have been proposed, including SMILES[24] and SELFIES[25] strings, molecular graphs,[26] fingerprints,[27] and descriptor-based feature vectors.[28,29] While these encodings capture varying levels of structural, electronic, or topological information, they are often high-dimensional and not guaranteed to align with the underlying property function landscape – especially when training data is limited and model overfitting becomes a concern. Recent efforts to address these issues have focused on data-driven embedding methods such as variational autoencoders (VAEs),[30,31] normalizing flows,[32] and deep kernel learning (DKL).[33–35] These models learn continuous molecular embeddings that enable the use of Bayesian optimization (BO)[36–38] for molecular design. However, they present practical limitations: training can be brittle and sensitive to hyperparameters; the

[a]The Ohio State University, Department of Chemical and Biomolecular Engineering, Columbus, OH, 43210, USA. E-mail: paulson.82@osu.edu

[b]University of California, Berkeley, Department of Chemical and Biomolecular Engineering, Berkeley, CA, 94720, USA

latent space may not reflect smooth changes in the property of interest; and the learned representation is often fixed and, thus, unable to adapt as new data becomes available.

An alternative to learning an embedding space is to apply BO directly on fixed molecular representations using specialized similarity kernels. For example, recent work has employed Gaussian processes (GPs) with Tanimoto fingerprint kernels or graph kernels for molecular structures.[39] While these methods avoid the need for learning a separate (typically highly parametrized) encoder, they assume that molecules deemed similar by the kernel will exhibit similar properties – an assumption that may not hold when the kernel structure is not adapted to the target task.

In this work, we present the Molecular Descriptors and Actively Identified Subspaces (MolDAIS) framework, illustrated in Fig. 1, which enables efficient and interpretable molecular property optimization using descriptor-based representations. MolDAIS builds upon the recently proposed Sparse Axis-Aligned Subspace BO (SAASBO) method,[40] adapting it to operate over large, chemically informed descriptor libraries. Rather than learning a new molecular embedding, MolDAIS leverages precomputed descriptors and performs adaptive feature selection using sparsity-inducing techniques. This allows the surrogate model to automatically and adaptively identify low-dimensional, property-relevant subspaces during the course of optimization.

The core variant of MolDAIS uses the SAAS prior within a fully Bayesian GP model to induce axis-aligned sparsity in the input space. Importantly, we also introduce two new (more scalable) screening alternatives based on mutual information (MI) and the maximal information coefficient (MIC), which provide runtime advantages while retaining interpretability and adaptivity. These screening variants represent a novel contribution beyond existing SAASBO implementations and offer a practical alternative for subspace selection when full Bayesian inference (using, *e.g.*, Hamiltonian Monte Carlo) is computationally prohibitive.

We evaluate MolDAIS on a wide range of benchmark and real-world MPO problems, including single- and multi-objective optimization tasks. Across these settings, MolDAIS consistently outperforms state-of-the-art baselines based on molecular graphs, SMILES strings, and latent embeddings, particularly when the number of available evaluations is fewer than 100. In addition to its strong empirical performance, MolDAIS offers
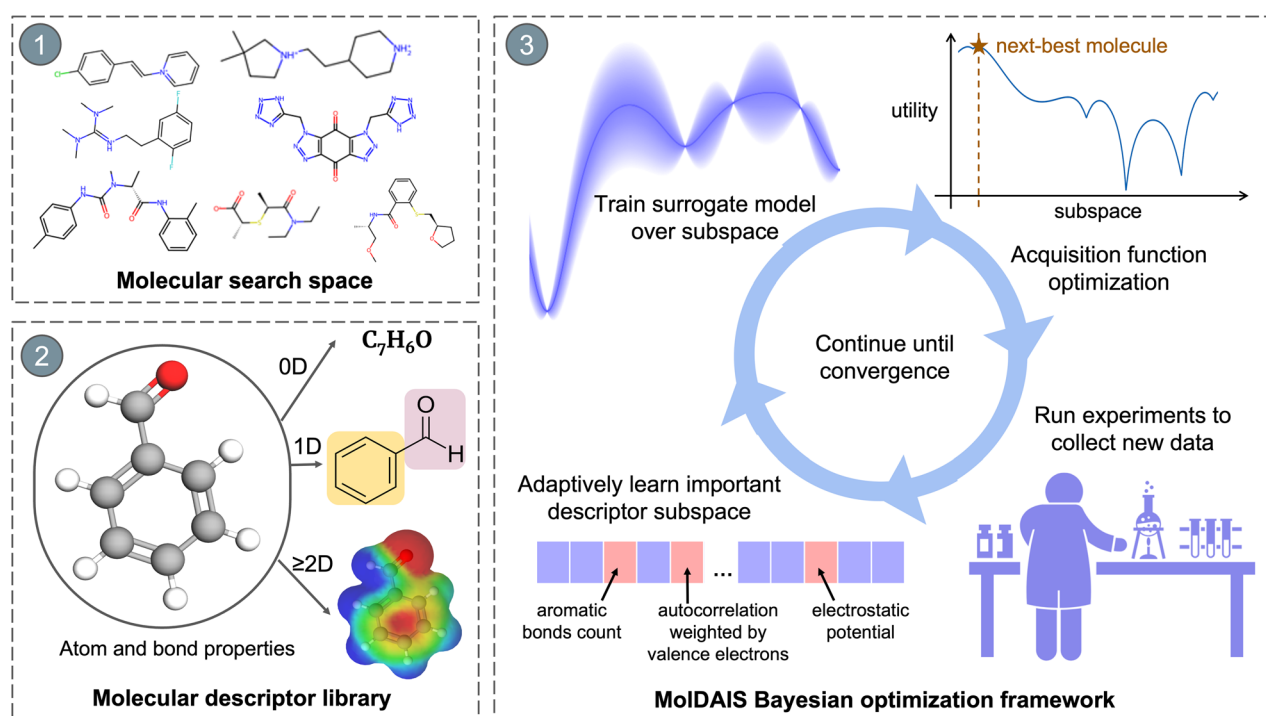
Fig. 1 Schematic of the proposed MolDAIS framework for Bayesian optimization of molecular properties. (1) The process begins by defining a discrete molecular search space, which serves as the optimization design domain. (2) Each molecule is then featurized using a comprehensive library of molecular descriptors. These descriptors may range from simple atom-level counts to complex graph-derived or quantum-informed features. Importantly, there is no restriction on descriptor type, but the assumption is that at least some are informative for the target property or constraints. (3) MolDAIS proceeds in a closed-loop Bayesian optimization cycle: a surrogate model is trained on existing data, an acquisition function is optimized to identify the next candidate molecule, and new property measurements are acquired *via* (typically expensive) experiments or simulations. The distinguishing feature of MolDAIS lies in how the surrogate model is constructed. Rather than relying on a fixed input representation, we impose a sparsity-inducing prior over the descriptor space, enabling the model to learn a compact, property-relevant subspace. This subspace is updated iteratively, allowing the model to revise its hypothesis about which features matter as new data is acquired, leading to improved sample efficiency and robustness in low-data regimes.

practical advantages: it avoids deep learning infrastructure, requires minimal tuning, and can be applied out-of-the-box to any descriptor-featurized molecular dataset. These features make it especially well suited for deployment by domain scientists who may not have significant ML expertise. Taken together, these contributions position MolDAIS as a flexible and extensible framework for data-efficient MPO, opening new opportunities for efficient, targeted, and accessible exploration of chemical space.

2 Preliminaries

A molecular property optimization (MPO) problem can be formally posed as a global optimization task of the form

$$m^{\star} = \underset{m \in \mathcal{M}}{\arg\max} F(m), \tag{1}$$

where m is a molecule from a discrete set $\mathcal{M}$ that defines the molecular search space and $F : \mathcal{M} \to \mathbb{R}$ is the black-box objective function that maps a molecule to its corresponding property value. The goal is to identify the optimal molecule $m^{\star}$ such that $F(m^{\star}) \geq F(m)$ for all $m \in \mathrm{M}$. While this is trivial for small sets of molecules, it quickly becomes intractable when the number of molecules in the set approaches $\sim 10^4$ or more, which is increasingly common in practical settings. Compounding this difficulty is the fact that evaluations of F typically require expensive simulations or wet-lab experiments, often involving noise. Hence, MPO problems require sample-efficient, intelligent search strategies over high-dimensional, structured, and discrete spaces.

2.1 Bayesian optimization

Bayesian optimization (BO) offers a principled framework for solving global optimization problems like (1) when the objective is expensive, black-box, and possibly noisy.[36–38] BO constructs a probabilistic surrogate model of F to guide the search process. The predictions from the surrogate model are combined with an acquisition function, which measures the potential future benefit of querying the objective at a candidate point, to select the next input to actually sample the true function.

2.1.1 Predictive model. At the core of BO is a probabilistic model that approximates F from the available (noisy) data. Among the various model choices (*e.g.*, random forests, Bayesian neural networks), Gaussian processes (GPs)[41] are the most popular due to their flexibility, uncertainty quantification, and analytical tractability. A GP defines a distribution over functions $F \sim \mathrm{GP}(\mu, k)$ with a mean function $\mu(m) : \mathrm{M} \to \mathbb{R}$ and covariance or kernel function $k(m, m') : \mathrm{M} \times \mathrm{M} \to \mathbb{R}$.

Given a dataset $\mathrm{D}_n = (m_i, y_i)_{i=1}^{n}$ of n observed property values $y_i = F(m_i) + \varepsilon_i$ for some (heteroskedastic) zero-mean Gaussian noise $\varepsilon_i \sim \mathrm{N}(0, \lambda_i)$, the posterior predictive distribution at a new point m is Gaussian with the following mean and variance[41]

$$\mu_n(m) = \mu(m) + \mathbf{k}_n(m)^{\top}(\mathbf{K}_n + \Lambda_n)^{-1}(\mathbf{y}_n - \mathbf{u}_n), \tag{2a}$$

$$\sigma_n^2(m) = \mathbf{k}(m, m) - \mathbf{k}_n(m)^{\top}(\mathbf{K}_n + \Lambda_n)^{-1}\mathbf{k}_n(m), \tag{2b}$$

where $\mathbf{k}_n(m) = (k(m, m_1), \ldots, k(m, m_n))$ is the vector of covariances between m and the training points $m_{1:n}$, $\mathbf{K}_n$ is the covariance matrix between the training points, $\mathbf{y}_n = (y_1, \ldots, y_n)$ is the vector of observed values, $\mathbf{u}_n = (\mu(m_1), \ldots, \mu(m_n))$ is the vector of mean values at the training points, and $\Lambda_n = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ is a diagonal matrix of measurement noise variances at the training points. GPs are known to automatically trade off model complexity *via* marginal likelihood maximization, avoiding the need for intensive hyperparameter tuning in most situations (especially when optimizing over simple continuous spaces). They are also capable of capturing fairly complex function behavior with a relatively small number of tunable hyperparameters.

2.1.2 Acquisition function. The acquisition function $\alpha(m|\mathrm{D}_n)$ guides the next evaluation by quantifying the utility of querying F at m. Common acquisition functions include expected improvement (EI),[42] upper confidence bound (UCB),[43] and probability of improvement (PI).[44] BO proceeds *via* the iterative strategy:

$$m_{n+1} = \underset{m \in \mathcal{M}}{\arg\max} \alpha(m|\mathrm{D}_n), \tag{3}$$

where $\mathrm{D}_{n+1} = \mathrm{D}_n \cup (m_{n+1}, y_{n+1})$ is the updated dataset. Acquisition functions are typically designed to balance exploration (regions with high uncertainty) and exploitation (regions with high objective values), with the aim of making BO sample-efficient. For the aforementioned acquisition functions, we can derive closed-form expressions for $\alpha(m|\mathrm{D}_n)$, making it much easier to optimize than F itself. Thus, in MPO settings, the computational overhead associated with training the surrogate and optimizing $\alpha(\cdot)$ is often negligible compared to running new experiments or simulations.

2.2 Traditional molecular representations

While BO traditionally assumes continuous input domains, MPO presents a unique challenge: molecules are inherently discrete objects that must be mapped to a numerical representation $\phi(m)$ suitable for modeling and optimization. The choice of representation significantly influences the success of BO, as it determines the structure of the surrogate model and the kernel's ability to define meaningful similarity measures. Below, we briefly review several major categories of molecular representations, each of which defines a different function for ϕ.

2.2.1 Graphs. Molecules can be represented as undirected graphs $G = (\mathcal{V}, \mathcal{E})$, where atoms correspond to nodes and bonds to edges. Graph encodings are highly expressive, capturing both topological connectivity and relational information. Atom- and bond-level attributes (*e.g.*, element, formal charge, hybridization, bond order, or aromaticity) are typically stored as node and edge labels or attribute tensors.[45,46] Despite their expressivity, graph representations pose two challenges: (i) invariance to node permutations demands permutation-equivariant models (*e.g.*, graph kernels or graph neural networks) and (ii) many graph similarity measures scale super-linearly with molecular size, leading to non-trivial computational cost. Moreover, a single molecule can admit multiple

valid graph encodings (*e.g.*, due to resonance or tautomerism), complicating featurization and downstream optimization.

2.2.2 Strings. Line notations such as SMILES (Simplified Molecular-Input Line-Entry System) and SELFIES (SELF-referencIng Embedded Strings)[25] map molecular graphs to character sequences, usually produced by a depth-first traversal. SMILES is compact and ubiquitous but syntactically fragile: minor token changes often yield invalid structures. SELFIES guarantees that every syntactically valid string decodes to a chemically valid molecule, improving robustness in generative tasks. String encodings interface naturally with language models and sequence kernels (*e.g.*, *n*-gram or substring kernels), yet suffer from non-uniqueness and limited structural interpretability, which can hinder generalization across chemical space.[47]

2.2.3 Fingerprints. Molecular fingerprints are fixed-length binary or count vectors that indicate the presence or frequency of predefined substructures. Structural fingerprints (*e.g.*, MACCS[48]) represent specific functional groups, while extended-connectivity fingerprints (ECFPs) iteratively hash atom-centered neighborhoods to capture local chemical environments.[49] Although ECFPs offer a balance between interpretability and predictive power, they are limited by their local scope and can suffer from sparsity and information loss due to hashing collisions. Fragment count vectors, which enumerate chemotypes such as ring systems, donors/acceptors, and pharmacophores, provide complementary global structural information. Recent "fragprint" representations combine ECFPs with these fragment counts to form hybrid encodings that capture both local topology and broader molecular context.[50] This mitigates some of the limitations of traditional fingerprints and can reveal functional similarities between structurally diverse molecules. Despite their advantages, fingerprint-based representations remain high-dimensional and relatively coarse/lossy summaries of molecular structure.

2.2.4 Descriptors. Molecular descriptors are numerical features that encode structural, physicochemical, or electronic properties of molecules based on their symbolic or graph-based representations. These features may be continuous, discrete, or categorical, and are widely used in cheminformatics due to their compatibility with classical ML models and ease of interpretation. Traditional topological descriptors are computed from 2D molecular graphs and capture properties such as atom counts, connectivity, surface area, and various constitutional and geometrical indices. Libraries like Mordred,[51] PaDEL,[52] and Dragon[53] compute thousands of such descriptors spanning diverse chemical categories. These features are fast to compute and chemically meaningful but often redundant or weakly informative in bulk, motivating the need for techniques capable of estimating the most relevant features for a given task (*e.g.*, through regularization).

More recent work has proposed expressive, theory-driven descriptor families that aim to overcome some of these limitations. Tensor algebra-based methods use multilinear forms and spatial (dis)similarity matrices to encode higher-order relationships among atoms, capturing geometric and relational structure beyond pairwise interactions.[54–56] Graph derivative

descriptors augment molecular graphs with experimentally derived node and/or edge attributes (*e.g.*, NMR shifts, bond energies), producing chemically grounded encodings that better reflect molecular structure.[57,58] Information-theoretic descriptors treat molecules as symbolic sequences and use entropy-based measures to quantify bonding patterns and structural complexity, often without requiring 3D coordinates.[59,60]

Each representation $\phi(m)$ enables the definition of a corresponding kernel $k(\phi(m), \phi(m'))$ that can be directly integrated into the GP modeling paradigm. Graphs require specially designed graph kernels; strings typically use substring-based similarity; fingerprints are often used with a Tanimoto kernel; and descriptors typically rely on standard continuous kernels like those from the Matérn class. The effectiveness of these kernels is tightly coupled to how well the underlying representation captures the key structural or physicochemical relationships relevant to the property of interest. In particular, descriptor-based representations, while chemically rich, can be highly redundant and heterogeneous. As a result, it is often difficult to define a globally smooth/stationary kernel over the full descriptor space. This motivates the need for adaptive strategies that can selectively identify and refine task-relevant subspaces as more information becomes available – a direction we pursue in this work.

MolDAIS, presented formally in Section 3, is agnostic to the specific choice of descriptors and can be readily applied to either classical topological descriptors or newer theory-guided families. Although we focus on traditional descriptors in this study, the modularity of our framework enables future extensions that incorporate more expressive or domain-specific encodings to further enhance model performance and interpretability.

2.3 Data-driven continuous molecular representations

To overcome the challenges of optimizing over discrete molecular structures, recent work has explored generative models, such as VAEs,[30,31] normalizing flows,[32] and generative adversarial networks (GANs),[61] that learn continuous latent embeddings of molecules. These models define an encoder $\phi(m)$ that maps a molecule to a latent vector z, and a decoder D that attempts to reconstruct the original molecule, $m \approx D(z)$. This formulation enables the use of continuous optimization techniques like BO over the latent space $z \in \mathcal{Z}$.

However, generative models present several limitations in practice. First, they are typically trained in an unsupervised fashion using loss functions that prioritize reconstruction accuracy over task relevance. Thus, the resulting embeddings may not exhibit smoothness with respect to the target property F, limiting their utility for optimization. Moreover, the encoder-decoder architecture is usually trained offline and fixed during optimization, preventing the representation from adapting to newly acquired labeled data. Second, latent spaces are often still high-dimensional, posing challenges for BO due to the curse of dimensionality: surrogate models become harder to train, uncertainty estimates degrade, and acquisition functions

become harder to optimize. While recent approaches attempt to fine-tune the generative model using property data and restricting optimization to local regions of the latent space using trust-region BO,[33] they typically require large numbers of evaluations, limiting their effectiveness in low-data regimes.

A related class of encoder-only models, such as DKL,[62] offer an alternative by removing the decoder. In DKL, a neural network $\phi_\theta(m)$ maps molecules to a continuous embedding space used as input to a GP. This allows joint learning of the representation and surrogate model. While DKL avoids the reconstruction problem, it introduces many trainable parameters and is prone to overfitting in low-data settings. Its performance is also sensitive to architectural choices and hyperparameters, which must be carefully tuned.[63]

In contrast to these data-driven approaches, our work returns to chemically grounded molecular descriptors, which are specifically designed to encode structural and physicochemical information. Although these descriptors often correlate with important molecular properties, their use in BO has been limited by the high dimensionality of the descriptor space and the difficulty of identifying which features are most relevant. We propose an alternative approach: rather than learning a full embedding from scratch, we adaptively identify low-dimensional, task-relevant subspaces from a precomputed descriptor library. By updating this subspace as new data is acquired, our method supports interpretable, sample-efficient optimization in low-data regimes. In this way, we aim to bridge the gap between data-driven representation learning and domain-informed molecular design.

3 Methodology

3.1 Adaptive subspace learning over descriptors

We focus on the descriptor representation of molecules, which defines a mapping $x = \phi(m) = (x_1, \ldots, x_D)$ from molecular space M to a continuous feature space $X \subseteq \mathbb{R}^D$. Assuming that this mapping is injective, *i.e.*, every molecule has a unique descriptor vector, we can equivalently pose the MPO problem as an optimization over the descriptor space:

$$x^\star = \operatorname*{argmax}_{x \in X} f(x), \quad \text{where} \quad f(x) = F(\phi^{-1}(x)). \tag{4}$$

The optimal molecule can then be recovered *via* $m^\star = \phi^{-1}(x^\star)$. In this work, we primarily use the Mordred descriptor library,[51] which span thousands of structural and physicochemical features. We normalize each feature to the unit interval so that $X = [0,1]^D$ with $D \approx 2000$; this space can be even larger if other descriptor libraries are considered and/or combined with the Mordred library.

Given this continuous vector representation, we can define a kernel over the descriptor space to use in the standard GP modeling approach summarized in Section 2.1. We adopt the Matérn 5/2 kernel due to its balance of expressiveness and smoothness:

$$k_\psi\left(x, x'\right) = \sigma_f^2\left(1 + \sqrt{5}\,r_\rho + \frac{5}{3}r_\rho^2\right)\exp\left(-\sqrt{5}\,r_\rho\right), \tag{5}$$

$$r_\rho = \sqrt{\sum_{i=1}^{D} \rho_i (x_i - x_i')^2}, \tag{6}$$

where ρ_i is the inverse squared lengthscale for dimension i and σ_f^2 is an output variance (or scale) parameter. The set of kernel hyperparameters is $\psi = \{\rho_1, \ldots, \rho_D, \sigma_f^2\}$. Intuitively, a large ρ_i implies that the GP varies sharply in the ith dimension, meaning the feature x_i is important for predicting $f(x)$, whereas small ρ_i indicates the feature is not particularly important.

These hyperparameters are typically estimated by maximizing the log marginal likelihood (LML) of the GP given current data D_n.[41] While LML-based training allows the model to adapt as data accumulates, it is also sensitive to noise and overfitting, especially in high-dimensional settings where the number of hyperparameters grows linearly with D. In effect, standard GPs implicitly assume that all input features could be relevant, which yields an overly broad hypothesis space and can degrade performance when limited amounts of data are available.

A natural way to regularize this space is to assume that only a sparse subset of the descriptor dimensions are relevant for the property of interest. These relevant dimensions may differ across properties, suggesting an adaptive approach to subspace identification. Following the SAASBO framework proposed by Eriksson and Jankowiak,[40] we encode this preference using the sparse axis-aligned subspace (SAAS) prior, which imposes a strong inductive bias toward sparse solutions by placing a half-Cauchy prior over each inverse lengthscale ρ_i (concentrating its mass near zero). Dimensions are only allowed to "escape" this prior (*i.e.*, take on larger values) if the data provides sufficient evidence that they are relevant. As more data is collected, more dimensions can be activated, leading to increasingly rich surrogate models. This mechanism provides an elegant Bayesian formulation for discovering low-dimensional subspaces in a fully data-driven way.

As in SAASBO, we infer a posterior over the kernel hyperparameters $\{\psi_l\}_{l=1}^L$ using Hamiltonian Monte Carlo (HMC). These samples define a marginal predictive distribution over the target property and, crucially, capture uncertainty over which descriptor dimensions are relevant by exploring different activated subspaces across samples. This uncertainty can be accounted for in the acquisition function, as described in Section 3.3. Fig. 2 provides a high-level overview of the surrogate modeling workflow underlying MolDAIS, which enables flexible and efficient learning of task-relevant descriptors for the property of interest.

3.2 Efficient alternatives to Bayesian subspace inference

While SAAS provides a principled Bayesian framework for discovering sparse, task-relevant subspaces, its computational cost can be prohibitive due to the need for posterior sampling of the kernel hyperparameters using, *e.g.*, HMC. As a practical alternative, we explore lightweight, data-driven feature screening methods that approximate the subspace selection process with substantially lower overhead. In principle, any
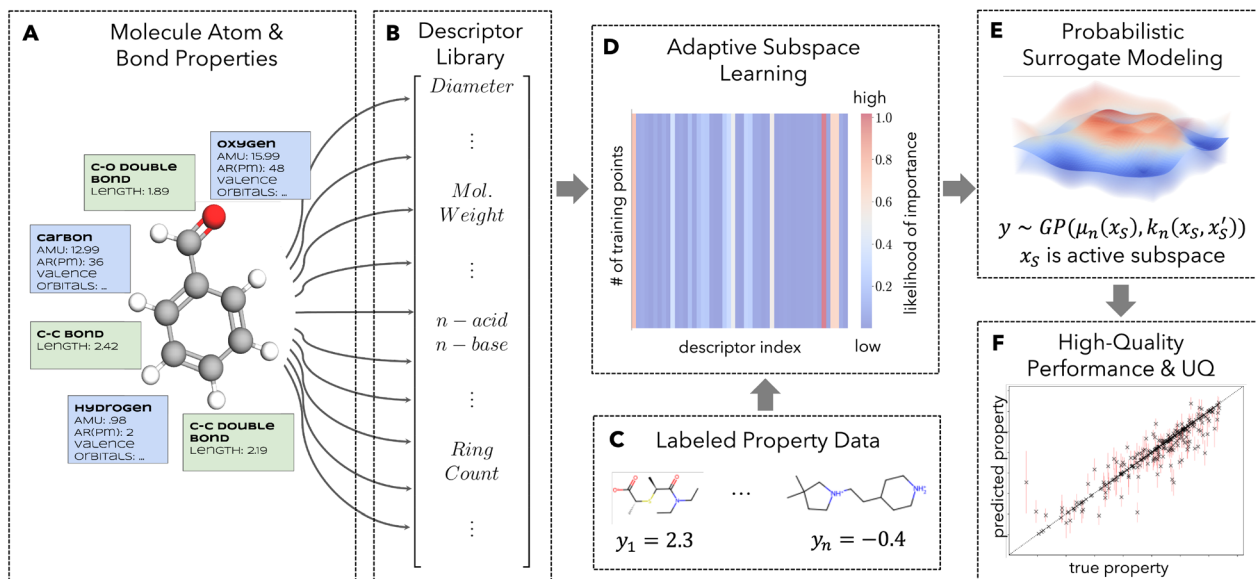
Fig. 2 Overview of the proposed surrogate modeling framework for molecular property prediction. (A) An example molecule is represented in terms of atomic and bond-level features. (B) These features are used to compute a rich library of molecular descriptors capturing structural, electronic, and topological properties. (C) A labeled dataset is assembled by pairing molecules with known property values from experiments or simulations. (D) A sparse subspace is inferred by placing a SAAS prior over the inverse lengthscales in the GP kernel, yielding a posterior distribution over feature relevance (measured in terms of the magnitude of the inverse lengthscales). (E) A Bayesian ensemble of surrogate models is constructed by marginalizing over posterior samples of the kernel hyperparameters. (F) The resulting surrogate provides both accurate predictions and calibrated uncertainty estimates for unseen "test" molecules. By identifying and modeling only a small (property-relevant) subspace of descriptors, this approach enables data-efficient learning even in high-dimensional descriptor spaces.

feature selection method could be used to guide subspace identification. In this work, we focus on two widely used, non-parametric approaches: mutual information (MI)[64,65] and the maximal information coefficient (MIC).[66,67] These methods offer several attractive properties: they are computationally inexpensive, make few distributional assumptions, and can detect non-linear dependencies between individual features x_i and the observed property values $y = f(x) + \varepsilon$.

MI quantifies the amount of information shared between a single feature and the target variable, offering a model-agnostic way to identify relevant descriptors. In our implementation, we compute MI scores using the `mutual_info_regression` function from `scikit-learn`,[68] which estimates mutual information *via* k-nearest neighbors. This estimator is non-parametric and well-suited for continuous variables. MIC, on the other hand, is designed to capture a broader range of functional relationships (including linear and nonlinear associations) by searching over grids in the joint space of two variables and identifying the grid that maximizes a normalized MI estimate. The resulting score ranges from 0 to 1, with higher values indicating stronger associations. MIC is particularly robust in settings where dependencies may be complex, heterogeneous, or obscured by noise. We compute MIC values using the `minepy` (ref. 69) Python package.

Both MI and MIC can be interpreted as imposing a hard prior over the kernel hyperparameters: only the top-ranked features are retained, while all others are effectively discarded by setting their inverse lengthscales to zero. The resulting GP model is then trained by maximizing the LML over this reduced

subspace. While these screening-based methods are less expressive than the fully Bayesian SAAS approach, they offer a nice balance between computational efficiency and predictive performance. We find that even these simple variants yield substantial improvements over standard GP models trained on the full descriptor space, particularly in low-data regimes where overfitting is a concern. Conceptually, this strategy can be viewed as an approximate counterpart to the adaptive subspace learning step shown in Fig. 2D, with the learned feature relevance distribution replaced by a deterministic feature selection procedure.

In our experiments, we default to selecting the top $K = 5$ features. This choice was guided by preliminary SAAS runs on two property prediction tasks, which consistently identified five or fewer highly relevant features. These supporting results are included in the SI. We then further tested this setting on a third benchmark (FreeSolv; see Fig. 4) and found that it maintained strong predictive performance. This gave us confidence in using a consistent K across all benchmarks, avoiding the need for additional tuning. Our choice of a small K also aligns with prior quantitative structure–activity relationship (QSAR) modeling literature that emphasizes sparsity and interpretability.[70,71] While adaptive schemes that grow K with data or optimize it dynamically could further improve performance, we leave these directions to future work.

3.3 Acquisition function optimization

By defining a prior over the kernel hyperparameters, our GP surrogate model becomes fully Bayesian, and the acquisition

function must be appropriately marginalized over the posterior distribution of these hyperparameters. In practice, the marginal expectation is typically approximated *via* Monte Carlo by averaging the acquisition function over L samples $\{\psi_l\}_{l=1}^{L}$ from the posterior:

$$m_{n+1} = \phi^{-1}(x_{n+1}) \text{ with } x_{n+1} = \arg\max_{x \in \mathscr{X}} \frac{1}{L} \sum_{l=1}^{L} \alpha(x|\mathscr{D}_n, \psi_l). \quad (7)$$

For the SAAS-GP method, we use the samples drawn from the posterior using HMC, while for the MI- or MIC-based (approximate) approaches, $L = 1$ and the posterior is approximated as a Dirac delta centered at the selected subspace.

In this work, we focus on the expected improvement (EI) acquisition function due to its widespread empirical success and simple analytical form,[42] which allows for efficient gradient-based optimization. All acquisition optimization routines are implemented using the `BoTorch` (ref. 72) Python library. An advantage of this is that we can easily consider "batch acquisition functions", where a collection of B points $X_{n+1} = \{x_{n+1}^{(1)}, \ldots, x_{n+1}^{(B)}\}$ are jointly selected at each iteration. This is highly relevant in experimental molecular discovery settings where it is common to have parallel evaluation capabilities. For example, in high-throughput chemical screening, one might evaluate a full batch of molecules in a single round using microtiter plates or multiplexed assays. Jointly optimizing a batch acquisition function allows the BO algorithm to account for redundancy and diversity in the selected molecules, significantly reducing the time-to-discovery in practical applications.

3.4 Extension to constrained, multi-objective problems

A key strength of our framework is its modularity. Since our contribution lies in how the surrogate model is constructed – namely, through adaptive subspace identification over a molecular descriptor library – our approach can be seamlessly extended to more complex MPO formulations. In particular, constrained and multi-objective optimization problems frequently arise in practical molecular discovery and can be addressed using well-established techniques from the BO literature.

In the constrained setting, the goal is typically to maximize a desired performance metric (*e.g.*, activity or selectivity) subject to one or more feasibility constraints (*e.g.*, toxicity, synthetic accessibility, or stability). Modern constrained BO algorithms can broadly be categorized into two families: (i) safe BO methods, which aim to ensure feasibility throughout the optimization process[73–75] and (ii) asymptotically feasible methods, which permit early constraint violations but require that the final solution satisfy all constraints.[76,77] Our surrogate modeling approach is compatible with both formulations. However, due to the strength of the (approximate) SAAS prior, uncertainty estimates may be less reliable in early iterations, making asymptotically feasible methods generally more practical in low-data regimes.

Similarly, many real-world applications involve trade-offs between competing objectives. For instance, in the context of designing new battery materials, one would typically want to maximize energy density and minimize degradation (or similarly maximize cycle life). In such cases, the goal is not to identify a single optimal molecule, but rather to characterize the Pareto frontier, *i.e.*, the set of non-dominated solutions for which no objective can be improved without degrading another. Our framework integrates naturally with multi-objective BO algorithms, including those based on hypervolume improvement, which quantify the expected gain in dominated volume in objective space.

It is worth noting that a key advantage of our proposed approach in both constrained and multi-objective settings is that each property or constraint can be modeled independently in different subspaces. As a result, the subspace learned for each target can adapt to the most relevant molecular features for that specific property. This flexibility is particularly useful when optimizing for multiple, physically distinct phenomena – enabling the surrogate model to allocate modeling capacity where it is most needed.

4 Results and discussion

4.1 Predictive accuracy and uncertainty quantification

We begin by evaluating the surrogate modeling performance of MolDAIS on four well-established small-data regression tasks from the GAUCHE benchmark suite:[39,78] lipophilicity (4200 molecules; $\log D$ at pH 7.4), ESOL (1128 molecules; aqueous solubility), FreeSolv (642 molecules; hydration free energy), and Photoswitch (392 molecules; $\pi \rightarrow \pi^*$ transition wavelength). For each dataset, we perform ten random 5/95 train–test splits to simulate the data-scarce conditions typical of early-stage molecular discovery and optimization.

We compare MolDAIS to one representative baseline from each major molecular representation class (descriptors, fingerprints, strings, and graphs described in Section 2.2) selected based on their GAUCHE performance. Two ablation variants of MolDAIS are included to assess the effect of adaptive subspace identification and unsupervised dimensionality reduction. All models are trained as GPs using the BoTorch library.[72] For all baselines, kernel hyperparameters are optimized *via* L-BFGS-B.[79] MolDAIS is trained using HMC with the No-U-Turn Sampler (NUTS),[80] using default BoTorch settings (512 warmup steps, thinning of 16, and 256 samples).

The six modeling approaches evaluated are:

● MolDAIS: a fully Bayesian GP with a Matérn 5/2 kernel and a SAAS prior over inverse lengthscales, allowing for adaptive identification of a task-relevant subspace of Mordred descriptors.

● MD-Mat: standard GP with a Matérn 5/2 kernel over the full Mordred molecular descriptor space.

● MD-Mat-PCA: standard GP with Matérn 5/2 kernel applied to a PCA-reduced embedding of the descriptor space (retaining 99% of variance); represents a fixed, unsupervised dimensionality reduction baseline.

● FP–TM: a fingerprint-based GP using molecular fragprints[50] with a Tanimoto kernel. Fragprints are hybrid representations that concatenate ECFPs with fragment count vectors,

capturing both local atom-centered neighborhoods and global structural features (*e.g.*, presence of ring systems).

- SMILES-Str: SMILES-based GP using a bag-of-characters string kernel based on Tanimoto similarity.[81]
- Graph-WL: graph-based GP using the Weisfeiler–Lehman kernel,[82] which encodes molecular similarity *via* subtree pattern counts.

Fig. 3 shows the distribution of test RMSE values across splits for each model–dataset combination. MolDAIS consistently achieves the lowest median RMSE and narrowest spread, demonstrating its robustness in low-data regimes. While Fragprints–TM and SMILES-Str perform well on select tasks, no baseline offers consistent performance across all problems. In contrast, MolDAIS adapts to each task by identifying a property-specific subspace, enabling broad generalization.

In sparse-data settings, accurate uncertainty estimates are as important as predictive accuracy for driving effective acquisition in BO. We evaluate uncertainty quantification (UQ) performance using two metrics introduced by Rasmussen *et al.*:[83] (i) the R^2 correlation between predicted root-mean variance (RMV) and observed prediction error (RMSE), and (ii) the miscalibration area, which quantifies deviation between empirical and nominal confidence intervals. Ideal models

should achieve R^2 near 1 and minimal miscalibration area. As shown in Table 1, MolDAIS consistently achieves high RMSE-RMV correlation and low miscalibration area across all datasets. Descriptor-based baselines (MD-Mat, MD-Mat-PCA) consistently underperform, suggesting that naive use of high-dimensional features (without sparsity or property-based adaptation) leads to poorly calibrated models. FP–TM performs well on the lipophilicity problem, but exhibits degraded calibration on other tasks, indicating fragility across domains. Although graph-WL yields high R^2 on ESOL, its calibration error remains relatively large, and it ranks lowest in RMSE across all four datasets (Fig. 3). Overall, these results suggest that MolDAIS offers more reliable uncertainty estimates, which are critical for sample-efficient optimization.

To further highlight the benefits of adaptive subspace learning, we examine how MolDAIS refines its descriptor space as more training data becomes available. Fig. 4 shows the distribution of median inverse lengthscales across Mordred descriptors at different training sizes (10, 30, 60, 90 samples) on the FreeSolv dataset. These inverse lengthscales serve as feature importance scores: higher values indicate greater relevance. As training size increases, the distribution progressively sharpens, revealing clear separation between relevant and irrelevant
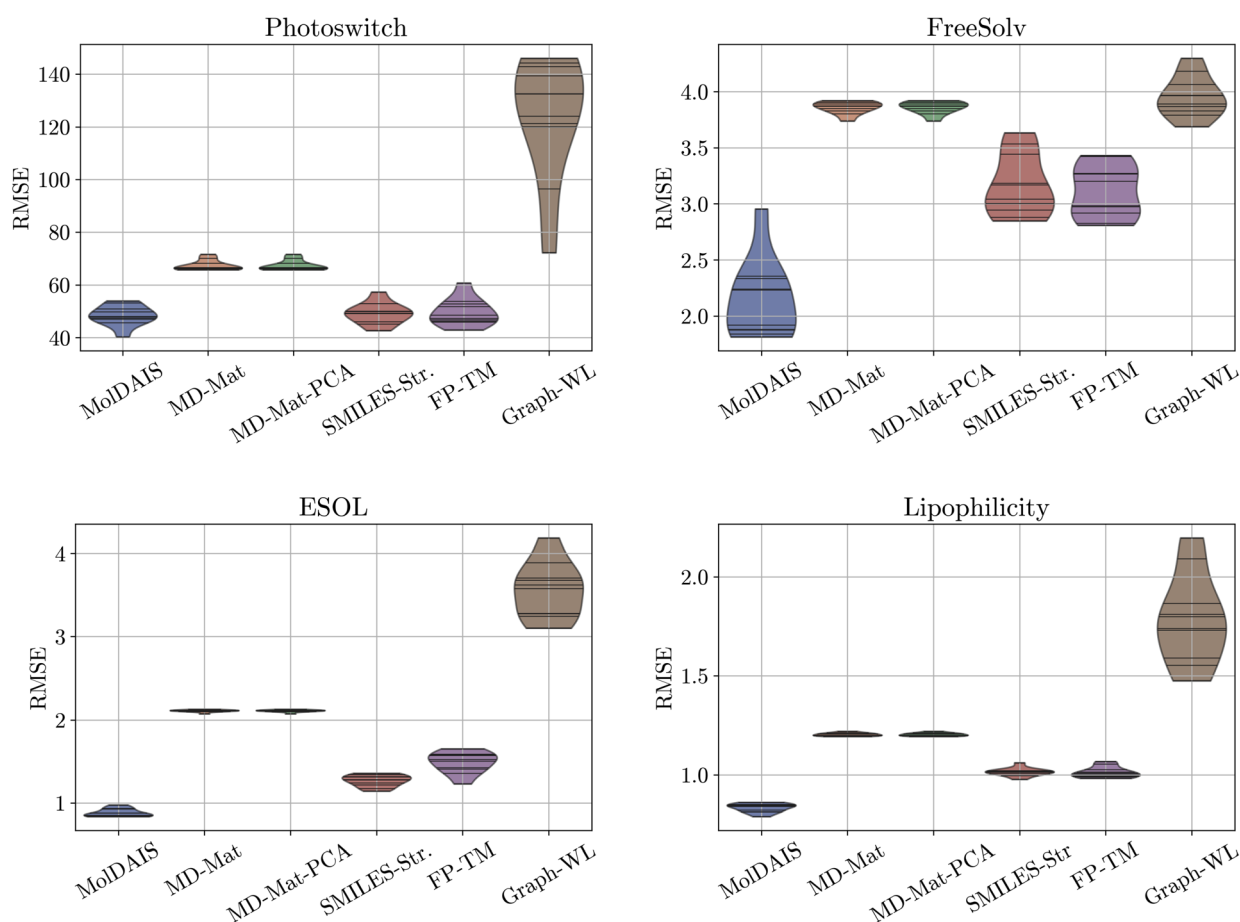

Fig. 3 Test RMSE over ten random 5/95 train−test splits for four small-data regression tasks: Photoswitch (top−left), FreeSolv (top−right), ESOL (bottom−left), and lipophilicity (bottom−right). The six methods include MolDAIS and one representative from each other major representation family (fingerprint−, string−, and graph−based), plus two ablations. MolDAIS consistently achieves the lowest median RMSE and narrowest spread.

Table 1 Uncertainty quantification analysis on the four small-data regression tasks. We report the R^2 for the RMSE vs. RMV correlation and miscalibration area values for ten random 5/95 train–test splits. Higher R^2 and lower miscalibration area indicate better calibration

RMSE vs. RMV R^2					Miscalibration area			
Method	ESOL	FreeSolv	Photoswitch	Lipophilicity	ESOL	FreeSolv	Photoswitch	Lipophilicity
MolDAIS	0.61 ± 0.20	**0.70 ±0.25**	**0.64 ±0.10**	0.79 ± 0.08	**0.12 ±0.05**	**0.07 ±0.04**	**0.09 ±0.05**	0.17 ± 0.07
MD-Mat	0.09 ± 0.09	0.02 ± 0.03	0.21 ± 0.19	0.04 ± 0.04	0.29 ± 0.00	0.36 ± 0.00	0.50 ± 0.00	0.18 ± 0.01
MD-Mat-PCA	0.08 ± 0.07	0.03 ± 0.02	0.09 ± 0.10	0.04 ± 0.04	0.29 ± 0.00	0.36 ± 0.00	0.50 ± 0.00	0.18 ± 0.01
SMILES-Str	0.48 ± 0.16	0.49 ± 0.11	0.57 ± 0.17	0.67 ± 0.10	0.19 ± 0.02	0.36 ± 0.02	0.49 ± 0.01	0.14 ± 0.03
FP–TM	0.37 ± 0.23	0.31 ± 0.08	0.43 ± 0.15	**0.85 ±0.05**	0.17 ± 0.03	0.29 ± 0.02	0.49 ± 0.01	**0.05 ±0.01**
Graph-WL	**0.96 ±0.01**	0.05 ± 0.05	0.51 ± 0.01	0.65 ± 0.02	0.24 ± 0.06	0.11 ± 0.14	0.19 ± 0.10	0.07 ± 0.02
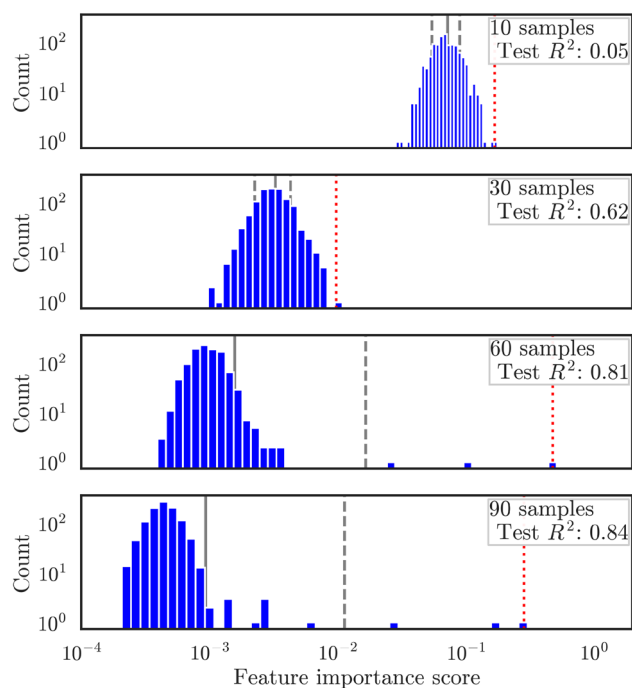
Fig. 4 Distribution of feature importance scores (measured as the median inverse lengthscale across 1060 Mordred descriptors) for different training sizes on the FreeSolv dataset. As data accumulates, MolDAIS progressively sharpens its subspace and improves test performance (R^2 on random held-out test set of 482 molecules). Grey solid and dashed lines represent the mean ± 1 standard deviation and the red dotted line denotes the largest feature importance score.

descriptors. At 30 samples, a few features begin to emerge as informative; by 60 samples, a distinct sparse subspace is apparent. Interestingly, even between 60 and 90 samples, the model continues to adjust feature relevance, illustrating MolDAIS's flexibility to update its hypotheses as new data is collected. This refinement corresponds to substantial improvements in predictive performance (R^2 increases from 0.05 at 10 samples to 0.84 at 90), and is a key indicator of potential strong performance when included in iterative optimization pipelines such as BO.

4.2 Optimization performance

We next assess how surrogate model quality translates into efficiency of the BO method. We retain the two strongest

methods from the model study, MolDAIS and FP–TM and the two descriptor abalations (MD-Mat and MD-Mat-PCA). We add two new efficient approximations to MolDAIS, mainly MolDAIS-MI and MolDAIS-MIC, to study the impact of the type of adaptive subspace identification method. Both variants use $K = 5$ selected features, as motivated in Section 3.2. We use this same, fixed value across all benchmarks to keep the methods simple and reproducible. We also add Random search as a baseline. For one of the benchmarks, we also compare against LADDER,[84] which is a recent latent-space BO method that develops a novel structure-coupled kernel. It thus combines a pre-trained junction tree variational autoencoder (JT-VAE) with a fingerprint similarity kernel and reports state-of-the-art performance for some MPO problems.

• Penalized log P:[30,31] a standard benchmark in the MPO literature, consisting of 250 000 molecules randomly selected from the ZINC database. The objective is a penalized octanol–water partition coefficient (log P), adjusted for synthetic accessibility and ring size, simulating a drug-likeness optimization problem.
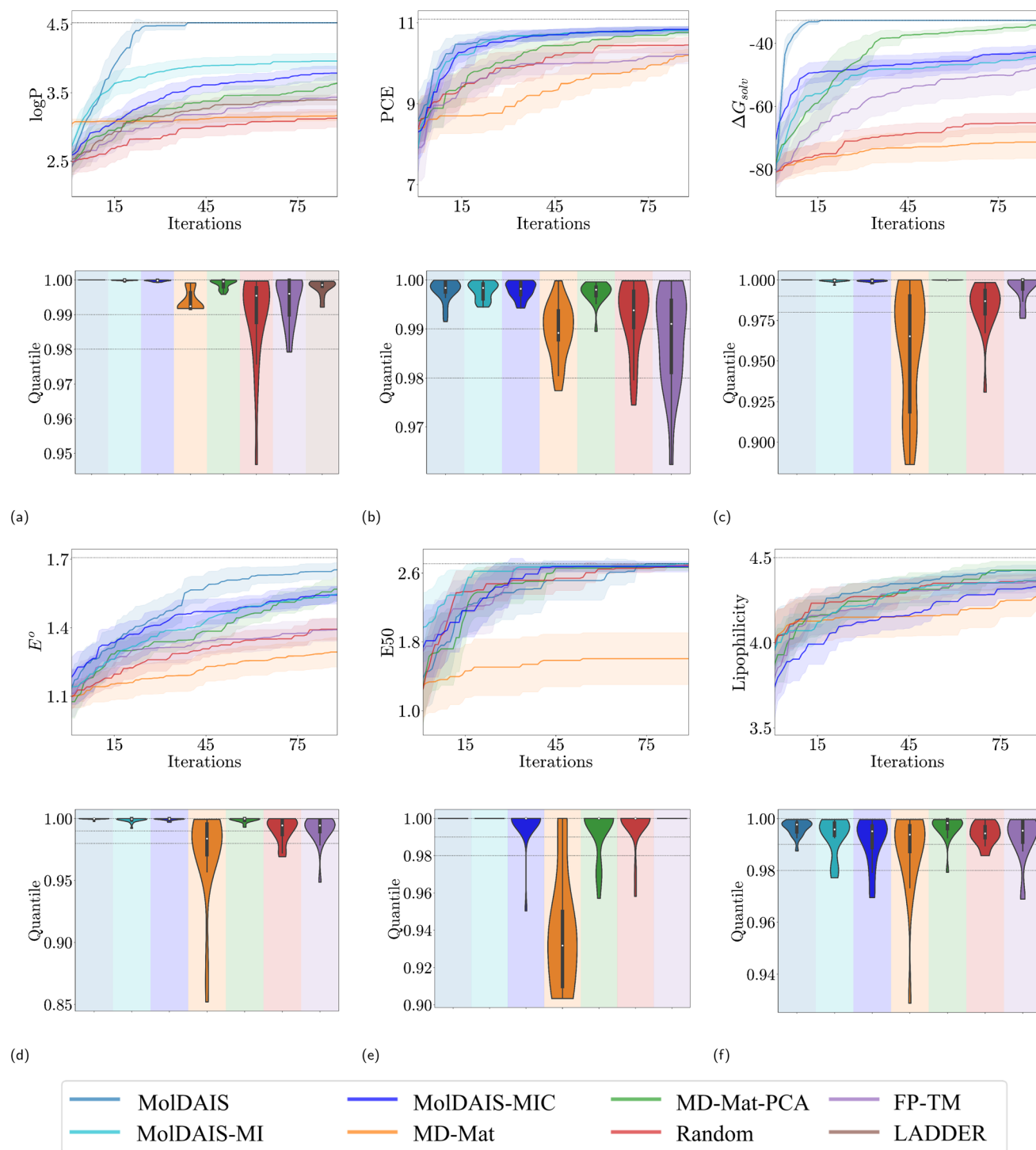
• Power Conversion Efficiency (PCE):[85] contains 29 978 organic photovoltaic molecules with computed PCE values, representing the efficiency of solar-to-electric energy conversion. This dataset serves as a proxy for high-throughput materials discovery in renewable energy applications.

• Antimalarial activity (EC50):[86] comprises 9998 organic molecules with measured half-maximal effective concentration (EC50) values against a sulfide-resistant strain of *Plasmodium falciparum*. The EC50 value reflects the concentration required to inhibit 50% of parasite.

• Solvation Free Energy (ΔG_{solv}) & Redox Potential (E^0):[87] contains over 103 000 quinone derivatives with the two properties of interest (ΔG_{solv} and E^0) being computed with density functional theory. These properties are relevant to energy storage and redox chemistry, and we use them to explore single- and multi-objective optimization scenarios.

We conduct a series of controlled optimization experiments using these datasets. Note that, since LADDER requires a pre-trained generative model that requires a very large molecular search space, we only apply it to the penalized log P benchmark wherein we use an established accurate generator. To ensure a fair comparison, we closely followed the official LADDER implementation and reproduced the protocol described in its original publication. This includes using the same pre-trained

Fig. 5 Bayesian optimization performance across six molecular property optimization tasks. Each panel corresponds to a specific dataset described in Section 4.2: (a) penalized log P, (b) power conversion efficiency (PCE), (c) solvation free energy ΔG_{solv}, (d) redox potential E^0, (e) antimalarial activity (EC50), and (f) lipophilicity. (Top row) Best-found property value *versus* number of tested molecules (iterations), averaged over 20 independent replicates of 10 random initial selections. Shaded regions represent 95% confidence intervals. (Bottom row) Distribution of final quantile scores, defined as the fraction of molecules in the search space with lower or equal property values than the best molecule found in each trial. Higher values and narrower distributions indicate better and more consistent performance. Results compare eight methods described in the text, including the proposed MolDAIS variants, standard descriptor- and fingerprint-based baselines, random search, and the recently developed LADDER method. The three dashed lines in the quantile plots denote the 100th, 99th, and 98th quantile of the dataset.

JT-VAE model and kernel configuration/hyperparameter tuning strategy. The only modification was to the initial batch of molecules, which we standardized across all baselines in our study for reproducibility. All methods follow the same procedure and are initialized with the same settings. Each optimization run begins with 10 randomly selected molecules from the search space M, and is given a budget of 90 additional evaluations, resulting in a total of 100 tested molecules per run. We use the expected improvement (EI) acquisition function for all methods except the random baseline, which selects molecules uniformly at random.

To account for variability in the initial data, each experiment is repeated across 20 independent trials, each with a different random seed. All methods share the same set of initializations to ensure comparability. Results are reported in Fig. 5. For each method, we report two performance metrics. The top row of each panel shows the mean evolution of the best-found objective value over the course of the optimization, averaged across the 20 replicates. Shaded regions indicate 95% confidence intervals. This metric captures how quickly each method identifies high-performing molecules. The bottom row of each panel shows the distribution of final quantile scores achieved by each method. For each replicate, we compute the quantile of the best-found molecule relative to the full search space M. A quantile score of $q \in [0, 1]$ indicates that the identified molecule outperforms $q \times 100\%$ of all possible candidates in the search space. We visualize the full distribution of quantile scores across replicates using violin plots, which highlight both median performance and variability across runs.

Across all case studies, MolDAIS demonstrates consistently strong sample efficiency. The fully Bayesian variant achieves the best overall performance on most tasks, with the approximate versions (MolDAIS-MI and MolDAIS-MIC) performing competitively and often closely tracking the SAAS version. Notably, MolDAIS-MI slightly outperforms MolDAIS-MIC in most cases, suggesting that mutual information may be a slightly more effective metric for subspace selection in this context. In contrast, both MD-Mat and Random underperform consistently, highlighting the difficulty of modeling and optimizing in high-dimensional descriptor spaces without subspace regularization.

The alternative baselines MD-Mat-PCA and FP–TM perform reasonably on select tasks (e.g., MD-Mat-PCA for ΔG_{solv} in Fig. 5c and FP–TM for EC50 in Fig. 5e), but show inconsistent behavior across datasets and tend to plateau earlier or exhibit high variance. These trends are further illustrated in the quantile scores, where MolDAIS consistently yields narrow, high-scoring values across all replicates, whereas other methods often exhibit greater spread or even failure cases. For instance, FP–TM identifies the global optimum in all EC50 trials, but fails to do so in PCE – occasionally even underperforming compared to naive random search. Such volatility highlights the importance of adapting the representation to the property being optimized. By automatically adapting the "active" descriptor subspace to specific tasks, MolDAIS enables more reliable optimization performance across a diverse range of properties.

A closer look at the penalized $\log P$ task (Fig. 5a) further illustrates this point. MolDAIS achieves the highest (best-found) molecular property values and is the only method that successfully identifies the globally optimal molecule in all 20 replicates, doing so after querying less than 0.04% of the candidate search space. While MolDAIS-MI converges more slowly, it consistently reaches top-performing candidates, as evidenced by the high quantile score with minimal spread. In contrast, LADDER is unable to locate the global optimum in any replicate and underperforms all three MolDAIS variants, both in terms of convergence speed and final quantile scores. This further demonstrates the value of MolDAIS's adaptive, descriptor-focused modeling approach in delivering accuracy and robustness across molecular optimization tasks.

4.3 Descriptor library choice and physical interpretability

A central feature of MolDAIS is its ability to adaptively identify sparse, informative subsets of molecular descriptors, which can offer interpretable insights into the physicochemical factors that drive property variation. However, the quality of these insights strongly depends on the starting descriptor library. In this section, we assess how descriptor representation affects both predictive performance and physical interpretability by comparing three widely used libraries (Mordred,[51] PaDEL,[52] and ChemBERTa[88]) for the prediction of ΔG_{solv}. ChemBERTa generates learned embeddings *via* pre-training a transformer architecture on large molecular datasets, capturing potentially rich but opaque chemical patterns. In contrast, Mordred and PaDEL are physics-inspired calculators that encode chemically meaningful features based on molecular graph and electronic structure heuristics.

We fix the feature selector to the MolDAIS-MI variant (with $K = 5$) and evaluate each descriptor library using a training set of 40 molecules and a test set of 100 held-out molecules. Fig. 6 shows the correlation matrix among the selected features across the three descriptor libraries. Descriptions of the interpretable descriptors are also provided in Table 2. Substantial alignment is observed between Mordred and PaDEL: three features (GATS1p, AATSC2i, and GATS1i) are shared across both libraries and correspond to autocorrelation measures of polarizability and ionization potential. These features are consistent with the established physics of solvation (that depends on charge distribution and polarizability). The remaining two descriptors in each library also show strong correlations; AETA_beta_s and ETA_BetaP_s are perfectly correlated ($R^2 = 1.0$), and AATS2s and PNSA-1 are moderately correlated ($R^2 = 0.80$), indicating convergence on related physical information. In contrast, the top ChemBERTa features are largely uncorrelated with the interpretable descriptors.

We find predictive performance mirrors this pattern; Fig. 7 shows parity plots of the predicted *versus* true ΔG_{solv} values. The model using Mordred descriptors achieves the best performance ($R^2 = 0.518$, miscalibration area = 0.069), followed closely by PaDEL ($R^2 = 0.447$, miscalibration = 0.074). ChemBERTa performs significantly worse ($R^2 = 0.203$, miscalibration = 0.086), highlighting the advantage of descriptors that are not
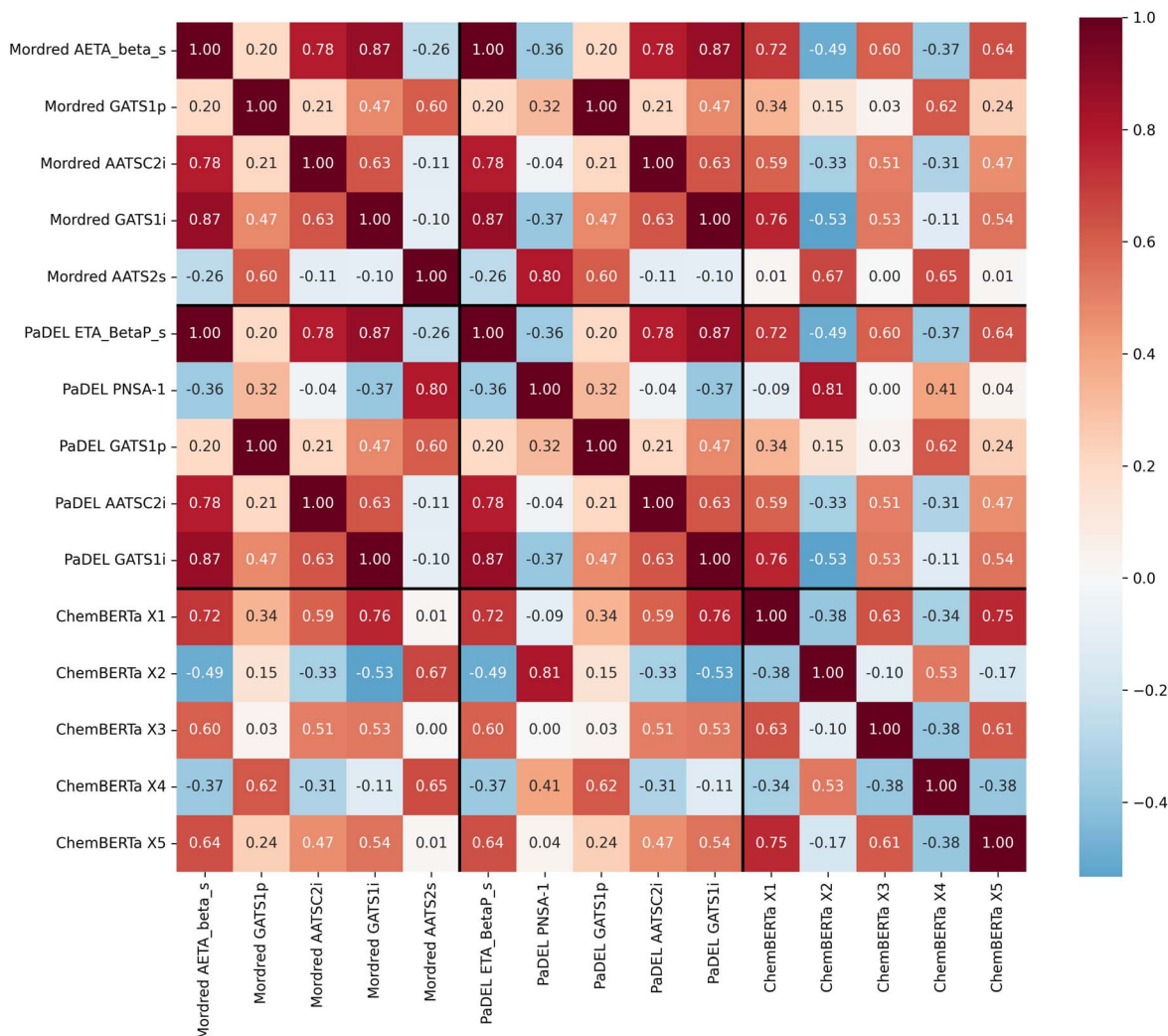
Fig. 6 Pairwise correlations among the five descriptors selected by MolDAIS-MI ($K = 5$) for ΔG_{solv} across three descriptor libraries. Mordred and PaDEL yield highly overlapping and strongly correlated features, while ChemBERTa embeddings (denoted X1 to X5) are mostly uncorrelated with those from physics-based descriptors. Definitions for Mordred and PaDEL features are provided in Table 2.

only informative but also grounded in physical intuition. Mordred's superior performance likely reflects its inclusion of detailed topological and electronic descriptors that are relevant for capturing solute–solvent interactions, particularly for redox-active molecules. It is worth noting similar (useful) trends were found for the PCE case study, which are provided in the SI, though the nature of the selected descriptors greatly differs due to the distinct physical mechanisms involved.

Table 2 provides a detailed overview of the selected descriptors, including their formal definitions, physical interpretation, and relevance to solvation behavior. All features shared between Mordred and PaDEL encode polarizability and ionization potential trends, which are two physicochemical quantities known to impact solvation behavior by influencing solute–solvent electrostatic and dispersion interactions.[89–91] Together, these results illustrate how the descriptor library choice has an important and meaningful impact on the transparency and scientific value/interpretability of MolDAIS.

4.4 Multi-objective optimization of organic electrodes

To illustrate the flexibility of MolDAIS beyond single-objective problems, we apply it to a multi-objective optimization task motivated by the discovery of high-performance organic electrode materials (OEMs). OEMs offer a promising path toward sustainable energy storage due to their structural tunability, biodegradability, and potential for low-cost manufacturing.[92–95] Unlike conventional electrodes that rely on scarce or toxic transition metals (*e.g.*, lithium, nickel, or cobalt), OEMs avoid environmental concerns associated with mining and enable safer chemistries. Aqueous zinc-ion batteries have emerged as an attractive platform for such materials, as they combine non-flammable electrolytes with high ionic conductivity and favorable zinc redox kinetics.[96–98]

However, identifying suitable OEMs remains challenging due to the need to balance multiple competing criteria, such as energy density and long-term cycling stability.[19] Here, we focus on two molecular properties that serve as proxies for these

Table 2 Description and physicochemical meaning of the top five interpretable descriptors selected by MolDAIS−MI ($K = 5$) for ΔG_{solv} from the Mordred and PaDEL libraries

Descriptor	Family[a]	Formal description	Physicochemical meaning & link to solvation	Package
GATS1p	GATS	Geary autocorrelation of atomic polarizability at topological lag 1	Measures polarizability contrast between bonded atoms; higher values indicate uneven electronic softness, which enhances induced dipole–solvent dispersion interactions	Mordred, PaDEL
GATS1i	GATS	Geary autocorrelation of atomic ionization potential at topological lag 1	Captures bond-level polarity due to ionization potential mismatch; reflects spatial heterogeneity in local electronegativity that governs dipole-solvent and hydrogen-bond interactions	Mordred, PaDEL
AATSC2i	ATS	Centered Moreau–Broto autocorrelation of atomic ionization potential at lag 2, averaged per atom	Quantifies how ionization potential trends persist over two-bond paths; indicates delocalized electron-withdrawing effects, which shape solvation shell structure and stabilization	Mordred, PaDEL
AETA_beta_s	ETA	Averaged extended topochemical atom index β^s (sigma-electron contribution)	Encodes density of electronegative atoms (*e.g.*, O, N) within the sigma-bond network; higher values indicate greater potential for electrostatic and hydrogen-bond interactions with polar solvents	Mordred
ETA_BetaP_s	ETA	Normalized extended topochemical beta index β^s (relative to molecular size)	Reflects concentration of polar atoms per unit size; higher values signal high polarity density, favoring compact, strongly interacting solvation shells	PaDEL
AATS2s	ATS	Averaged Moreau–Broto autocorrelation at lag 2 weighted by intrinsic state (electrotopological index)	Measures how electronic influence from polar atoms propagates over 2-bond paths; higher values indicate inductive effects that extend solute polarity, enhancing polarization-based solvation	Mordred
PNSA-1	CPSA	Sum of solvent-accessible surface area over atoms with negative partial charge	Quantifies hydrogen-bond acceptor surface area; larger values support stronger solute–solvent electrostatic stabilization *via* polar interactions with water	PaDEL

[a] GATS = geary autocorrelation; ATS = averaged topological autocorrelation; ETA = extended topochemical atom; CPSA = charged partial surface area.

targets: redox potential (E^0) (which relates to achievable voltage) and solvation free energy (ΔG_{solv}) (which governs stability and ion transport in aqueous environments). The goal is to identify molecules that jointly maximize E^0 while minimizing ΔG_{solv}, corresponding to maximization of the Pareto frontier in this two-objective space.

We adopt a multi-objective BO framework using the expected hypervolume improvement (EHVI) acquisition function[99] to sequentially select molecules. As before, we use the 103 000-molecule quinone dataset with DFT-computed property values.[87] Each algorithm is initialized with 10 randomly selected molecules and allowed 90 additional evaluations. We compare three methods: MolDAIS-MIC, FP–TM, and Random, and repeat each trial 20 times with independent random seeds. We report results only for MolDAIS-MIC among our variants, as it performs similarly to the original MolDAIS while having lower computational cost.

Fig. 8 summarizes the results. The top panel shows the average log hypervolume of the discovered Pareto front as a function of the number of iterations. MolDAIS-MIC achieves substantially higher hypervolume than FP–TM and Random, with consistently faster growth across replicates. The bottom panel visualizes the final sample distribution and learned Pareto fronts for the median replicate of the final hypervolume value over the 20 trials. MolDAIS-MIC successfully discovers a wide range of diverse tradeoff OEM designs that closely approximate the true Pareto frontier, while FP–TM and Random recover only a narrow band of suboptimal candidates. Notably, nearly 20 points on the MolDAIS-MIC front strictly dominate the FP–TM set, highlighting the benefits of learning property-specific subspaces that evolve as more property data is collected. Further inspection of the selected subspaces reveals minimal overlap between those learned for E^0 and ΔG_{solv}, supporting the hypothesis that modeling each property in a tailored
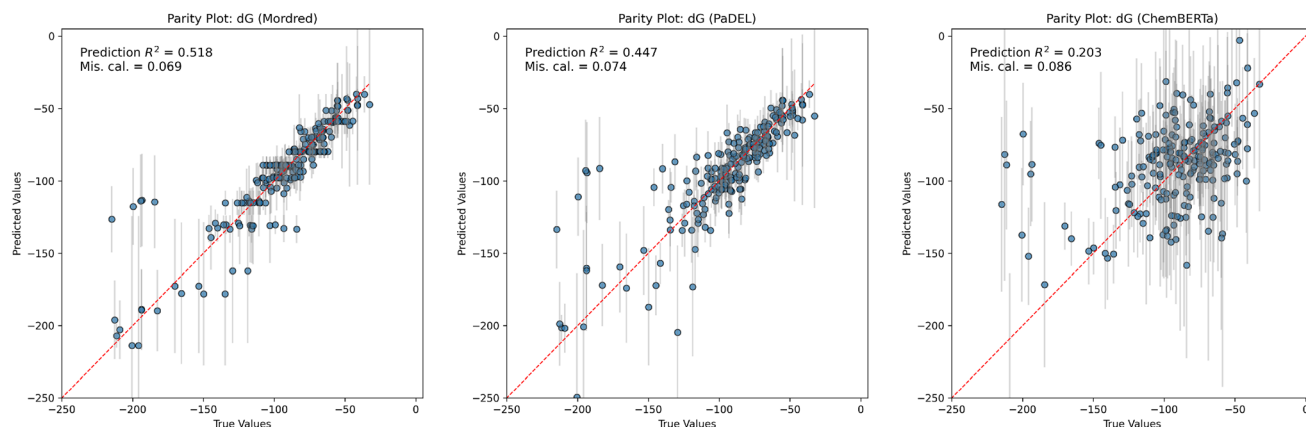
Fig. 7 Prediction performance for ΔG_{solv} using MolDAIS-MI ($K = 5$) for the various descriptor libraries. Left to right: Mordred, PaDEL, Chem-BERTa. Each plot shows the parity between predicted and ground truth values on the test set. R^2 values (higher is better) and miscalibration areas (lower is better) are also reported in the top left corner of each plot.
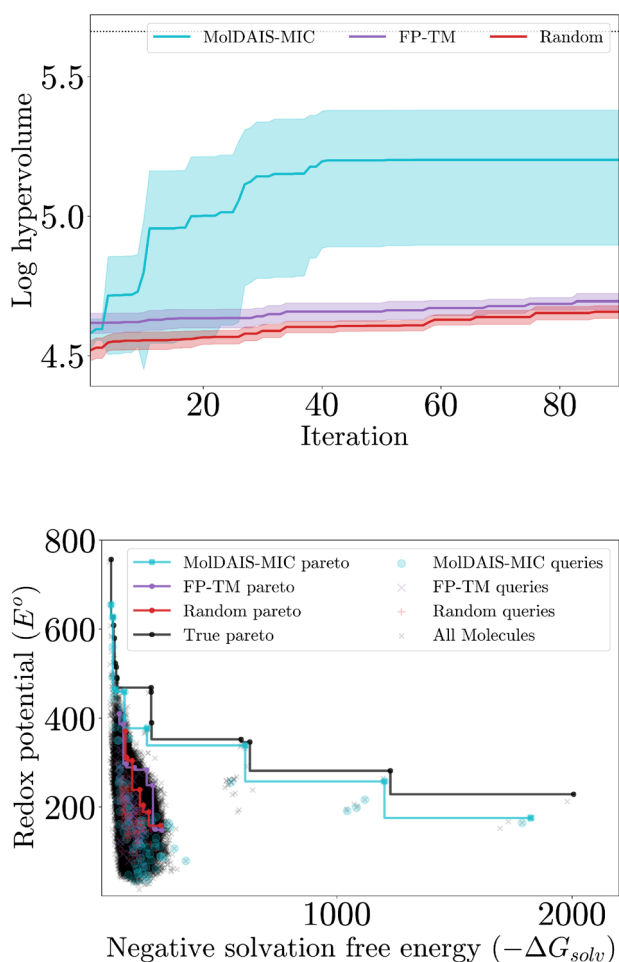

Fig. 8 Multi-objective Bayesian optimization of organic electrode materials. (Top) Mean log hypervolume of the Pareto front as a function of iteration across 20 random initializations; shaded regions indicate 95% confidence intervals. (Bottom) Plot of the sampled redox potential and solvation free energy points for each algorithm at the final iteration for the median replicate. MolDAIS-MIC discovers a broad, high-quality Pareto frontier that closely matches the true Pareto set, whereas FP–TM and Random discover fewer and less competitive molecules that tradeoff between redox potential and solvation free energy.

latent space is critical to navigating tradeoffs in complex molecular design problems.

5 Conclusions

We presented MolDAIS, a flexible framework for molecular property optimization (MPO) that combines a large descriptor library with adaptive subspace identification to enable sample-efficient molecule discovery in low-data settings. MolDAIS builds on the recently proposed sparse axis-aligned subspace (SAAS) prior for GP models, applying it to descriptor-based molecular representations to automatically identify property-relevant features that flexibly evolve as more data is collected. In addition to validating this approach, we developed simple yet effective approximations of the SAAS prior that reduce computational cost while preserving its key benefits. By integrating these modeling approaches into a molecular Bayesian optimization (BO) framework, MolDAIS efficiently balances exploration and exploitation to discover high-performing molecules with minimal data – making it applicable even when the properties of interest can only be evaluated for tens to hundreds of molecules during the optimization campaign.

Across a diverse suite of benchmark and real-world MPO tasks, including both modeling and optimization, MolDAIS consistently outperforms state-of-the-art methods based on molecular graphs, SMILES strings, and fingerprints. Importantly, it requires minimal hyperparameter tuning and avoids the need to train large-scale, highly parameterized models, making it a practical and accessible tool for experimental researchers without deep expertise in machine learning. While this work focused on single- and multi-objective BO, the modularity of MolDAIS makes it readily extensible to more complex design scenarios. For example, it can support multi-fidelity optimization, where data from simulation and experiment are combined, or novelty-driven discovery, where the goal is to explore underrepresented regions of chemical space or uncover diverse (atypical) molecular candidates.

There remain several interesting opportunities for extending the MolDAIS framework. In this work, we examined three sparsity-inducing strategies that assume axis-aligned relevance of individual descriptors. A natural next step is to account for correlated descriptor interactions – particularly in cases where properties vary along curved or entangled manifolds in the feature space. One potential approach would be to integrate nonlinear dimensionality reduction methods, such as kernel PCA, within the BO loop, while still applying sparsity-inducing priors in the resulting latent space. Alternatively, structured kernel priors that encode pairwise or hierarchical relationships between descriptors could help capture higher-order dependencies. In addition, while we focused on the Mordred descriptor library due to its broad applicability and ease of use, MolDAIS is fully compatible with any descriptor set. Incorporating richer, domain-specific descriptors (such as quantum-derived features for catalysis or electronic reactivity indices for redox chemistry) could further enhance performance and interpretability in specialized applications. In practice, such extensions may involve tailoring descriptor sets to specific tasks or materials domains, enabling domain-informed optimization without sacrificing generality. MolDAIS is also compatible with learned or parameterized descriptor sets, such as those generated by evolutionary approaches like AExOp-DCS,[100] which could further enrich the feature space and thus enhance performance on specific MPO problems.

Finally, our results suggest a broader insight: even when the "right" descriptors are not known *a priori*, it is still possible to perform competitive MPO, provided that a large and diverse library of candidate features is available and appropriate regularization is applied. By learning task-relevant subspaces in a data-driven fashion, MolDAIS demonstrates how sparsity and adaptability can bridge the gap between generic descriptors and tailored predictive performance. A particularly interesting direction for future work is to systematically evaluate how well this framework generalizes across different core chemical scaffolds (an issue commonly studied with scaffold-based train/test splits in benchmarks such as MoleculeNet[101]), which is important for, *e.g.*, "scaffold hopping" in drug discovery.[102] Overall, we believe this work establishes a flexible and probabilistically grounded framework for low-data MPO, with clear pathways for future development and deployment in data-driven chemistry and materials science.

Author contributions

FS: conceptualization (co-lead), methodology (lead), data curation (lead), software (lead), investigation (lead), visualization (lead), formal analysis (lead), writing – original draft (lead), writing – review & editing (lead) TB: methodology (supporting), data curation (supporting), software (supporting), investigation (supporting), visualization (supporting), formal analysis (supporting), writing – review & editing (supporting) JAP: conceptualization (co-lead), methodology (supporting), investigation (supporting), writing – review & editing (supporting), supervision (lead), resources (lead), funding acquisition (lead).

Conflicts of interest

There are no conflicts to declare.

Data availability

The MolDAIS code is openly available at **https://github.com/PaulsonLab/MolDAIS** and archived on Zenodo (DOI: **https://zenodo.org/records/16944671**), with version 1.0.0 used in this study. The repository includes optimization results for all replicates across all problems, except for log *P*. Optimization results for the log *P* problem are available at **https://doi.org/10.6084/m9.figshare.25506292.v1**. Datasets for the PCE, log *P*, EC50, and Lipophilicity problems are available at **https://doi.org/10.6084/m9.figshare.25506295.v1**. Datasets for the Photoswitch, FreeSolv, and ESOL problems are available at **https://doi.org/10.6084/m9.figshare.29983504**. Datasets for the quinone redox potential and solvation free energy benchmarks are available at **https://doi.org/10.6084/m9.figshare.29602217.v1**.

Supplementary information is available that shows the molecular property distributions for the benchmark problems, additional details on the hyperparameter selection for the approximate SAAS variants, and further analysis on the PCE case study. See DOI: **https://doi.org/10.1039/d5dd00188a**.

Acknowledgements

Notes and references

1 L. Cheng, R. S. Assary, X. Qu, A. Jain, S. P. Ong, N. N. Rajput, K. Persson and L. A. Curtiss, *J. Phys. Chem. Lett.*, 2015, **6**, 283–291.

2 G. Agarwal, H. A. Doan, L. A. Robertson, L. Zhang and R. S. Assary, *Chem. Mater.*, 2021, **33**, 8133–8144.

3 Q. Zhang, A. Khetan, E. Sorkun, F. Niu, A. Loss, I. Pucher and S. Er, *Energy Storage Mater.*, 2022, **47**, 167–177.

4 A. Jain, I. A. Shkrob, H. A. Doan, L. A. Robertson, L. Zhang and R. S. Assary, *Digital Discovery*, 2023, **2**, 1197–1208.

5 G. Huang, F. Huang and W. Dong, *Chem. Eng. J.*, 2024, 152294.

6 I. Khanna, *Drug Discovery Today*, 2012, **17**, 1088–1102.

7 G. Schneider, *Nat. Rev. Drug Discovery*, 2018, **17**, 97–113.

8 L. Patel, T. Shukla, X. Huang, D. W. Ussery and S. Wang, *Molecules*, 2020, **25**, 5277.

9 S. Dara, S. Dhamercherla, S. S. Jadav, C. M. Babu and M. J. Ahsan, *Artif. Intell. Rev.*, 2022, **55**, 1947–1999.

10 G. J. Hutchings, *J. Mater. Chem.*, 2009, **19**, 1222–1235.

11 B. R. Goldsmith, J. Esterhuizen, J.-X. Liu, C. J. Bartel and C. Sutton, *AIChE J.*, 2018, **64**, 2311–2323.

12 A. K. Cheetham, R. Seshadri and F. Wudl, *Nat. Synth.*, 2022, **1**, 514–520.

13 L. Kavalsky, V. I. Hegde, B. Meredig and V. Viswanathan, *Digital Discovery*, 2024, **3**, 999–1010.

14 H. Haick and D. Cahen, *Acc. Chem. Res.*, 2008, **41**, 359–366.

15 L. Sun, Y. A. Diaz-Fernandez, T. A. Gschneidtner, F. Westerlund, S. Lara-Avila and K. Moth-Poulsen, *Chem. Soc. Rev.*, 2014, **43**, 7378–7411.

16 P. Friederich, A. Fediai, S. Kaiser, M. Konrad, N. Jung and W. Wenzel, *Adv. Mater.*, 2019, **31**, 1808256.

17 J. Kang, J. B.-H. Tok and Z. Bao, *Nat. Electron.*, 2019, **2**, 144–150.

18 J.-N. Kim, J. Lee, H. Lee and I.-K. Oh, *Nano Energy*, 2021, **82**, 105705.

19 J. Park, F. Sorourifar, M. R. Muthyala, A. M. Houser, M. Tuttle, J. A. Paulson and S. Zhang, *J. Am. Chem. Soc.*, 2024, **146**, 31230–31239.

20 J. M. Blacquiere, *ACS Catal.*, 2021, **11**, 5416–5437.

21 S. Liu, J. Zeng, Z. Wu, H. Hu, A. Xu, X. Huang, W. Chen, Q. Chen, Z. Yu, Y. Zhao, R. Zhang, Z. Zhang, P. Zhou and G. Liu, *Nat. Commun.*, 2023, **14**, 7655.

22 R. Pollice, G. dos Passos Gomes, M. Aldeghi, R. J. Hickman, M. Krenn, C. Lavigne, M. Lindner-D'Addario, A. Nigam, C. T. Ser, Z. Yao and A. Aspuru-Guzik, *Acc. Chem. Res.*, 2021, **54**, 849–860.

23 C. Bilodeau, W. Jin, T. Jaakkola, R. Barzilay and K. F. Jensen, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2022, **12**, e1608.

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

25 M. Krenn, F. Häse, A. Nigam, P. Friederich and A. Aspuru-Guzik, *Mach. Learn.: Sci. Technol.*, 2020, **1**, 045024.

26 O. Wieder, S. Kohlbacher, M. Kuenemann, A. Garon, P. Ducrot, T. Seidel and T. Langer, *Drug Discovery Today: Technol.*, 2020, **37**, 1–12.

27 A. Cereto-Massagué, M. J. Ojeda, C. Valls, M. Mulero, S. Garcia-Vallvé and G. Pujadas, *Methods*, 2015, **71**, 58–63.

28 F. Sorourifar, T. Banker and J. A. Paulson, Accelerating Black-Box Molecular Property Optimization by Adaptively Learning Sparse Subspaces, *arXiv*, 2024, arXiv:2401.0139, DOI: **10.48550/arXiv.2401.01398**.

29 T. Morishita and H. Kaneko, *ACS Omega*, 2023, **8**, 33032–33038.

30 M. J. Kusner, B. Paige and J. M. Hernandez-Lobato, *International Conference on Machine Learning*, 2017, pp. 1945–1954.

31 R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams and A. Aspuru-Guzik, *ACS Cent. Sci.*, 2018, **4**, 268–276.

32 A. Tripp, E. Daxberger and J. M. Hernández-Lobato, *Adv. Neural Inf. Process. Syst.*, 2020, 11259–11272.

33 N. T. Maus, H. T. Jones, J. S. Moore, M. J. Kusner, J. Bradshaw and J. R. Gardner, *Proceedings of the 36th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2024.

34 O. Ronen, A. I. Humayun, R. Balestriero, R. Baraniuk and B. Yu, Mitigating over-exploration in latent space optimization using LES, *arXiv*, 2024, arXiv:2406.09657, DOI: **10.48550/arXiv.2406.09657**.

35 S. Singh and J. M. Hernández-Lobato, *Commun. Chem.*, 2024, **7**, 136.

36 B. Shahriari, K. Swersky, Z. Wang, R. P. Adams and N. De Freitas, *Proc. IEEE*, 2015, **104**, 148–175.

37 P. I. Frazier, *A Tutorial on Bayesian Optimization*, 2018.

38 J. A. Paulson and C. Tsay, *Curr. Opin. Green Sustainable Chem.*, 2024, 100983.

39 R.-R. Griffiths, L. Klarner, H. Moss, A. Ravuri, S. Truong, Y. Du, S. Stanton, G. Tom, B. Rankovic, A. Jamasb, et al., *Adv. Neural Inf. Process. Syst.*, 2023, **36**, 76923–76946.

40 D. Eriksson and M. Jankowiak, *Uncertainty in Artificial Intelligence*, 2021, pp. 493–503.

41 C. K. Williams and C. E. Rasmussen, *Gaussian processes for machine learning*, MIT Press, Cambridge, MA, 2006, vol. 2.

42 D. R. Jones, M. Schonlau and W. J. Welch, *J. Global Optim.*, 1998, **13**, 455–492.

43 N. Srinivas, A. Krause, S. M. Kakade and M. Seeger, *arXiv*, 2009, preprint, arXiv:0912.3995, DOI: **10.48550/arXiv.0912.3995**.

44 H. J. Kushner, *J. Fluids Eng.*, 1964, **86**, 97–106.

45 A. Fout, J. Byrd, B. Shariat and A. Ben-Hur, *Adv. Neural Inf. Process. Syst.*, 2017, 6533–6542.

46 P. Reiser, M. Neubert, A. Eberhard, L. Torresi, C. Zhou, C. Shao, H. Metni, C. van Hoesel, H. Schopmans, T. Sommer and P. Friederich, *Commun. Mater.*, 2022, **3**, 93.

47 P. Schwaller, D. Probst, A. C. Vaucher, V. H. Nair, D. Kreutter, T. Laino and J.-L. Reymond, *Nat. Mach. Intell.*, 2021, **3**, 144–152.

48 J. L. Durant, B. A. Leland, D. R. Henry and J. G. Nourse, *J. Chem. Inf. Comput. Sci.*, 2002, **42**, 1273–1280.

49 D. Rogers and M. Hahn, *J. Chem. Inf. Model.*, 2010, **50**, 742–754.

50 R.-R. Griffiths, J. L. Greenfield, A. R. Thawani, A. R. Jamasb, H. B. Moss, A. Bourached, P. Jones, W. McCorkindale, A. A. Aldrick, M. J. Fuchter and A. A. Lee, *Chem. Sci.*, 2022, **13**, 13541–13551.

51 H. Moriwaki, Y.-S. Tian, N. Kawashita and T. Takagi, *J. Cheminf.*, 2018, **10**, 4.

52 C. W. Yap, *J. Comput. Chem.*, 2011, **32**, 1466–1474.

53 A. Mauri, V. Consonni, M. Pavan and R. Todeschini, *Match*, 2006, **56**, 237–248.

54 J. E. Terán, Y. Marrero-Ponce, E. Contreras-Torres, C. R. García-Jacas, R. Vivas-Reyes, E. Terán and F. J. Torres, *Sci. Rep.*, 2019, **9**, 11391.

55 C. R. García-Jacas, Y. Marrero-Ponce, R. Vivas-Reyes, J. Suárez-Lezcano, F. Martinez-Rios, J. E. Terán and L. Aguilera-Mendoza, *J. Comput. Chem.*, 2020, **41**, 1209–1227.

56 R. Izquierdo, R. Zadorosny, M. Rosales, Y. Marrero-Ponce and N. Cubillan, *ACS Omega*, 2025, **10**, 18312–18331.

57 Y. Marrero-Ponce, O. M. Santiago, Y. M. López, S. J. Barigye and F. Torrens, *J. Comput.-Aided Mol. Des.*, 2012, **26**, 1229–1246.

58 B. Zhang, X. Zhang, W. Du, Z. Song, G. Zhang, G. Zhang, Y. Wang, X. Chen, J. Jiang and Y. Luo, *Proc. Natl. Acad. Sci. U. S. A.*, 2022, **119**, e2212711119.

59 S. J. Barigye, Y. Marrero-Ponce, V. Alfonso-Reguera and F. Pérez-Giménez, *Chem. Phys. Lett.*, 2013, **570**, 147–152.

60 R. Guha and D. Velegol, *J. Cheminf.*, 2023, **15**, 54.

61 Y. J. Lee, H. Kahng and S. B. Kim, *Mol. Inf.*, 2021, **40**, 2100045.

62 A. G. Wilson, Z. Hu, R. Salakhutdinov and E. P. Xing, *Artificial Intelligence and Statistics*, 2016, pp. 370–378.

63 S. W. Ober, C. E. Rasmussen and M. van der Wilk, *Uncertainty in Artificial Intelligence*, 2021, pp. 1206–1216.

64 L. F. Kozachenko and N. N. Leonenko, *Probl. Inf. Transm.*, 1987, **23**, 9–16.

65 B. C. Ross, *PLoS One*, 2014, **9**, e87357.

66 D. N. Reshef, Y. A. Reshef, H. K. Finucane, S. R. Grossman, G. McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher and P. C. Sabeti, *Science*, 2011, **334**, 1518–1524.

67 J. B. Kinney and G. S. Atwal, *Proc. Natl. Acad. Sci. U. S. A.*, 2014, **111**, 3354–3359.

68 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and É. Duchesnay, *J. Mach. Learn. Res.*, 2011, **12**, 2825–2830.

69 D. Albanese, M. Filosi, R. Visintainer, S. Riccadonna, G. Jurman and C. Furlanello, *Bioinformatics*, 2013, **29**, 407–408.

70 J. G. Topliss, *J. Med. Chem.*, 1972, **15**, 1006–1011.

71 A. Cherkasov, E. N. Muratov, D. Fourches, A. Varnek, I. I. Baskin, M. Cronin, J. Dearden, P. Gramatica, Y. C. Martin, R. Todeschini, V. Consonni, V. E. Kuz'min, R. Cramer, R. Benigni, C. Yang, J. Rathman, L. Terfloth, J. Gasteiger, A. M. Richard and A. Tropsha, *J. Med. Chem.*, 2014, **57**, 4977–5010.

72 M. Balandat, B. Karrer, D. Jiang, S. Daulton, B. Letham, A. G. Wilson and E. Bakshy, *Adv. Neural Inf. Process. Syst.*, 2020, **33**, 21524–21538.

73 D. Krishnamoorthy and F. J. Doyle, *IEEE Control Syst. Lett.*, 2022, **6**, 2834–2839.

74 F. Berkenkamp, A. Krause and A. P. Schoellig, *Mach. Learn.*, 2023, **112**, 3713–3747.

75 K. J. Chan, J. A. Paulson and A. Mesbah, *2023 62nd IEEE Conference on Decision and Control (CDC)*, 2023, pp. 4106–4111.

76 C. Lu and J. A. Paulson, *IFAC-PapersOnLine*, 2022, **55**, 895–902.

77 W. Xu, Y. Jiang, B. Svetozarevic and C. Jones, *International Conference on Machine Learning*, 2023, pp. 38485–38498.

78 L. J. Klarner, GAUCHE: Gaussian Process Benchmarking Library for Chemistry, https://github.com/leojklarner/gauche, 2025, Accessed: 2025-05-08.

79 C. Zhu, R. H. Byrd, P. Lu and J. Nocedal, *ACM Trans. Math Software*, 1997, **23**, 550–560.

80 M. D. Hoffman, A. Gelman, et al., *J. Mach. Learn. Res.*, 2014, **15**, 1593–1623.

81 D.-S. Cao, J.-C. Zhao, Y.-N. Yang, C.-X. Zhao, J. Yan, S. Liu, Q.-N. Hu, Q.-S. Xu and Y.-Z. Liang, *SAR QSAR Environ. Res.*, 2012, **23**, 141–153.

82 N. Shervashidze, P. Schweitzer, E. J. van Leeuwen, K. Mehlhorn and K. M. Borgwardt, *J. Mach. Learn. Res.*, 2011, **12**, 2539–2561.

83 M. H. Rasmussen, C. Duan, H. J. Kulik and J. H. Jensen, *J. Cheminf.*, 2023, **15**, 121.

84 A. Deshwal and J. Doppa, *Adv. Neural Inf. Process. Syst.*, 2021, 8185–8200.

85 J. Hachmann, R. Olivares-Amaya, S. Atahan-Evrenk, C. Amador-Bedolla, R. S. Sánchez-Carrera, A. Gold-Parker, L. Vogt, A. M. Brockway and A. Aspuru-Guzik, *J. Phys. Chem. Lett.*, 2011, **2**, 2241–2251.

86 F.-J. Gamo, L. M. Sanz, J. Vidal, C. de Cozar, E. Alvarez, J.-L. Lavandera, D. E. Vanderwall, D. V. S. Green, V. Kumar, S. Hasan, J. R. Brown, C. E. Peishoff, L. R. Cardon and J. F. Garcia-Bustos, *Nature*, 2010, **465**, 305–310.

87 D. P. Tabor, R. Gómez-Bombarelli, L. Tong, R. G. Gordon, M. J. Aziz and A. Aspuru-Guzik, *J. Mater. Chem. A*, 2019, **7**, 12833–12841.

88 S. Chithrananda, G. Grand and B. Ramsundar, *arXiv*, 2020, preprint, arXiv:2010.09885, DOI: 10.48550/arXiv.2010.09885.

89 C. Amovilli, *Chem. Phys. Lett.*, 1994, **229**, 244–249.

90 C. Curutchet, M. Orozco, F. J. Luque, B. Mennucci and J. Tomasi, *J. Comput. Chem.*, 2006, **27**, 1769–1780.

91 A. W. Milne and M. Jorge, *J. Chem. Theory Comput.*, 2018, **15**, 1065–1078.

92 Y. Liang, Y. Jing, S. Gheytani, K.-Y. Lee, P. Liu, A. Facchetti and Y. Yao, *Nat. Mater.*, 2017, **16**, 841–848.

93 D. J. Kim, D.-J. Yoo, M. T. Otley, A. Prokofjevs, C. Pezzato, M. Owczarek, S. J. Lee, J. W. Choi and J. F. Stoddart, *Nat. Energy*, 2019, **4**, 51–59.

94 P. Poizot, J. Gaubicher, S. Renault, L. Dubois, Y. Liang and Y. Yao, *Chem. Rev.*, 2020, **120**, 6490–6557.

95 X. Yu and A. Manthiram, *Adv. Energy Sustainability Res.*, 2021, **2**, 2000102.

96 Y. Gao, G. Li, F. Wang, J. Chu, P. Yu, B. Wang, H. Zhan and Z. Song, *Energy Storage Mater.*, 2021, **40**, 31–40.

97 G. Li, L. Sun, S. Zhang, C. Zhang, H. Jin, K. Davey, G. Liang, S. Liu, J. Mao and Z. Guo, *Adv. Funct. Mater.*, 2024, **34**, 2301291.

98 Y. Ran, F. Dong, S. Sun and Y. Lei, *Adv. Energy Mater.*, 2025, 2406139.

99 S. Daulton, M. Balandat and E. Bakshy, *Adv. Neural Inf. Process. Syst.*, 2020, 9851–9864.

100 L. A. García-González, Y. Marrero-Ponce, C. A. Brizuela and C. R. García-Jacas, *Mol. Inf.*, 2023, **42**, 2200227.

101 Z. Wu, B. Ramsundar, E. N. Feinberg, J. Gomes, C. Geniesse, A. S. Pappu, K. Leswing and V. Pande, *Chem. Sci.*, 2018, **9**, 513–530.

102 H. Sun, G. Tawa and A. Wallqvist, *Drug Discovery Today*, 2012, **17**, 310–324.