A 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 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

April 7, 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 the relevant adsorption properties in the candidate NPMs, would enable the matching of NPMs with adsorption tasks. However, in practice the NPM-property matrix is only partially observed (incomplete); (i) many properties of any given NPM have not been measured and (ii) any given property has not been measured for all NPMs.

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. A low rank model, trained on 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). We find the performance of the COF recommendation system varies for different adsorption tasks and diminishes precipitously when the fraction of missing entries exceeds 60 %. The concepts in our COF recommendation system can be applied broadly 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 post-synthetic modifiability [15], make 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 Fig. 1.

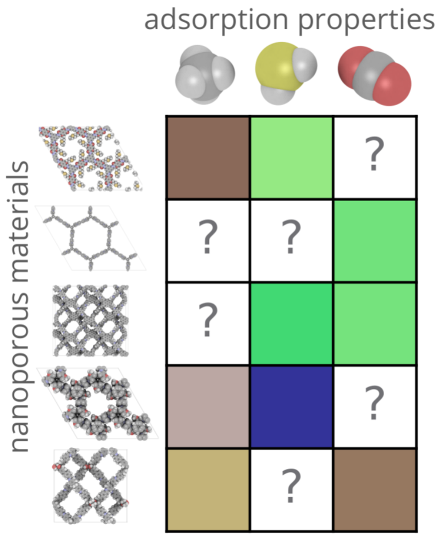


Figure 1: A recommendation system for nanoporous materials (NPMs). In this toy NPM–adsorption-property matrix, entry (m, p) represents the value of adsorption property p of NPM m. Many entries are unobserved ("?") because measurements are missing. The goal of our NPM recommendation system is to use the 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 recommend customer-specific products to 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 towards 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. A low rank structure implies 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 (NPM, property) values [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 ratings matrix (rows: movies, columns: users, entries: ratings) [23]. The movie-user ratings 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 ratings 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], that recommends COFs for various gas adsorption tasks. The COF–gas-adsorption-property matrices pertain to 560 experimentally-reported COFs [25] and the simulated uptakes 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 valvues. From the observed (COF, gas adsorption) values, we machine-learn a low rank model, giving low-dimensional latent vector representations of both the COFs and the adsorption properties, allowing (i) accurate imputation of the missing values of the adsorption properties and (ii) the embedding of the COFs, represented as points, into a “map” wherein 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 [26–32]. Supervised machine learning models have been widely used to predict the adsorption properties of NPMs [33–41] from vectors of hand-crafted structural features [42, 43] or from a graph representation [44]. Unsupervised machine learning methods have been used to embed NPMs into a low-dimensional “material space” [45] and cluster together NPMs with similar structures [46–48]. Genetic algorithms [49–51] and Monte Carlo tree search [52] have been used to more efficiently search for the NPM(s) with an optimal adsorption property. Finally, recently [53] an autoencoder enabled inverse design [54, 55] 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 [56–58] of (i) crystal structure models of NPMs [25, 59–63], (ii) simulated [17, 62, 64, 65] and experimentally measured [19] adsorption properties of NPMs, and (iii) electronic properties of NPMs [66, 67], 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 [68–70].

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 explicitly hand-crafted features of the NPMs, and (ii) performing multi-task prediction and transferring knowledge between tasks while handling missing values in the target vectors associated with the NPMs. N.b. recommendation systems have been built for use in the chemical sciences to impute missing gas permeabilities in polymers [71], antiviral activities of molecules [72], and stabilities of inorganic materials [73, 74].

2 The 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 the missing (material, property) values and (ii) draw a map of the materials, wherein materials with similar properties congregate.

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

The objective. The objective is to complete the material-property matrix by predicting the missing entries, A_{mp} for $(m, p) \in \{1, 2, ..., M\} \times \{1, 2, ..., P\} \setminus \Omega$.

The low-rank model. From an element perspective, the low-rank model 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 is positive (negative) if $\mathbf{m}_m$ and $\mathbf{p}_p$ point in roughly the same (opposite) direction. The magnitude of the interaction term depends on both the angle between $\mathbf{m}_m$ and $\mathbf{p}_p$ and their norms. 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. See Koren et al. [23].

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 Fig. 2. The dimensionality of the latent space, $k < M, P$, imposes the constraint $\text{rank}(\mathbf{M}^\intercal \mathbf{P}) \leq k$, hence eqn. 2 is a low rank approximation of the matrix $\mathbf{A}$.

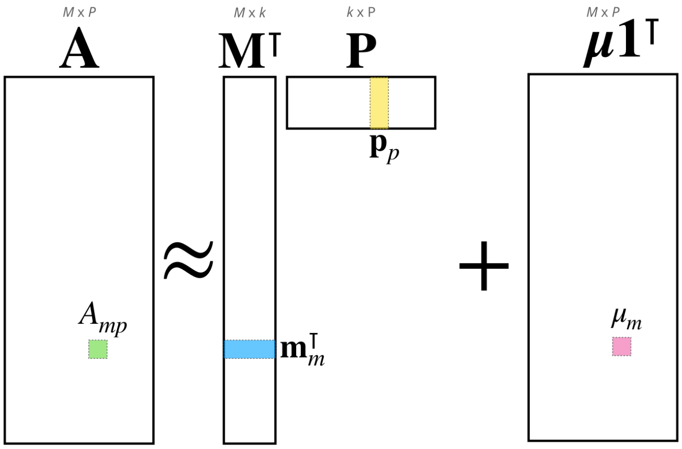


Figure 2: The low rank model of the matrix $\mathbf{A} \approx \mathbf{M}^\intercal \mathbf{P} + \boldsymbol{\mu}\mathbf{1}^\intercal$. The columns of $\mathbf{M}$ and $\mathbf{P}$ contain the latent representations of the M materials and P properties, respectively, which lie in a k-dimensional space. The vector $\boldsymbol{\mu}$ contains the M material biases.

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 eqn. 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 eqn. 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 low-dimensional embedding techniques [75], 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 good but still suboptimal performance, (iii) selecting diverse materials to efficiently explore material space in an experimental design strategy, 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 $\mathbf{p}_p$'s, that tend to take on similar (dissimilar) values in NPMs congregate (separate).

As a consequence of the dot product $\mathbf{m}_m^\intercal \mathbf{p}_p$ in eqn. 1, the magnitude and directions of a pair of latent material and property vectors $(\mathbf{m}_m, \mathbf{p}_p)$, taken together, indicate the affinity/aversion for each other, since $\mathbf{m}_m^\intercal \mathbf{p}_p = ||\mathbf{m}_m||_2 ||\mathbf{p}_p||_2 \cos\phi$, with ϕ the angle between $\mathbf{m}_m$ and $\mathbf{p}_p$.

Machine-learning the low rank model. We learn the latent representations of the materials and properties and the material biases by balancing (i) the matching of the observed values of the matrix by the model given in eqn. 1 and (ii) the complexity 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 $\ell = \ell(\mathbf{M}, \mathbf{P}, \boldsymbol{\mu})$:

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

The first term is the approximation error, measured over all observed (m, p) pairs. 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 by the number of elements in the sum to properly weigh regularization of the latent material and property vectors.

Either stochastic gradient descent or alternating minimization can be used to find the $(\mathbf{M}, \mathbf{P}, \boldsymbol{\mu})$ that minimize ℓ. The latter alternates between fixing $\mathbf{M}$ and optimizing $\mathbf{P}$ and fixing $\mathbf{P}$ and optimizing $\mathbf{M}$. See Refs. [22, 23].

There are two hyperparameters in the low rank model: (1) $k \in \{0, 1, ..., \min(M, P)\}$, 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 complexity 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, the materials are COFs, and the 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 The dataset

We leverage an open data set of simulated gas adsorption properties in $M = 560$ experimentally reported, structurally optimized, and porous COF materials [17, 25]. We selected $P = 16$ simulated adsorption properties of interest, comprised of 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 that apply to different gas storage and separation applications. We $\log_{10}$-transformed the Henry coefficients because of the relatively long tail of their distributions. The resulting COF–adsorption-property matrix, $\mathbf{A}_{\text{complete}} \in \mathbb{R}^{560\times 16}$ is fully observed, allowing us to study the effect of the fraction of missing entries on the performance of the low rank model.

Fig. 3 displays the distribution of the [standardized] properties and the pairwise relationships between the adsorption properties (see Fig. S1 for a pairwise correlation matrix). Some properties are strongly correlated, e.g., CH_4 uptake at (298 K, 65 bar) and H_2 uptake at (77 K, 100 bar), while others, such as Xe and H_2O Henry coefficients, are not. The low rank model exploits these correlations between properties to learn low-dimensional representations of the materials and properties.

3.2 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 $M = 560 \times P = 16$) with a fraction θ of missing entries. We construct $\mathbf{A}^{(\theta)}$ by (uniform) randomly sampling, without replacement, $(1-\theta)MP$ entries to ablate (change to `missing`) from the MP entries of $\mathbf{A}_{\text{complete}}$. Fig. 4 visualizes a resulting incomplete COF–adsorption-property matrix $\mathbf{A}^{(0.4)}$ with a fraction $\theta = 0.4$ missing entries.

3.3 Standardization of adsorption properties

We standardize the adsorption properties (the columns of $\mathbf{A}^{(\theta)}$) to have mean zero and unit variance using only the observed training examples. Standardization accounts for the different scales of the different properties and prevents properties with a larger variance from dominating the loss function in eqn. 3. See Ref. [22] for theoretical arguments for standardization. The entries in Fig. 4 are standardized, hence the diverging colormap.

3.4 Training, hyperparameter tuning, and testing

We use `LowRankModels.jl` [22] in the Julia programming language [76] to train our low rank models of the form in eqn. 2. `LowRankModels.jl` implements an alternating proximal gradient descent [22] to minimize the loss in eqn. 3.

For training and hyperparameter (k, λ) tuning, we randomly partitioned the [simulated] observed entries of $\mathbf{A}^{(\theta)}$ into an 80/20 % training/validation set. The loss ℓ in eqn. 3 is minimized over the

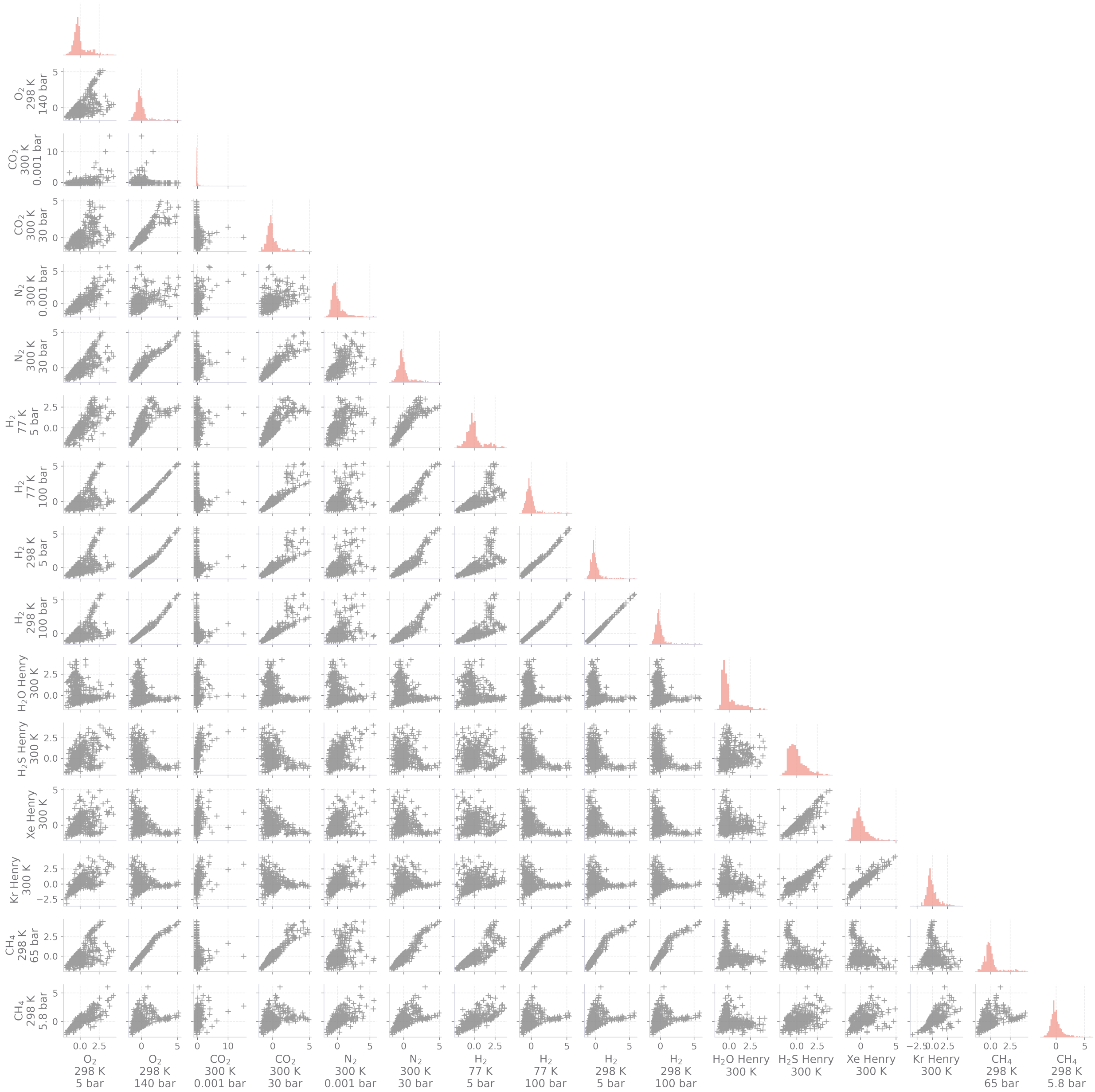


Figure 3: The distribution of (diagonal) and pairwise relationships between (off-diagonal) the simulated gas adsorption properties of the COFs (data from Ref. [17]). Each point represents a COF. Each property was standardized to have zero mean and unit variance.

training set, while the validation set is used to select optimal hyperparameters. The remaining, [simulated] unobserved entries serve as test data to estimate the generalization error of the low rank model for matrix completion.

To determine the optimal hyperparameter tuple $(k_{opt}^{(\theta)}, \lambda_{opt}^{(\theta)})$ for a given $\mathbf{A}^{(\theta)}$, we perform a hyperparameter sweep over a (k, λ) grid, training one low rank model for each (k, λ). We select $(k_{opt}^{(\theta)}, \lambda_{opt}^{(\theta)})$ as the hyperparameter tuple whose low rank model produces the lowest approximation error over

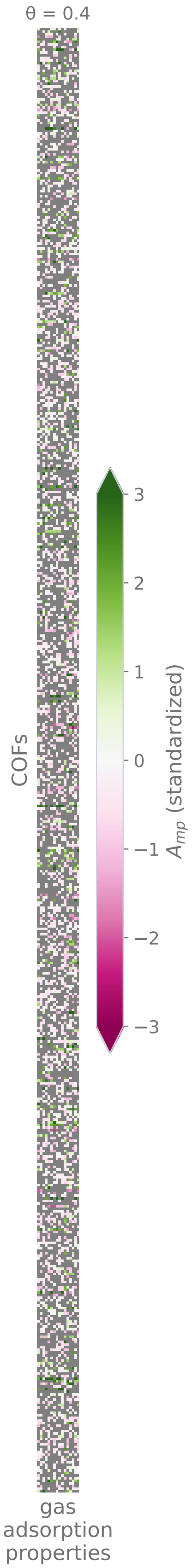


Figure 4: The 560×16 COF–adsorption-property matrix $\mathbf{A}^{(0.4)}$ with $\theta = 0.4$ the fraction of observed entries. Each column (adsorption property) is standardized to have zero mean and unit variance. Unobserved entries are gray. The value of the observed entries are depicted by the colors in the colorbar.

the validation set. The grid is the Cartesian product of (1) $k \in \{1, 2, ..., 15\}$ and (2) 25 values of λ ranging from 10 to 1000 and evenly spaced on a log-scale.

The *deployment* low rank model is then a new low rank model (a) with hyperparameter tuple $(k_{opt}^{(\theta)}, \lambda_{opt}^{(\theta)})$ and (b) trained on all [simulated] observed entries (train + validation data) of $\mathbf{A}^{(\theta)}$. 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 evaluation is done on the standardized and, in the case of Henry coefficients, $\log_{10}$-transformed values.

3.5 Results for observed fraction $\theta = 0.4$

We now demonstrate the utility of a low rank model, using a particular instance of a COF–adsorption-property matrix with a fraction $\theta = 0.4$ observed entries (shown in Figs. 4), for (i) imputing missing entries and (ii) drawing of a map of COFs and adsorption properties. Figs. 5-8 all correspond to the same deployment low rank model trained using the instance of $\mathbf{A}^{(0.4)}$ shown in Fig. 4, where the hyperparameter sweep found $k_{opt}^{(0.4)} = 3$, $\lambda_{opt}^{(0.4)} = 215.44$.

3.5.1 Imputing missing entries

We judge the performance of the low rank model for imputing the missing entries of the COF–adsorption-property matrix $\mathbf{A}^{(0.4)}$ 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 Fig. 5 shows the joint distribution of predicted and actual values of the (standardized and, in the case of Henry coefficients, $\log_{10}$-transformed) adsorption properties in the test data set—the simulated unobserved entries of $\mathbf{A}^{(0.4)}$. The density is greatest along the diagonal line of equality, indicating that the recommendation system is providing predictive value. The RMSE and Spearman's rank correlation coefficient on the test data is 0.6 and 0.77, respectively.

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 adsorption property by the deployment low rank model and its actual value (from the test set) is shown for each adsorption property in Fig. 6. With the exception of H_2O Henry coefficients, the recommendation system ranks the 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 Fig. S1).

As a baseline to judge the 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 eqn. 1. This material bias model, equivalent to the low rank model in eqn. 2 when $k = 0$, gives $A_{mp} \approx \mu_m$ and only considers 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 material bias model

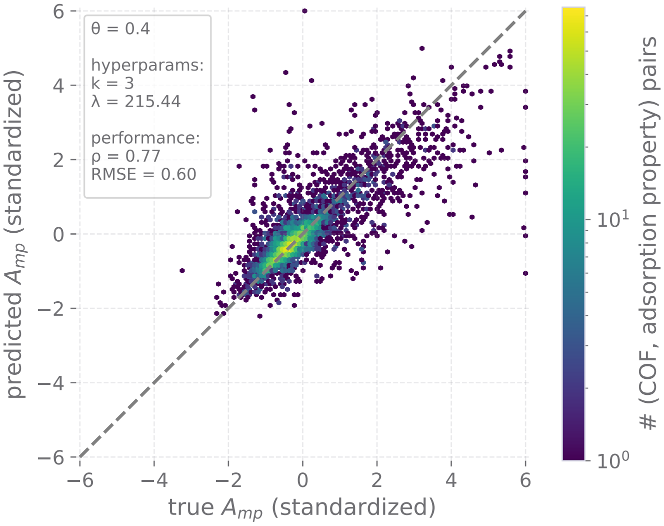


Figure 5: A parity plot showing the joint distribution, over the test data set (simulated unobserved entries in $\mathbf{A}^{(0.4)}$), of (y-axis) the predictions of the missing adsorption properties by the deployment low rank model and (x-axis) the actual value of the missing entries. The diagonal line represents perfect prediction.

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 Fig. 6 show Spearman's rank correlation coefficients between the value of the missing property (from the test set) and the prediction of the missing property by the benchmark material bias model. Indeed, the interaction term enhances the ability of the recommendation system to rank COFs according to their adsorption properties, 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 the 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).

3.5.2 The COF biases

The learned material bias of COF m, $\mu_m \in \mathbb{R}$, in eqn. 1 roughly describes the typical value of the (standardized) gas adsorption properties of COF m. Visualization of $\boldsymbol{\mu}$ can give us an idea of which COFs tend to exhibit the largest and smallest values of the gas adsorption properties. Fig. 7 visualizes the extremes of $\boldsymbol{\mu}$ from the deployment low rank model trained on $\mathbf{A}^{(0.4)}$ and displays the COF structures with the lowest and highest material biases. CCOF-2 (COF-LZU8) has the largest (smallest) μ_m, indicating that CCOF-2 (COF-LZU8) tends to exhibit the highest (lowest) values of the (standardized) gas adsorption properties among the COFs. Given a new gas adsorption task, the high material

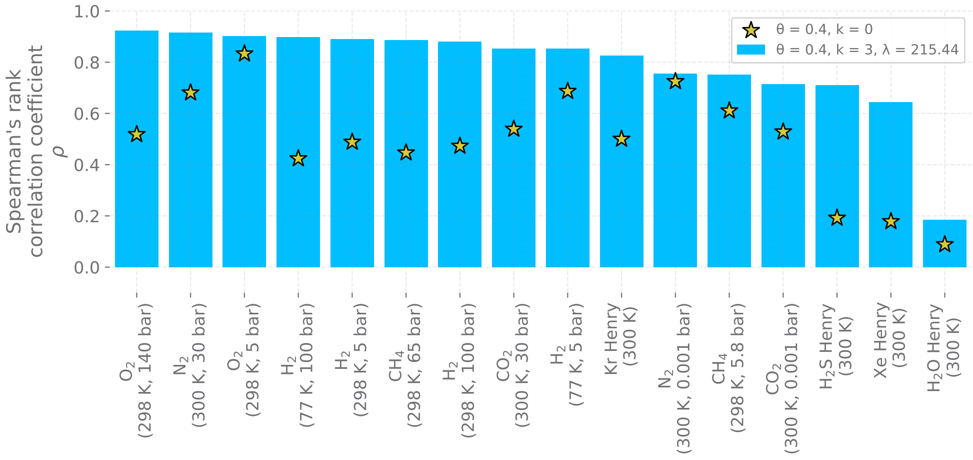


Figure 6: For each adsorption property, the height of the bar shows Spearman's rank correlation coefficient ρ between the prediction of a missing adsorption property in $\mathbf{A}^{(0.4)}$ (in the test set) by the deployment low rank model and its actual value. For comparison, the stars show ρ for the benchmark, material offset model where $A_{mp} \approx \mu_m$ and the interaction term is excluded.

bias μ_m of CCOF-2 makes it a good candidate for measurements, in the absence of any other information; in the analogy of movie recommendation systems (see Box), CCOF-2 is like a movie that is widely liked.

3.5.3 The learned map of COFs and gas adsorption properties

The learned latent representation of COF m, $\mathbf{m}_m \in \mathbb{R}^k$, encodes its adsorption properties into a compressed, low-dimensional vector. The locations of the COF representations in the latent space of COFs provide a "map" of the COFs. Within this map, COFs with similar (dissimilar) adsorption properties congregate (separate). Similarly, $\mathbf{p}_p \in \mathbb{R}^k$ is a latent representation of gas adsorption property p. Because the interaction term in eqn. 1 is $\mathbf{m}_m^\intercal \mathbf{p}_p$, the latent property vectors $\mathbf{p}_p$ indicate which regions of latent space tend to contain COFs with high values of those adsorption properties.

To visualize the map of COFs and adsorption properties, we resort to a dimension reduction method, Uniform Manifold Approximation and Projection (UMAP) [75], which embeds the latent representations of the COFs and properties, contained in the columns of $\mathbf{M}$ and columns of $\mathbf{P}$ respectively, into a 2D space. N.b. we apply UMAP on the horizontally concatenated matrix $\mathbf{M} \parallel \mathbf{P}$, as opposed to $\mathbf{M}$ and $\mathbf{P}$ separately, so that the latent representations of the material and property vectors are comparable. Fig. 8 shows the map of COFs and adsorption properties from the deployment low rank model for $\mathbf{A}^{(0.4)}$.

In the map of COFs in Figs. 8a, each point represents a COF, colored by (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 this map tend to exhibit similar values of the 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

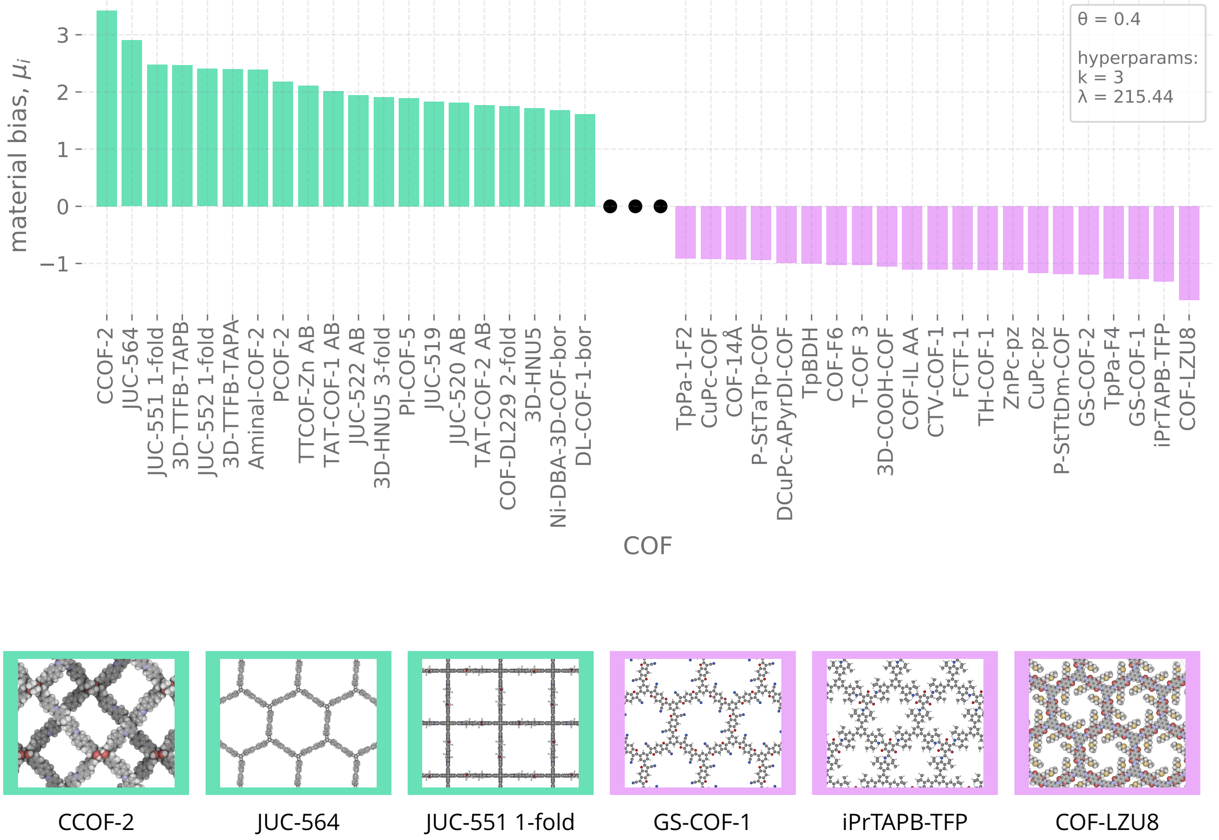
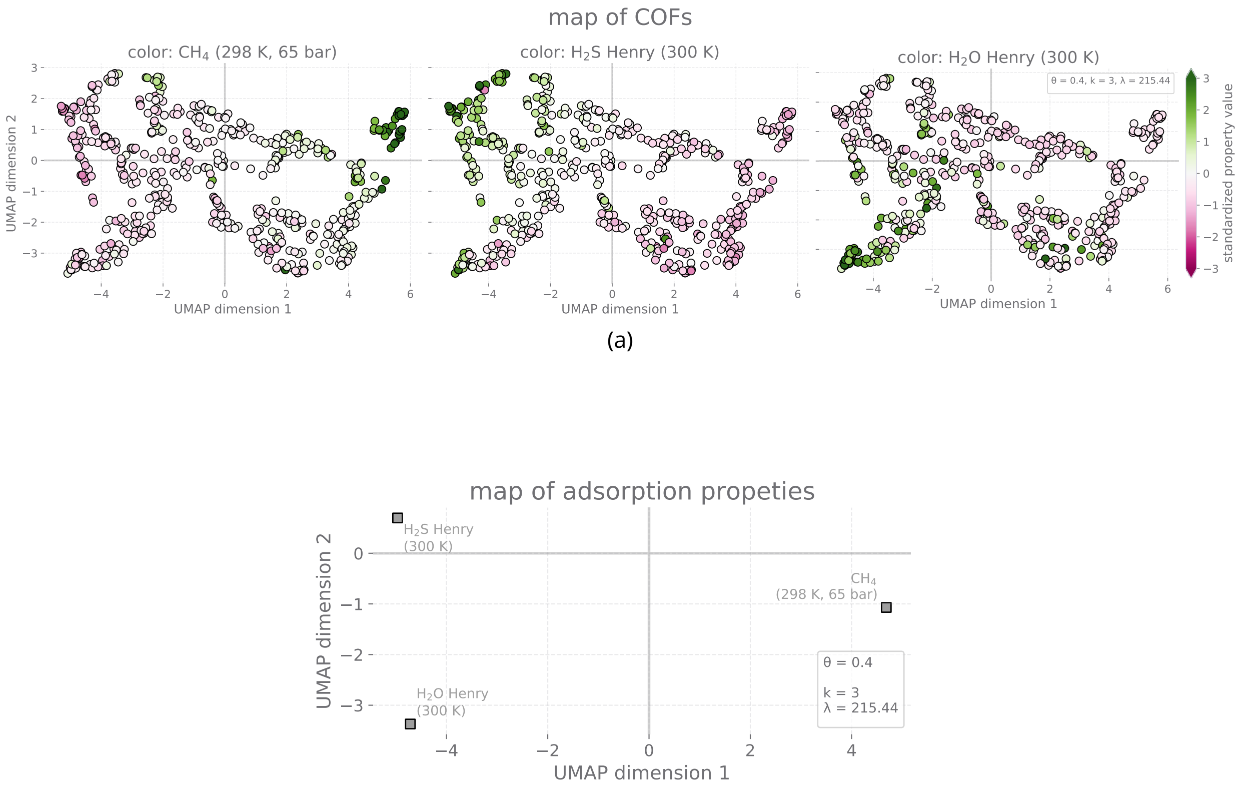


Figure 7: The ranked COF biases, $\{\mu_m\}$, for the $\theta = 0.4$ deployment low rank model. The COFs with the lowest and highest μ_m are shown, with the corresponding top and bottom three COF structures visualized below.

300 K, respectively, tend to lie on the right, top left, and bottom left of the latent COF space. Fig. 8b displays the locations of the latent representations of these three adsorption properties. Comparing Fig. 8a and Fig. 8b, the latent vectors of COFs with large (small) values of a property are oriented in the same (opposite) direction of the latent vector representation of that property—consistent with the interaction term in eqn. 1 as the dot product $\mathbf{m}_m^\intercal \mathbf{p}_p$. [As a cautionary note, UMAP embeddings do not preserve angles between vectors in the higher-dimensional space.]

Fig. S4 shows the COF map colored by the other adsorption properties, and Fig. S5 shows the complete adsorption property map.

In summary, Fig. 8 illustrates that the recommendation system 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. Such a 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 for other adsorption tasks, where the latent representations can serve as feature vectors for the COFs.



(a)

(b)

Figure 8: Learned map of COFs and gas adsorption properties: UMAP [75] embeddings of the latent representations of the (a) COFs, $\{\mathbf{m}_m\}$, and (b) a subset of adsorption properties, $\{\mathbf{p}_p\}$, into a 2D plane. (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. (b) Each point represents an adsorption property.

3.6 The effect of observed fraction θ on performance

Because the COF–adsorption-property $\mathbf{A}_{\text{complete}}$ from Ref. [17] is in reality complete, we have the luxury of studying the impact of the fraction of observed entries, θ, on the performance of the recommendation system. This investigation is important to address the practical question: how complete must the COF–adsorption-property matrix be for the recommendation system to reliably rank COFs according to their adsorption properties?

For a fraction of observed values $\theta \in \{0.1, 0.2, ..., 0.9\}$, 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, retrained a deployment model on all observed entries, then tested the deployment model on the

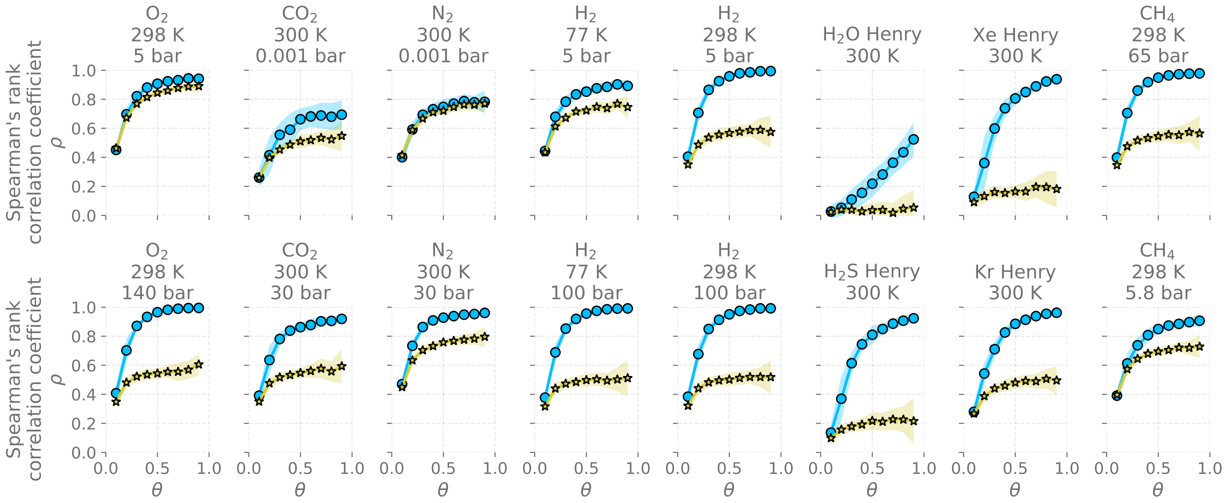


Figure 9: The effect of the fraction of observed values θ on the performance of the recommendation system for missing adsorption property imputation. Blue circles show the mean (over 50 simulations of data collection) Spearman's rank correlation coefficient between the prediction of the missing adsorption property (from test data) and its true value, as a function of θ, and for each adsorption property. Yellow stars shows the mean Spearman's rank correlation coefficient for the benchmark material bias model. Shaded bands signify the standard deviation.

unobserved (missing) entries serving as test data. Fig. S2 shows the distribution (among the simulations of data collection) of optimal hyperparameters $(k_{opt}^{(\theta)}, \lambda_{opt}^{(\theta)})$ for each θ. Fig. 9 shows Spearman's rank correlation coefficient, ρ, between the prediction of the missing adsorption property by the deployment low rank model and its actual value, for each adsorption property, as θ varies. The bands show the standard deviation over the 50 simulations of data collection. As in Fig. 6, ρ for the H_2O Henry coefficient is much lower than for other properties $\forall\theta$ 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 ρ), until θ is reduced below $\theta = 0.4$, where the fraction of missing values is too large to provide accurate predictions owing to a paucity of training examples. In conclusion, at least 40 % of the values of the COF–adsorption-property matrix must be observed for the recommendation system to reliably rank COFs according to their adsorption properties. For comparison, ρ for the baseline material offset model is also shown in Fig. 6 as the dashed line; the interaction term provides significant predictive value, with the exception of N_2 adsorption at (300 K, 0.001 bar) and O_2 adsorption at (298 K, 5 bar).

4 Conclusion and Discussion

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 values, 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. The (material, property) values are mathematically analogous to (product, customer) ratings in commercial recommendation systems.

We demonstrated a COF recommendation system for different gas adsorption applications. Our COF–adsorption-property matrix was composed of the simulated uptake of several gases at different conditions in 560 COF structures by Ongari et al. [17]. We simulated the process of data observation by artificially introducing missing values into the matrix. The (simulated) unobserved entries served as test data to assess the performance of the data imputation by the recommendation system. To both (i) impute the missing adsorption properties and (ii) machine-learn a "map" of COFs, wherein COFs with similar adsorption properties congregate, we trained a low rank matrix model [22] of the COF–adsorption-property matrix that had missing entries. The recommendation system was able to rank COFs according to their adsorption properties reasonably well (Spearman's rank correlation coefficient > 0.6), with the exception of water Henry coefficients. Moreover, coloring of the learned map of COFs by the adsorption properties indicated that, indeed, COFs with similar (dissimilar) adsorption properties clustered together (separated) in the map. The imputation performance of the recommendation system precipitously drops once the fraction of missing entries exceeds 60 %, though this figure does not necessarily generalize to other data sets.

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

The success of a recommendation system for NPMs is, however, predicated on structured, open databases of NPMs and their adsorption properties. One such database is the NIST/ARPA–E Database of Novel and Emerging Adsorbent Materials [19] (NIST-ISODB) that has collected and compiled gas adsorption measurements in NPMs from the literature, for both experimental and simulation sources, for a variety of gases at a wide range of conditions. We originally set out to develop a recommendation system using Henry coefficients extracted from this experimental data, but we found the resultant recommendation system was unable to reliably rank NPMs according to their adsorption properties. Particularly, we found the recommendation system with an interaction term included could not outperform the baseline material offset model. We propose four explanations for the poor performance of our recommendation system based on an NPM–adsorption-property matrix from NIST-ISODB; the explanations include both data-centric and model-centric concerns. First, the Henry coefficient matrix we constructed based on NIST-ISODB was only $\sim$20 % complete (i.e., too many missing values), which may have limited the success of recommendations for the remainder of the matrix. (Recall that the COF–adsorption-property matrix needed to be at least 40 % complete to satisfactorily make recommendations for the remainder of the matrix. This benchmark for the COF recommender system may not generalize to a broad set of NPMs and properties, but is nonetheless informative.) Second, the NPM–adsorption-property matrix may have included too much noise for a successful recommendation. It is known from meta-analyses of isotherms catalogued in NIST-ISODB [77, 78] that experimentally measured gas adsorption isotherms in NPMs exhibit high variance; this variance is ultimately manifested as noise that limits the success of the recommendation system. Third, the accuracy and reliability of Henry coefficients obtained from isotherms in NIST-ISODB are naturally limited by the source of data in the database itself. NIST-ISODB is primarily constructed from manual extraction of isotherm data from graphical figures in literature articles, since it is not common practice in the adsorption community to provide gas adsorption measurements as raw tabular data in publications. (As of the time of writing this work, only 1.3 % of isotherms in NIST-ISODB were from tabular data sources.) Consequently, the adsorption isotherm data loses precision first when the data is plotted graphically and second by human error when the graphical figure is digitized back to numerical data. This loss of precision particularly affects the generation of Henry coefficients from figures, as those coefficients are especially dependent on low-pressure data, which is often difficult to extract from isotherms plotted on a linear pressure scale. Fourth, our low rank model in eqn. 1 is linear; a non-linear model of the matrix [79, 80] may be able to capture relationships between the adsorption properties and achieve better performance. The third issue can be addressed by community adoption of the practice of releasing raw adsorption isotherm data in standardized, structured formats (cf. the CIF standard for crystallography [81]), which has been discussed previously [82–84].

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. Future work entails (a) creating a model for the selection bias in selecting materials for measurements of properties and (b) accounting for the selection bias in the recommendation system [85].

Another interesting direction for future work is to determine what (material, property) measurements should be made next to most improve the recommendation system, in an active learning strategy [86, 87].

As a remark, recommendation systems suffer from the *cold start* [24] problem: if a new material is reported, but none of its properties have been observed, the recommendation system is unable to make a prediction about any of its properties because we do not have data to learn the latent vector of this material, $\mathbf{m}_{M+1}$.

To (i) improve the performance of the recommendation system and (ii) alleviate the cold start problem, we propose to include structural and chemical properties of the materials that contribute to the prediction, in addition to the observed adsorption properties. For example, we could include in the model other information about the NPM structures, such as the void fraction, surface area, percent carbon atoms, etc. In the analogy with movie recommendation systems, this is analogous to including features about the movies, such as the genre, directors, year of production, and actors. These features could be added as additional (fully observed) columns in the material-property matrix, $\mathbf{A}$.

The material recommendation system is practically useful for recommending (i, application-led material search [1]) a material that optimizes a specific property or (ii, material-led application search [17]) an application for a given material. To motivate an experimental measurement in the lab, it may be necessary to quantify the uncertainty associated with a property imputed by the recommendation system. We remark that one could achieve this through bootstrapping and training an ensemble of recommendation systems on the bootstrap samples of observed (material, property) values or more advanced matrix completion methods designed to quantify uncertainty [88, 89].

Acknowledgements

A.S., X.F., and C.M.S. acknowledge the National Science Foundation for support under grant No. 1920945. We thank Madeleine Udell for tips on `LowRankModels.jl`.

Copyright and Disclaimer

Certain commercially available items may be identified in this paper. This identification does not imply recommendation by NIST, nor does it imply that it is the best available for the purposes described.

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] Eric J Gosselin, Casey A Rowland, and Eric D Bloch. Permanently microporous metal–organic polyhedra. *Chemical Reviews*, 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*, 2019.

[22] Madeleine Udell, Corinne Horn, Reza Zadeh, and Stephen Boyd. Generalized low rank models. *arXiv preprint arXiv:1410.0342*, 2014.

[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*, 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] 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.

[34] 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.

[35] 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*, 2020.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] 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. *arXiv preprint arXiv:2012.03723*, 2020.

[45] Connor W Coley. Defining and exploring chemical spaces. *Trends in Chemistry*, 2020.

[46] 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.

[47] 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.

[48] 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.

[49] 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.

[50] 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.

[51] 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.

[52] 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.

[53] 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, and Alán Aspuru-Guzik. Inverse design of nanoporous crystalline reticular materials with deep generative models. *Nature Machine Intelligence*, January 2021.

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

[55] 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.

[56] 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.

[57] 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.

[58] 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.

[59] 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.

[60] Yongchul G Chung, Emmanuel Haldoupis, Benjamin J Bucior, Maciej Haranczyk, Seulchan Lee, Hongda Zhang, Konstantinos D Vogiatzis, Marija Milisavljevic, Sanliang Ling, Jeffrey S Camp, et al. Advances, updates, and analytics for the computation-ready, experimental metal–organic framework database: CoRE MOF 2019. *Journal of Chemical & Engineering Data*, 2019.

[61] 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.

[62] 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.

[63] 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.

[64] 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.

[65] 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.

[66] Andrew Rosen, Shaelyn Iyer, Debmalya Ray, Zhenpeng Yao, Alan Aspuru-Guzik, Laura Gagliardi, Justin Notestein, and Randall Q Snurr. Machine learning the quantum-chemical properties of metal–organic frameworks for accelerated materials discovery with a new electronic structure database. *ChemRxiv*, 2020.

[67] 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.

[68] 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.

[69] 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.

[70] 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.

[71] 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.

[72] 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.

[73] 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.

[74] 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.

[75] Leland McInnes, John Healy, and James Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[76] Jeff Bezanson, Stefan Karpinski, Viral B Shah, and Alan Edelman. Julia: A fast dynamic language for technical computing. *arXiv preprint arXiv:1209.5145*, 2012.

[77] 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.

[78] Lukas W. Bingel, Andrew Chen, Mayank Agrawal, and David S. Sholl. Experimentally verified alcohol adsorption isotherms in nanoporous materials from literature meta-analysis. *Journal of Chemical & Engineering Data*, 65(10):4970–4979, 2020.

[79] Jicong Fan and Tommy WS Chow. Non-linear matrix completion. *Pattern Recognition*, 77:378–394, 2018.

[80] Jicong Fan and Madeleine Udell. Online high rank matrix completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8690–8698, 2019.

[81] S. R. Hall, F. H. Allen, and I. D. Brown. The crystallographic information file (CIF): a new standard archive file for crystallography. *Acta Crystallographica Section A*, 47(6):655–685, 1991.

[82] 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.

[83] N. Scott Bobbitt, Benjamin J. Bucior, Haoyuan Chen, Nathaniel Tracy-Amoroso, Zhao Li, Daniel W. Siderius, and Randall Q. Snurr. MOFdb: An accessible online database of computational adsorption data for nanoporous materials. *Journal of Chemical & Engineering Data*, 2021. In preparation.

[84] Jack D Evans, Volodymyr Bon, Irena Senkovska, and Stefan Kaskel. A universal standard archive file for adsorption data. *ChemRxiv*, 2021.

[85] 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.

[86] 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.

[87] 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.

[88] 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.

[89] Yuxuan Zhao and Madeleine Udell. Matrix completion with quantified uncertainty through low rank gaussian copula. *arXiv preprint arXiv:2006.10829*, 2020.