Recommendation System to Predict Missing Adsorption Properties of Nanoporous Materials

Arni Sturluson[1], Ali Raza[2], Grant D. McConachie[1], Daniel W. Siderius[3], Xiaoli Z. Fern[*2], and Cory M. Simon[*1]

[1]School of Chemical, Biological, and Environmental Engineering. Oregon State University. Corvallis, OR.
[2]School of Electrical Engineering and Computer Science. Oregon State University. Corvallis, OR, USA
[3]National Institute of Standards and Technology. Chemical Sciences Division. Gaithersburg, MD, USA.
[*]{xiaoli.fern, cory.simon}@oregonstate.edu

October 12, 2021

Abstract

Nanoporous materials (NPMs) selectively adsorb and concentrate gases into their pores and thus could be used to store, capture, and sense many different gases. Modularly synthesized classes of NPMs, such as covalent organic frameworks (COFs), offer a large number of candidate structures for each adsorption task. A complete NPM-property table, containing measurements of relevant adsorption properties in candidate NPMs, would enable the matching of NPMs with adsorption tasks. However, in practice, the NPM-property matrix is only partially observed (incomplete); many different properties of many different NPMs have not been measured. The idea in this work is to leverage the observed (NPM, property) values to impute the missing ones. Similarly, commercial recommendation systems impute missing entries in an incomplete product–customer ratings matrix to recommend products to customers. We demonstrate a COF recommendation system to match COFs with adsorption tasks by training a low-rank model of an incomplete COF–adsorption-property matrix constructed from simulated uptakes of CH_4, H_2O, H_2S, Xe, Kr, CO_2, N_2, O_2, and H_2 at various conditions. A low-rank model of the COF–adsorption-property matrix, fit to the observed (COF, adsorption property) values, provides (i) predictions of the missing (COF, adsorption property) values and (ii) a map of COFs, wherein COFs, represented as points, with similar (dissimilar) adsorption properties congregate (separate). The COF recommendation system is able to rank COFs reasonably well for most of the adsorption properties, but imputation performance diminishes precipitously when the fraction of missing entries exceeds 60 %. The concepts in our COF recommendation system can be applied broadly to impute missing data pertaining to many different materials and properties.

1 Introduction

Nanoporous materials (NPMs) [1] often exhibit permanent porosity and possess large-area internal surfaces [2] decorated with functional groups. This enables them to (selectively) adsorb and concentrate gases in their pores [3–5]. As a result, NPMs have applications in storing [5], separating [4, 6], and sensing [7] gases, as well as in catalysis [8].

Advanced families of NPMs, such as metal-organic frameworks (MOFs) [9], covalent organic frameworks (COFs) [10], porous polymer networks (PPNs) [11], porous organic cages (POCs) [12], and metal-organic polyhedra (MOPs) [13, 14], are constructed modularly from molecular building blocks. The copiousness of compatible building blocks within many topologies, together with postsynthetic modifiability [15], makes the number of possible NPM structures extremely large [16].

Thus, we have a large list of candidate NPMs and a list of their adsorption properties we wish to know for their many applications. If this NPM-property data table were complete, both searching for (i) the optimal NPM for a given application [1] and (ii) the optimal application for a given NPM [17] would be trivial look-up problems. However, in practice, the NPM-property data table, whether constructed from experimental adsorption measurements [18, 19] or molecular simulations of gas adsorption in libraries of NPMs [17], is likely incomplete because many (NPM, property) values have not been observed, i.e., (i) for any given NPM, only a proportion of its adsorption properties have been measured, and (ii) for any given adsorption property, it has been measured in only a proportion of the NPMs (see Figure 1).

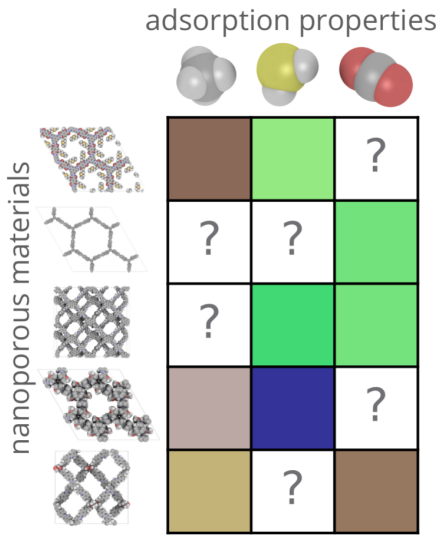


Figure 1: Recommendation system for nanoporous materials (NPMs). In this toy NPM–adsorption-property matrix, entry (m, p) represents the value of adsorption property p of NPM m. Many entries are unobserved ("?") because measurements are missing. The goal of our NPM recommendation system is to use the observed entries (values depicted by color) to impute the unobserved entries, allowing recommendation of NPMs for various adsorption tasks (requiring a certain adsorption property). This is analogous to commercial recommendation systems that aim to make product recommendations tailored for specific customers, with NPM:product::adsorption property:customer.

The idea in this work is to leverage the observed (NPM, property) values to predict the missing ones, i.e., to impute the missing values of, or complete, the NPM-property matrix. A machine learning strategy to complete the NPM-property matrix is much less expensive and time consuming than experimentally measuring or computationally simulating these missing properties. The machine-completed NPM-property matrix is valuable because it can be used to direct higher fidelity but more expensive (experimental or simulated) measurements toward the most promising materials, thereby using less resources in the search for the optimal NPM for a given application.

Our hypothesis, which would permit accurate matrix completion, is that the NPM-property matrix exhibits a low-rank structure [20,21], owing to underlying structural and chemical similarities among both NPMs and gases that dictate their interactions. Specifically, we assume a low-rank structure where both NPMs and adsorption properties can be represented by low-dimensional vectors that together express the affinity between a (NPM, property) pair. These latent representations can be jointly machine learned using the observed (NPM, property) values, then used to impute the missing ones [22,23].

Our goal to "fill in" the missing values in the NPM-property matrix—primarily to recommend NPMs for specific adsorption tasks—is analogous to the goal of a commercial recommendation system that recommends products to customers. For example, consider a movie recommendation system at Netflix [23]. Movie ratings by Netflix users are stored in a movie-user rating matrix (rows: movies, columns: users, entries: rating) [23]. The movie-user rating matrix is incomplete; most entries are missing because (i) each user rated only a small proportion of the movies and (ii) each movie is rated by only a small proportion of the users. A movie recommendation system leverages the observed (movie, user) ratings (perhaps, in addition to features of the movies and users) to impute the missing ones [24]. The machine-completed movie-user rating matrix is then used to make user-specific recommendations of movies. Thus, the (material, property) values in our material recommendation system are analogous to (product, customer) ratings in commercial recommendation systems.

Herein, we demonstrate a prototype recommendation system, based on a low-rank matrix model [22,23], that recommends COFs for various gas adsorption tasks. The COF–gas-adsorption-property matrices pertain to 572 experimentally reported COFs [25] and the simulated adsorption of CH_4, H_2O, H_2S, Xe, Kr, CO_2, N_2, O_2, and H_2 in those COFs at various conditions [17] relevant to different gas storage and separation applications. Advantageously, this COF–gas-adsorption-property matrix is in reality complete, allowing us to ablate different fractions of the entries and investigate how imputation performance depends on the fraction of missing values. From the observed (COF, gas adsorption) values, we machine learn a low-rank model of the COF–gas-adsorption-property matrix, giving low-dimensional latent vector representations of both the COFs and the adsorption properties. This low-rank model provides: (i) predictions of the missing (COF, gas adsorption property) values and (ii) a "map," wherein COFs are represented as points and COFs with similar (dissimilar) adsorption properties congregate (separate). Such a map of COFs is useful for experimental design to explore COF space and for the optimization of promising but still suboptimal "lead" COFs.

1.1 Review of Previous Work

Machine learning plays an important role in the discovery and deployment of NPMs [26–33]. Supervised machine learning models have been widely used to predict the adsorption properties of NPMs [34–43] from vectors of hand-crafted structural features [44, 45] or a graph representation [46, 47]. Unsupervised machine learning methods have been used to embed NPMs into a low-dimensional "material space" [48] and cluster together NPMs with similar structures [49–52]. Genetic algorithms [53–56], Monte Carlo tree search [57], and Bayesian optimization [58,59] have been used to more efficiently search for the NPM(s) with an optimal adsorption property. Finally, recently [60] an autoencoder enabled inverse design [61,62] of NPMs, where one specifies a desired adsorption

property, and the machine learning model generates a NPM structure with that property. To enable machine learning approaches to NPM discovery, several open, structured databases [63–65] of (i) crystal structure models of NPMs [25, 66–70], (ii) simulated [17, 69, 71, 72] and experimentally measured [19] adsorption properties of NPMs, and (iii) electronic properties of NPMs [73, 74] have been curated. Text mining and natural language processing could be used to extract data and knowledge from the literature for machine learning studies as well [75–77].

Our material recommendation system deviates from previous data-driven approaches to predict properties of NPMs by: (i) as a latent variable model, embedding materials into a latent space, negating the need for explicit, hand-crafted features of the NPMs, and (ii) performing multitask prediction while (a) transferring knowledge between tasks and (b) handling missing values in the target vectors associated with NPMs. Loosely related, meta-learning has been used to predict an adsorption property of materials at different conditions by learning an intermediate representation of the material based only on available adsorption data [78]. N.b., recommendation systems have been built for use in chemical sciences to impute missing gas permeabilities in polymers [79], antiviral activities of molecules [80], and stabilities of inorganic materials [81, 82].

2 Material Recommendation System

Here, we formulate the general problem of material-property matrix completion.

A material recommendation system jointly machine learns, from observed (material, property) values, low-dimensional latent vector representations of the materials and properties that express (material, property) affinities. These learned representations allow us to (i) impute missing (material, property) values and (ii) draw a map of the materials, wherein materials with similar properties congregate.

2.1 Material—Property Matrix

Suppose we have M candidate materials with P properties of interest. Entry (m, p) of the $M \times P$ material-property matrix $\mathbf{A}$, $A_{mp} \in \mathbb{R}$, represents the value of property p of material m.

2.2 Observations (Data)

We have observations of A_{mp} for $(m, p) \in \Omega \subset \{1, 2, \ldots, M\} \times \{1, 2, \ldots, P\}$, which defines Ω as the set of ordered pairs describing the entries in $\mathbf{A}$ that are observed. That is, the material-property matrix $\mathbf{A}$ is not complete; some entries are missing ($|\Omega| < MP$).

2.3 Objective

The objective is to leverage the observations $\{A_{mp}\}_{(m,p)\in\Omega}$ to complete the material-property matrix, i.e. to predict the missing entries, $\{A_{mp}\}_{(m,p)\in\{1,2,\ldots,M\}\times\{1,2,\ldots,P\}\setminus\Omega}$.

2.4 The low-rank matrix model

From an element perspective, the low-rank model [22, 23] assumes that each element of the matrix, A_{mp}, decomposes into

$$A_{mp} \approx \mathbf{m}_m^\intercal \mathbf{p}_p + \mu_m, \tag{1}$$

where $\mathbf{m}_m \in \mathbb{R}^k$ and $\mathbf{p}_p \in \mathbb{R}^k$ are low-dimensional ($k < M, P$), latent vector representations of material m and property p, respectively, and $\mu_m \in \mathbb{R}$ is a bias for material m. The material-property interaction term, the dot product $\mathbf{m}_m^\intercal \mathbf{p}_p$, represents the "affinity" (if positive) or "aversion" (if negative) of material m for property p. Geometrically, the interaction term (i) depends on both the norm of the vectors $\mathbf{m}_m$ and $\mathbf{p}_p$ and the angle between them and (ii) is positive (negative) if the vectors point in roughly the same (opposite) directions. The material bias μ_m reflects variation of the values of the properties of material m independent of interactions; some materials may simply tend to have higher or lower values of the properties.

From a matrix perspective, the low-rank model factorizes the material-property matrix $\mathbf{A}$ as:

$$\mathbf{A} \approx \mathbf{M}^\intercal \mathbf{P} + \boldsymbol{\mu}\mathbf{1}^\intercal \tag{2}$$

with the columns of matrices $\mathbf{M} \in \mathbb{R}^{k \times M}$ and $\mathbf{P} \in \mathbb{R}^{k \times P}$ containing the latent representations of materials and properties, respectively; the entries of the column vector $\boldsymbol{\mu} \in \mathbb{R}^M$ containing the material biases; and $\mathbf{1} \in \mathbb{R}^P$ a column vector of ones. See Figure 2. The dimensionality of the latent space, $k < M, P$, imposes the constraint $\text{rank}(\mathbf{M}^\intercal \mathbf{P}) \leq k$, hence eq 2 is a low-rank approximation of the matrix $\mathbf{A}$.

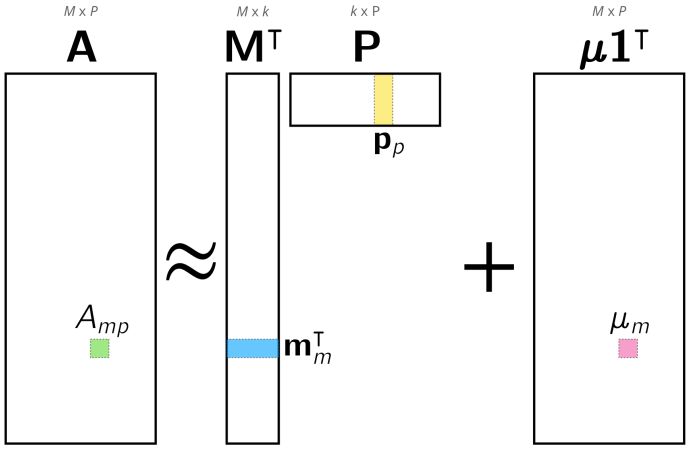


Figure 2: The low-rank model of the material-property matrix $\mathbf{A} \approx \mathbf{M}^\intercal \mathbf{P} + \boldsymbol{\mu}\mathbf{1}^\intercal$. The columns of $\mathbf{M}$ and $\mathbf{P}$ contain the latent representations of the M materials and P properties, respectively, which lie in a k-dimensional space. The vector $\boldsymbol{\mu}$ contains the M material biases. Entry (m, p) of $\mathbf{A}$ is modeled as $A_{mp} \approx \mathbf{m}_m^\intercal \mathbf{p}_p + \mu_m$.

2.5 The utility of the low-rank model

The low-rank model of the materials-property matrix is useful for two purposes [22].

(1) Imputation of missing entries. The decomposition in eq 1 holds for both observed and unobserved (material, property) values. Thus, once we learn $\mathbf{M}$, $\mathbf{P}$, and $\boldsymbol{\mu}$ from the *observed* entries, we can predict the unobserved entries, as is clear from eq 2.

(2) Construction of a low-dimensional map of the materials and properties. The rows of a fully observed version of $\mathbf{A}$, which lie in a P-dimensional vector space, can be viewed as feature vectors of the materials. In this view, each material is represented by a list of its properties. The set of latent vector

representations of the materials, in the rows of $\mathbf{M}^\intercal$, are embeddings/ compressions of the rows of $\mathbf{A}$ into a lower ($k < P$) dimensional vector space. [22] Within this latent space, materials, represented by $\{\mathbf{m}_m\}$, that tend to have similar (dissimilar) properties congregate (separate). Using dimension reduction techniques, we can visualize the scatter of the materials in the low-dimensional space to draw a "map" of materials. The latent representations of the materials, $\{\mathbf{m}_m\}$, and the map that visualizes them are useful for: (i) grouping together/organizing materials with similar properties, (ii) lead-optimization, where we search the map for materials nearby a "lead" material with a good but still suboptimal property, (iii) selecting diverse materials to efficiently explore material space in an experimental design, and (iv) training supervised machine learning models for other prediction tasks, as $\mathbf{m}_m$ is a feature vector for material m.

Similarly, the columns of a complete version of $\mathbf{A}$ can be viewed as vector representations of the properties, and the columns of $\mathbf{P}$, the latent vector representations of the properties, are embeddings/compressions of them. Within this latent space, properties, represented by $\{\mathbf{p}_p\}$, that tend to take on similar (dissimilar) values in NPMs congregate (separate).

The geometric interpretation of the dot product $\mathbf{m}_m^\intercal \mathbf{p}_p$ in eq 1 makes a comparison between the map of materials and the map of properties useful. The magnitudes of and the angle between a pair of latent material and property vectors $(\mathbf{m}_m, \mathbf{p}_p)$, taken together, indicate the affinity/aversion of material m with property p.

2.6 Machine Learning the low-rank model

We learn the latent representations of the materials and properties and the material biases from the observed (material, property) values by balancing (i) the matching of the observed entries of the matrix by the model given in eq 1 and (ii) the simplicity of the latent vector representations, to avoid overfitting. Specifically, we aim to choose the $\mathbf{M}$, $\mathbf{P}$, and $\boldsymbol{\mu}$ that minimize the loss $\ell = \ell(\mathbf{M}, \mathbf{P}, \boldsymbol{\mu})$:

$$\ell(\mathbf{M}, \mathbf{P}, \boldsymbol{\mu}) = \sum_{(m,p)\in\Omega} [A_{mp} - (\mathbf{m}_m^\intercal \mathbf{p}_p + \mu_m)]^2 + \lambda \left(\frac{1}{M} \sum_{m=1}^{M} ||\mathbf{m}_m||_2^2 + \frac{1}{P} \sum_{p=1}^{P} ||\mathbf{p}_p||_2^2 \right). \quad (3)$$

The first term is the approximation error, measured over all observations. The second term provides L2 regularization of the latent vector representations of the materials and properties to prevent overfitting and improve generalization, where $\lambda > 0$ is the regularization parameter. The sums are normalized to properly weigh regularization of the latent material and property vectors.

Stochastic gradient descent or alternating minimization are commonly used to find a $(\mathbf{M}, \mathbf{P}, \boldsymbol{\mu})$ that (locally [22]) minimize ℓ [22, 23]. In alternating minimization, we alternate between optimizing $\mathbf{M}$ with $\mathbf{P}$ fixed and optimizing $\mathbf{P}$ with $\mathbf{M}$ fixed.

Two hyperparameters are involved in fitting a low-rank matrix model to the observed entries of $\mathbf{A}$: (1) $k \in \{0, 1, \ldots, \min(M, P)\}$, the dimensionality of the vector space containing the latent representations of the materials and properties and (2) $\lambda \in [0, \infty)$, the regularization parameter that trades off prediction accuracy on the training data and the simplicity of the latent vector representations.

3 Case Study: A COF Recommendation System

We now demonstrate a material recommendation system based on a low-rank matrix model. Here, materials are COFs, and properties are the equilibrium uptakes of a variety of gases at different conditions, obtained from molecular simulations. The Julia code to reproduce all of our work is available at `github.com/SimonEnsemble/material_recommendation_system`.

3.1 Dataset

We construct the COF–adsorption-property matrix using an open data set (v9 on Materials Cloud [64]) of simulated gas adsorption properties in $M = 572$ experimentally reported, porous COF materials [17,25]. We selected $P = 16$ simulated adsorption properties—the uptake [units: mmol/g] and Henry coefficients [mmol/(g·bar)] of a variety of gases (CH_4, H_2O, H_2S, Xe, Kr, CO_2, N_2, O_2 and H_2) at various conditions pertinent to gas storage and separation applications (see Table 1). We $\log_{10}$-transformed the Henry coefficients because of the relatively long tail of their distributions. The resulting COF–adsorption-property matrix, $\mathbf{A}_{\text{complete}} \in \mathbb{R}^{572\times 16}$ is fully observed, allowing us to study how imputation performance of the low-rank model depends on the completeness of the matrix.

Table 1: List of Equilibrium Adsorption Properties Included in Our COF Recommendation System.

adsorption property	thermodynamic condition	units
O_2 uptake	298 K, 5 bar	mmol/g
O_2 uptake	298 K, 140 bar	mmol/g
CO_2 uptake	300 K, 0.001 bar	mmol/g
CO_2 uptake	300 K, 30 bar	mmol/g
N_2 uptake	300 K, 0.001 bar	mmol/g
N_2 uptake	300 K, 30 bar	mmol/g
H_2 uptake	77 K, 5 bar	mmol/g
H_2 uptake	77 K, 100 bar	mmol/g
H_2 uptake	298 K, 5 bar	mmol/g
H_2 uptake	298 K, 100 bar	mmol/g
CH_4 uptake	298 K, 65 bar	mmol/g
CH_4 uptake	298 K, 5.8 bar	mmol/g
H_2O Henry coefficient	300 K	mmol/(g·bar)
H_2S Henry coefficient	300 K	mmol/(g·bar)
Xe Henry coefficient	300 K	mmol/(g·bar)
Kr Henry coefficient	300 K	mmol/(g·bar)

Figure 3 displays the distribution of and pairwise relationships between [standardized] adsorption properties, and Figure S1 displays a pairwise correlation matrix of properties. Some properties are strongly correlated, e.g., CH_4 uptake at (298 K, 65 bar) and O_2 uptake at (298 K, 140 bar), while others, such as H_2 uptake at (77 K, 5 bar) and H_2S Henry coefficients at 300 K, are not. The low-rank model exploits these correlations between properties to learn low-dimensional representations of

materials and properties.

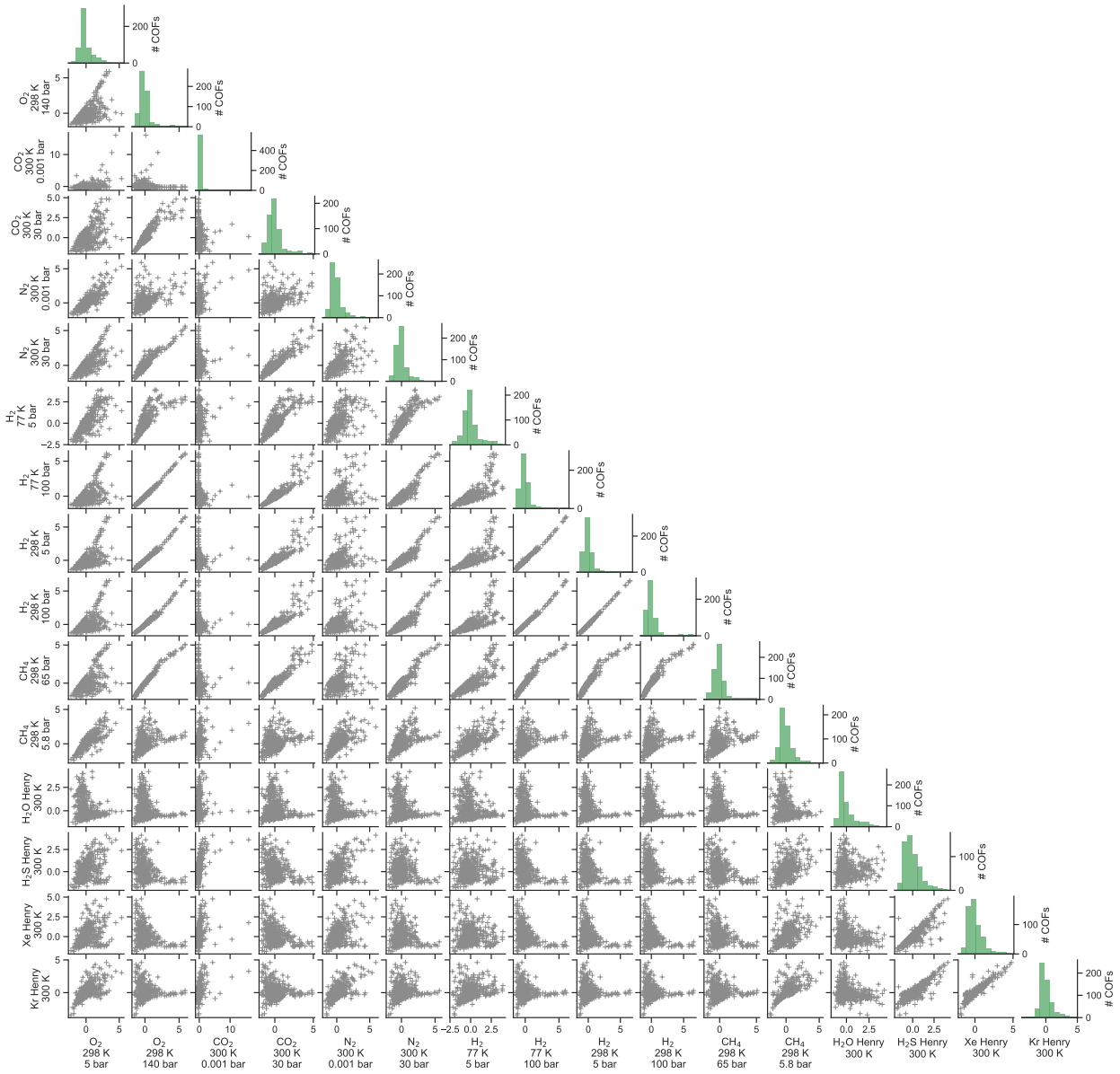


Figure 3: The distribution of (diagonal) and pairwise relationships between (off-diagonal; each point represents a COF) the simulated gas adsorption properties of the COFs (data from Ref. [17]). Each property was z-score standardized to have zero mean and unit variance.

3.2 Methods

3.2.1 Simulating the Process of Data Collection

We simulate the stochastic process of incomplete data collection to construct an incomplete COF–adsorption-property matrix $\mathbf{A}^{(\theta)}$ (still $M = 572 \times P = 16$) with a fraction $\theta \in [0, 1]$ of observed entries. We construct $\mathbf{A}^{(\theta)}$ by (uniform) randomly sampling, without replacement, $(1 - \theta)MP$ en-

tries to ablate (change to `missing`) from the MP entries of $\mathbf{A}_{\text{complete}}$. Figure S2 in the Supporting Information visualizes an instance of an incomplete COF–adsorption-property matrix $\mathbf{A}^{(\theta=0.4)}$.

3.2.2 Standardization of Adsorption Properties

In accordance with standard practice [22], we z-score standardize each adsorption property, i.e., we subtract from each column of $\mathbf{A}^{(\theta)}$ the mean of the corresponding adsorption property and then divide by its standard deviation (each computed using only the observed, training examples). We z-score standardize the adsorption properties for two reasons. First, standardization prevents properties with a larger range/variance from dominating the loss function in eq 3. Second, z-score standardization centers the latent vectors at the origin and is consistent with underlying model assumptions, e.g., a probabilistic interpretation of (closely related) principal component analysis [83] is that the data matrix is observed with normally distributed errors, justifying z-score standardization as opposed to Min–Max normalization [22].

3.2.3 Training, Hyperparameter Tuning, and Testing

Given an incomplete COF–adsorption-property matrix $\mathbf{A}^{(\theta)}$, we outline how we arrive at a low-rank model for deployment and test its imputation performance. Naturally, the [simulated] observed entries of $\mathbf{A}^{(\theta)}$ are used for training and hyperparameter tuning, while the [simulated, but actually known] unobserved entries are used as test data to evaluate the generalization error of the final deployment low-rank model.

3.2.4 Fitting

To fit a low-rank matrix model to training observations, we use `LowRankModels.jl` [22] in the Julia programming language [84]. `LowRankModels.jl` implements alternating proximal gradient descent [22] to minimize the loss in eq 3 over the training observations.

3.2.5 Hyperparameter Selection

We aim to determine the optimal hyperparameters $(k_{opt}^{(\theta)}, \lambda_{opt}^{(\theta)})$ for the low-rank model. We randomly partition the [simulated] observed entries of $\mathbf{A}^{(\theta)}$ into an 80/20 % training/validation set. We then fit a set of low-rank models, each with different hyperparameters, to the training data. The optimal hyperparameters follow from those of the model that gives the lowest imputation error (RMSE) over the validation set. The hyperparameter search is conducted over a regular 2D grid of (1) $k \in \{1, 2, \ldots, 15\}$ and (2) 25 values of λ ranging from 10 to 1000 and evenly spaced on a log scale.

3.2.6 Deployment of Low-Rank Model and Testing

Finally, the *deployment* low-rank model is a new low-rank model fit to all [simulated] observed entries in $\mathbf{A}^{(\theta)}$ with hyperparameters $(k_{opt}^{(\theta)}, \lambda_{opt}^{(\theta)})$. We evaluate the performance of the deployment model by comparing its predictions of the missing entries to the actual values of the missing entries that comprise the test data.

N.b., the loss and performance metrics are computed on the standardized and, in the case of Henry coefficients, $\log_{10}$-transformed values.

3.3 Results for Observed Fraction $\theta = 0.4$

We now demonstrate the utility of a low-rank matrix model, using a particular instance of an incomplete COF–adsorption-property matrix $\mathbf{A}^{(\theta=0.4)}$, for (i) imputing missing (COF, adsorption property) values and (ii) drawing of a map of COFs and adsorption properties. Figures 4–6 all pertain to the same deployment low-rank model trained using the instance of $\mathbf{A}^{(\theta=0.4)}$ visualized in Figure S2. Our hyperparameter sweep found $k_{opt}^{(0.4)} = 9$, $\lambda_{opt}^{(0.4)} = 100$.

3.3.1 Imputing the Missing Entries

We judge the performance of the deployment low-rank matrix model for imputing the missing entries of the COF–adsorption-property matrix $\mathbf{A}^{(\theta=0.4)}$ by comparing the predictions of the missing entries to the actual values in the test data set, composed of the [simulated] unobserved entries.

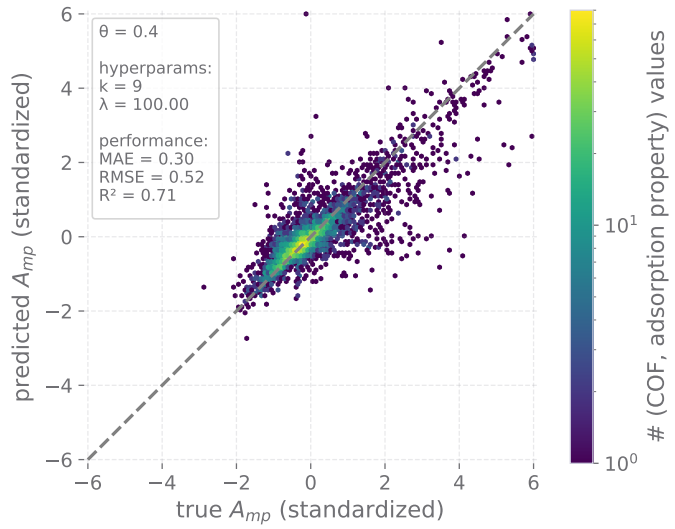
The parity plot in Figure 4a shows the joint distribution of the predictions for and actual values of the missing (COF, adsorption property) entries in the test data set. The density is greatest along the diagonal line of equality. The mean absolute error (MAE), root mean square error (RMSE), and coefficient of determination (R^2) are 0.3, 0.52, and 0.71, respectively. The magnitude of the RMSE and MAE are directly interpretable as units of a standard deviation in the adsorption property because they are computed on the z-score standardized properties. These metrics indicate that the low-rank model imputes the missing entries of the matrix with reasonable accuracy.

The ultimate utility of the recommendation system is to rank COFs according to specific properties (for specific applications). Spearman's rank (here, a ranking of COFs) correlation coefficient, ρ, between the prediction of a missing (COF, adsorption property) value by the deployment low-rank model and its actual value (from the test set), grouped by adsorption property, is shown in Figure 4b (blue bars). If the recommendation system were to *randomly* rank the COFs, ρ would be zero, and the search for the optimal COF orchestrated by the recommendation system would be equivalent to a random, trial-and-error search; a perfectly accurate ranking of the COFs would result in $\rho = 1$. With the exception of H_2O Henry coefficients, the recommendation system ranks COFs according to their properties reasonably well, with $\rho > 0.6$. The relatively poor ranking of COFs by H_2O Henry coefficient is explained by its very weak correlation with the other properties (see Figure S1).

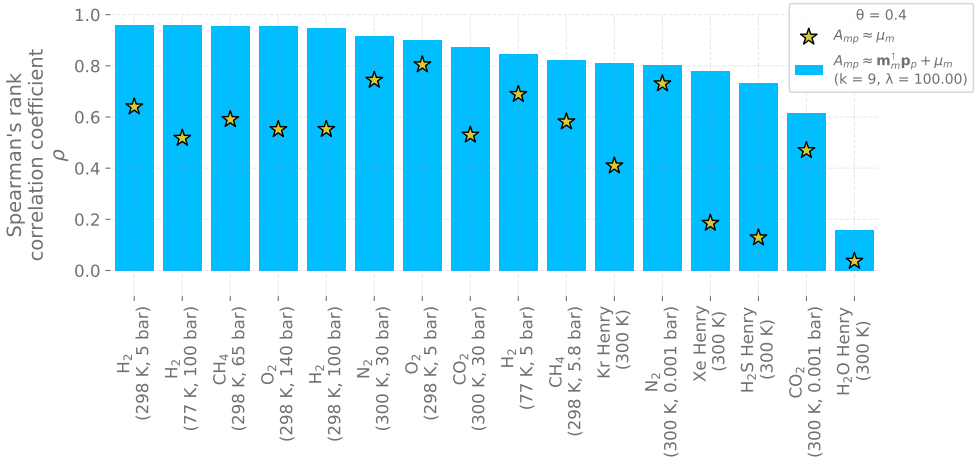
Figure S3 shows the correlation between and the distributions of the material biases $\{\mu_m\}$ and the interaction terms $\{\mathbf{m}_m^\intercal \mathbf{p}_p\}$, for the unobserved entries, in the deployment low-rank model. The distributions are centered around approximately zero and exhibit a similar range, indicating that both play a role in the prediction of the missing adsorption properties.

3.3.2 Comparing Imputation Performance to a Benchmark Model $A_{mp} \approx \mu_m$.

As a baseline to judge the imputation performance of our recommendation system, we also train and test (on the same data) a benchmark model that excludes the interaction term $\mathbf{m}_m^\intercal \mathbf{p}_p$ in eq 1



(a)



(b)

Figure 4: Performance of the deployment low-rank model in the task of imputing the missing (COF, adsorption property) values in $\mathbf{A}^{(\theta=0.4)}$ comprising the test set. (a) Parity plot showing the joint distribution of the predictions for and actual values of the missing (COF, adsorption property) values. The diagonal line represents perfect prediction. (b) Bar plot showing Spearman's rank correlation coefficient ρ between the predictions for and actual values of the missing (COF, adsorption property) values, grouped by adsorption property. For comparison, the stars show ρ for the benchmark, material bias model where $A_{mp} \approx \mu_m$ and the interaction term is excluded.

($\implies k = 0$). This material bias model $A_{mp} \approx \mu_m$ [naively] considers only whether the COF in question tends to exhibit high or low values of the properties (reflected in μ_m) when predicting A_{mp}. By comparing the imputation performance of this $k = 0$ material bias model with the $k > 0$ low-rank model, we quantify the extent to which the interactions between the COFs and the gas adsorption properties—encoded in $\mathbf{m}_m^\intercal \mathbf{p}_p$ terms for $k > 0$—are useful in the recommendation

system for imputing the missing values.

For each adsorption property, the stars in Figure 4b show Spearman's rank correlation coefficients between the predictions of the missing (COF, adsorption property) values made by the benchmark material bias model $A_{mp} \approx \mu_m$ and the actual values. Indeed, the interaction term enhances the ability of the recommendation system to rank COFs according to each adsorption property, though by different margins depending on the property. O_2 adsorption at (298 K, 5 bar) and N_2 adsorption at (300 K, 0.001 bar) are two properties where the interaction term is playing only a marginal role. Overall, this indicates that our recommendation system is (i) learning interactions between COFs and the adsorption properties and (ii) more likely to suggest high-performing COFs for an application than a simpler strategy that selects COFs purely based on how they perform on average (as in the material bias model).

Unsurprisingly, the material biases μ_m in the $k > 0$ low-rank model in eq 1 are strongly correlated with the material biases in the benchmark material bias model $A_{mp} \approx \mu_m$ (see Figure S4).

3.3.3 The COF biases

The learned material bias of COF m, μ_m in eq 1, roughly describes the typical value of the (standardized) gas adsorption properties of COF m. Figure 5 shows a partial, sorted bar plot of the material biases $\{\mu_m\}$ of the COFs and displays the COF structures with the lowest and highest material biases. Py-1P-quasi-AB (COF-LZU8) has the largest (smallest) μ_m, indicating that Py-1P-quasi-AB (COF-LZU8) tends to exhibit the highest (lowest) values of the (standardized) gas adsorption properties among the COFs. Given a new gas adsorption task for which we seek a maximal uptake, the high material bias μ_m of Py-1P-quasi-AB makes it a good candidate for measurements, in the absence of any other information; in the analogy of movie recommendation systems, Py-1P-quasi-AB is like a movie that is widely liked. Figure S3 shows the distribution of $\{\mu_m\}$.

3.3.4 Learned Map of COFs and Gas Adsorption Properties

We now visualize and analyze the learned latent vector representations of the COFs, $\{\mathbf{m}_m\}$, and of the adsorption properties, $\{\mathbf{p}_p\}$, in the deployment low-rank matrix model of $\mathbf{A}^{(\theta=0.4)}$. We resort to principal component analysis (PCA) to project $\{\mathbf{m}_m\}$ and $\{\mathbf{p}_p\}$, contained in the columns of $\mathbf{M}$ and $\mathbf{P}$, respectively, onto the 2D subspace spanned by the first two principal components of $\mathbf{M} \| \mathbf{P}$, where $\cdot \| \cdot$ denotes horizontal concatenation. The dimension reduction incurred a relative reconstruction error of 63 % based on the Frobenius norm.

The resulting map of COFs in Figure 6a visualizes the organization of COF vectors $\{\mathbf{m}_m\}$ in the latent space. Because the learned latent representation of COF m, $\mathbf{m}_m$, encodes its adsorption properties, COFs with similar (dissimilar) adsorption properties are expected to congregate (separate) in the map. To illustrate, each COF vector in Figure 6a is colored by the (left) CH_4 adsorption at (298 K, 65 bar), (middle) H_2S Henry coefficient at 300 K, and (right) H_2O Henry coefficient at 300 K. Indeed, nearby COFs in the map tend to exhibit similar values of a given adsorption property: COFs with the highest CH_4 uptake at (298 K, 65 bar), H_2S Henry coefficients at 300 K, and H_2O Henry coefficients at 300 K, respectively, tend to lie on the bottom left, top, and top right of the latent COF space. Figure S7

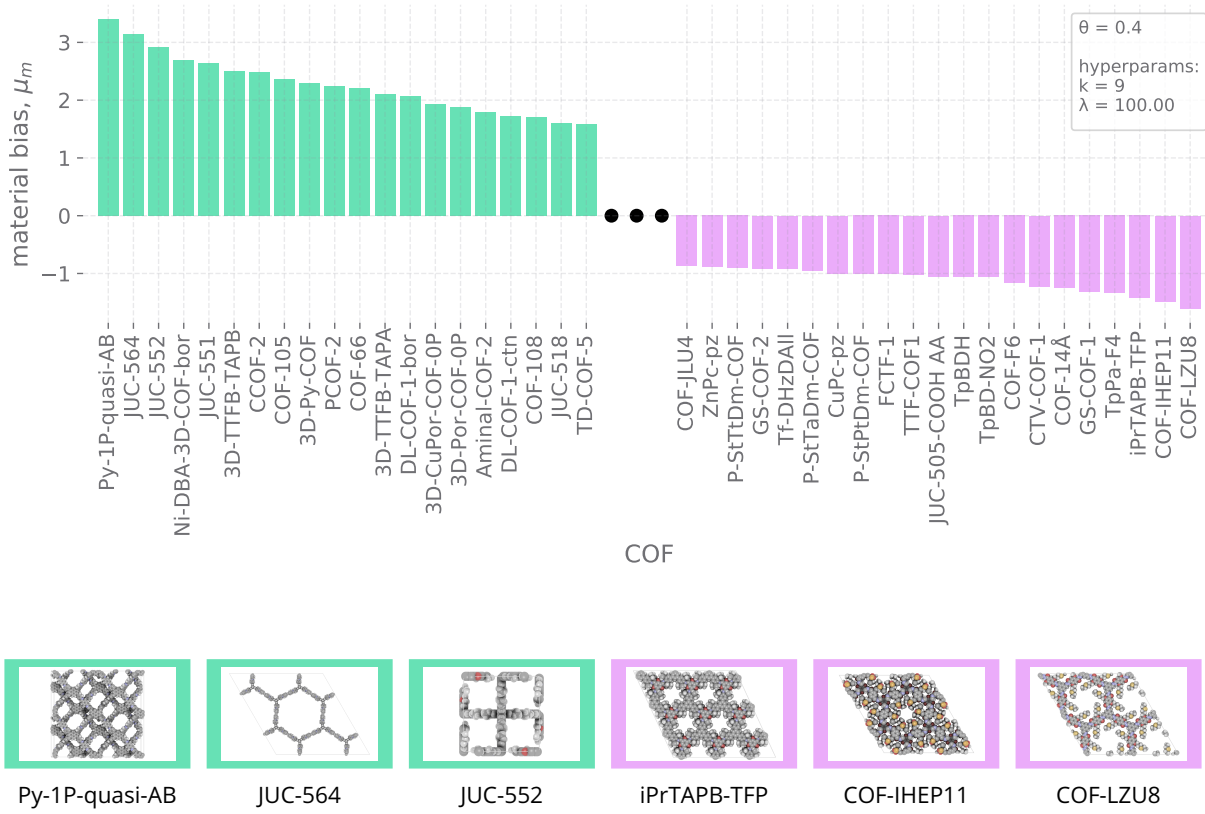


Figure 5: Ranked COF biases, $\{\mu_m\}$, from the deployment low-rank model. Only the COFs with the lowest and highest μ_m are shown. The bottom row visualizes the three top- and bottom-ranked COF structures.

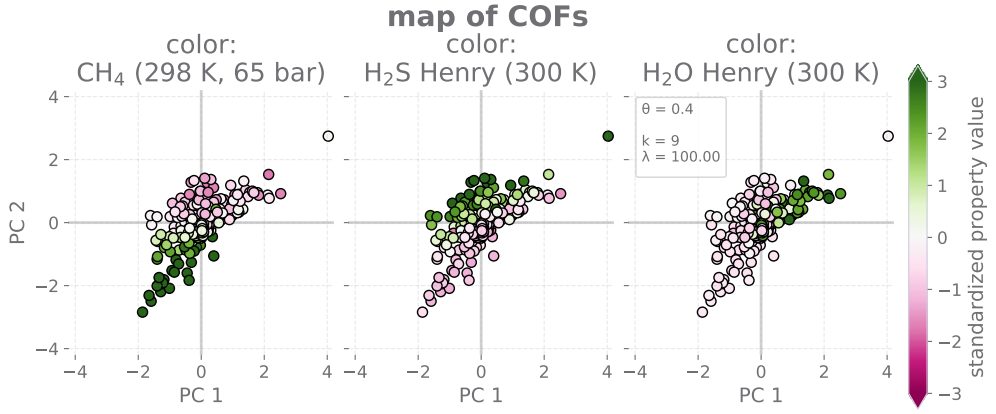
shows the COF map colored by the remaining adsorption properties.

Figure S8 shows the complete map of adsorption properties–i.e., the organization of the adsorption property vectors $\{\mathbf{p}_p\}$ in the latent space. Because the interaction term in eq 1 is the dot product $\mathbf{m}_m^\intercal \mathbf{p}_p$, the location of the latent property vector $\mathbf{p}_p$ indicates the region of latent space that tends to contain COFs with high values of adsorption property p. To illustrate, compare (i) the location of the latent property vectors $\{\mathbf{p}_p\}$ pertaining to the CH_4 adsorption at (298 K, 65 bar), H_2S Henry coefficient at 300 K, and H_2O Henry coefficient at 300 K in Figure 6b and (ii) the location of the COFs with the highest and lowest values of these properties in Figure 6a. The COFs whose latent vectors are oriented in the same (opposite) direction of the latent vector of a property tend to have large (small) values of that property.

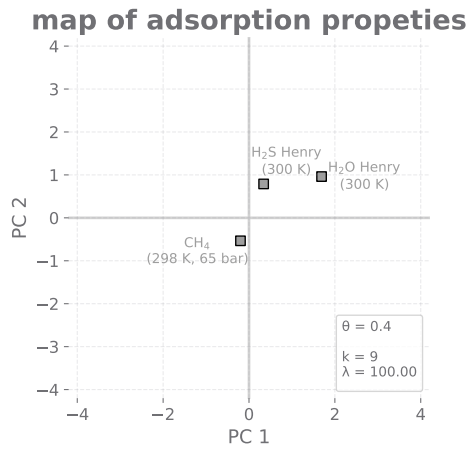
In summary, Figure 6 illustrates that a low-rank model of an incomplete COF–adsorption-property matrix machine learns a "map" of COFs, wherein COFs with similar adsorption properties congregate. These latent representations were learned from the observed values in an *incomplete* COF–adsorption-property matrix. The map of COFs, wherein proximity implies similarity of adsorption properties, is practically useful for: (1) lead optimization, where we search the latent space for nearest neighbors of a lead COF with good but insufficient performance, (2) selecting diverse sets of COFs

in an experimental design strategy to efficiently explore COF space, and (3) building supervised machine learning models to predict other properties of COFs, where the latent representations of the COFs can serve as feature vectors.

In movie recommendation systems, the latent variables of the movies, contained in the vector representations of the movies learned via a low-rank model from the observed (movie, user) ratings, may correspond to interpretable, intuitive features of movies such as genre or orientation toward children; however, they may also be uninterpretable [23]. It would be interesting to inspect the relationship between the learned latent variables in the vectors $\{\mathbf{m}_m\}$ and the structures of the COFs they represent, which were not explicitly input to the recommendation system.



(a)



(b)

Figure 6: Learned map of COFs and gas adsorption properties: the latent representations of (a) COFs, $\{\mathbf{m}_m\}$, and (b) a subset of the adsorption properties, $\{\mathbf{p}_p\}$, projected onto the 2D subspace spanned by the first two principal components (PCs) of $\mathbf{M} \parallel \mathbf{P}$. (a) Each point represents a COF, colored by (left) CH_4 adsorption at (298 K, 65 bar), (middle) H_2S Henry coefficients at 300 K and (right) H_2O Henry coefficients at 300 K. (Outlier: Py-1P-quasi-AB.) (b) Each point represents an adsorption property.

3.4 Effect of Observed Fraction θ on Performance

How complete must the COF–adsorption-property matrix be for the recommendation system to reliably rank COFs according to their adsorption properties? Because the COF–adsorption-property matrix is in reality complete, we have the luxury of studying the impact of the fraction of observed entries, θ, on the imputation performance of the recommendation system.

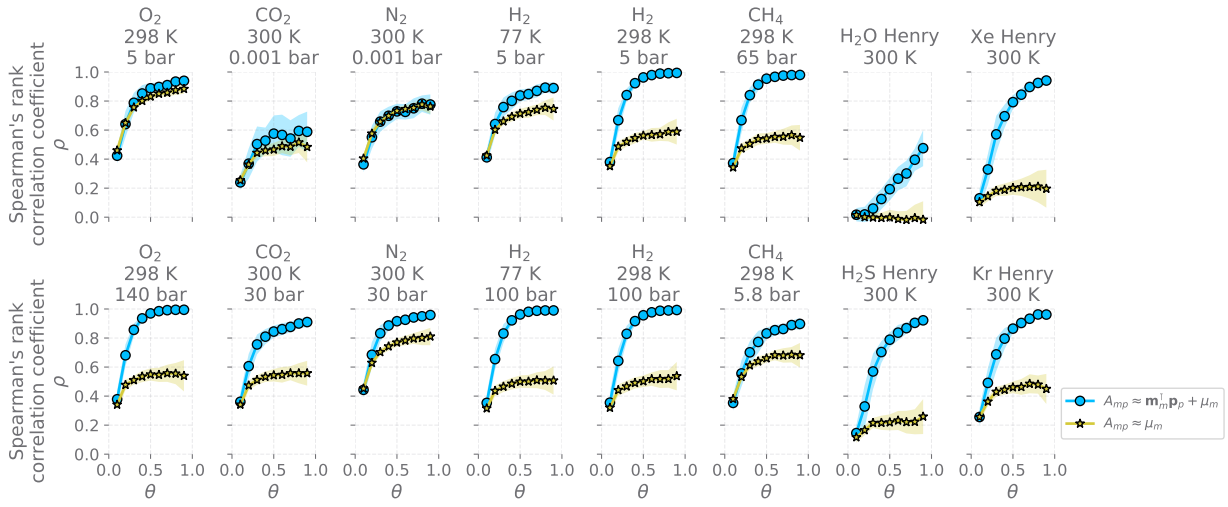


Figure 7: Effect of the fraction of observed values θ on the performance of the recommendation system for imputing missing (COF, adsorption property) values. Blue circles show the mean (over 50 simulations of data collection) Spearman's rank correlation coefficient between the prediction of the missing (COF, adsorption property) value and its true value, as a function of θ, grouped by adsorption property. Yellow stars show the mean Spearman's rank correlation coefficient for the benchmark material bias model. Shaded bands signify the standard deviation.

For fraction of observed values $\theta \in \{0.1, 0.2, \ldots, 0.9\}$, we sampled an ensemble of COF–adsorption property matrices $\mathbf{A}^{(\theta)}$ (50 simulations of data collection for each θ). For each instance of $\mathbf{A}^{(\theta)}$, we conducted a hyperparameter sweep using a training/validation split of the observed entries, trained a deployment model on all observed entries using the optimal hyperparameters, and then tested the imputation performance of the deployment model on the unobserved entries serving as test data. Figure S5 shows the distribution (among the simulations of data collection) of optimal hyperparameters $(k_{opt}^{(\theta)}, \lambda_{opt}^{(\theta)})$ for each θ. Figure 7 shows Spearman's rank correlation coefficient, ρ, between the prediction of the missing (COF, adsorption property) value by the deployment low-rank model and its actual value, grouped by adsorption property, as θ varies. The bands show the standard deviation over the 50 simulations of data collection. As in Figure 4b, ρ for the H_2O Henry coefficient is unsatisfactory $\forall\theta$ owing to its poor correlation with the other properties. For the majority of the gas adsorption properties, the recommendation system ranks COFs according to the property reasonably well (i.e., a reasonably high $\rho > 0.7$), until θ is reduced below $\theta = 0.4$. With such a paucity of training examples for $\theta < 0.4$, the low-rank model cannot learn useful latent representations of COFs and adsorption properties, resulting in diminished imputation performance. For comparison, ρ for the baseline material bias model is also shown in Figure 4b; the interaction term provides sig-

nificant predictive value, with the exception of N_2 adsorption at (300 K, 0.001 bar) and O_2 adsorption at (298 K, 5 bar).

In conclusion, the imputation performance of the low-rank matrix model rapidly diminishes when more than 60 % of the entries in the COF–adsorption-property matrix are missing. As the fraction of observed entries θ in the matrix increases, the recommendation system ranks COFs according to their adsorption properties more reliably.

3.5 Uncertainty Quantification

Suppose our material recommendation system predicts some missing (material, property) value to be optimal for an adsorption-based engineering application. To motivate an experimental measurement in the lab, we may wish to quantify the uncertainty associated with this prediction. One approach to quantify the uncertainties associated with the (material, property) values imputed by the low-rank model in eq 2 is through bootstrap resampling of the observed entries.

We demonstrate uncertainty quantification via bootstrapping for our COF recommendation system based on an instance of an incomplete COF–adsorption-property matrix, $\mathbf{A}^{(\theta=0.4)}$. First, we select 100 bootstrap samples of the observed (COF, adsorption property) values (bootstrap sample = random sample of the observations Ω, with replacement, of size $|\Omega|$). Then, we fit a low-rank model to each bootstrap sample of observations, giving us an ensemble of 100 different low-rank models. The point prediction of a missing (material, property) value follows from the mean prediction among the ensemble of low-rank models; the associated uncertainty follows from the standard deviation of the predictions. Figure S9a shows bootstrap confidence intervals overlaid on the true values of the missing (COF, adsorption property) values. The width of each bootstrap confidence interval reflects the uncertainty in the prediction. To assess whether the uncertainty estimates via bootstrap resampling indeed capture uncertainty, we show the sharpness and calibration/honesty of the uncertainty estimates [85] in Figures S9b and S9c. Indeed, the uncertainty estimates are satisfactorily honest, demonstrating that bootstrap resampling is an easy-to-implement, but computationally expensive means to quantify uncertainty in a recommendation system.

4 Discussion

4.1 Challenges

The success of a recommendation system for NPMs is predicated on structured, open, and high-fidelity databases of NPMs and their adsorption properties [86–88]. The NIST/ARPA–E Database of Novel and Emerging Adsorbent Materials [19] (NIST-ISODB) is a collection of compiled gas adsorption measurements in NPMs from the literature, from both experimental and simulation sources, for a variety of gases at a wide range of conditions. We initially set out to develop a recommendation system using Henry coefficients extracted from experimental data in the NIST-ISODB, but we found the resultant recommendation system to perform poorly. We discuss this further and provide possible explanations in Section S7.

Our material recommendation system suffers from the *cold start* [24] problem. Suppose a new material is reported, but none of its properties have been observed. Then, the model in eqn 1 is unable to make a prediction about any of this material's properties because we do not have data to learn its latent representation, $\mathbf{m}_{M+1}$.

Some material properties of interest may be derived from properties that appear in the material–property matrix. Owing to error propagation, the ranking by the recommendation system according to a derived property may be less accurate than according to a property that directly appears in a column of the matrix. We illustrate this in Section S9 for dilute Xe/Kr selectivity in the COFs, which follows from the ratio of the Xe and Kr Henry coefficients that appear in two separate columns. For this reason, it may be beneficial to include the derived property of interest as a separate column in the matrix; however, if it is derived from many other properties, this column will be sparse.

Likely, new (material, property) observations, perhaps pertaining to new materials and new properties, will be continually acquired. To update the recommendation system in light of newly acquired data, we do not need to retrain the low-rank model from scratch. The loss in eq 3 (with Ω augmented) can be minimized on-line with, e.g., alternating minimization. Given new observations pertaining to existing materials and properties, the current low-rank model serves as a warm start for minimizing the loss, and a few more iterations of alternating minimization will update the latent vectors of the materials and properties and the material biases. Given new observations pertaining to a new material (property), we can learn the latent vector of the new material (property) by treating the latent vectors of the properties (materials) as static, much like conducting an iteration of alternating minimization [22]. That said, the run time of training a new low-rank model from scratch is likely to be insignificant given the typical size of (material, property) data sets. For example, in this work, the run time to fit a low-rank model with `LowRankModels.jl`, on the 572×16 COF–adsorption-property matrix, is on the order of seconds.

We introduced missing entries in the (in reality, fully observed) COF–adsorption-property matrix by (uniform) randomly selecting entries to ablate. In practice, however, (i) some properties are more commonly measured than others, (ii) some materials are more commonly studied than others owing to e.g., ease of synthesis, and (iii) there are likely correlations between and temporal trends with the binary random variables that represent whether the (material, property) values are observed. To expand on (iii), for example, a material with a superior (inferior) value of a desired property may become popular (unpopular) for measurements of other properties. In Section S8, we show the imputation performance of a low-rank model fit to an incomplete COF–adsorption-property matrix where entries were observed in a bias manner; 10 % of the COFs were popular for measurements and 10 % of the COFs were unpopular. Unsurprisingly, the imputation performance of the recommendation system was enhanced and diminished among the popular and unpopular COFs, respectively.

Below, we propose future research directions to address each of these challenges and improve the imputation performance of a material recommendation system.

4.2 Future Research Directions

4.2.1 Include Structural and Chemical Features of the Materials

Most importantly, we propose to build a recommendation system that leverages structural and chemical features of the materials, in addition to the observed adsorption properties, to impute the missing entries in the material-property matrix. A strength of the low-rank model in eq 1 is that vector representations of the materials are not needed to impute the missing entries of the material-property matrix; rather, the vector representations of the materials are learned from their observed properties. However, a recommendation system that uses chemical and structural features of the materials to learn structure-property relationships will boost imputation performance and solve the cold start problem for new materials or properties.

A list of hand-crafted features of the materials, such as the void fraction, surface area, pore size, percent carbon atoms, etc., could be input to the recommendation system for it to learn their relationships with the properties. In the analogy with movie recommendation systems, this is like including features about the movies, such as the genre, directors, year of production, and actors. A simple approach to include hand-crafted features of the materials is to append each feature as a (fully observed) column to the material-property matrix. However, then the latent material vectors will encode the structural features of the materials in addition to the (more relevant to applications) properties.

4.2.2 Address Selection Bias

To characterize selection bias, it would be interesting to construct a temporal model for the process by which materials and properties are selected for experimental investigation. More, we can work to debias a recommendation system trained on experimental data with selection bias [89].

4.2.3 Active Learning and Bayesian Optimization in a Low-Rank Model Framework

In an active learning strategy, we aim to decide which missing (material, property) values to observe/measure in order to reduce the uncertainty in the predictions by the recommendation system [90, 91]. In Bayesian optimization [92], we aim to find the material with the optimal property in a column of the material-property matrix using the fewest experiments—by revealing the fewest missing entries in the column. The property of interest could be a property currently covered by the recommendation system, or a new, unseen property. Both active learning and Bayesian optimization will require a matrix completion method that quantifies uncertainty in its predictions.

4.2.4 Uncertainty Quantification

We demonstrated bootstrap resampling of the observations for quantifying uncertainty in the predictions of the recommendation system. However, training many low-rank matrix models, one for each bootstrap sample, is computationally expensive. More advanced, probabilistic matrix completion methods are designed to quantify uncertainties in the imputed entries of the matrix [93, 94].

4.2.5 Nonlinear Models

The low-rank model in eq 1 models property p of material m as a linear model in terms of the material vector input: $A_{mp} \approx f_p(\mathbf{m}_m) = \mathbf{p}_p^\intercal \mathbf{m}_m + \mu_m$ (view $\mathbf{p}_p$ as parameters of the function $f_p(\cdot)$). In contrast, a non-linear version of matrix factorization [95] that treats $f_p(\cdot)$ as a non-linear function may grant us more expressiveness for capturing complex relationships between materials and adsorption properties, thereby improving imputation performance.

5 Conclusions

In materials science, we are often interested in many different properties of many different materials. The corresponding material-property matrix often, in practice, has many missing entries since every property of every material has not been measured. The idea of a material recommendation system is to leverage the observed (material, property) values to impute the missing ones. Interestingly, the (material, property) values in a material recommendation system are mathematically analogous to (product, customer) ratings in commercial recommendation systems. A material recommendation system is useful for recommending (i, application-led material search [1]) a material that optimizes a specific property or (ii, material-led application search [17]) an optimal application for a given material.

A material recommendation system constituted by a low-rank model of the incomplete material-property matrix, fit to the observed (material, property) values, (i) provides predictions of the missing entries in the material-property matrix and (ii) generates a map of materials, wherein materials with similar properties congregate.

We demonstrated a recommendation system that recommends COFs for different gas adsorption tasks. We constructed incomplete COF–adsorption-property matrices from 16 simulated gas adsorption properties of 572 COF structures from Ongari et al. [17]. Then, we trained low-rank models [22] of the incomplete COF–adsorption-property matrices and (i) assessed their performance on the task of predicting the missing entries and (ii) inspected their learned map of COFs. Given that fewer than 60 % of the entries of the matrix were missing, the recommendation system was able to rank the COFs according to their (missing) adsorption properties reasonably well (Spearman's rank correlation coefficient > 0.6), with the exception of H_2O Henry coefficients. Imputation performance diminishes rapidly when more than 60 % of the entries in the material-property matrix are missing. Though the 60 % figure does not necessarily generalize to other data sets, this demonstrates that the success of a material recommendation system is predicated on having a sufficient amount of training data. Finally, we drew the map of COFs and colored each COF vector by the adsorption properties to find that, indeed, COFs with similar (dissimilar) adsorption properties clustered together (separated) in the map.

We conclude that material recommendation systems, if sufficient training data is available, could be widely useful for leveraging existing measurements of properties of materials to fill in missing measurements. In turn, this could accelerate the matching of materials for specific applications.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.chemmater.1c01201.

Supporting Information contains correlations between COF adsorption properties, a visualization of the COF–adsorption-property matrix, a correlation between and distributions of material biases and interaction terms, the distribution of optimal hyperparameters among the low-rank models, the latent space of COFs colored by other adsorption properties, the latent space of properties, uncertainty quantification via bootstrap resampling, preliminary work on the NIST-ISODB, and simulations of selection bias in data collection.

Acknowledgements

A.S., X.F., and C.M.S. acknowledge the National Science Foundation for support under grant No. 1920945. The authors thank Madeleine Udell for tips on `LowRankModels.jl` and Nikolas Christian Achatz and Tristan McInteer Gavin for querying the latest COF data from Materials Cloud.

Copyright and Disclaimer

References

[1] Anna G Slater and Andrew I Cooper. Porous materials. function-led design of new porous materials. *Science*, 348(6238):aaa8075–aaa8075, 2015.

[2] Omar K. Farha, Ibrahim Eryazici, Nak Cheon Jeong, Brad G. Hauser, Christopher E. Wilmer, Amy A. Sarjeant, Randall Q. Snurr, SonBinh T. Nguyen, A. Özgür Yazaydın, and Joseph T. Hupp. Metal–organic framework materials with ultrahigh surface areas: Is the sky the limit? *Journal of the American Chemical Society*, 134(36):15016–15021, August 2012.

[3] Russell E Morris and Paul S Wheatley. Gas storage in nanoporous materials. *Angewandte Chemie International Edition*, 47(27):4966–4981, 2008.

[4] Jian-Rong Li, Ryan J Kuppler, and Hong-Cai Zhou. Selective gas adsorption and separation in metal–organic frameworks. *Chemical Society Reviews*, 38(5):1477–1504, 2009.

[5] Alexander Schoedel, Zhe Ji, and Omar M Yaghi. The role of metal–organic frameworks in a carbon-neutral energy cycle. *Nature Energy*, 1(4):1–13, 2016.

[6] Zoey R Herm, Eric D Bloch, and Jeffrey R Long. Hydrocarbon separations in metal–organic frameworks. *Chemistry of Materials*, 26(1):323–338, 2013.

[7] Lauren E Kreno, Kirsty Leong, Omar K Farha, Mark Allendorf, Richard P Van Duyne, and Joseph T Hupp. Metal–organic framework materials as chemical sensors. *Chemical Reviews*, 112(2):1105–1125, 2012.

[8] David Farrusseng, Sonia Aguado, and Catherine Pinel. Metal–organic frameworks: opportunities for catalysis. *Angewandte Chemie International Edition*, 48(41):7502–7513, 2009.

[9] Hiroyasu Furukawa, Kyle E Cordova, Michael O'Keeffe, and Omar M Yaghi. The chemistry and applications of metal–organic frameworks. *Science*, 341(6149):1230444, 2013.

[10] Christian S Diercks and Omar M Yaghi. The atom, the molecule, and the covalent organic framework. *Science*, 355(6328), 2017.

[11] Weigang Lu, Daqiang Yuan, Dan Zhao, Christine Inge Schilling, Oliver Plietzsch, Thierry Muller, Stefan Brase, Johannes Guenther, Janet Blumel, Rajamani Krishna, Zhen Li, and Hong-Cai Zhou. Porous polymer networks: Synthesis, porosity, and applications in gas storage/separation. *Chemistry of Materials*, 22(21):5964–5972, November 2010.

[12] Andrew I Cooper. Porous molecular solids and liquids. *ACS Central Science*, 3(6):544–553, 2017.

[13] Aeri J Gosselin, Casey A Rowland, and Eric D Bloch. Permanently microporous metal–organic polyhedra. *Chemical Reviews*, 120(16):8987–9014, 2020.

[14] David J Tranchemontagne, Zheng Ni, Michael O'Keeffe, and Omar M Yaghi. Reticular chemistry of metal–organic polyhedra. *Angewandte Chemie International Edition*, 47(28):5136–5147, 2008.

[15] Sukhendu Mandal, Srinivasan Natarajan, Prabu Mani, and Asha Pankajakshan. Post-synthetic modification of metal–organic frameworks toward applications. *Advanced Functional Materials*, page 2006291, 2020.

[16] Peter G Boyd, Yongjin Lee, and Berend Smit. Computational development of the nanoporous materials genome. *Nature Reviews Materials*, 2(8):1–15, 2017.

[17] Daniele Ongari, Leopold Talirz, and Berend Smit. Too many materials and too many applications: An experimental problem waiting for a computational solution. *ACS Central Science*, 6(11):1890–1900, 2020.

[18] Paul Iacomi and Philip L Llewellyn. Data mining for binary separation materials in published adsorption isotherms. *Chemistry of Materials*, 32(3):982–991, 2020.

[19] D.W. Siderius, V.K. Shen, R.D. Johnson III, and R.D. van Zee, editors. *NIST/ARPA-E Database of Novel and Emerging Adsorbent Materials*. National Institute of Standards and Technology, Gaithersburg, MD, 20899, 2014.

[20] Madeleine Udell and Alex Townsend. Why are big data matrices approximately low rank? *SIAM Journal on Mathematics of Data Science*, 1(1):144–160, 2019.

[21] Madeleine Udell. Big data is low rank. *SIAG/OPT Views and News*, 27(1), 2019.

[22] Madeleine Udell, Corinne Horn, Reza Zadeh, and Stephen Boyd. Generalized low rank models. *Foundations and Trends® in Machine Learning*, 9(1):1–118, 2016.

[23] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

[24] Charu C Aggarwal. *Recommender systems*. Springer, 2016.

[25] Daniele Ongari, Aliaksandr V Yakutovich, Leopold Talirz, and Berend Smit. Building a consistent and reproducible database for adsorption evaluation in covalent–organic frameworks. *ACS Central Science*, 5(10):1663–1675, 2019.

[26] Seyed Mohamad Moosavi, Kevin Maik Jablonka, and Berend Smit. The role of machine learning in the understanding and design of materials. *Journal of the American Chemical Society*, 142(48):20273–20287, 2020.

[27] Steven Bennett, Andrew Tarzia, Martijn A Zwijnenburg, and Kim E Jelfs. Artificial intelligence applied to the prediction of organic materials. *Machine Learning in Chemistry*, 17:280, 2020.

[28] Siwar Chibani and François-Xavier Coudert. Machine learning approaches for the prediction of materials properties. *APL Materials*, 8(8):080701, 2020.

[29] Sanggyu Chong, Sangwon Lee, Baekjun Kim, and Jihan Kim. Applications of machine learning in metal-organic frameworks. *Coordination Chemistry Reviews*, 423:213487, 2020.

[30] Zenan Shi, Wenyuan Yang, Xiaomei Deng, Chengzhi Cai, Yaling Yan, Hong Liang, Zili Liu, and Zhiwei Qiao. Machine-learning-assisted high-throughput computational screening of high performance metal–organic frameworks. *Molecular Systems Design & Engineering*, 5(4):725–742, 2020.

[31] Kevin Maik Jablonka, Daniele Ongari, Seyed Mohamad Moosavi, and Berend Smit. Big-data science in porous materials: materials genomics and machine learning. *Chemical reviews*, 120(16):8066–8129, 2020.

[32] Seyed Mohamad Moosavi, Aditya Nandy, Kevin Maik Jablonka, Daniele Ongari, Jon Paul Janet, Peter G. Boyd, Yongjin Lee, Berend Smit, and Heather J. Kulik. Understanding the diversity of the metal-organic framework ecosystem. *Nature Communications*, 11(1), August 2020.

[33] Krishnendu Mukherjee and Yamil J. Colón. Machine learning and descriptor selection for the computational discovery of metal-organic frameworks. *Molecular Simulation*, pages 1–21, April 2021.

[34] Michael Fernandez, Tom K Woo, Christopher E Wilmer, and Randall Q Snurr. Large-scale quantitative structure–property relationship (QSPR) analysis of methane storage in metal-organic frameworks. *The Journal of Physical Chemistry C*, 117(15):7681–7689, 2013.

[35] Benjamin J Bucior, N Scott Bobbitt, Timur Islamoglu, Subhadip Goswami, Arun Gopalan, Taner Yildirim, Omar K Farha, Neda Bagheri, and Randall Q Snurr. Energy-based descriptors to rapidly predict hydrogen storage in metal–organic frameworks. *Molecular Systems Design & Engineering*, 4(1):162–174, 2019.

[36] Eun Hyun Cho, Xuepeng Deng, Changlong Zou, and Li-Chiang Lin. Machine learning-aided computational study of metal–organic frameworks for sour gas sweetening. *The Journal of Physical Chemistry C*, 124(50):27580–27591, 2020.

[37] Cory M Simon, Rocio Mercado, Sondre K Schnell, Berend Smit, and Maciej Haranczyk. What are the best materials to separate a xenon/krypton mixture? *Chemistry of Materials*, 27(12):4459–4475, 2015.

[38] Ryther Anderson, Jacob Rodgers, Edwin Argueta, Achay Biong, and Diego A. Gómez-Gualdrón. Role of pore chemistry and topology in the CO_2 capture capabilities of MOFs: From molecular simulation to machine learning. *Chemistry of Materials*, 30(18):6325–6337, August 2018.

[39] Giorgos Borboudakis, Taxiarchis Stergiannakos, Maria Frysali, Emmanuel Klontzas, Ioannis Tsamardinos, and George E Froudakis. Chemically intuited, large-scale screening of mofs by machine learning techniques. *npj Computational Materials*, 3(1):40, 2017.

[40] Maryam Pardakhti, Ehsan Moharreri, David Wanik, Steven L Suib, and Ranjan Srivastava. Machine learning using combined structural and chemical descriptors for prediction of methane adsorption performance of metal organic frameworks (MOFs). *ACS Combinatorial Science*, 19(10):640–645, 2017.

[41] Eun Hyun Cho and Li-Chiang Lin. Nanoporous material recognition via 3D convolutional neural networks: Prediction of adsorption properties. *The Journal of Physical Chemistry Letters*, pages 2279–2285, March 2021.

[42] Ruimin Ma, Yamil J. Colón, and Tengfei Luo. Transfer learning study of gas adsorption in metal–organic frameworks. *ACS Applied Materials & Interfaces*, 12(30):34041–34048, July 2020.

[43] Alauddin Ahmed and Donald J. Siegel. Predicting hydrogen storage in mofs via machine learning. *Patterns*, 2(7):100291, 2021.

[44] Thomas F. Willems, Chris H. Rycroft, Michaeel Kazi, Juan C. Meza, and Maciej Haranczyk. Algorithms and tools for high-throughput geometry-based analysis of crystalline porous materials. *Microporous and Mesoporous Materials*, 149(1):134–141, February 2012.

[45] Aditi S. Krishnapriyan, Maciej Haranczyk, and Dmitriy Morozov. Topological descriptors help predict guest adsorption in nanoporous materials. *The Journal of Physical Chemistry C*, 124(17):9360–9368, April 2020.

[46] Hiroshi Ohno and Yusuke Mukae. Machine learning approach for prediction and search: Application to methane storage in a metal–organic framework. *The Journal of Physical Chemistry C*, 120(42):23963–23968, 2016.

[47] Ali Raza, Faaiq Waqar, Arni Sturluson, Cory Simon, and Xiaoli Fern. Towards explainable message passing networks for predicting carbon dioxide adsorption in metal-organic frameworks. *Machine Learning for Molecules Workshop at NeurIPS 2020*, 2020.

[48] Connor W. Coley. Defining and exploring chemical spaces. *Trends in Chemistry*, 3(2):133–145, February 2021.

[49] Arni Sturluson, Melanie T Huynh, Arthur HP York, and Cory M Simon. Eigencages: Learning a latent space of porous cage molecules. *ACS Central Science*, 4(12):1663–1676, 2018.

[50] Thomas C Nicholas, Andrew L Goodwin, and Volker L Deringer. Understanding the geometric diversity of inorganic and hybrid frameworks through structural coarse-graining. *Chemical Science*, 46, 2020.

[51] Yongjin Lee, Senja D. Barthel, Paweł Dłotko, S. Mohamad Moosavi, Kathryn Hess, and Berend Smit. Quantifying similarity of pore-geometry in nanoporous materials. *Nature Communications*, 8(1), May 2017.

[52] Thomas C. Nicholas, Eugeny V. Alexandrov, Vladislav A. Blatov, Alexander P. Shevchenko, Davide M. Proserpio, Andrew L. Goodwin, and Volker L. Deringer. Visualization and quantification of geometric diversity in metal–organic frameworks. `https://doi.org/10.33774/chemrxiv-2021-bdkwx`, 2021. *ChemRxiv*. accessed 2021-09-23.

[53] Yongchul G. Chung, Diego A. Gómez-Gualdrón, Peng Li, Karson T. Leperi, Pravas Deria, Hongda Zhang, Nicolaas A. Vermeulen, J. Fraser Stoddart, Fengqi You, Joseph T. Hupp, Omar K. Farha, and Randall Q. Snurr. In silico discovery of metal–organic frameworks for precombustion CO_2 capture using a genetic algorithm. *Science Advances*, 2(10):e1600909, 2016.

[54] Yi Bao, Richard L Martin, Cory M Simon, Maciej Haranczyk, Berend Smit, and Michael W Deem. In silico discovery of high deliverable capacity metal–organic frameworks. *The Journal of Physical Chemistry C*, 119(1):186–195, 2015.

[55] Sean P. Collins, Thomas D. Daff, Sarah S. Piotrkowski, and Tom K. Woo. Materials design by evolutionary optimization of functional groups in metal-organic frameworks. *Science Advances*, 2(11), 2016.

[56] Nicole Beauregard, Maryam Pardakhti, and Ranjan Srivastava. In silico evolution of high-performing metal organic frameworks for methane adsorption. *Journal of Chemical Information and Modeling*, July 2021.

[57] Xiangyu Zhang, Kexin Zhang, and Yongjin Lee. Machine learning enabled tailor-made design of application-specific metal–organic frameworks. *ACS Applied Materials & Interfaces*, 12(1):734–743, 2019.

[58] Gael Donval, Calum Hand, James Hook, Emiko Dupont, Malena Sabaté Landman, Malina Freitag, Matthew Lennox, and Tina Düren. Autonomous exploration and identification of high performing adsorbents using active learning. `https://doi.org/10.26434/chemrxiv.14555706.v1`, 2021. *ChemRxiv*. accessed 2021-09-23.

[59] Aryan Deshwal, Cory Simon, and Janardhan Rao Doppa. Bayesian optimization of nanoporous materials. `https://doi.org/10.33774/chemrxiv-2021-4624n-v2`, 2021. *ChemRxiv*. accessed 2021-09-23.

[60] Zhenpeng Yao, Benjamín Sánchez-Lengeling, N Scott Bobbitt, Benjamin J Bucior, Sai Govind Hari Kumar, Sean P Collins, Thomas Burns, Tom K Woo, Omar K Farha, Randall Q Snurr, et al. Inverse design of nanoporous crystalline reticular materials with deep generative models. *Nature Machine Intelligence*, 3(1):76–86, 2021.

[61] Benjamin Sanchez-Lengeling and Alán Aspuru-Guzik. Inverse molecular design using machine learning: Generative models for matter engineering. *Science*, 361(6400):360–365, 2018.

[62] R. Pollice, Gabriel dos Passos Gomes, Matteo Aldeghi, Riley J. Hickman, Mario Krenn, Cyrille Lavigne, Michael Lindner-D'Addario, AkshatKumar Nigam, Cher Tian Ser, Zhenpengand Yao, and Alan Aspuru-Guzik. Data-driven strategies for accelerated materials design. *Accounts of Chemical Research*, 54(4):849–860, 2021.

[63] Guillaume Fraux, Siwar Chibani, and François-Xavier Coudert. Modelling of framework materials at multiple scales: current practices and open questions. *Philosophical Transactions of the Royal Society A*, 377(2149):20180220, 2019.

[64] Leopold Talirz, Snehal Kumbhar, Elsa Passaro, Aliaksandr V. Yakutovich, Valeria Granata, Fernando Gargiulo, Marco Borelli, Martin Uhrin, Sebastiaan P. Huber, Spyros Zoupanos, Carl S. Adorf, Casper Welzel Andersen, Ole Schütt, Carlo A. Pignedoli, Daniele Passerone, Joost VandeVondele, Thomas C. Schulthess, Berend Smit, Giovanni Pizzi, and Nicola Marzari. Materials cloud, a platform for open computational science. *Scientific Data*, 7(1), September 2020.

[65] François-Xavier Coudert. Materials databases: the need for open, interoperable databases with standardized data and rich metadata. *Advanced Theory and Simulations*, 2(11):1900131, 2019.

[66] Yongchul G Chung, Jeffrey Camp, Maciej Haranczyk, Benjamin J Sikora, Wojciech Bury, Vaiva Krungleviciute, Taner Yildirim, Omar K Farha, David S Sholl, and Randall Q Snurr. Computation-ready, experimental metal–organic frameworks: A tool to enable high-throughput screening of nanoporous crystals. *Chemistry of Materials*, 26(21):6185–6192, 2014.

[67] Yongchul G. Chung, Emmanuel Haldoupis, Benjamin J. Bucior, Maciej Haranczyk, Seulchan Lee, Hongda Zhang, Konstantinos D. Vogiatzis, Marija Milisavljevic, Sanliang Ling, Jeffrey S. Camp, Ben Slater, J. Ilja Siepmann, David S. Sholl, and Randall Q. Snurr. Advances, updates, and analytics for the computation-ready, experimental metal–organic framework database: CoRE MOF 2019. *Journal of Chemical & Engineering Data*, 64(12):5985–5998, November 2019.

[68] Minman Tong, Youshi Lan, Qingyuan Yang, and Chongli Zhong. Exploring the structure-property relationships of covalent organic frameworks for noble gas separations. *Chemical Engineering Science*, 168:456–464, 2017.

[69] Peter G. Boyd, Arunraj Chidambaram, Enrique García-Díez, Christopher P. Ireland, Thomas D. Daff, Richard Bounds, Andrzej Gładysiak, Pascal Schouwink, Seyed Mohamad Moosavi, M. Mercedes Maroto-Valer, Jeffrey A. Reimer, Jorge A. R. Navarro, Tom K. Woo, Susana Garcia, Kyriakos C. Stylianou, and Berend Smit. Data-driven design of metal–organic frameworks for wet flue gas CO2 capture. *Nature*, 576(7786):253–256, December 2019.

[70] Peyman Z Moghadam, Aurelia Li, Seth B Wiggin, Andi Tao, Andrew GP Maloney, Peter A Wood, Suzanna C Ward, and David Fairen-Jimenez. Development of a cambridge structural database subset: a collection of metal–organic frameworks for past, present, and future. *Chemistry of Materials*, 29(7):2618–2625, 2017.

[71] Rocío Mercado, Rueih-Sheng Fu, Aliaksandr V Yakutovich, Leopold Talirz, Maciej Haranczyk, and Berend Smit. In silico design of 2D and 3D covalent organic frameworks for methane storage applications. *Chemistry of Materials*, 30(15):5069–5086, 2018.

[72] Benjamin J. Bucior, Andrew S. Rosen, Maciej Haranczyk, Zhenpeng Yao, Michael E. Ziebel, Omar K. Farha, Joseph T. Hupp, J. Ilja Siepmann, Alán Aspuru-Guzik, and Randall Q. Snurr. Identification schemes for metal–organic frameworks to enable rapid search and cheminformatics analysis. *Crystal Growth & Design*, 19(11):6682–6697, September 2019.

[73] Andrew S. Rosen, Shaelyn M. Iyer, Debmalya Ray, Zhenpeng Yao, Alán Aspuru-Guzik, Laura Gagliardi, Justin M. Notestein, and Randall Q. Snurr. Machine learning the quantum-chemical properties of metal–organic frameworks for accelerated materials discovery. *Matter*, 4(5):1578–1597, May 2021.

[74] Dalar Nazarian, Jeffrey S Camp, and David S Sholl. A comprehensive set of high-quality point charges for simulations of metal–organic frameworks. *Chemistry of Materials*, 28(3):785–793, 2016.

[75] Sanghoon Park, Baekjun Kim, Sihoon Choi, Peter G Boyd, Berend Smit, and Jihan Kim. Text mining metal–organic framework papers. *Journal of Chemical Information and Modeling*, 58(2):244–251, 2018.

[76] Vahe Tshitoyan, John Dagdelen, Leigh Weston, Alexander Dunn, Ziqin Rong, Olga Kononova, Kristin A Persson, Gerbrand Ceder, and Anubhav Jain. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature*, 571(7763):95–98, 2019.

[77] Scott Spangler, Angela D Wilkins, Benjamin J Bachman, Meena Nagarajan, Tajhal Dayaram, Peter Haas, Sam Regenbogen, Curtis R Pickering, Austin Comer, Jeffrey N Myers, et al. Automated hypothesis generation based on mining scientific literature. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1877–1886, 2014.

[78] Yangzesheng Sun, Robert F. DeJaco, Zhao Li, Dai Tang, Stephan Glante, David S. Sholl, Coray M. Colina, Randall Q. Snurr, Matthias Thommes, Martin Hartmann, and J. Ilja Siepmann. Fingerprinting diverse nanoporous materials for optimal hydrogen storage conditions using meta-learning. *Science Advances*, 7(30):eabg3983, July 2021.

[79] Qi Yuana, Mariagiulia Longob, Aaron Thorntonc, Neil B McKeownd, Bibiana Comesaña-Gándarad, Johannes C Jansenb, and Kim E Jelfsa. Imputation of missing gas permeability data for polymer membranes using machine learning. *Journal of Membrane Science*, page 119207, 2021.

[80] Ekaterina A Sosnina, Sergey Sosnin, Anastasia A Nikitina, Ivan Nazarov, Dmitry I Osolodkin, and Maxim V Fedorov. Recommender systems in antiviral drug discovery. *ACS Omega*, 5(25):15039–15051, 2020.

[81] Atsuto Seko, Hiroyuki Hayashi, Hisashi Kashima, and Isao Tanaka. Matrix-and tensor-based recommender systems for the discovery of currently unknown inorganic compounds. *Physical Review Materials*, 2(1):013805, 2018.

[82] Hiroyuki Hayashi, Katsuyuki Hayashi, Keita Kouzai, Atsuto Seko, and Isao Tanaka. Recommender system of successful processing conditions for new compounds based on a parallel experimental data set. *Chemistry of Materials*, 31(24):9984–9992, November 2019.

[83] Michael E Tipping and Christopher M Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999.

[84] Jeff Bezanson, Alan Edelman, Stefan Karpinski, and Viral B Shah. Julia: A fresh approach to numerical computing. *SIAM Review*, 59(1):65–98, 2017.

[85] Kevin Tran, Willie Neiswanger, Junwoong Yoon, Qingyang Zhang, Eric Xing, and Zachary W Ulissi. Methods for comparing uncertainty quantifications for material property predictions. *Machine Learning: Science and Technology*, 1(2):025006, 2020.

[86] Arni Sturluson, Melanie T Huynh, Alec R Kaija, Caleb Laird, Sunghyun Yoon, Feier Hou, Zhenxing Feng, Christopher E Wilmer, Yamil J Colón, Yongchul G Chung, et al. The role of molecular modelling and simulation in the discovery and deployment of metal-organic frameworks for gas storage and separation. *Molecular Simulation*, 45(14-15):1082–1121, 2019.

[87] Jack D Evans, Volodymyr Bon, Irena Senkovska, and Stefan Kaskel. A universal standard archive file for adsorption data. *Langmuir*, 37(14):4222–4226, 2021.

[88] Jongwoo Park, Joshua D. Howe, and David S. Sholl. How reproducible are isotherm measurements in metal–organic frameworks? *Chemistry of Materials*, 29(24):10487–10495, 2017.

[89] Jiawei Chen, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, and Xiangnan He. Bias and debias in recommender system: A survey and future directions. *arXiv preprint arXiv:2010.03240*, 2020.

[90] Neil Rubens, Dain Kaplan, and Masashi Sugiyama. Active learning in recommender systems. In P.B. Kantor, F. Ricci, L. Rokach, and B. Shapira, editors, *Recommender Systems Handbook*, pages 735–767. Springer, 2011.

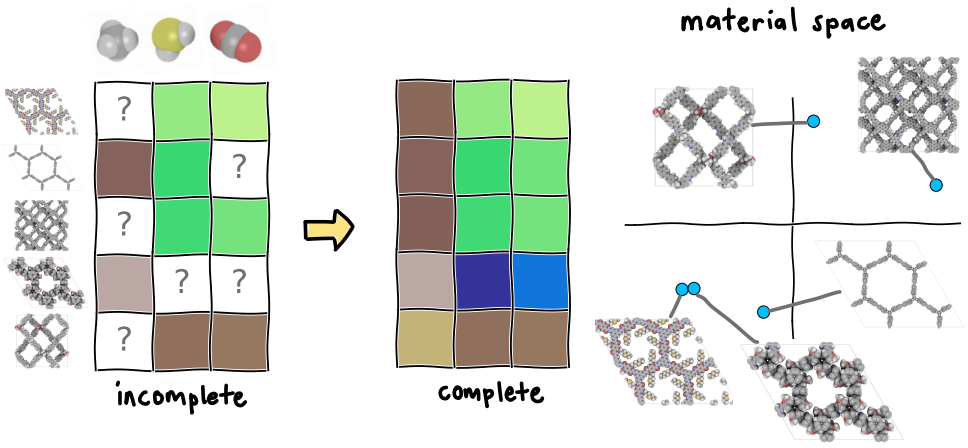
[91] Chengrun Yang, Yuji Akimoto, Dae Won Kim, and Madeleine Udell. Oboe: Collaborative filtering for automl model selection. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1173–1183, 2019.

[92] Bobak Shahriari, Kevin Swersky, Ziyu Wang, Ryan P Adams, and Nando De Freitas. Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE*, 104(1):148–175, 2015.

[93] Yuxin Chen, Jianqing Fan, Cong Ma, and Yuling Yan. Inference and uncertainty quantification for noisy matrix completion. *Proceedings of the National Academy of Sciences*, 116(46):22931–22937, October 2019.

[94] Yuxuan Zhao and Madeleine Udell. Matrix completion with quantified uncertainty through low rank gaussian copula. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[95] Neil D Lawrence and Raquel Urtasun. Non-linear matrix factorization with gaussian processes. In *Proceedings of the 26th annual international conference on machine learning*, pages 601–608, 2009.



TOC graphic