

Multi-fidelity Bayesian Optimization of Covalent Organic Frameworks for Xenon/Krypton Separations

Nickolas Gantzler,[†] Aryan Deshwal,[‡] Janardhan Rao Doppa,^{*,‡} and Cory M. Simon^{*,¶}

[†]*Department of Physics. Oregon State University. Corvallis, OR, USA.*

[‡]*School of Electrical Engineering and Computer Science. Washington State University. Pullman, WA, USA.*

[¶]*School of Chemical, Biological, and Environmental Engineering. Oregon State University, Corvallis, OR, USA.*

E-mail: jana.doppa@wsu.edu; cory.simon@oregonstate.edu

Abstract

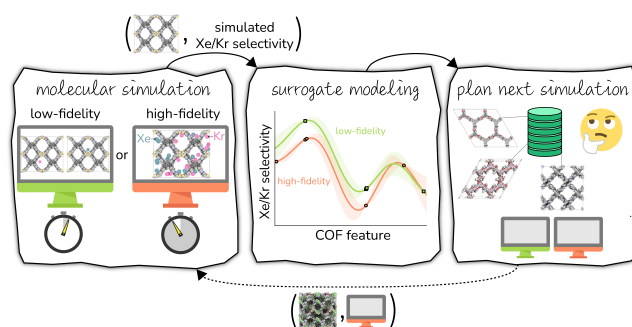
Our objective is to search a large candidate set of covalent organic frameworks (COFs) for the one with the largest equilibrium adsorptive selectivity for xenon (Xe) over krypton (Kr) at room temperature. To predict the Xe/Kr selectivity of a COF structure, we have access to two molecular simulation techniques: (1) a higher-fidelity, binary grand canonical Monte Carlo simulation and (2) a lower-fidelity Henry coefficient calculation that (a) approximates the adsorbed phase as dilute and, consequently, (b) incurs a smaller computational runtime than the higher-fidelity simulation.

To efficiently search for the COF with the largest high-fidelity Xe/Kr selectivity, we employ a multi-fidelity Bayesian optimization (MFBO) approach. MFBO constitutes a sequential, automated feedback loop of (1) conduct a low- or high-fidelity molecular simulation of Xe/Kr adsorption in a COF, (2) use the simulation data gathered thus far to train a surrogate model that cheaply predicts, with quantified uncertainty, the low- and high-fidelity simulated Xe/Kr selectivity of COFs from their structural/chemical features, and then (3) plan the next simulation (i.e., choose the next COF and fidelity) in consideration of balancing

exploration, exploitation, and cost.

We find that MFBO acquires the optimal COF among the candidate set of 609 structures using only 38 low-fidelity and nine high-fidelity simulations, incurring only 2.5%, 5% on average, and 18% on average of the computational runtime of an exhaustive, random, and single-fidelity BO search, respectively.

TOC image:



Introduction

Bayesian optimization for materials discovery

The discovery and development of new materials is vital for both sustaining and technologically-advancing our society. Computational meth-

ods, including electronic structure calculations, molecular simulations, and materials informatics/machine learning, can predict the properties of materials and thus be employed to optimize, screen, and design new materials rapidly and cost-effectively—accelerating the rate of materials optimization and discovery.^{1–5}

Bayesian optimization (BO)^{6–9} combines supervised machine learning, uncertainty quantification, and decision-making algorithms to automatically and efficiently design a sequence of experiments, in the lab or a computer simulation, to find materials with an optimal property for some application.¹⁰ Given (i) a pool or space¹¹ of candidate materials and (ii) an experimental protocol (in the lab or a simulation) to measure/evaluate/predict the relevant property of a material, BO iteratively designs experiments (i.e., chooses materials to subject to an experiment) to find the optimal material with the fewest (costly) experiments. The two ingredients of BO for automated experiment planning are:

- a *surrogate¹ model*, a supervised machine learning model that computationally predicts—inexpensively, and with quantified uncertainty—the property of any material from its compositional, chemical, and/or structural features.
- an *acquisition function*, which uses the surrogate model to score each material according to its utility for the next experiment. The acquisition function is designed to balance (i) exploitation (“acquire a material with the optimal predicted property”) to greedily pursue the material we believe may be optimal with the limited information we currently possess and (ii) exploration (“acquire a material whose predicted property is highly uncertain”) to gather more information about the structure-property relationship.

The “experiment-analysis-plan” feedback loop¹² that constitutes BO (see Fig. 1) iterates through (i) conduct an experiment to obtain a (material,

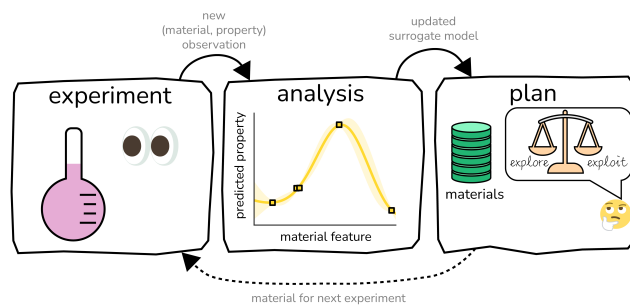


Figure 1: Standard **Bayesian optimization (BO) of materials** constitutes a feedback loop of (i) conduct an experiment, (ii) analyze the data collected thus far to construct a surrogate model of the experiment, and (iii) plan the next experiment in consideration of balancing exploration and exploitation.

property) observation, (ii) update the surrogate model in light of this new experimental data, then (iii) select the next material for an experiment by maximizing the acquisition function. Compared to random search, BO leverages the surrogate model to make principled decisions balancing the exploration-exploitation trade-off so as to uncover the optimal material early in the sequential search. Because the acquisition function and optimization algorithm negate the need for *humans* to design the experiments inside the experiment-analysis-plan feedback loop, BO can orchestrate autonomous, “self-driving” labs^{12–18} that employ automated instrumentation and/or robots to conduct a sequence of experiments with the goal of resource-efficient materials discovery and optimization.

BO has been deployed for the optimization and discovery of many different materials^{19–22} in the lab or a computer simulation, including nanoporous materials,^{23–26} nanoparticles,²⁷ light emitting diodes,²⁸ carbon nanotubes,²⁹ photovoltaics,^{30–32} additively manufactured structures,³³ polymers,^{34–38} thermoelectrics,³⁹ antimicrobial active surfaces,⁴⁰ quantum dots,⁴¹ luminescent materials,⁴² catalysts,^{43–45} thin films,⁴⁶ and solid chemical propellants.⁴⁷ More, BO has been used to optimize processes to synthesize materials and chemicals^{48–51,51} or to employ materials for an industrial-scale task.⁵²

¹“surrogate” for the experiment

Multi-fidelity Bayesian optimization for materials discovery

Often, we have multiple options of different experiments to measure/evaluate/predict the relevant property of the material—experiments that trade (1) fidelity, i.e. the extent to which the experiment faithfully measures/evaluates/predicts the property of the material, for (2) affordability. For example, a computer simulation is usually a low-fidelity and -cost estimation of the material property compared to a high-fidelity and -cost measurement of the material property in the laboratory.

Multi-fidelity Bayesian optimization (MFBO)^{9,53} takes advantage of multiple types of experiments that trade-off fidelity and affordability to search for a material with an optimal property while incurring the minimal cost.⁵⁴ MFBO modifies the experiment-analysis-plan loop of standard BO in Fig. 1 by extending: (i) the surrogate model, to (a) predict the property of materials according to experiments of *all fidelities* and (b) capture the correlations between the material properties according to each experimental fidelity, enabling observed outcomes of low-fidelity experiments to inform predicted outcomes of high-fidelity experiments, and (ii) the acquisition function, to pick the next material *and* the next experimental fidelity, while balancing exploration, exploitation, *and* the cost of the different experiments. In turn, MFBO leverages low-fidelity experiments to cheaply scope out which regions of materials space contain (i) poor-performing materials, to avoid wasting resources on high-fidelity experiments there, and (ii) high-performing materials, to focus high-fidelity experiments there. MFBO (or its parent, multi-information-source BO⁵⁵) has been scarcely applied to materials discovery.^{54,56–58}

Our contribution

In this work, we employ MFBO to search a pool of ~600 covalent organic framework (COF) crystal structures⁵⁹ for the one with the highest simulated xenon/krypton selectivity at room temperature, while incurring the minimal computational

expense. We are armed with two molecular simulation methods to predict the Xe/Kr selectivity of a COF: (higher-fidelity & -cost) Markov-chain Monte Carlo simulation of the binary grand-canonical ensemble, where the COF hosts multiple adsorbates (both Xe and Kr) during the simulation; and, (lower-fidelity & -cost) Monte Carlo integration to calculate the Xe and Kr Henry coefficients in the COF, which makes the dilute approximation, so the COF hosts only a single adsorbate during the simulation. Our MFBO routine employs (i) a multi-fidelity Gaussian process (GP)⁶⁰ surrogate model to predict the simulated Xe/Kr selectivity of a COF from its structural and chemical features and (ii) a cost-aware, multi-fidelity expected improvement⁶¹ acquisition function to design the next simulation. MFBO acquires the COF with the largest high-fidelity simulated Xe/Kr selectivity using only 38 low- and nine high-fidelity simulations, incurring only 2.5%, 5% on average, and 18% on average of the computational run time of a high-fidelity exhaustive search, random search, and single-fidelity BO, respectively. More, MFBO robustly out-performs single-fidelity BO, over randomly chosen COFs used to initialize the surrogate model. Our results demonstrate the promise of MFBO to cost-effectively discover materials for a variety of applications when in possession of multiple options of laboratory experiments and/or computer simulations, that trade fidelity for affordability, to measure/evaluate/predict the property of materials.

Box 1: COFs for Xe/Kr Separations

Xe/Kr separations. The noble gases xenon (Xe) and krypton (Kr) have many uses/applications (e.g. lighting, insulation in multi-pane windows, propellant for ion thrusters, anesthesia, and imaging).^{62,63} The majority of Xe and Kr production is via their isolation from air (abundance: Xe, 0.09 ppm, Kr, 1.1 ppm⁶²) via distillation at cryogenic temperatures. Particularly, the production of pure O₂ and N₂ from air via cryogenic distillation produces a byproduct stream enriched with both Xe and Kr; this mixture is then

subject to an additional cryogenic distillation to obtain pure Xe and Kr.^{62,63} Note, distillation exploits the difference in boiling points of Xe and Kr, -108.1°C and -153.2°C , respectively, to separate them.⁶⁴

COFs. Covalent organic frameworks (COFs) are nanoporous, crystalline materials composed of organic molecules linked by covalent bonds to form an extended (2D or 3D) network. COFs tend to exhibit high internal surface areas and chemical and thermal stability.^{65,66} More, the modular nature of COF synthesis, as well as their post-synthetic modifiability, enables a vast number of different COF structures to be realized.

COFs for Xe/Kr separations. As opposed to energy-intensive cryogenic distillation, nanoporous materials, such as COFs, could be used to more efficiently separate Xe from Kr, at room temperature, via selective adsorption.^{64,67} Much research is focused on (i) experimentally synthesizing^{68–70} or (ii) computationally designing,^{71–82} using molecular simulations of adsorption, nanoporous materials for Xe/Kr separations—i.e., materials with high Xe/Kr selectivity, Xe capacity, stability, and fast adsorption kinetics.

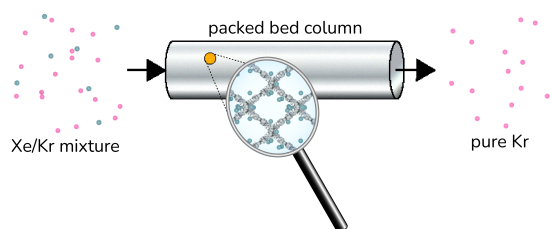


Illustration of an idealized COF-based Xe/Kr separation. A column is packed with COF adsorbent material. The Xe/Kr mixture is fed to the column. The COF selectively adsorbs the Xe, letting the Kr pass through the column. After the adsorbent is saturated with Xe, heating or pulling vacuum desorbs the Xe in the COF and regenerates it for another cycle of adsorption.



Results

Problem setup

We possess a candidate set \mathcal{X} of 609 experimentally-reported covalent organic frameworks (COFs)⁵⁹ for the task of Xe/Kr separations.

Our objective is to find the COF $\mathbf{x}^* \in \mathcal{X}$ that exhibits the highest equilibrium adsorptive Xe/Kr selectivity ($:= y$) when immersed in a 20 mol%/80 mol% Xe/Kr mixture at 1 bar and 298 K.



To computationally predict the Xe/Kr selectivity of a COF, we are armed with two different molecular simulation techniques. Each molecular simulation employs Lennard-Jones interatomic potentials (parameters from Universal Force Field⁸³) to describe the potential energy of a configuration of a rigid COF hosting Xe and/or Kr adsorbate(s). Given a COF, our choice of which simulation to perform to predict its Xe/Kr selectivity involves a trade-off between fidelity and computational run-time.

  High-fidelity (fidelity parameter $\ell := \frac{2}{3}$) simulation. Run-time: ca. 230 min.

The high-fidelity simulation constitutes a Markov chain Monte Carlo (MC) simulation of the COF in the binary grand-canonical (BGC) ensemble. During the molecular simulation, generally the COF hosts both and multiple Xe and Kr adsorbates; these adsorbates enter/leave the COF from/to the gas phase and move around in the pores of the COF to explore configurations. The key measurable during the BGC simulation is the average number of adsorbates in the COF system, $\langle \mathbf{n} \rangle$, with $\mathbf{n} := [n_{\text{Xe}}, n_{\text{Kr}}]$. Our high-fidelity prediction of the adsorptive Xe/Kr selectivity of the COF is then

$$y^{(2/3)} = \frac{\langle n_{\text{Xe}} \rangle / \langle n_{\text{Kr}} \rangle}{p_{\text{Xe}} / p_{\text{Kr}}}, \quad (1)$$

with partial pressures in the gas phase $p_{\text{Kr}} = 0.8$ bar and $p_{\text{Xe}} = 0.2$ bar.

  Low-fidelity ($\ell := \frac{1}{3}$) simulation. Run-time: ca. 15 min. The low-fidelity

prediction of the Xe/Kr selectivity of a COF relies on the dilute approximation in the BGC ensemble and models adsorption in the COF with Henry's law

$$\langle \mathbf{n} \rangle = \begin{bmatrix} H_{\text{Xe}} & 0 \\ 0 & H_{\text{Kr}} \end{bmatrix} \mathbf{p}, \quad (2)$$

with $\mathbf{p} := [p_{\text{Xe}}, p_{\text{Kr}}]$. We compute the Henry coefficients of Xe and Kr in the COF, H_{Xe} and H_{Kr} , via two separate ordinary MC integrations. The dilute approximation assumes the density of adsorbed gas in the COF is sufficiently small (i.e., small \mathbf{p}) to justify neglecting adsorbate-adsorbate interactions; consequently, the COF hosts only a single adsorbate during each Henry coefficient simulation—making it computationally cheaper than a BGCMC simulation. Our low-fidelity prediction of the Xe/Kr selectivity of the COF, then, is the ratio of the Henry coefficients

$$y^{(1/3)} = \frac{H_{\text{Xe}}}{H_{\text{Kr}}}, \quad (3)$$

which follows from eqn. 1 when Henry's law in eqn. 2 holds.

See Methods for details about both molecular simulation techniques.

Given access to (only) these two molecular simulation techniques that trade fidelity and computational runtime, we reframe the objective as:

🎯 find the COF $\mathbf{x}^* \in \mathcal{X}$ with the highest adsorptive Xe/Kr selectivity according to the high-fidelity BGCMC simulation, $y^{(2/3)}$, while incurring the minimal computational cost, measured by the sum of run times of the (both low- and high-fidelity) simulations we conduct to find \mathbf{x}^* .

Multi-fidelity Bayesian optimization (MFBO) of COFs for Xe/Kr separations

We provide an overview of multi-fidelity Bayesian optimization (MFBO) to efficiently find the COF with the largest high-fidelity Xe/Kr selectivity.

Defining COF space (Fig. 2)

For surrogate modeling, we must define a space in which we mathematically represent each COF as a point in a continuous space.¹¹ Inspired by several computational studies revealing the structure-property relationships of porous materials for Xe/Kr separations,^{71,75,79,84} we elected to represent each COF with a vector $\mathbf{x} \in \mathbb{R}^{14}$ that lies in a continuous space, listing its following structural (computed from Zeo++⁸⁵) and compositional features derived from its crystal structure: density, gravimetric surface area, void fraction, largest included sphere diameter, and elemental fractions of metals, halogens, phosphorus, sulfur, nitrogen, silicon, hydrogen, carbon, oxygen, and boron. See Fig. 2. We min-max normalized the features.

An equation-free overview of MFBO (Fig. 3)

MFBO constitutes a simulation-analysis-plan feedback loop and results in a machine-curated sequence of high- and low-fidelity molecular simulations of Xe/Kr adsorption in candidate COFs. Fig. 3 illustrates the feedback loop. The algorithms inside the loop are designed to minimize the computational runtime expended until we find the COF with the largest high-fidelity simulated Xe/Kr selectivity.

1 **Simulation.** We conduct either a low- or high-fidelity simulation of Xe/Kr adsorption in a COF structure to obtain its predicted Xe/Kr selectivity. This generates a new data point—a COF structure “labeled” with its simulated Xe/Kr selectivity under that fidelity.

2 **Analysis.** We use this new data point to update our *surrogate model* of the simulations. This surrogate model is a supervised machine learning model that can, with negligible computational runtime, predict both the low- and high-fidelity simulated Xe/Kr selectivity of a COF not simulated before—and quantify uncertainty the uncertainty in this prediction. The inputs to the surrogate model for its prediction about a COF are (cheaply computed) structural and chemical features of its crystal structure. The surrogate model is trained on all labeled data—i.e.,

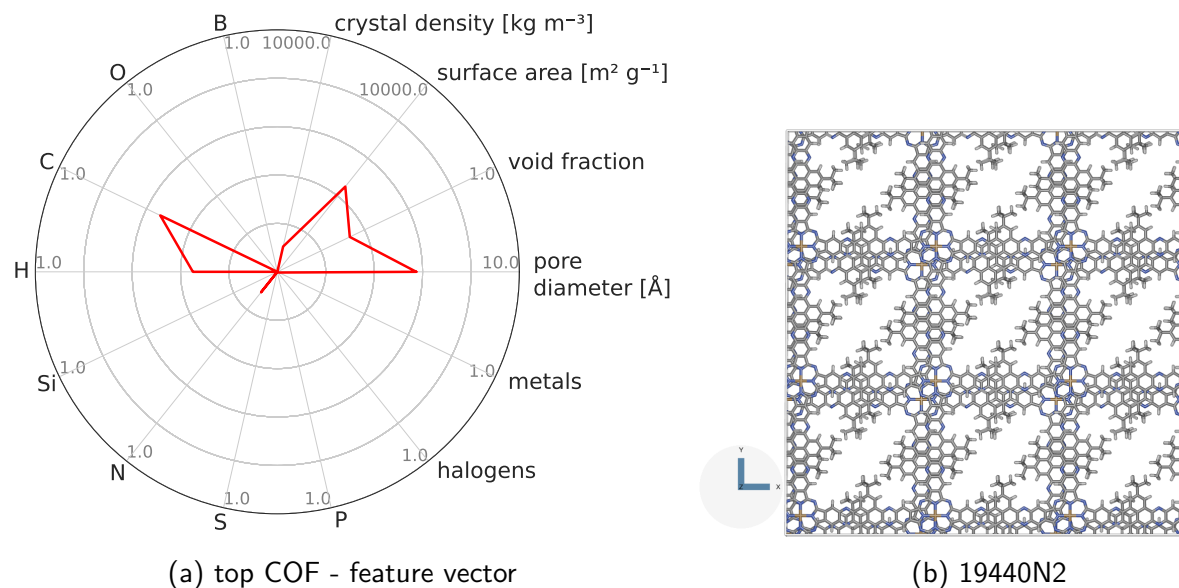


Figure 2: **Defining COF space.** We represent each COF with a vector of four structural and ten compositional features. For example, the radar plot in (a) visualizes the raw feature vector \mathbf{x} of the COF whose crystal structure is in (b).

all (COF features, simulated Xe/Kr selectivity) pairs—gathered from simulations we have conducted thus far in the search. Thus, the surrogate model summarizes our knowledge, thus far in the search, about (i) the relationship between (a) the structural and chemical features of the COFs and (b) their simulated Xe/Kr selectivity and (ii) correlations between the low- and high-fidelity simulated Xe/Kr selectivities.

3 Plan. Completing the loop, we judiciously select the (a) COF and (b) fidelity for the next simulation. An acquisition function relies on the surrogate model to score each (COF, fidelity) pair according to its appeal for the next simulation; the plan for the new simulation follows from the (COF, fidelity) pair with the maximal score. The acquisition function is designed to balance three often competing desires: (i) exploitation, to select a COF that the surrogate model predicts to have a large high-fidelity simulated Xe/Kr selectivity; (ii) exploration, to select a COF with a high-fidelity simulated Xe/Kr selectivity about which the surrogate model is highly uncertain; and (iii) cost reduction, which incentivizes choosing a low-fidelity simulation that provides useful but incomplete information about the high-fidelity selectivity.

● In practice, we cannot know for certain when we have recovered the optimal COF. Possible strategies to terminate the iterative MFBO search include when: (i) computational resources are exhausted, (ii) a COF with a sufficiently large high-fidelity Xe/Kr selectivity has been recovered, or (iii) a large runtime has elapsed since we last discovered a COF with an improved Xe/Kr selectivity over those COFs we have acquired thus far.

The multi-fidelity surrogate model

Our multi-fidelity surrogate model treats the fidelity- $\ell \in \{\frac{1}{3}, \frac{2}{3}\}$ simulated Xe/Kr selectivity of a COF represented by \mathbf{x} , $y^{(\ell)} \in \mathbb{R}$, as a realization of a random variable $Y^{(\ell)}(\mathbf{x})$. The surrogate model specifies a probability density for $Y^{(\ell)}(\mathbf{x})$. Under this Bayesian perspective, the posterior probability density of $Y^{(\ell)}(\mathbf{x}) \mid \mathcal{D}_{[n]}$ at iteration n of the MFBO search reflects our beliefs, grounded by the simulation data collected thus far in the search,

$$\mathcal{D}_{[n]} := \{([\mathbf{x}_{[1]}, \ell_{[1]}], y_{[1]}], \dots, ([\mathbf{x}_{[n]}, \ell_{[n]}], y_{[n]})\}, \quad (4)$$

about the fidelity- ℓ simulated Xe/Kr selectivity of the COF represented by \mathbf{x} . This density concentrates in the region of the line where we believe the selectivity of the COF lies, and the spread of this

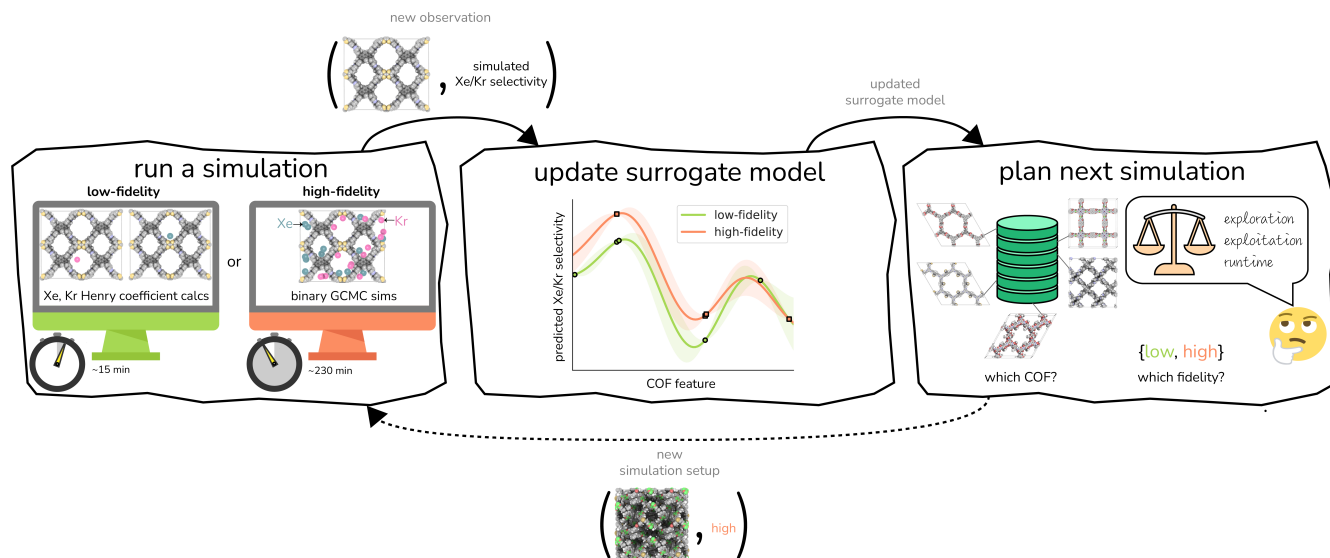


Figure 3: **Multi-fidelity Bayesian optimization of COFs for Xe/Kr separations** constitutes an iterative, machine-orchestrated feedback loop of (i) molecular simulation, (ii) updating the multi-fidelity surrogate model of the simulations, and (iii) planning the next simulation.

density reflects our uncertainty about its selectivity. Intuitively, the mean of the posterior density of the conditional random variable $Y^{(\ell)}(\mathbf{x}) \mid \mathcal{D}_{[n]}$ is a point-prediction of the fidelity- ℓ Xe/Kr selectivity of COF \mathbf{x} , and the variance of it is a measure of our uncertainty about the predicted selectivity.

We adopt a multi-fidelity Gaussian process (GP)^{60,86,87} surrogate model:

$$Y^{(\ell)}(\mathbf{x}) \sim \mathcal{GP}(0, k([\mathbf{x}, \ell], [\mathbf{x}', \ell'])) \quad (5)$$

with a kernel function between two simulation setups (\mathbf{x}, ℓ) and (\mathbf{x}', ℓ') as a scaled (by factor α , a hyperparameter) product of a material and fidelity kernel function:

$$k([\mathbf{x}, \ell], [\mathbf{x}', \ell']) = \alpha k_{\text{mat}}(\mathbf{x}, \mathbf{x}') k_{\text{fid}}(\ell, \ell'), \quad (6)$$

with

$$k_{\text{mat}}(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{1}{2}\|\mathbf{x} - \mathbf{x}'\|^2/\gamma^2\right) \quad (7)$$

$$k_{\text{fid}}(\ell, \ell') = c + (1 - \ell)^{1+\delta}(1 - \ell')^{1+\delta}. \quad (8)$$

- The material kernel function $k_{\text{mat}} : \mathbb{R}^{14} \times \mathbb{R}^{14} \rightarrow \mathbb{R}$ is a squared exponential kernel with a length-scale hyperparameter γ . Roughly, k_{mat} quantifies the similarity between any pair of COFs. If two COFs are

nearby in COF space, they are declared to be similar by the kernel.

- The fidelity kernel function $k_{\text{fid}} : \{\frac{1}{3}, \frac{2}{3}\} \times \{\frac{1}{3}, \frac{2}{3}\} \rightarrow \mathbb{R}$ is a down-sampling kernel^{60,88} with offset and power hyperparameters c and δ . Roughly, k_{fid} quantifies the similarity between any pair of simulation fidelities. It can take on only three distinct values—expressing the low-low, high-high, and low-high fidelity simulation similarities.

Empirically, GPs tend to be effective surrogate models for Bayesian optimization of molecules in the small-data regime.⁸⁹

In Methods, we explain the meaning behind the notation of the multi-fidelity GP in eqn. 5, following the Bayesian paradigm⁹⁰ of (i) specifying a prior distribution, (ii) collecting the simulation data, then (iii) updating the prior to a posterior distribution. The resulting posterior distribution is Gaussian

$$Y^{(\ell)}(\mathbf{x}) \mid \mathcal{D}_{[n]} \sim \mathcal{N}(\mu_{[n]}(\mathbf{x}, \ell), \sigma_{[n]}^2(\mathbf{x}, \ell)) \quad (9)$$

with mean

$$\mu_{[n]}(\mathbf{x}, \ell) = \mathbf{k}_{\mathcal{D}_{[n]}}^T (\mathbf{K}_{\mathcal{D}_{[n]}} + \sigma^2 \mathbf{I})^{-1} \mathbf{y}_{\mathcal{D}_{[n]}} \quad (10)$$

and variance

$$\sigma_{[n]}^2(\mathbf{x}, \ell) = k([\mathbf{x}, \ell], [\mathbf{x}, \ell]) - \mathbf{k}_{\mathcal{D}_{[n]}}^T (\mathbf{K}_{\mathcal{D}_{[n]}} + \sigma^2 \mathbf{I})^{-1} \mathbf{k}_{\mathcal{D}_{[n]}} \quad (11)$$

written in terms of

- $\mathbf{y}_{\mathcal{D}_{[n]}}$: the vector of simulated Xe/Kr selectivities of COFs we observed thus far in $\mathcal{D}_{[n]}$ (see eqn. 31)
- $\mathbf{k}_{\mathcal{D}_{[n]}}$: the vector giving the kernel between (i) the COF \mathbf{x} and fidelity ℓ in question and (ii) the COFs and fidelities $(\mathbf{x}_{[i]}, \ell_{[i]})$'s in the simulation data $\mathcal{D}_{[n]}$ (see eqn. 30)
- $\mathbf{K}_{\mathcal{D}_{[n]}}$: the matrix giving the kernel between the COFs and fidelities $(\mathbf{x}_{[i]}, \ell_{[i]})$'s in the simulation data $\mathcal{D}_{[n]}$ (see eqn. 29)
- σ^2 : the variance of the noise contaminating the outcome of our simulations (see eqn. 26).

Intuitively:

- the mean $\mu_{[n]}(\mathbf{x}, \ell)$ in eqn. 10 is a weighted combination of the observed simulated Xe/Kr selectivities $\mathbf{y}_{\mathcal{D}_{[n]}}$, with the similarity between the simulation in question (\mathbf{x}, ℓ) and the previously conducted simulations in $\mathcal{D}_{[n]}$ involved in forming the weights.
- the variance $\sigma_{[n]}^2(\mathbf{x}, \ell)$ in eqn. 11 is that of the prior reduced according to the similarity between the simulation in question (\mathbf{x}, ℓ) and the previously conducted simulations in $\mathcal{D}_{[n]}$.

The subscript $[n]$ in our notation emphasizes that the surrogate model changes with iteration n ; we expect the surrogate model to improve its predictions as the search progresses and the simulation data $\mathcal{D}_{[n]}$ grows in size.

💡 The GP in eqn. 5 is designed to (i) incorporate our domain knowledge that COFs with similar pore size, surface area, composition, etc. will tend to exhibit similar Xe/Kr selectivities and (ii) learn, from the simulation data $\mathcal{D}_{[n]}$, (a) the relationship between the structural and compositional features of COFs in \mathbf{x} and simulated Xe/Kr selectivity $y^{(\ell)}$

and (b) through the fidelity kernel, correlations between the low- and high-fidelity simulations, allowing outcomes of low-fidelity simulations to inform us about the high-fidelity Xe/Kr selectivity we ultimately wish to maximize. Fig. 3, middle panel, visualizes a toy multi-fidelity GP for a one-dimensional COF space: the dark lines show the mean function $\mu(\mathbf{x}, \ell)$; the shaded bands highlight $\mu(\mathbf{x}, \ell) \pm \sigma(\mathbf{x}, \ell)$, quantifying uncertainty by showing a credible interval for each predicted selectivity; and the points show the multi-fidelity data $\mathcal{D}_{[n]}$ on which the toy GP is trained.

Automated simulation planning

At the plan stage, the MFBO algorithm judiciously selects the next simulation setup, completing the closed loop. This simulation plan constitutes two choices: (i) the COF $\mathbf{x}_{[n+1]}$ in which to conduct simulations of Xe/Kr adsorption, and (ii) the fidelity $\ell_{[n+1]}$ of the molecular simulation. The plan is judicious because it employs (i) the surrogate model—particularly, the posterior in eqn. 9—and (ii) running estimates of the computational runtimes of the low- and high-fidelity simulations, $\{\tau_{[n]}^{(1/3)}, \tau_{[n]}^{(2/3)}\}$, to design the next simulation setup, $(\mathbf{x}_{[n+1]}, \ell_{[n+1]})$, so as to balance exploration, exploitation, and cost.

Particularly, we rely on an augmented, cost-aware expected improvement acquisition function to score the appeal of each setup (\mathbf{x}, ℓ) for the next simulation. The simulation plan follows from maximizing the acquisition function:

$$(\mathbf{x}_{[n+1]}, \ell_{[n+1]}) = \arg \max_{(\mathbf{x}, \ell) \in \mathcal{X} \times \{\frac{1}{3}, \frac{2}{3}\}} \mathbb{E} \left[\max[0, Y^{(2/3)}(\mathbf{x}) \mid \mathcal{D}_{[n]} - \hat{y}_{[n]}^{(2/3)*}] \cdot \text{corr}[Y^{(\ell)}(\mathbf{x}) \mid \mathcal{D}_{[n]}, Y^{(2/3)}(\mathbf{x}) \mid \mathcal{D}_{[n]}] \cdot \left(\frac{\tau_{[n]}^{(2/3)}}{\tau_{[n]}^{(\ell)}} \right) \right] \quad (12)$$

The acquisition function being maximized is a product of three terms:

- **Expected improvement (EI)**: the amount that the high-fidelity simulated Xe/Kr se-

lectivity of COF \mathbf{x} is expected to improve upon the largest high-fidelity Xe/Kr selectivity we observed thus far, $\hat{y}_{[n]}^{(2/3)*}$. Owing to the $\max[0, \cdot]$ operator, the integral constituting this expectation \mathbb{E} has a contribution only for predicted high-fidelity Xe/Kr selectivity $y^{(2/3)}$ greater than $\hat{y}_{[n]}^{(2/3)*}$. Because both (a) a large posterior variance $\sigma_{[n]}^2(\mathbf{x}, \frac{2}{3})$ (reflecting uncertainty) and (b) a large mean $\mu_{[n]}(\mathbf{x}, \frac{2}{3})$ will contribute density to this region, maximizing this EI term balances exploitation and exploration, by favoring COFs whose predicted high-fidelity selectivity is large and/or uncertainty.

- **Correlation with the high-fidelity selectivity:** the correlation between the simulated Xe/Kr selectivity of the COF \mathbf{x} under (i) the fidelity- ℓ simulation and (ii) a high-fidelity simulation. If $\ell = 1/3$ and this term is small (large), this simulation setup is downgraded (upgraded) because the outcome of this low-fidelity simulation cannot (can) inform us about the high-fidelity selectivity we ultimately wish to optimize.
- **Cost ratio.** The ratio of the runtime of a high-fidelity simulation to the fidelity- ℓ simulation, to promote low-fidelity simulations owing to their smaller runtime.

Owing to these three components, maximizing the acquisition function at each iteration gives a simulation plan $(\mathbf{x}_{[n+1]}, \ell_{[n+1]})$ for the next iteration with a high utility per cost for our objective of finding the COF with the largest high-fidelity Xe/Kr selectivity soon.

Maximizing the acquisition function. Because (i) the acquisition function is computationally cheap to evaluate and (ii) we are searching over a relatively small, finite set of COFs \mathcal{X} ($|\mathcal{X}| = 609$), we find $(\mathbf{x}_{[n+1]}, \ell_{[n+1]})$ at each iteration via exhaustive search.

The acquired set of COFs. We refer to the set of COFs in $\mathcal{D}_{[n]}$ at iteration n , automatically chosen by sequentially maximizing the acquisition function, as the set of *acquired* COFs.

The state of MFBO performance

We judge the quality of the MFBO search at iteration n by the largest high-fidelity simulated Xe/Kr selectivity among the acquired set of COFs in $\mathcal{D}_{[n]}$:

$$\hat{y}_{[n]}^{(2/3)*} := \max_{1 \leq i \leq n: \ell_{[i]}=2/3} y_{[i]}. \quad (13)$$

Initialization


We initiate the MFBO loop at the plan stage with a surrogate model trained on six data points $\mathcal{D}_{[6]}$: three diverse COFs “labeled” with their simulated, both low- and high-fidelity, Xe/Kr selectivities. We select the initial COF as the one closest to the center of the (normalized) COF feature space. For the two subsequent COFs, we select (2) the COF most distal in COF space from the initial COF then (3) the COF with the maximal minimum distance to the first two COFs.

MFBO performance

We now execute the MFBO loop in Fig. 3 to iteratively search for the COF with the largest high-fidelity simulated Xe/Kr selectivity.

MFBO search efficiency curve (Fig. 4)

Fig. 4 shows the search efficiency of MBFO by visualizing, as the MFBO search progresses, (i, top panel) the largest high-fidelity Xe/Kr selectivity among the COFs in which we’ve simulated with high-fidelity Xe/Kr adsorption so far thus far— $\hat{y}_{[n]}^{(2/3)*}$ in eqn. 13, and (ii, bottom panel) the accumulated computational runtime (see Methods for our compute hardware specifications). The gray region highlights the $n = 6$ simulations used to initialize the surrogate model.

 The MFBO algorithm acquires the COF \mathbf{x}^* (19440N2 = CuPc-pz COF⁹¹) with the largest high-fidelity Xe/Kr selectivity $y^{(2/3)*}$ (18.53) after conducting only 47 molecular simulations—nine high-fidelity, 38 low-fidelity—incurring a computational runtime of 58 hr. Recall, there are 609 COF candidates; thus, we recovered the top COF while circumventing many wasteful molecular simulations in non-optimal COFs.

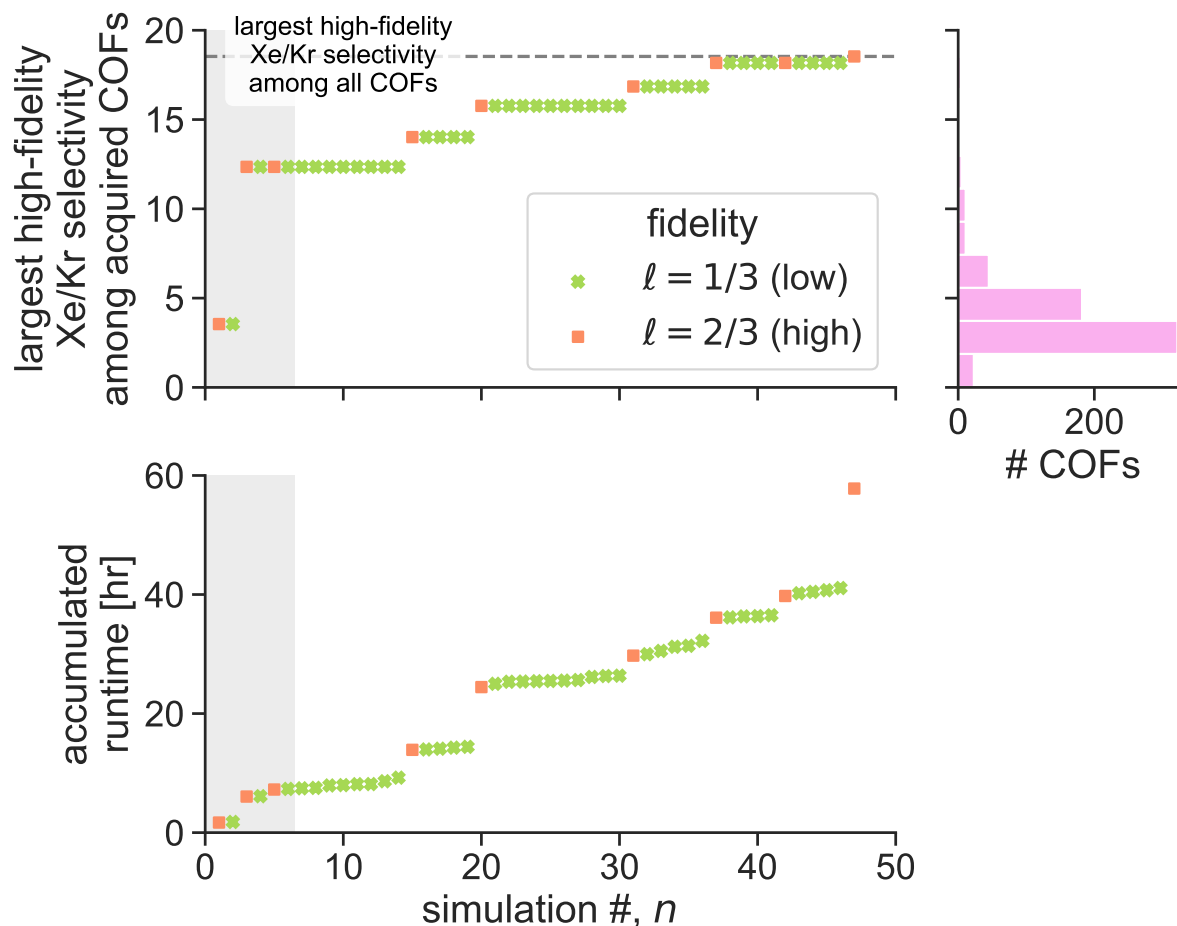


Figure 4: **MFBO search efficiency.** As the MFBO search progresses, (top) the maximum high-fidelity Xe/Kr selectivity among the *acquired* COFs and (bottom) the accumulated runtime. Different markers are used to delineate between low- and high-fidelity simulations. The gray region highlights the initialization stage. The dashed line (top panel) indicates the maximum high-fidelity selectivity over *all* COFs. For context, the histogram (top right) shows the distribution of high-fidelity selectivity over all COFs.

For context, the distribution of high-fidelity Xe/Kr selectivities for all COFs, shown in Fig. 4 (top right), is skewed right. This shows that MBFO acquired the optimal COF \mathbf{x}^* from the thin tail of the distribution.

Note the most dramatic increases in accumulated runtime owe to high-fidelity simulations. Despite that the majority of simulations performed were low-fidelity, the low-fidelity simulations only account for $\sim 15\%$ of the accumulated runtime to find the optimal COF \mathbf{x}^* .

As evidence that MFBO is allocating computational resources intelligently, (1) several low-fidelity simulations precede each high-fidelity simulation and (2) five out of six of the MFBO-

acquired COFs for high-fidelity simulations resulted in an improvement of the largest high-fidelity Xe/Kr selectivity observed.

(At the iteration preceding the acquirement of the optimal COF, Fig. S1 shows the predictivity of the surrogate model, and Fig. S2 shows the observed correlation between low- and high-fidelity selectivities.)

! Of course, in practice, we cannot know when we have recovered the optimal COF \mathbf{x}^* . For the purposes of benchmarking MFBO, for this article, we *actually* conducted both low- and high-fidelity molecular simulations in *all* COFs, so we know when we have acquired the optimal COF \mathbf{x}^* .

MFBO acquisition dynamics (Fig. 5)

To gain insight into the acquisition dynamics of MFBO, Fig. 5 visualizes the scatter of all COFs in feature space and marks the acquired set of COFs in $\mathcal{D}_{[n]}$ at six different stages of the search. Different marks are used for low- and high-fidelity simulations.


We used principal component analysis (PCA) to reduce the dimensionality of the COF feature vectors $\{\mathbf{x}_1, \dots, \mathbf{x}_{609}\}$ from 14 to two, for visualization. Each panel in Fig. 5 shows the first two principal components of COF feature space; each point represents a COF, colored according to its high-fidelity Xe/Kr selectivity. Note, the COFs with the largest high-fidelity Xe/Kr selectivities tend to concentrate in the upper-right region of COF PC space.

Judging by the location of the acquired set of COFs in PC COF space, MFBO explores diverse regions of COF space, yet concentrates its COF acquires in the regions containing the highest performers. Interestingly, each high-fidelity simulation in a COF was preceded by a low-fidelity simulation in the same COF. This suggests that the MFBO algorithm is cautious to conduct expensive high-fidelity simulations and conservatively utilizes the low-fidelity simulations to explore COF space.

Comparing MFBO with baseline sequential search methods (Fig. 6)


We compare the search efficiency of MFBO with single-fidelity (SF) BO, random search, exhaustive search, and a two-stage screening.

Exhaustive search. An exhaustive search runs a high-fidelity simulation of Xe/Kr adsorption in each of the 609 COFs in \mathcal{X} . While guaranteed to find the optimal COF \mathbf{x}^* , an exhaustive search incurs a high cost because it fails to exploit (i) the cheap, low-fidelity simulations available and (ii) the information contained in the simulation data $\mathcal{D}_{[n]}$, about the relationship between the Xe/Kr selectivity of the COFs and their structural and compositional features in \mathbf{x} , as the search proceeds.


 The runtime of the exhaustive search was ~ 2332 hr. By comparison, MFBO incurred 2.5%

of the runtime of the exhaustive search.

Two-stage screening. A two-stage search (1) runs a low-fidelity simulation of Xe/Kr adsorption in each of the 609 COFs in \mathcal{X} , then (2) (a) sorts the COFs according to their low-fidelity simulated Xe/Kr selectivity, in descending order, then (b) queues this list of COFs for high-fidelity simulations of Xe/Kr adsorption, working down the list. This search strategy leverages the cheap, low-fidelity simulations available in stage (1) to recover the optimal COF early in the sequence of stage (2). However, this strategy still fails to leverage the information contained in the simulation data $\mathcal{D}_{[n]}$ as the search proceeds to (i) avoid running low-fidelity simulations in every COF during stage (1) and (ii) adjust the sequence of high-fidelity simulations as high-fidelity simulation data is collected in stage (2).

 This two-stage search incurs a runtime of ~ 189 hr to find the optimal COF \mathbf{x}^* , still more than MFBO (58 hr).

Random search with the high-fidelity simulations. A random search sequentially chooses a COF at random (without replacement) for a high-fidelity simulation of Xe/Kr adsorption. We conduct 1000 random searches and show the mean and two standard deviations of the search efficiency curves in Fig. 6a. MFBO recovers the optimal COF \mathbf{x}^* with much less accumulated runtime compared to a typical random search.

 The average run time incurred during by a random search to acquire the optimal COF is 1152 hr. By comparison MFBO incurred 5% of the average runtime of the random search.

Single-fidelity Bayesian Optimization (SFBO). Finally, we assess the performance of single-fidelity (SF) BO of COFs for Xe/Kr separations—standard Bayesian optimization with the high-fidelity simulation of Xe/Kr adsorption using, for a controlled comparison to MFBO, (i) the same three COFs for initialization², (ii) the expected improvement acquisition function,

²Note that the initialization cost of MFBO is higher than that of its SFBO counterpart due to the inclusion

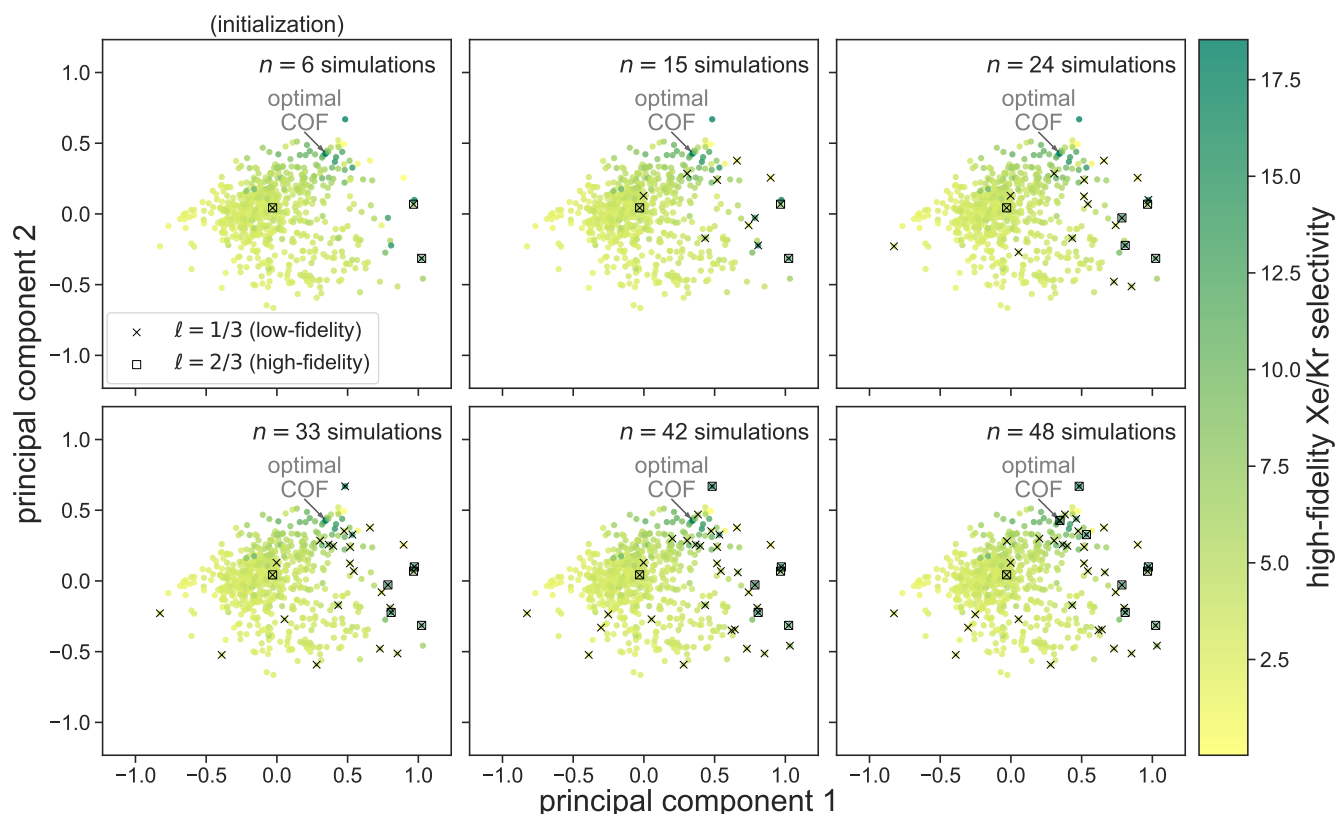


Figure 5: **Visualizing MFBO acquisition dynamics** by showing the location of COFs acquired by MFBO in COF space as the search proceeds. Each panel shows the first two principal components of COF space and corresponds to a different iteration of the MBFO search. Each point represents a COF, colored according to its high-fidelity Xe/Kr selectivity. Up to that iteration, the acquired set of COFs in $\mathcal{D}_{[n]}$ are marked; COFs subject to low- vs. high-fidelity simulations are distinguished by marker type. The arrow points to the optimal COF with largest high-fidelity Xe/Kr selectivity.

and (iii) a GP surrogate model with an identical material kernel.

🕒 Fig. 6a shows the search efficiency curve of SFBO compared to MFBO. SFBO incurred a runtime of ~ 315 hr, more than five times that of MFBO (58 hr).

Robustness of MFBO performance to initialization.

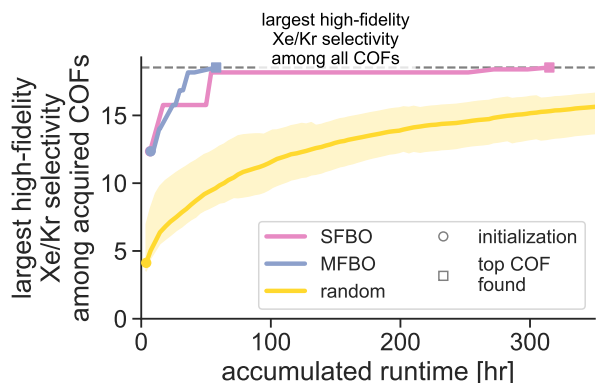
How robust is the MFBO performance to different initialization schemes? We conducted 100 MFBO and SFBO searches whose surrogate model is initialized with training data from simulations in three COFs: the first *randomly* selected (as opposed to the COF nearest the center), the next two chosen via max-min distance for diversity. Fig. 6b shows the distribution of accumulated

of the additional low-fidelity simulations. We include the runtime incurred for initialization.

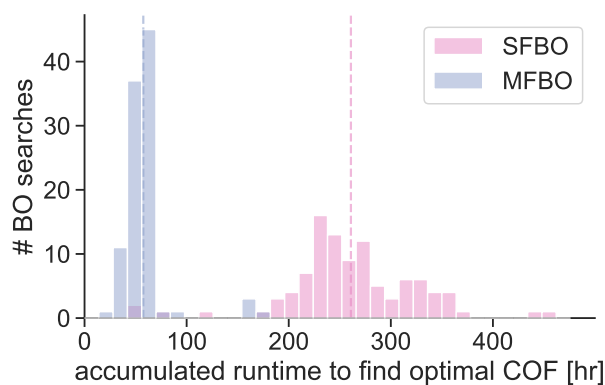
runtimes to find the optimal COF \mathbf{x}^* over these random initializing sets of COFs. (Each individual search efficiency trace is shown in Fig. S6.) While there is significant variance in the runtimes (standard deviations: 63 hr for SFBO, 25 hr for MFBO), the distribution of the runtime of MFBO is shifted far to the left of that of SFBO (means: 261 hr for SFBO, 58 hr for MFBO).

Conclusions

Our goal was to efficiently search a database of ~ 600 COFs for the one exhibiting the largest adsorptive Xe/Kr selectivity. We had access to two molecular simulations of Xe/Kr adsorption to predict the selectivity of a COF: a high-fidelity binary grand-canonical Monte Carlo simulation and a low-fidelity Henry coefficient calculation with



(a)



(b)

Figure 6: **Comparing the search efficiency of MFBO to random search and single-fidelity (SF) BO.** (a) The largest high-fidelity Xe/Kr selectivity among acquired COFs as a function of the computational runtime incurred, as each search progresses. The bands on the random search curve show two standard deviations. (b) The distribution of computational runtimes to find the COF with the largest high-fidelity Xe/Kr selectivity, over random selections of the COF that initializes the search. Vertical dashed lines show the average.

a smaller runtime. We employed multi-fidelity Bayesian optimization (MFBO) to orchestrate the sequential search for the COF with the largest high-fidelity Xe/Kr selectivity. MFBO constituted an iterative feedback loop of (1) conduct a low- or high-fidelity simulation of Xe/Kr adsorption in a COF, (2) use the simulation data gathered so far to train a surrogate model that predicts the selectivity of COFs, according to both low-

and high-fidelity simulations, based on their structural and chemical features, with quantified uncertainty, then (3) choose the COF and fidelity for the next simulation via maximizing an acquisition function that balances exploration, exploitation, and cost. MFBO acquired the optimal COF among the ~ 600 candidates, having the largest high-fidelity Xe/Kr selectivity, using only 38 low-fidelity and nine high-fidelity simulations— incurring only 5% and 22% of the average runtime to find the top COF via random sequential search and single-fidelity BO, respectively. Visualizing the location of the acquired COFs in the design space as the search proceeds reveals that MFBO judiciously planned the sequence of simulations to balance exploration, exploitation, and the cost of the two types of simulations.

Similar in spirit to multi-fidelity machine learning and two-stage search, the cheap-, low-fidelity calculations of dilute adsorption properties could serve as features (inputs) to a supervised machine learning model to predict the high-fidelity adsorption property.⁹²

Discussion

Though taking place in a computer simulation and pertaining to the specific task of discovering COFs for Xe/Kr separations, our study hints at the potential for multi-fidelity Bayesian optimization to reduce the time and cost to discover new materials in a variety of contexts. While the cost in our work was computational runtime, costs will be much more significant in the bona fide laboratory, involving lab space, equipment/instrumentation, reagents, salaries of operators, etc.

Generally, the performance of MFBO for materials discovery is predicated on the surrogate model and the acquisition function. The surrogate model must (i) be fed features of the materials that are informative about the property (this relies on domain knowledge) and (ii) give well-calibrated/honest uncertainty estimates^{89,93} while having a small number of training data points. The acquisition function for experimental planning must balance exploration, exploitation, and cost. Another factor for MFBO performance is

the accuracy of the measurements of the material property; a challenge in the bona fide laboratory is that measurement noise is greater than in a computer simulation³.

Future work includes (1) inventing new (a) predictive, uncertainty-calibrated, and data-efficient multi-fidelity surrogate models and (b) exploration-, exploitation-, and cost-balancing multi-fidelity acquisition functions; (2) extending MFBO to (a) the batch setting, where experiments can be done in parallel and multiple materials are selected at each iteration^{6,97} and (b) the multi-objective setting,⁹⁸ where we seek the Pareto-optimal set of COFs.

Methods

The COF crystal structures

We obtained the crystal structures of the 609 COF candidates from the Clean, Uniform, Refined with Automatic Tracking from Experimental Database (CURATED).⁵⁹

The two molecular simulation techniques to predict the Xe/Kr selectivity of a COF

The binary grand-canonical ensemble

The binary grand-canonical ensemble concerns a crystalline COF immersed in and in thermodynamic equilibrium with an 20 mol%/80 mol% Xe/Kr gas mixture at $T = 298$ K at $P = 1$ bar. The system volume Ω comprises a replicated unit cell of the COF that hosts Xe and Kr adsorbates. The volume V , chemical potential of Xe and Kr $\boldsymbol{\mu} = [\mu_{Xe}, \mu_{Kr}]$, and temperature T of the system are fixed, whereas the number of adsorbates $\mathbf{n} = [n_{Xe}, n_{Kr}]$ hosted in the system and potential energy E of the system fluctuate as it exchanges adsorbates and heat with the bulk Xe/Kr gas mixture.

³E.g., see Refs. ^{94,95} pertaining to reproducibility issues for adsorption in porous materials across labs. Self-driving labs can alleviate reproducibility issues.⁹⁶

The chemical potential $\boldsymbol{\mu}$ is set by the Xe/Kr gas mixture; the ideal gas law gives $\boldsymbol{\mu}$ in terms of the temperature T and partial pressures of Xe and Kr, $\mathbf{p} = [p_{Xe}, p_{Kr}]$:

$$\mu_g = k_B T \log[\beta p_g \Lambda_g^3] \text{ for } g \in \{Xe, Kr\}, \quad (14)$$

with Λ_g the de Broglie wavelength of adsorbate g , k_B the Boltzmann constant, and $\beta := (k_B T)^{-1}$.

A microstate of the system is defined by (i) the number of adsorbates \mathbf{n} and (ii) their positions

$$\mathbf{R}^{(\mathbf{n})} := [\mathbf{r}_{Xe,1} \cdots \mathbf{r}_{Xe,n_{Xe}} \mathbf{r}_{Kr,1} \cdots \mathbf{r}_{Kr,n_{Kr}}] \quad (15)$$

in the system ($\mathbf{R}^{(\mathbf{n})} \in \mathbb{R}^{3 \times (n_{Xe} + n_{Kr})}$). Approximating the COF as rigid, the positions of the atoms of the COF are fixed.

Let $E = E(\mathbf{n}, \mathbf{R}^{(\mathbf{n})})$ be the potential energy of a microstate $(\mathbf{n}, \mathbf{R}^{(\mathbf{n})})$. Of course, $E = E(\mathbf{n}, \mathbf{R}^{(\mathbf{n})})$ is COF-dependent.

In the BGC ensemble, the partition function is a sum/integral over microstates^{99–101}

$$\Xi(\boldsymbol{\mu}, V, T) = \sum_{\mathbf{n} \in \mathbb{N}_{\geq 0}^2} e^{\beta \boldsymbol{\mu} \cdot \mathbf{n}} \prod_{g \in \{Xe, Kr\}} \frac{1}{n_g! \Lambda_g^{3n_g}} \int_{\Omega} e^{-\beta E(\mathbf{n}, \mathbf{R}^{(\mathbf{n})})} d\mathbf{R}^{(\mathbf{n})}, \quad (16)$$

and the probability of a microstate is

$$\pi(\mathbf{n}, \mathbf{R}^{(\mathbf{n})}) \propto e^{-\beta E(\mathbf{n}, \mathbf{R}^{(\mathbf{n})})} \prod_{g \in \{Xe, Kr\}} \frac{V^{n_g}}{n_g! \Lambda_g^{3n_g}} e^{\beta \boldsymbol{\mu}_g n_g}. \quad (17)$$

In each molecular simulation technique below, the ultimate goal is to predict the expected number of adsorbates in the system under the BGC ensemble:

$$\langle \mathbf{n} \rangle = \left(\frac{\partial \log \Xi}{\partial (\beta \boldsymbol{\mu})} \right)_{\beta, V}, \quad (18)$$

from which the Xe/Kr adsorptive selectivity follows.

The atomistic model

We model the potential energy $E = E(\mathbf{n}, \mathbf{R}^{(\mathbf{n})})$ of the system in microstate $(\mathbf{n}, \mathbf{R}^{(\mathbf{n})})$ by treating the

adsorbate-COF and adsorbate-adsorbate interactions as pairwise additive and described by 12-6 Lennard-Jones interatomic potentials (parameters from the Universal Force Field,⁸³ Lorentz-Berthelot combining rules,¹⁰¹ truncated to neglect interactions beyond 14 Å). We apply periodic boundary conditions to mimic the crystalline COF.

Binary grand-canonical Monte Carlo simulation

The high-fidelity simulation constitutes a Markov chain Monte Carlo (MC) simulation of the system under the BGC ensemble governed by the probability distribution in eqn. 17. Our microstate transition proposals include random adsorbate insertions and deletions, translations, reinsertions, and identity swaps, with acceptance rules dictated by Metropolis-Hastings. Our BGCMC simulation constitutes 500 Monte Carlo cycles (defined as x microstate transition proposals, with $x = \max(20, n_{Xe} + n_{Kr})$) per Å³ volume of the system. We discard the first half of the cycles for burn-in.

Henry coefficient calculations

Henry's law, valid under dilute conditions, follows from eqn. 18 if we approximate the sum in Ξ in eqn. 16 by including only the dominant terms $\mathbf{n} \in \{[0, 0], [1, 0], [0, 1]\}$ at dilute conditions, giving Henry's law in eqn. 2 with Henry coefficients

$$H_{Xe} = \beta \int_{\Omega} e^{-\beta E([1,0], \mathbf{r}_{Xe})} d\mathbf{r}_{Xe} \quad (19)$$

$$H_{Kr} = \beta \int_{\Omega} e^{-\beta E([0,1], \mathbf{r}_{Kr})} d\mathbf{r}_{Kr}. \quad (20)$$

For the low-fidelity prediction of Xe/Kr selectivity, we compute H_{Xe} and H_{Kr} of a COF from two ordinary Monte Carlo integrations (500 insertions/Å³), i.e. Widom particle insertions.¹⁰⁰

Comparing runtimes

The computational cost, measured in run time, of a high-fidelity BGCMC simulation of Xe/Kr adsorption in a given COF is greater than the sum

of the costs of the two low-fidelity Henry coefficient calculations, i.e. $\tau^{(2/3)} > \tau^{(1/3)}$. First, a single Monte Carlo state transition tends to be more computationally expensive for the BGCMC simulation because, in contrast to the Henry coefficient calculations, generally (i) multiple adsorbates are present in the system and (ii) adsorbate-adsorbate interactions are included. Second, the high-fidelity BGCMC simulation must explore a more voluminous state space than the Henry coefficient calculation in order to compute a reliable average.

Of course, this cost comparison depends on the number of MC cycles/insertions dedicated to each simulation; we allocated 500 cycles/insertions per Å³ volume of the system in an attempt to grant each simulation with reasonably comparable errors in the average $\langle \mathbf{n} \rangle$.

N.b., with further approximation, the computational expense of the Henry coefficient calculations can be reduced by biasing the samples of adsorbate configurations to lie nearby the internal surface (pore walls) of the COF.¹⁰²

Remark on high- vs. low-fidelity

We refer to the BGCMC simulation as providing a "high-fidelity" estimate of the Xe/Kr selectivity of a COF, but only *relative* to the lower-fidelity Henry coefficient calculation. First, arguably, *the* high-fidelity measurement of the adsorptive Xe/Kr selectivity of a COF constitutes synthesizing and characterizing it in the lab, then taking mixed-gas adsorption measurements.¹⁰³ Second, even higher-fidelity simulations of Xe/Kr adsorption are possible by (i) calculating the potential energy of a configuration $E = E(\mathbf{n}, \mathbf{R}^{(\mathbf{n})})$ using a machine learning model trained on energy calculations based on a higher level of theory (e.g. density functional theory),^{104,105} (ii) modeling the flexibility of the COF,¹⁰⁶ and/or (iii) modeling crystalline defects in the COF,¹⁰⁷ etc. If "high-fidelity" instead refers to performance in the real-world application, we must also consider competing adsorbates such as CO₂ and H₂O as well as other objectives such as stability,¹⁰⁸ thermal conductivity,¹⁰⁹ and adsorption kinetics.¹¹⁰

Software


We implemented the BGCMC and Henry coefficient calculations in `PorousMaterials.jl`.

Hardware

To put our reported computational runtimes in perspective, the hardware specifications for the compute nodes on which we ran our (serial) simulations are listed in Tab. 1. We assigned each simulation to a random core based on its availability. Though the high- and low-fidelity simulations for a given COF are not guaranteed to run on the same core, the specifications of each core are similar for a reasonable comparison of runtimes.

The multi-fidelity Gaussian process surrogate model

We explain our multi-fidelity GP in the context of the Bayesian paradigm of (i) impose a prior distribution, (ii) collect data, then (iii) in light of the data, update the prior distribution to a posterior distribution.

 For more understanding about GPs, see Refs.^{86,87}

The prior distribution of \mathbf{Y}

The *prior* distribution of the $2X$ ($X = 609$) random variables of interest for our problem,

$$\mathbf{Y} := \begin{bmatrix} \mathbf{Y}^{(1/3)} \\ \mathbf{Y}^{(2/3)} \end{bmatrix} := \begin{bmatrix} Y^{(1/3)}(\mathbf{x}_1) \\ \vdots \\ Y^{(1/3)}(\mathbf{x}_X) \\ Y^{(2/3)}(\mathbf{x}_1) \\ \vdots \\ Y^{(2/3)}(\mathbf{x}_X) \end{bmatrix}, \quad (21)$$

expresses our beliefs about the simulated Xe/Kr selectivities of the COFs under each fidelity *before* any molecular simulations are conducted—i.e., before we obtain any simulation data on which to base our beliefs.

The joint prior distribution expressed by the GP in eqn. 5 is a Gaussian distribution with (i) a mean

of the zero-vector and (ii) a covariance matrix exhibiting a block structure:

$$\mathbf{Y} \sim \mathcal{N} \left(\mathbf{0}, \alpha \begin{bmatrix} k_{\text{fid}}\left(\frac{1}{3}, \frac{1}{3}\right) \mathbf{K}_{\text{mat}} & k_{\text{fid}}\left(\frac{1}{3}, \frac{2}{3}\right) \mathbf{K}_{\text{mat}} \\ k_{\text{fid}}\left(\frac{2}{3}, \frac{1}{3}\right) \mathbf{K}_{\text{mat}} & k_{\text{fid}}\left(\frac{2}{3}, \frac{2}{3}\right) \mathbf{K}_{\text{mat}} \end{bmatrix} \right), \quad (22)$$

where $\mathbf{K}_{\text{mat},ij} = k_{\text{mat}}(\mathbf{x}_i, \mathbf{x}_j)$ is the COF similarity matrix.

We elucidate the assumption behind eqn. 22 and the intuition behind the kernel functions by inspecting the *marginal* prior distribution of

- the fidelity- ℓ simulated Xe/Kr selectivity of a COF \mathbf{x} ,

$$Y^{(\ell)}(\mathbf{x}) \sim \mathcal{N} \left(0, \alpha [c + (1 - \ell)^{2(1+\delta)}] \right). \quad (23)$$

Apparently, the hyperparameters c and δ forming the variance express our fidelity-dependent, COF-independent prior uncertainty about the simulated Xe/Kr selectivity of any given COF.

- a pair of simulated Xe/Kr selectivities, $Y^{(\ell)}(\mathbf{x})$ and $Y^{(\ell')}(\mathbf{x}')$, whose covariance is given by the kernel function k in eqn. 6:

$$\text{cov}[Y^{(\ell)}(\mathbf{x}), Y^{(\ell')}(\mathbf{x}')] = \alpha k_{\text{mat}}(\mathbf{x}, \mathbf{x}') k_{\text{fid}}(\ell, \ell'). \quad (24)$$

With the kernel functions quantifying our notion of "similarity", our prior belief is that the simulated selectivity of two COFs will be similar (dissimilar) for (i) two similar (dissimilar) COFs under (ii) two similar (dissimilar) simulation fidelities. Importantly, the material kernel function in eqn. 7 paired with our design of COF space captures our domain knowledge that COFs with closeby composition, pore size, surface area, etc. tend to exhibit similar adsorption properties.^{71,75,79,84} Note, for $\ell \neq \ell'$ but $\mathbf{x} = \mathbf{x}'$, it is apparent that the hyperparameters c and δ of the fidelity kernel function also capture the correlation between the high- and low-fidelity Xe/Kr selectivities for a given COF. This allows observed low-fidelity simulated Xe/Kr selectivities to appropriately inform the predictions about the high-fidelity selectivities we ultimately wish to maximize.

Table 1: Hardware specifications for the computational resources used for our simulations.

nodes 1-4	Model	Dell PowerEdge R740
	Processor	2x 10-core 2.20 GHz Intel Xeon Silver 4114 w/ 16896 KB cache
	Memory	128 GB RAM @2666 MT/s
nodes 5-8	Model	Dell PowerEdge R740
	Processor	2x 22-core 2.10 GHz Intel Xeon Gold 6152 w/ 30976 KB cache
	Memory	128 GB RAM @2666 MT/s

Collecting the simulation data.

At iteration n of the MFBO search, we have collected simulation data

$$\mathcal{D}_{[n]} := \{([\mathbf{x}_{[1]}, \ell_{[1]}], y_{[1]}), \dots, ([\mathbf{x}_{[n]}, \ell_{[n]}], y_{[n]})\}. \quad (25)$$

I.e., $\mathbf{x}_{[i]}$ is the vector representation of the COF, $\ell_{[i]}$ is the fidelity, and $y_{[i]}$ is the observed Xe/Kr selectivity of the simulation conducted at iteration i . In light of this simulation data $\mathcal{D}_{[n]}$, we wish to *update* our prior distribution in eqn. 22.

We view each observed fidelity- ℓ simulated Xe/Kr selectivity $y^{(\ell)}$ of a COF represented by \mathbf{x} as a noisy evaluation of a black-box function $f(\mathbf{x}, \ell)$ that represents the relationship between the fidelity- ℓ Xe/Kr selectivity of a COF and its features \mathbf{x} . Particularly, we assume

$$y^{(\ell)} = f(\mathbf{x}, \ell) + \epsilon, \quad (26)$$

where ϵ is a realization of un-observable noise drawn from a Gaussian distribution $E \sim \mathcal{N}(0, \sigma^2)$. The source of this noise is the inherent stochasticity involved in the Monte Carlo simulation; however, the noise may also have a contribution from the lack of information contained about the selectivity within the COF features \mathbf{x} .

The posterior distribution of $Y^{(\ell)} \mid \mathcal{D}_{[n]}$

The *posterior* distribution of $Y^{(\ell)}(\mathbf{x})$ expresses our beliefs about the fidelity- ℓ simulated Xe/Kr selectivity of a COF with features \mathbf{x} *in light of the simulation data* $\mathcal{D}_{[n]}$. The posterior is an update to our prior distribution, obtained by conditioning the prior distribution in eqn. 22 on the observations $\{Y^{(\ell_{[i]})}(\mathbf{x}_{[i]}) = y_{[i]}\}_{i=1}^n$ in the data $\mathcal{D}_{[n]}$.

We find the marginal posterior distribution of

$Y^{(\ell)}(\mathbf{x}) \mid \mathcal{D}_{[n]}$ by first writing the marginal prior distribution, following from eqn. 22, of (i) the fidelity- ℓ simulated Xe/Kr selectivity of COF represented by \mathbf{x} and (ii) the observed (i.e., noise-contaminated) selectivities in the simulations we have already done in $\mathcal{D}_{[n]}$:

$$\begin{bmatrix} Y^{(\ell)}(\mathbf{x}) \\ \mathbf{Y}_{\mathcal{D}_{[n]}} \end{bmatrix} \sim \mathcal{N} \left(\mathbf{0}, \begin{bmatrix} k([\mathbf{x}, \ell], [\mathbf{x}, \ell]) & \mathbf{k}_{\mathcal{D}_{[n]}}^T \\ \mathbf{k}_{\mathcal{D}_{[n]}} & \mathbf{K}_{\mathcal{D}_{[n]}} + \sigma^2 \mathbf{I} \end{bmatrix} \right), \quad (27)$$

written in terms of (1) the vector of random variables denoting the simulated Xe/Kr selectivities of the COFs in the acquired set at those specific fidelities:

$$\mathbf{Y}_{\mathcal{D}_{[n]}} := \begin{bmatrix} Y_{[1]} \\ \vdots \\ Y_{[n]} \end{bmatrix}, \quad (28)$$

(2) the kernel matrix between the simulation setups in the data $\mathcal{D}_{[n]}$

$$\mathbf{K}_{\mathcal{D}_{[n]}} := \begin{bmatrix} k([\mathbf{x}_{[1]}, \ell_{[1]}], [\mathbf{x}_{[1]}, \ell_{[1]}) & \cdots & k([\mathbf{x}_{[1]}, \ell_{[1]}], [\mathbf{x}_{[n]}, \ell_{[n]}) \\ \vdots & \ddots & \vdots \\ k([\mathbf{x}_{[n]}, \ell_{[n]}], [\mathbf{x}_{[1]}, \ell_{[1]}) & \cdots & k([\mathbf{x}_{[n]}, \ell_{[n]}], [\mathbf{x}_{[n]}, \ell_{[n]}) \end{bmatrix}, \quad (29)$$

and (3) the kernel vector between the simulation setup of interest $[\mathbf{x}, \ell]$ and those in the data $\mathcal{D}_{[n]}$

$$\mathbf{k}_{\mathcal{D}_{[n]}} := \begin{bmatrix} k([\mathbf{x}, \ell], [\mathbf{x}_{[1]}, \ell_{[1]}) \\ \vdots \\ k([\mathbf{x}, \ell], [\mathbf{x}_{[n]}, \ell_{[n]}) \end{bmatrix}. \quad (30)$$

We obtain the posterior distribution of $Y^{(\ell)}(\mathbf{x})$ by conditioning the prior in eqn. 27 on the *ob-*

served simulated Xe/Kr selectivities of the COFs in the data \mathcal{D}_n :

$$\mathbf{Y}_{\mathcal{D}_{[n]}} = \mathbf{y}_{\mathcal{D}_{[n]}} := \begin{bmatrix} y_{[1]} \\ \vdots \\ y_{[n]} \end{bmatrix}. \quad (31)$$

Upon conditioning, the posterior distribution of $Y^{(\ell)}(\mathbf{x})$ is also a Gaussian distribution, given in eqn. 9.

Remarks

Sources of uncertainty. Uncertainty in the Xe/Kr selectivity of a COF may owe to (i) a lack of simulations on COFs in the neighborhood of COF space around \mathbf{x} , (ii) a lack of mutual information between outcomes of simulations of different fidelities, (iii) a lack of information about the selectivity contained in the features, and/or (iv) inherent variability/noise in the outcome of the Monte Carlo simulation.

Centering the outputs. For the zero-mean prior in eqn. 22 to be reasonable, we center the simulated Xe/Kr selectivities (the $y_{[i]}$'s) in the data $\mathcal{D}_{[n]}$ at each iteration.

Hyperparameters. The kernel function in eqn. 6 contains four hyperparameters: α , γ , c , and δ . And, we have the noise hyperparameter σ from eqn. 26. At each iteration, these hyperparameters are tuned to maximize the marginal likelihood of the data \mathcal{D}_n .

Function space view of a GP. For our problem of searching a fixed pool of COFs, we are only interested in the joint distribution of the random variables listed in \mathbf{Y} in eqn. 22. However, an alternative view of the GP in eqn. 5 is that it specifies a (prior and posterior) distribution over functions $f(\mathbf{x}, \ell)$ that aim to approximate the black-box input (COF \mathbf{x} , fidelity ℓ) – output (simulated Xe/Kr selectivity, $y^{(\ell)}$) relationship under-lurking the simulations—the black-box function $f(\mathbf{x}, \ell)$ in eqn. 26. This perspective is illustrated in the middle panel of Fig. 3, where the dark line shows the

posterior mean function $\mu_{[n]}(\mathbf{x}, \ell)$ and the bands show a posterior credible "interval" for these functions, the region $\mu_{[n]}(\mathbf{x}, \ell) \pm \sigma_{[n]}(\mathbf{x}, \ell)$.

GP implementation. We use the implementation of the multi-fidelity GP in the BoTorch¹¹¹ library in Python, which builds upon GPyTorch.¹¹²

Conflicts of Interest

None to declare.

Open code and data

All code and simulation data to reproduce our results is available at <https://github.com/SimonEnsemble/multi-fidelity-BO-of-COFs-for-Xe-Kr-seps>.

Acknowledgements

For funding and support, N.G. and C.M.S. acknowledge the U.S. Department of Defense (DoD) Defense Threat Reduction Agency (HDTRA-19-31270) and A.D. and J.D. acknowledge the National Science Foundation Grants IIS-1845922 and OAC-1910213. C.M.S. and N.G. thank the Oregon State University College of Engineering High-Performance Computing Cluster manager Robert Yelle.

References

- (1) Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; Persson, K. A. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *Applied Physics Letters Materials* **2013**, *1*, 011002.
- (2) Pyzer-Knapp, E. O.; Suh, C.; Gómez-Bombarelli, R.; Aguilera-Iparraguirre, J.; Aspuru-Guzik, A. What is high-throughput

- virtual screening? A perspective from organic materials discovery. Annual Review of Materials Research **2015**, 45, 195–216.
- (3) Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine learning for molecular and materials science. Nature **2018**, 559, 547–555.
- (4) Jablonka, K. M.; Ongari, D.; Moosavi, S. M.; Smit, B. Big-data science in porous materials: materials genomics and machine learning. Chemical Reviews **2020**, 120, 8066–8129.
- (5) Keith, J. A.; Vassilev-Galindo, V.; Cheng, B.; Chmiela, S.; Gastegger, M.; Müller, K.-R.; Tkatchenko, A. Combining machine learning and computational chemistry for predictive insights into chemical systems. Chemical Reviews **2021**, 121, 9816–9872.
- (6) Shahriari, B.; Swersky, K.; Wang, Z.; Adams, R. P.; De Freitas, N. Taking the human out of the loop: A review of Bayesian optimization. Proceedings of the IEEE **2015**, 104, 148–175.
- (7) Agnihotri, A.; Batra, N. Exploring Bayesian Optimization. Distill **2020**, <https://distill.pub/2020/bayesian-optimization>.
- (8) Frazier, P. I. A tutorial on Bayesian optimization. arXiv preprint arXiv:1807.02811 **2018**,
- (9) Garnett, R. Bayesian Optimization; Cambridge University Press, 2023.
- (10) Liang, Q.; Gongora, A. E.; Ren, Z.; Tiihonen, A.; Liu, Z.; Sun, S.; Deneault, J. R.; Bash, D.; Mekki-Berrada, F.; Khan, S. A.; Hippalgaonkar, K.; Maruyama, B.; Brown, K. A.; III, J. F.; Buonassisi, T. Benchmarking the performance of Bayesian optimization across multiple experimental materials science domains. Nature Partner Journals Computational Materials **2021**, 7, 188.
- (11) Coley, C. W. Defining and exploring chemical spaces. Trends in Chemistry **2021**, 3, 133–145.
- (12) Stach, E. et al. Autonomous experimentation systems for materials development: A community perspective. Matter **2021**, 4, 2702–2726.
- (13) Häse, F.; Roch, L. M.; Aspuru-Guzik, A. Next-Generation Experimentation with Self-Driving Laboratories. Trends in Chemistry **2019**, 1, 282–291.
- (14) Burger, B.; Maffettone, P. M.; Gusev, V. V.; Aitchison, C. M.; Bai, Y.; Wang, X.; Li, X.; Alston, B. M.; Li, B.; Clowes, R.; Rankin, N.; Harris, B.; Sprick, R. S.; Cooper, A. I. A mobile robotic chemist. Nature **2020**, 583, 237–241.
- (15) Nambiar, A. M. K.; Breen, C. P.; Hart, T.; Kulesza, T.; Jamison, T. F.; Jensen, K. F. Bayesian Optimization of Computer-Proposed Multistep Synthetic Routes on an Automated Robotic Flow Platform. American Chemical Society Central Science **2022**, 8, 825–836.
- (16) Baird, S. G.; Sparks, T. D. What is a minimal working example for a self-driving laboratory? Matter **2022**, 5, 4170–4178.
- (17) Jiang, Y.; Salley, D.; Sharma, A.; Keenan, G.; Mullin, M.; Cronin, L. An artificial intelligence enabled chemical synthesis robot for exploration and optimization of nanomaterials. Science Advances **2022**, 8, eabo2626.
- (18) Pomberger, A.; Jose, N.; Walz, D.; Meissner, J.; Holze, C.; Kopczynski, M.; Müller-Bischof, P.; Lapkin, A. Automated pH Adjustment Driven by Robotic Workflows and Active Machine Learning. Chemical Engineering Journal **2023**, 451, 139099.
- (19) Arróyave, R.; Khatamsaz, D.; Vela, B.; Couperthwaite, R.; Molkeri, A.; Singh, P.;

- Johnson, D. D.; Qian, X.; Srivastava, A.; Allaire, D. A perspective on Bayesian methods applied to materials discovery and design. *MRS Communications* **2022**, *1–13*.
- (20) Wang, K.; Dowling, A. W. Bayesian optimization for chemical products and functional materials. *Current Opinion in Chemical Engineering* **2022**, *36*, 100728.
- (21) Packwood, D. *Bayesian Optimization for Materials Science*; Springer, 2017.
- (22) Comlek, Y.; Pham, T. D.; Snurr, R.; Chen, W. Rapid Design of Top-Performing Metal-Organic Frameworks with Qualitative Representations of Building Blocks. arXiv preprint arXiv:2302.09184 **2023**,
- (23) Deshwal, A.; Simon, C. M.; Doppa, J. R. Bayesian optimization of nanoporous materials. *Molecular Systems Design & Engineering* **2021**, *6*, 1066–1086.
- (24) Taw, E.; Neaton, J. B. Accelerated Discovery of CH₄ Uptake Capacity Metal–Organic Frameworks Using Bayesian Optimization. *Advanced Theory and Simulations* **2022**, *5*, 2100515.
- (25) Tang, H.; Jiang, J. Active learning boosted computational discovery of covalent–organic frameworks for ultrahigh CH₄ storage. *American Institute of Chemical Engineers Journal* **2022**, *68*, e17856.
- (26) Pyzer-Knapp, E. O.; Chen, L.; Day, G. M.; Cooper, A. I. Accelerating computational discovery of porous solids through improved navigation of energy–structure–function maps. *Science Advances* **2021**, *7*, eabi4763.
- (27) Vaddi, K.; Chiang, H. T.; Pozzo, L. D. Autonomous retrosynthesis of gold nanoparticles via spectral shape matching. *Digital Discovery* **2022**, *1*, 502–510.
- (28) Rouet-Leduc, B.; Barros, K.; Lookman, T.; Humphreys, C. J. Optimisation of GaN LEDs and the reduction of efficiency droop using active machine learning. *Scientific Reports* **2016**, *6*, 1–6.
- (29) Chang, J.; Nikolaev, P.; Carpena-Núñez, J.; Rao, R.; Decker, K.; Islam, A. E.; Kim, J.; Pitt, M. A.; Myung, J. I.; Maruyama, B. Efficient closed-loop maximization of carbon nanotube growth rate using bayesian optimization. *Scientific Reports* **2020**, *10*, 9040.
- (30) Herbol, H. C.; Hu, W.; Frazier, P.; Clancy, P.; Poloczek, M. Efficient search of compositional space for hybrid organic–inorganic perovskites via Bayesian optimization. *Nature Partner Journals Computational Materials* **2018**, *4*, 51.
- (31) Sun, S. et al. A data fusion approach to optimize compositional stability of halide perovskites. *Matter* **2021**, *4*, 1305–1322.
- (32) Zhang, Y.; Apley, D. W.; Chen, W. Bayesian optimization for materials design with mixed quantitative and qualitative variables. *Scientific Reports* **2020**, *10*, 1–13.
- (33) Gongora, A. E.; Snapp, K. L.; Whiting, E.; Riley, P.; Reyes, K. G.; Morgan, E. F.; Brown, K. A. Using simulation to accelerate autonomous experimentation: A case study using mechanics. *iScience* **2021**, *24*, 102262.
- (34) Langner, S.; Häse, F.; Perea, J. D.; Stubhan, T.; Hauch, J.; Roch, L. M.; Heumueller, T.; Aspuru-Guzik, A.; Brabec, C. J. Beyond Ternary OPV: High-Throughput Experimentation and Self-Driving Laboratories Optimize Multi-component Systems. *Advanced Materials* **2020**, *32*, 1907801.
- (35) Ramesh, P. S.; Patra, T. K. Polymer sequence design via molecular simulation-based active learning. *Soft Matter* **2023**, *19*, 282–294.

- (36) Reis, M.; Gusev, F.; Taylor, N. G.; Chung, S. H.; Verber, M. D.; Lee, Y. Z.; Isayev, O.; Leibfarth, F. A. Machine-learning-guided discovery of 19F MRI agents enabled by automated copolymer synthesis. Journal of the American Chemical Society **2021**, *143*, 17677–17689.
- (37) Li, C.; Rubín de Celis Leal, D.; Rana, S.; Gupta, S.; Sutti, A.; Greenhill, S.; Slezak, T.; Height, M.; Venkatesh, S. Rapid Bayesian optimisation for synthesis of short polymer fiber materials. Scientific Reports **2017**, *7*, 1–10.
- (38) Tamasi, M. J.; Patel, R. A.; Borca, C. H.; Kosuri, S.; Mugnier, H.; Upadhya, R.; Murthy, N. S.; Webb, M. A.; Gormley, A. J. Machine Learning on a Robotic Platform for the Design of Polymer–Protein Hybrids. Advanced Materials **2022**, *34*, 2201809.
- (39) Seko, A.; Togo, A.; Hayashi, H.; Tsuda, K.; Chaput, L.; Tanaka, I. Prediction of low-thermal-conductivity compounds with first-principles anharmonic lattice-dynamics calculations and Bayesian optimization. Physical Review Letters **2015**, *115*, 205901.
- (40) Zhai, H.; Yeo, J. Computational Design of Antimicrobial Active Surfaces via Automated Bayesian Optimization. American Chemical Society Biomaterials Science & Engineering **2022**,
- (41) Epps, R. W.; Bowen, M. S.; Volk, A. A.; Abdel-Latif, K.; Han, S.; Reyes, K. G.; Amassian, A.; Abolhasani, M. Artificial Chemist: An Autonomous Quantum Dot Synthesis Bot. Advanced Materials **2020**, *32*, 2001626.
- (42) Kitamura, Y.; Toshima, H.; Inokuchi, A.; Tanaka, D. Bayesian optimization of the composition of the lanthanide metal-organic framework MIL-103 for white-light emission. Molecular Systems Design & Engineering **2023**,
- (43) Zhang, Y.; Peck, T. C.; Reddy, G. K.; Banerjee, D.; Jia, H.; Roberts, C. A.; Ling, C. Descriptor-Free Design of Multi-component Catalysts. American Chemical Society Catalysis **2022**, *12*, 10562–10571.
- (44) Pedersen, J. K.; Clausen, C. M.; Krysiak, O. A.; Xiao, B.; Batchelor, T. A. A.; Löffler, T.; Mints, V. A.; Banko, L.; Arenz, M.; Savan, A.; Schuhmann, W.; Ludwig, A.; Rossmeisl, J. Bayesian Optimization of High-Entropy Alloy Compositions for Electrocatalytic Oxygen Reduction. Angewandte Chemie **2021**, *133*, 24346–24354.
- (45) Rohr, B.; Stein, H. S.; Guevarra, D.; Wang, Y.; Haber, J. A.; Aykol, M.; Suram, S. K.; Gregoire, J. M. Benchmarking the acceleration of materials discovery by sequential learning. Chemical Science **2020**, *11*, 2696–2706.
- (46) MacLeod, B. P. et al. Self-driving laboratory for accelerated discovery of thin-film materials. Science Advances **2020**, *6*.
- (47) Baird, S. G.; Hall, J. R.; Sparks, T. D. Compactness matters: Improving Bayesian optimization efficiency of materials formulations through invariant search spaces. Computational Materials Science **2023**, *224*, 112134.
- (48) Xu, W.; Liu, Z.; Piper, R. T.; Hsu, J. W. P. Bayesian Optimization of photonic curing process for flexible perovskite photovoltaic devices. Solar Energy Materials and Solar Cells **2023**, *249*, 112055.
- (49) Torres, J. A. G.; Lau, S. H.; Anchuri, P.; Stevens, J. M.; Tabora, J. E.; Li, J.; Borovika, A.; Adams, R. P.; Doyle, A. G. A Multi-Objective Active Learning Platform and Web App for Reaction Optimization. Journal of the American Chemical Society **2022**, *144*, 19999–20007.
- (50) Shields, B. J.; Stevens, J.; Li, J.; Parasram, M.; Damani, F.; Alvarado, J. I. M.;

- Janey, J. M.; Adams, R. P.; Doyle, A. G. Bayesian reaction optimization as a tool for chemical synthesis. *Nature* **2021**, *590*, 89–96.
- (51) Schweidtmann, A. M.; Clayton, A. D.; Holmes, N.; Bradford, E.; Bourne, R. A.; Lapkin, A. A. Machine learning meets continuous flow chemistry: Automated optimization towards the Pareto front of multiple objectives. *Chemical Engineering Journal* **2018**, *352*, 277–282.
- (52) Ward, A.; Pini, R. Efficient Bayesian Optimization of Industrial-Scale Pressure-Vacuum Swing Adsorption Processes for CO₂ Capture. *Industrial & Engineering Chemistry Research* **2022**, *61*, 13650–13668.
- (53) Lam, R.; Allaire, D. L.; Willcox, K. E. Multifidelity Optimization using Statistical Surrogate Modeling for Non-Hierarchical Information Sources. 56th AIAA/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference. 2015; p 0143.
- (54) Fare, C.; Fenner, P.; Benatan, M.; Varsi, A.; Pyzer-Knapp, E. O. A multi-fidelity machine learning approach to high throughput materials screening. *Nature Partner Journals Computational Materials* **2022**, *8*, 257.
- (55) Herbol, H. C.; Poloczek, M.; Clancy, P. Cost-effective materials discovery: Bayesian optimization across multiple information sources. *Materials Horizons* **2020**, *7*, 2113–2123.
- (56) Tran, A.; Tranchida, J.; Wildey, T.; Thompson, A. P. Multi-fidelity machine-learning with uncertainty quantification and Bayesian optimization for materials design: Application to ternary random alloys. *The Journal of Chemical Physics* **2020**, *153*, 074705.
- (57) Foumani, Z. Z.; Shishehbor, M.; Yousefpour, A.; Bostanabad, R. Multi-fidelity cost-aware Bayesian optimization. *Computer Methods in Applied Mechanics and Engineering* **2023**, *407*, 115937.
- (58) Palizhati, A.; Aykol, M.; Suram, S.; Hummelshøj, J. S.; Montoya, J. H. Multi-fidelity Sequential Learning for Accelerated Materials Discovery. *ChemRxiv* **2021**,
- (59) Ongari, D.; Yakutovich, A. V.; Talirz, L.; Smit, B. Building a consistent and reproducible database for adsorption evaluation in Covalent-Organic Frameworks. *Materials Cloud Archive* **2021**,
- (60) Wu, J.; Toscano-Palmerin, S.; Frazier, P. I.; Wilson, A. G. Practical Multi-fidelity Bayesian Optimization for Hyperparameter Tuning. *Uncertainty in Artificial Intelligence*. 2020; pp 788–798.
- (61) Huang, D.; Allen, T. T.; Notz, W. I.; Miller, R. A. Sequential kriging optimization using multiple-fidelity evaluations. *Structural and Multidisciplinary Optimization* **2006**, *32*, 369–382.
- (62) Häussinger, P.; Glatthaar, R.; Rhode, W.; Kick, H.; Benkmann, C.; Weber, J.; Wunschel, H.-J.; Stenke, V.; Leicht, E.; Stenger, H. Noble Gases. *Ullmann's Encyclopedia of Industrial Chemistry* **2001**,
- (63) Banerjee, D.; Simon, C. M.; Elsaidi, S. K.; Haranczyk, M.; Thallapally, P. K. Xenon Gas Separation and Storage Using Metal-Organic Frameworks. *Chem* **2018**, *4*, 466–494.
- (64) Banerjee, D.; Cairns, A. J.; Liu, J.; Motkuri, R. K.; Nune, S. K.; Fernandez, C. A.; Krishna, R.; Strachan, D. M.; Thallapally, P. K. Potential of Metal-Organic Frameworks for Separation of Xenon and Krypton. *Accounts of Chemical Research* **2015**, *48*, 211–219.
- (65) Diercks, C. S.; Yaghi, O. M. The atom, the molecule, and the covalent organic framework. *Science* **2017**, *355*.

- (66) Côté, A. P.; Benin, A. I.; Ockwig, N. W.; O’Keeffe, M.; Matzger, A. J.; Yaghi, O. M. Porous, Crystalline, Covalent Organic Frameworks. *Science* **2005**, *310*, 1166–1170.
- (67) Wang, H.; Li, J. General strategies for effective capture and separation of noble gases by metal–organic frameworks. *Dalton Transactions* **2018**, *47*, 4027–4031.
- (68) Yuan, M.; Wang, X.; Chen, L.; Zhang, M.; He, L.; Ma, F.; Liu, W.; Wang, S. Tailoring Pore Structure and Morphologies in Covalent Organic Frameworks for Xe/Kr Capture and Separation. *Chemical Research in Chinese Universities* **2021**, *37*, 679–685.
- (69) Banerjee, D.; Simon, C. M.; Plonka, A. M.; Motkuri, R. K.; Liu, J.; Chen, X.; Smit, B.; Parise, J. B.; Haranczyk, M.; Thallapally, P. K. Metal–organic framework with optimally selective xenon adsorption and separation. *Nature Communications* **2016**, *7*, 1–7.
- (70) Jia, Z.; Yan, Z.; Zhang, J.; Zou, Y.; Qi, Y.; Li, X.; Li, Y.; Guo, X.; Yang, C.; Ma, L. Pore Size Control via Multiple-Site Alkylation to Homogenize Sub-Nanoporous Covalent Organic Frameworks for Efficient Sieving of Xenon/Krypton. *American Chemical Society Applied Materials & Interfaces* **2020**, *13*, 1127–1134.
- (71) Tong, M.; Lan, Y.; Yang, Q.; Zhong, C. Exploring the structure-property relationships of covalent organic frameworks for noble gas separations. *Chemical Engineering Science* **2017**, *168*, 456–464.
- (72) Ren, E.; Coudert, F.-X. Thermodynamic exploration of xenon/krypton separation based on a high-throughput screening. *Faraday Discussions* **2021**, *231*, 201–223.
- (73) Wang, J.; Zhou, M.; Lu, D.; Fei, W.; Wu, J. Virtual screening of nanoporous materials for noble gas separation. *American Chemical Society Applied Nano Materials* **2022**, *5*, 3701–3711.
- (74) Lin, W.-q.; Xiong, X.-l.; Liang, H.; Chen, G.-h. Multiscale Computational Screening of Metal–Organic Frameworks for Kr/Xe Adsorption Separation: A Structure–Property Relationship-Based Screening Strategy. *American Chemical Society Applied Materials & Interfaces* **2021**, *13*, 17998–18009.
- (75) Simon, C. M.; Mercado, R.; Schnell, S. K.; Smit, B.; Haranczyk, M. What Are the Best Materials to Separate a Xenon/Krypton Mixture? *Chemistry of Materials* **2015**, *27*, 4459–4475.
- (76) Cooley, I.; Efford, L.; Besley, E. Computational Predictions for Effective Separation of Xenon/Krypton Gas Mixtures in the MFM Family of Metal–Organic Frameworks. *The Journal of Physical Chemistry C* **2022**, *126*, 11475–11486.
- (77) Gantzler, N.; Kim, M.-B.; Robinson, A.; Terban, M. W.; Ghose, S.; Dinnebier, R. E.; York, A. H.; Tiana, D.; Simon, C. M.; Thallapally, P. K. Computation-informed optimization of Ni(PyC)₂ functionalization for noble gas separations. *Cell Reports Physical Science* **2022**, *3*, 101025.
- (78) Ryan, P.; Farha, O. K.; Broadbelt, L. J.; Snurr, R. Q. Computational Screening of Metal–Organic Frameworks for Xenon/Krypton Separation. *American Institute of Chemical Engineers Journal* **2010**, *57*, 1759–1766.
- (79) Sikora, B. J.; Wilmer, C. E.; Greenfield, M. L.; Snurr, R. Q. Thermodynamic analysis of Xe/Kr selectivity in over 137 000 hypothetical metal–organic frameworks. *Chemical Science* **2012**, *3*, 2217.
- (80) Parkes, M. V.; Staiger, C. L.; IV, J. J. P.; Allendorf, M. D.; Greathouse, J. A.

- Screening metal–organic frameworks for selective noble gas adsorption in air: effect of pore size and framework topology. *Physical Chemistry Chemical Physics* **2013**, *15*, 9093.
- (81) Chung, Y. G.; Haldoupis, E.; Bucior, B. J.; Haranczyk, M.; Lee, S.; Zhang, H.; Vogiatzis, K. D.; Milisavljevic, M.; Ling, S.; Camp, J. S.; Slater, B.; Siepmann, J. I.; Sholl, D. S.; Snurr, R. Q. Advances, Updates, and Analytics for the Computation-Ready, Experimental Metal–Organic Framework Database: CoRE MOF 2019. *Journal of Chemical & Engineering Data* **2019**, *64*, 5985–5998.
- (82) Gu, C.; Yu, Z.; Liu, J.; Sholl, D. S. Construction of an anion-pillared MOF database and the screening of MOFs suitable for Xe/Kr separation. *ACS Applied Materials & Interfaces* **2021**, *13*, 11039–11049.
- (83) Rappe, A. K.; Casewit, C. J.; Colwell, K. S.; Goddard, W. A.; Skiff, W. M. UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *Journal of the American Chemical Society* **1992**, *114*, 10024–10035.
- (84) Ren, E.; Coudert, F.-X. Thermodynamic exploration of xenon/krypton separation based on a high-throughput screening. *Faraday Discussions* **2021**, *231*, 201–223.
- (85) Willems, T. F.; Rycroft, C. H.; Kazi, M.; Meza, J. C.; Haranczyk, M. Algorithms and tools for high-throughput geometry-based analysis of crystalline porous materials. *Microporous and Mesoporous Materials* **2012**, *149*, 134–141.
- (86) Görtler, J.; Kehlbeck, R.; Deussen, O. A Visual Exploration of Gaussian Processes. *Distill* **2019**, <https://distill.pub/2019/visual-exploration-gaussian-processes>.
- (87) Rasmussen, C. E.; Williams, C. K. I. *Gaussian processes for machine learning; Adaptive computation and machine learning*; MIT Press, 2006.
- (88) Mikkola, P.; Martinelli, J.; Filstroff, L.; Kaski, S. Multi-Fidelity Bayesian Optimization with Unreliable Information Sources. *arXiv preprint arXiv:2210.13937* **2022**,
- (89) Tom, G.; Hickman, R. J.; Zinzuwadia, A.; Mohajeri, A.; Sanchez-Lengeling, B.; Aspuru-Guzik, A. Calibration and generalizability of probabilistic models on low-data chemical datasets with DIONYSUS. *Digital Discovery* **2023**,
- (90) van de Schoot, R.; Depaoli, S.; King, R.; Kramer, B.; Märten, K.; Tadesse, M. G.; Vannucci, M.; Gelman, A.; Veen, D.; Willemsen, J.; Yau, C. Bayesian statistics and modelling. *Nature Reviews Methods Primers* **2021**, *1*, 1–26.
- (91) Wang, M. et al. Unveiling Electronic Properties in Metal–Phthalocyanine-Based Pyrazine-Linked Conjugated Two-Dimensional Covalent Organic Frameworks. *Journal of the American Chemical Society* **2019**, *141*, 16810–16816.
- (92) Ren, E.; Coudert, F.-X. Gas Separation Selectivity Prediction Based on Finely Designed Descriptors. *ChemRxiv* **2023**,
- (93) Tran, K.; Neiswanger, W.; Yoon, J.; Zhang, Q.; Xing, E.; Ulissi, Z. W. Methods for comparing uncertainty quantifications for material property predictions. *Machine Learning: Science and Technology* **2020**, *1*, 025006.
- (94) Han, R.; Walton, K. S.; Sholl, D. S. Does chemical engineering research have a reproducibility problem? *Annual review of chemical and biomolecular engineering* **2019**, *10*, 43–57.
- (95) Park, J.; Howe, J. D.; Sholl, D. S. How reproducible are isotherm measurements in

- metal–organic frameworks? Chemistry of Materials **2017**, *29*, 10487–10495.
- (96) Seifrid, M.; Pollice, R.; Aguilar-Granda, A.; Morgan Chan, Z.; Hotta, K.; Ser, C. T.; Vestfrid, J.; Wu, T. C.; Aspuru-Guzik, A. Autonomous chemical experiments: Challenges and perspectives on establishing a self-driving lab. Accounts of Chemical Research **2022**, *55*, 2454–2466.
- (97) Ginsbourger, D.; Le Riche, R.; Carraro, L. Kriging is well-suited to parallelize optimization; Springer, 2010; Vol. 2; pp 131–162.
- (98) Belakaria, S.; Deshwal, A.; Doppa, J. R. Max-value Entropy Search for Multi-Objective Bayesian Optimization. Conference on Neural Information Processing Systems. 2019; pp 7823–7833.
- (99) Kalikmanov, V. I. Statistical physics of fluids: basic concepts and applications; Springer Science & Business Media, 2013.
- (100) Frenkel, D.; Smit, B. Understanding Molecular Simulation: From Algorithms to Applications; Computational science; Elsevier Science, 2001.
- (101) Dubbeldam, D.; Torres-Knoop, A.; Walton, K. S. On the inner workings of Monte Carlo codes. Molecular Simulation **2013**, *39*, 1253–1292.
- (102) Ren, E.; Coudert, F.-X. Rapid Adsorption Enthalpy Surface Sampling (RAESS) to Characterize Nanoporous Materials. Chemical Science **2023**,
- (103) Mason, J. A.; McDonald, T. M.; Bae, T.-H.; Bachman, J. E.; Sumida, K.; Dutton, J. J.; Kaye, S. S.; Long, J. R. Application of a High-throughput Analyzer in Evaluating Solid Adsorbents for Post-Combustion Carbon Capture via Multicomponent Adsorption of CO₂, N₂, and H₂O. Journal of the American Chemical Society **2015**, *137*, 4787–4803.
- (104) Vandenhoute, S.; Cools-Ceuppens, M.; DeKeyser, S.; Verstraelen, T.; Van Speybroeck, V. Machine learning potentials for metal-organic frameworks using an incremental learning approach. Nature Partner Journals Computational Materials **2023**, *9*, 19.
- (105) Yang, C.-T.; Pandey, I.; Trinh, D.; Chen, C.-C.; Howe, J. D.; Lin, L.-C. Deep learning neural network potential for simulating gaseous adsorption in metal–organic frameworks. Materials Advances **2022**, *3*, 5299–5303.
- (106) Heinen, J.; Dubbeldam, D. On flexible force fields for metal–organic frameworks: Recent developments and future prospects. Wiley Interdisciplinary Reviews: Computational Molecular Science **2018**, *8*, e1363.
- (107) Hossain, M. I.; Cunningham, J. D.; Becker, T. M.; Grabicka, B. E.; Walton, K. S.; Rabideau, B. D.; Glover, T. G. Impact of MOF defects on the binary adsorption of CO₂ and water in UiO-66. Chemical Engineering Science **2019**, *203*, 346–357.
- (108) Nandy, A.; Duan, C.; Kulik, H. J. Using Machine Learning and Data Mining to Leverage Community Knowledge for the Engineering of Stable Metal–Organic Frameworks. Journal of the American Chemical Society **2021**, *143*, 17535–17547.
- (109) Islamov, M.; Babaei, H.; Anderson, R.; Sezginel, K. B.; Long, J. R.; McGaughey, A. J. H.; Gomez-Gualdrón, D. A.; Wilmer, C. E. High-throughput screening of hypothetical metal-organic frameworks for thermal conductivity. Nature Partner Journals Computational Materials **2023**, *9*, 11.
- (110) Van Heest, T.; Teich-McGoldrick, S. L.; Greathouse, J. A.; Allendorf, M. D.;

Sholl, D. S. Identification of metal–organic framework materials for adsorption separation of rare gases: applicability of ideal adsorbed solution theory (IAST) and effects of inaccessible framework regions. The Journal of Physical Chemistry C **2012**, 116, 13183–13195.

- (111) Balandat, M.; Karrer, B.; Jiang, D. R.; Daulton, S.; Letham, B.; Wilson, A. G.; Bakshy, E. BoTorch: A Framework for Efficient Monte-Carlo Bayesian Optimization. *Advances in Neural Information Processing Systems* 33. 2020.
- (112) Gardner, J. R.; Pleiss, G.; Bindel, D.; Weinberger, K. Q.; Wilson, A. G. GPyTorch: Blackbox Matrix-Matrix Gaussian Process Inference with GPU Acceleration. *Advances in Neural Information Processing Systems*. 2018.