Machine learning-guided equations for super-fast prediction of methane storage capacities of COFs

Alauddin Ahmed

Mechanical Engineering Department, University of Michigan, Ann Arbor, MI 48109, United States

ABSTRACT: Covalent organic framework (COF) is a prominent class of nanoporous materials under consideration for vehicular methane storage. However, evaluating a COF for its methane capacity involves multiple experimental or computational steps, which is expensive and time consuming. Consequently, the discovery of high-capacity COFs for methane storage is very slow. Here we developed equations for super-fast prediction of deliverable methane capacities of COFs from a small number (3 to 7) of physically meaningful and measurable crystallographic features. We provided a set of equations with different fidelities for on-demand predictions based on the accessibility of crystallographic features. We found that an equation with only three crystallographic primary features, as variables, can predict deliverable capacities of 84,800 COFs with a root-mean-square error (RMSE) of 10 cm^3 (standard temperature and pressure, STP) cm^{-3} and mean absolute percentage error (MAPE) of 5%. However, the highest fidelity equation developed here contains seven crystallographic primary features of COFs with RMSE and MAPE of 8.1 cm^3 (STP) cm^{-3} and 4.2%, respectively. With that, we predicted methane storage capacities of 468,343 previously unexplored COFs using the highest fidelity equation and identified several hundred promising candidates with record-setting performance. CUBE_PBB_BA2, a hypothetical COF not yet synthesized, sets the new record of balancing gravimetric (0.396 g g-1) and volumetric (221 cm^3 (STP) cm^{-3}) deliverable methane storage capacities under the pressure swing between 65 and 5.8 bar at 298K. Also, 3D-HNU5, a previously synthesized COF, has shown the potential to achieve the gravimetric and volumetric methane storage U.S. Department of Energy target (0.5 g g^{-1} and 315 cm^3 (STP) cm^{-3}) simultaneously with uptakes of 0.755 g g^{-1} and 334 cm^3 (STP) cm^{-3} at 100 bar/270 K.

INTRODUCTION

Natural gas (NG) is the most abundant fossil fuel on earth.[1,2] It has been used as an alternative vehicular fuel for many years in many parts of the world. Also, nearly 25 % lower carbon footprint per unit energy and the lower price compared to its gasoline counterpart are considered to be the additional advantages of NG for on-board applications.[3,4] Chemically, NG is a mixture of different hydrocarbons of which nearly 95% is methane. Although the energy per unit mass (gravimetric energy density or specific energy) of methane is higher, the energy per unit volume (volumetric energy density) of methane is approximately 1000 times lower than that of conventional gasoline at ambient temperature (298 K) and pressure (1 atm). Consequently, achieving gasoline gallon equivalent (GGE) onboard methane storage, in an efficient and cost-effective manner, is one of the major necessities to the widespread adoption of NG as an alternative vehicular fuel.[4–9]

Covalent organic framework (COF),[10–12] a relatively novel class of porous solid-state adsorbents, is one of the pioneering candidates for on-board methane storage. A COF is formed by self-assembly of small molecules (i.e., chemical building blocks) in a compatible crystal net (or topology) via strong covalent bonds. This approach of COF synthesis is known as reticular synthesis.[10–12] Large surface areas per unit mass or volume, light-weights, high thermal and chemical stabilities, and low cost are the most attractive features of COFs as methane storage materials.[11]

The selection of a COF for methane storage depends on its high capacity at operating conditions (especially, temperature and pressure) suitable for on-board applications.[13] High pressure (65 bar) adsorption and low pressure (5.8 bar) desorption, at room temperature (298K), are widely acceptable conditions for methane storage in solid-state adsorbents including COFs.[14,15] The difference between high-pressure adsorption and low-pressure desorption is commonly known as the deliverable (aka, usable or working) methane storage capacity under a pressure swing (PS).

The current approach of synthesizing a COF for methane storage involves chemical intuition and domain expertise with an anticipation of high storage capacities. Although this traditional trial and error approach has been successful in many occasions, the pace of discovery is too slow to meet the fast-technological growth.[16–18] Evidently, only a modest number of (~500) experimental COF crystal structures[19–23]can be found in the literature to date since the first successful synthesis of a COF in 2005 by Yaghi and co-workers.[24]

Fortunately, scientists are now able to generate computational clones of experimentally synthesized COFs, imitating chemical building-block approach of COFs synthesis.[25–28] Also, assuming chemical building blocks and crystal nets (aka, topologies) as functional units of inheritance (i.e. genes) of COFs, scientists are now able to create new generations of computer-made COFs.[25–28] Such ability to design artificial (aka "hypothetical" or "in silico") COFs has created an opportunity to computationally generate thousands of COFs via incorporating domain knowledge, design rules, and desired functionalities.[25–28] In that spirit, Mercado et al.[29] generated 69,840 hypothetical COFs in search of a suitable adsorbent for methane storage. Mercado et al.[29] conducted a computational high-throughput screening of these COFs for methane storage capacities based on grand canonical Monte Carlo (GCMC) simulations. They identified 300 promising candidates with deliverable methane capacities surpassing HKUST-1 (190 cm^3 (STP) cm^{-3}), a

metal-organic framework (MOF) with record methane storage capacities. Recently, Lan et al.[26] reported a big database of 463,694 hypothetical COFs (from now on we will call this database "COF-genome-2018"), which has not yet been screened for methane capacities.

Unfortunately, the flexibility of COF design comes with the curse of nearly infinite possibilities, which can be imagined by the availability of 166.4 billion potential ligands or chemical building blocks (i.e., small molecules with 17 or fewer atoms) from the chemical space project.[30] The only restriction in this design space comes from the valency (i.e., connectivity of building units),[31] which is synonymous to 'jigsaw puzzle' for ligand self-assembly.

Over the years, high-throughput computing (HTC) has been successful in screening large databases of nanoporous materials including COFs for the discovery of high-capacity methane adsorbents.[16,32–38] Recently, hybrid methods involving HTC and machine learning (ML) are becoming handy while the number of hypothetical COFs/MOFs are on the rise.[33,39–44]The large datasets generated via HTC have created opportunities for developing ML models. The current focus, however, is to increase predictive accuracy of ML models by choosing different/sophisticated algorithms and/or generating novel input features.[44–46] Less attention has been given in generating features considering their accessibility, cost, measurability, and/or physical and chemical significance (i.e., interpretability).[47] Also, sometimes, feature generation could be daunting even for expert users. For example, Fanourgakis et al. could not use nearly 43% of a database because of the limitation of the code they used for generating features for their ML models.[48] Furthermore, feature generation for an unseen material could be quite expensive if it involves multiple experiments, high-level density functional theory (DFT) calculations, or new computer code.[47] However, oftentimes an user would like to know only methane capacity of a COF with a minimal information they have before conducting expensive experiments or computations.

Despite its applauding success in different disciplines, ML has rarely been used for formulating equations for the calculation of deliverable gas capacities of COFs/MOFs. Thus far, ML-based linear, polynomial, and multilinear regression models have been represented as the closest approximations of analytical equations for other nanoporous materials;[33,39,49] however, none exist for COFs. The predictive accuracy of linear or multilinear regression (MLR) models are often modest,[33,49] which severely compromise their applications in real-world problems. Also, the regression models trained with fictitious or unphysical features may not be directly deployable for unseen compounds (i.e., the compound not used in training and testing).[50] Furthermore, trained ML models may require special software, device, and data format for the predictions.

Recently, several methods – including genetic algorithm,[51–54] symbolic regression[55] (e.g., AI Feynman,[56]symbolic pregression[57]), SISSO[53,54,] (sure independence screening and sparsifying operator) – have been used for developing data-driven equations in science and engineering.[58–60]However, for our problem, SISSO[58] algorithm appears to have many advantages: lower training error compared to EUREQA[54] commercial software (see Fig. 2 and the Supplementary Information of Ref. [58]); options of generating multi-dimensional descriptors leading to single to multi-term equations; tunability of huge space of feature combinations; and open-source.

However, like many symbolic regression algorithms, prior applications of SISSO algorithm involved only datasets of modest sizes, on the order of tens to hundreds.[55,61–63] Fortunately, we possess an overwhelming dataset of tens of thousands of COFs, which limits our ability to use the SISSO algorithm effectively for this large dataset. In fact, there exists no prescribed recipe for handling large datasets in the literature using the SISSO algorithm.[58] Also, the predictability of SISSO-based equations on a completely unseen, diverse, and heterogeneous dataset consisting of hundreds of thousands compounds has yet to be examined.

Here we present a systematic approach of developing ML-guided equations for on-demand prediction of deliverable volumetric methane capacities of COFs. For this purpose, we selected an optimal set of primary features by first grouping based on their relevance and then developing ML models. We found that the feature subset consisting of eight crystallographic features could predict deliverable capacities of COFs with less than 3% mean absolute percentage errors (MAPE). Since crystallographic primary features are experimentally measurable quantities and can be calculated accurately and inexpensively using open-source codes,[64] we decided to use this subset for developing equations. We employed the sure independence screening and sparsifying operator (SISSO)[65] method for developing the equations for calculating deliverable methane capacities of COFs. To apply the SISSO algorithm on the large dataset of 84,800 COFs, we developed a multi-stage computational approach involving statistics and high-throughput computing. We developed a set of equations with different number of variables (i.e., primary features) and hence with different fidelities for on-demand prediction of deliverable methane capacities. Leveraging the high-fidelity equation predictions followed by GCMC verification, we screened a large database of 468,343 COFs for methane capacities. We identified hundreds of COFs that can potentially outperform state-of-the-art MOFs. Importantly, we identified a previously synthesized COF which has shown potential of meeting the U.S. Department of Energy methane storage gravimetric and volumetric targets (0.5 g g^{-1} and 315 cm^3 (STP) cm^3) simultaneously

METHODOLOGY

Database creation and selection: We created a super-database of in total 538,182 COF structures compiled from open-source databases (Table 1) reported to date. We adopted Berkeley-COFs-2018 dataset for developing ML-guided equations for the prediction of deliverable methane capacities. Following are the rationale behind this choice: open-source; diversity in ligands, topology, and dimensions of COFs; a rich set of features comprise categorical, continuous, and text data types; well-documented metadata regarding crystal nets and chemical building blocks; and GCMC calculated methane storage capacities at different operating (i.e., temperature and pressure) conditions.

Table 1. Super database of COFs.

Database identity	Type	Number of crystal structures
Berkeley-COFs-2018	Hypothetical	69,839
Berkeley-COFs-2014	Hypothetical	4,144
MG-COFs	Hypothetical	463,694
CURATED-COFs/CoRE COFs	Real	505
	Total	538,182

Primary feature design via data extraction and transformation: In Berkeley-COFs-2018 dataset, the 'dimension', 'bond type', 'linkerA', 'linkerB', and 'net' variables were originally provided in the text format. We transformed these four features into nominal categorical data types according to the schemes presented in Supplementary Tables 1-5. We extracted an additional seven features via manual text mining of the literature (Ref. 66) that reported the Berkeley-COFs-2018 dataset. These are: linker-1 termination type, linker-2 termination type, interpenetration, degree of interpenetration, linker-1 shape, linker-2 shape, and liner-1 & 2 combined shape. The numerical data assignment schemes of these seven categorical variables can be found in the Supplementary Tables 6-12.

Also, we augmented the features set by introducing two crystallographic properties of COFs, which are absent in the original Berkeley-COFs-2018 dataset. Volumetric surface areas (S_v) and pore volumes (V_p) of COFs were calculated from single crystal density (ρ_c), gravimetric surface area (S_g), and void fraction (F_v) via the relationships $S_v = \rho_c S_g$ (or, $S_v = F_v S_g / V_p$) and $V_p = F_v / \rho_c$, respectively. We compiled in total 40 primary features for the Berkeley-COFs-2018[25] dataset. A complete list of these primary features including their data types can be found in the Supplementary Table 13.

Grouping of primary features: We sorted 40 primary features compiled from the Berkeley-COFs-2018[25] dataset into 6 different groups based on their physical/chemical similarity, availability, and domain knowledge: modular, compositional, Euler, supercell, crystallographic, heat of desorption. The primary features under these groups are listed in the Supplementary Table 13.

Calculation of crystallographic properties of unseen COFs: We calculated $\rho_c, F_v, S_g, S_v, V_p$, largest included sphere diameter (D_i), largest free sphere diameter (D_f), and largest included sphere along free sphere path diameter (D_{if}) of in total 468,343 COFs from Berkeley-COFs-2014, MG-COFs, and CURATED-COFs/CoRE-COFs databases using Zeo++ code.[64] S_g and S_v of each COFs were calculated by 5000 Monte Carlo (MC) insertion of a fictitious spherical probe particle with a diameter 3.72 Å, which is equivalent to the approximate kinetic diameter of a N_2 molecule. V_p, free volume not occupied by the framework atoms, of each COF was calculated by 5000 MC insertions of a point particle with vanishing diameter (0 Å). F_v of each COF was determined from the ratio of V_p to the total volume of the unit cell.

Machine learning model development: We trained and tested in total 7 set of ML models based on Berkeley-COFs-2018[25] dataset with an augmented set of primary features. Among these, 6 sets of ML models were generated based on the 6 groups of primary features given in the Supplementary Table 13. The remaining set of ML models was developed based on all 40 features together.

We employed 14 different ML regression algorithms,[67–70] as implemented in the scikit-learn[71] package (Supplementary Table 14), for the selection of best ML algorithm. The entire dataset was first shuffled prior to developing any ML models via random permutations as implemented in the "shuffle" utility function of Scikit-learn. Mersenne Twister pseudo-random number generator as implemented in NumPy 'RandomState' instance via "numpy.random" function was used for generating the random number used in both shuffling and splitting.[72–75] However, 75% of the shuffled data were used for training and hyperparameter optimization and the rest (i.e., 25%) of the data were held-out as the test set. For all ML models, hyperparameters were optimized via 10-fold cross-validation method. The detailed workflow used here for the development of ML models can be found in the Supplementary Figure 1. The coefficient of determination (R^2), average unsigned error (AUE), root-mean-squared error (RMSE), and median absolute error (MAE) were used to assess the accuracy of ML model predictions compared to the actual data (here GCMC calculated usable capacities) of the test set. The details regarding these metrics can be found in the Supplementary Section S5 (Supplementary equations 1 to 4).

Development of equations for predicting methane capacities of COFs: ***Dataset augmentation and cleaning.*** Nearly 87% of our compiled super database (Table 1) contains COFs from different databases other than the dataset (Berkeley-COFs-2018) used for ML model development. Also, high methane capacity COFs are rare. Therefore, to increase the robustness of the developed equations, we augmented our ML dataset by adding 15,041 compounds selected from 450,526 unseen COFs from three other datasets of the super database (Table 1). Two-third of this additional dataset we selected based on their ML predicted high capacities and the rest randomly. We calculated deliverable methane capacities of this additional set using GCMC simulations as discussed in the following section. However, if the value of the crystallographic features of a COF is zero, we eliminated that compound from our final dataset for SISSO-based explorations. Therefore, we used a dataset of in total 84,800 COFs (Supplementary Text File 1) for the development and testing of equations using SISSO[58] algorithm.

Calculation of deliverable capacities of additional MOFs via GCMC simulations. For consistency with the Berkeley-COFs-2018 dataset, we used the same non-bonding interatomic potential parameters for representing methane molecules (Transferable Potentials for Phase Equilibria – United Atom, TraPPE-UA model)[76] and COF crystal structures (DREIDING force field)[77] with no internal bonding interactions. Also, non-bonding interactions between united-atom methane molecules and COF atoms were calculated via Lorentz–Berthelot mixing rules.[78,79] All interatomic potential calculations were truncated at 12 Å, which were later compensated by adding appropriate tail corrections.[80,81] The unit cell lengths smaller than twice of the truncation distance were replicated until at least 24 Å. Methane capacities were computed via GCMC[80,82–85] simulation method using the open-source RASPA[86] code. Methane capacity of COFs was calculated by averaging the number of methane molecules in the simulation

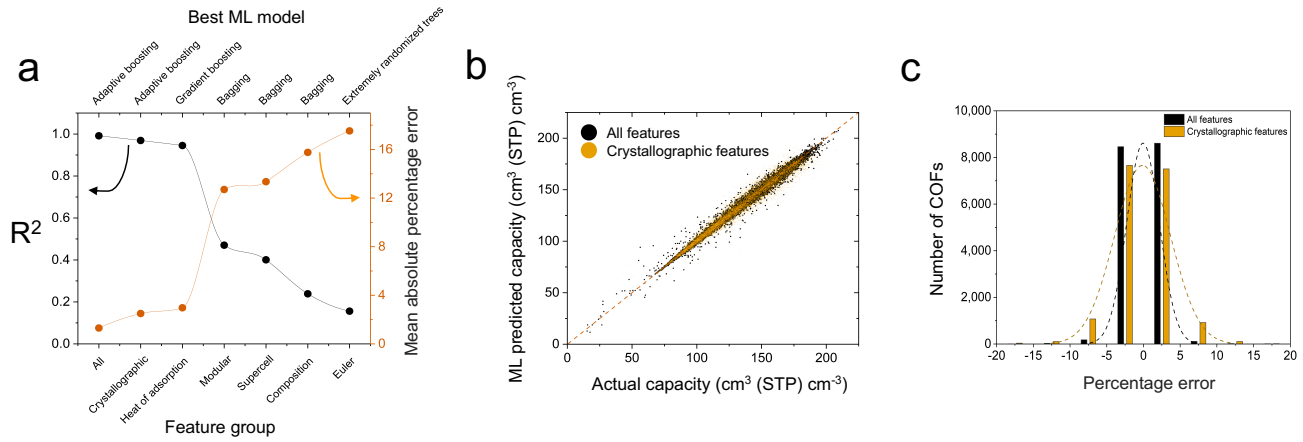

Figure 1. Performance of machine learning models in predicting deliverable methane storage capacity of COFs under the pressure swing between 65 and 5.8 bar at 298 K. (a) Predictability of ML models developed based on different features groups (bottom x-axis). The best ML algorithm, out of 14 examined, for each group of features is shown on the top x-axis. Performance of models were assessed based on R^2 (left y-axis) and mean absolute percentage error (right y-axis). (b) A comparison of correlations between actual and ML predicted deliverable methane storage capacities of COFs. Correlations are shown for ML models developed based on all features and crystallographic features. The read dashed line indicates the line of best fit for R^2 = 1.0. (c) A comparison of distributions of the numbers of COFs as a function of percentage error between actual and ML predicted deliverable methane storage capacities of COFs. Distributions are shown for ML models developed based on all features and crystallographic features. Dashed lines show the extent of Gaussian nature of distributions.

cell over 5000 GCMC production cycles, preceded by 5000 initialization cycles. At each cycle, translation, insertion and deletion Monte Carlo steps of methane molecules were performed with equal probabilities. Deliverable volumetric (DC_v) and gravimetric (DC_g)methane capacities at 298K were calculated by subtracting low pressure (5.8) methane adsorption from the high pressure (65 bar) adsorption. Methane fugacity as a function of temperature and pressure were calculated using the Peng-Robinson equation.[87]

General setup for the SISSO calculations. We employed the sure independence screening and sparsifying (SISSO)[58,59,88] algorithm to identify suitable mathematical combinations of primary features for the development of equations for predicting deliverable methane capacities of COFs. SISSO was allowed to conduct four binary addition, subtraction, multiplication, division) and seven unary operators (exponent, logarithm, square-root, cube-root, inverse operations, square, and cube) on the provided set of primary features. The SISSO algorithm ranked *composite* (i.e., secondary features generated by SISSO) features based on their correlations with the target output (here deliverable methane storage capacities (DC_v) in COFs). Up to three combinations of SISSO-generated *composite* features (i.e., secondary features generated by SISSO) were allowed to construct equations for predicting DC_v. Linear and multilinear regressions were employed to construct single and 2-and-3 term equations, respectively.

RESULTS AND DISCUSSION

Screening primary features. We developed ML models based on all (40) primary features and 6 disjoint subsets of these (Supplementary Table 13) for predicting deliverable methane storage capacities of COFs under the pressure swing between 65 and 5.8 bar at 298K. Based on 14 different regression algorithms (Supplementary Table 14) and 7 features group, we developed in total 98 ML models for the prediction of methane capacities.

Figure 1a illustrates the predictability (in term of R^2 and MAPE) of best performing ML models (top x-axis) developed based on different groups of features (bottom x-axis). Supplementary Table 15 presents three additional performance metrics (AUE, RMSE, and MAPE) for the ML models shown in Fig. 1a. Noticeably, the best performing ML regression algorithm is different for a different group of features. Also, the accuracy ML predictability can vary significantly depending on the choice of algorithms. This suggests that the random choice of an ML algorithm could lead to a poor ML model for the prediction of deliverable methane storage capacity of COFs, which agrees with the 'free lunch theorem'[89,90] of ML.

The predictability of an ML model trained with all 40 primary features using adaptive boosting (AB) algorithm with decision trees (DT) as a base estimator is the best with R^2 and AUE of 0.99 and 1.7 cm^3 (STP) cm^{-3}, respectively. Interestingly, an ML model trained with only 8 crystallographic primary features using the same AB(DT) method is the second best with R^2 and AUE 0.968 and 3.51 cm^3 (STP) cm^{-3}, respectively.

Figure 1b shows the correlations between actual (GCMC calculated) and ML model predicted methane capacities of COFs. This plot compares the predictability of ML models developed based on 40 (all) features and 8 crystallographic features. Apart from slight divergence of points for the crystallographic feature-based ML model predictions, no unusual pattern in data points is visible.[91] This is further made clear from the Gaussian distribution of percentage errors (Figure 1c) between actual and ML model predicted methane capacities. Both ML models can predict methane storage capacities of ~92% COFs with errors less than 5%. The prediction errors for the rest of MOFs fall within 10%.

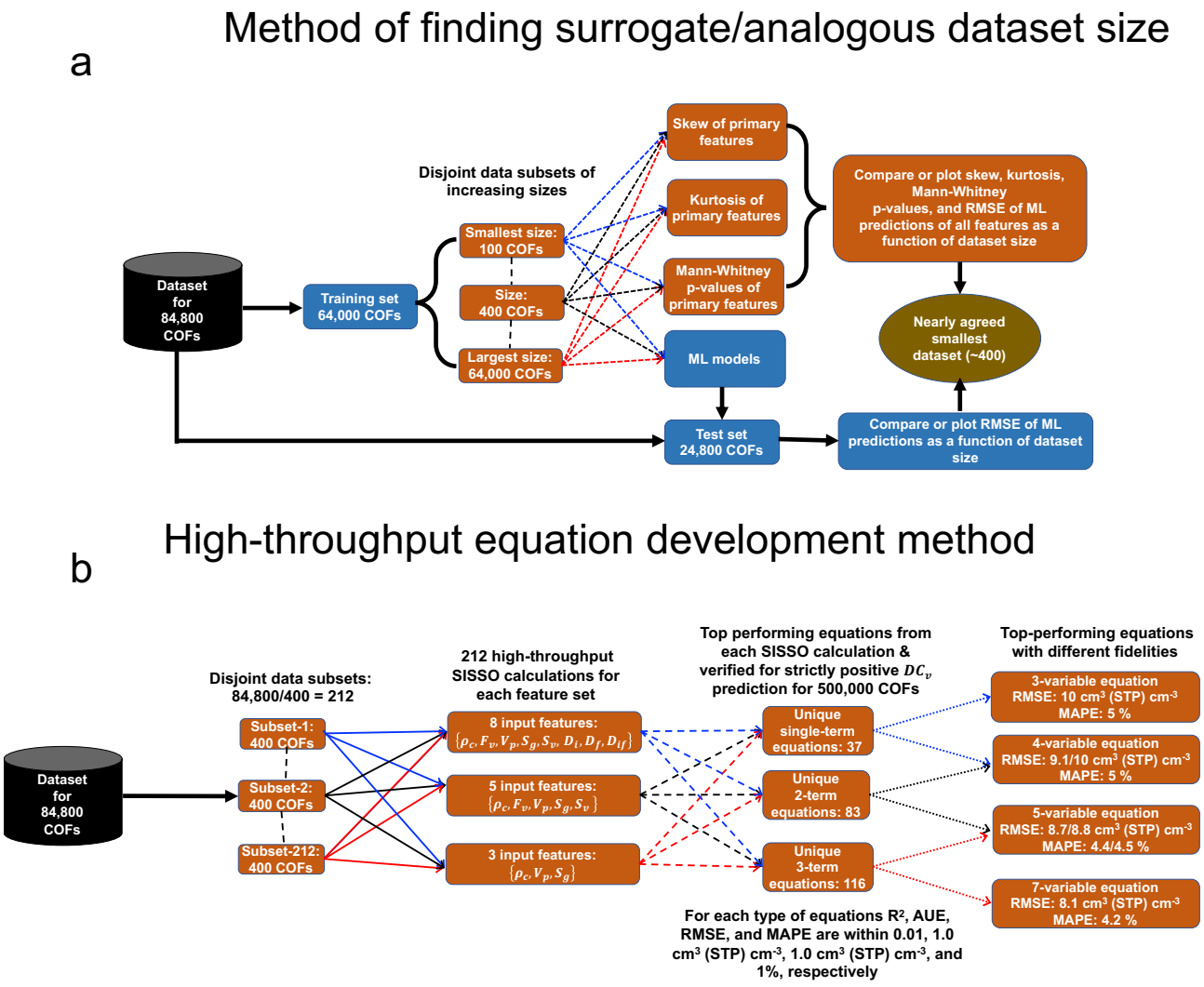

Figure 2. Computational method of developing equations based on a large dataset. (a) The method of identifying surrogate/analogous dataset that can represent the entire dataset of 84,800 COFs. (b) High-throughput computational protocol used for the development of equations for calculating usable methane capacities via SISSO algorithm. Coefficient of determinant (R^2), average unsigned error (AUE), root-mean-square error (RMSE), and mean absolute percentage error (MAPE)of top-performing equations were calculated against the entire dataset.

Construction and validation of composite features: *Finding the smallest surrogate/analogous dataset size.* Figure 2a shows different components of the workflow used in finding the smallest surrogate/analogous dataset size that we can consider as a being representative of the entire dataset. First, we split our dataset of 84,800 COFs into training and test sets of sizes 63,635 and 21,165, respectively. By randomly selecting COFs, we constructed 90 subsets of COFs based on the training set of 63,635 COFs. The smallest subset contains 100 COFs and the rest comprise increasing numbers of COFs up to 63,635 (largest subset) as shown in the Supplementary Table 16. Next, we calculated mean, median, skew (measure of deviation from normal distribution in terms of non-identical tails),[92–95] kurtosis (shape of the non-normal distribution in terms of height or flatness),[92–95] and the Mann-Whitney p-values (to test the likeliness of two subsets of data originated from the same population)[92–96] of 8 crystallographic primary features of all 90 subsets. Comparison of all these statistical measures (Supplementary Tables 17-24) of primary features suggested that a randomly selected subset containing ~400 COFs, at a minimum, could be a surrogate/analogous dataset, as a representative of the entire training set. To further verify this finding, we trained 90 ML models based on the subsets and evaluated their performances against a fixed test set of 21,165 COFs. The comparison of R^2, AUE, RMSE of 90 ML models (Supplementary Table 25) confirmed that a dataset of ~400 COFs could be a surrogate/analogous dataset without sacrificing accuracies drastically.

High-throughput equation development. Figure 2b shows major components of the high-throughput computational approach developed here using the SISSO algorithm. We constructed 212 disjoint (i.e., no COF in common) subsets of 84,800 COFs. Each subset comprised 8 primary features (input) and deliverable methane capacities (output) of 400 COFs. We call these subsets 8-feature subsets.

The crystallographic features of COFs such as ρ_c, S_g, and V_p are all routinely measurable properties. S_v and F_v of a COF are usually

calculated from S_g and V_p, respectively, multiplying by its ρ_c. However, measured values of all three pore geometric features (D_i, D_f, and D_{if}) together are rarely reported in the literature because of the experimental complexity and cost.[97] Therefore, an equation for the calculation of deliverable methane storage capacities based on ρ_c, S_g, V_p and/or their derivatives (S_v and F_v) is highly desirable.

To this end, we deleted three pore geometric primary features (i.e., D_i, D_f, and D_{if}) from 8-feature subsets and constructed 212 new 5-feature subsets. Similarly, by deleting F_v and S_v from 5-feature subsets, we constructed 212 new 3-feature subsets. In summary, we generated in total 636 data subsets of COFs. We carried out 636 independent SISSO calculations based on these data subsets in a high-throughput fashion for finding the best performing set of equations.

Screening top-performing equations: We compiled in total 1,908 (including duplicates) top-preforming single-term, 2-term, and 3-term equations from 636 separate SISSO calculations based on disjoint datasets, each consisting of 400 COFs. These equations resulted from a search of nearly 75 billion SISSO-constructed composite features. After deletion of the duplicates, we imposed following criteria for further screening of top-performing equations:

Criterion 1: Predictability of strictly positive deliverable capacities of 538,102 COFs: We eliminated all the equations that predicted negative deliverable methane capacities of a single COF using an in-house script written based on the SymPy[98] python library.

Criterion 2: Hierarchy of high-fidelity equations: We kept only the equations able to reproduce actual deliverable methane capacities of 84, 800 COFs with R^2, AUE, RMSE, and MAPE values simultaneously within 0.01, 1.0 cm^3 (STP) cm^{-3}, 1.0 cm^3 (STP) cm^{-3}, and 1%, respectively, of the highest-performing single-term (eqn. 1a/1b), 2-term (eqn. 2a), and 3-term equations (eqn. 3a), as shown in Table 2.

Single-term equations for calculating deliverable methane capacities: We identified eqn. 1a as the top-performing single-term equation via a successive screening based upon AUE, RMSE, and MAPE. This equation requires four primary features (ρ_c, F_v, S_g, S_v) as variables for the prediction of methane deliverable capacities under the PS condition. Since $S_v = \rho_c S_g$ and $F_v = \rho_c V_p$, we can transform eqn. 1a into eqns. 1b & 1c:

$$DC_v = 6.48\, F_v S_v S_g^{-1/2} exp(-\rho_c^2) + 38.7 \quad (1a)$$

$$DC_v = 6.48\, F_v S_g^{1/2} \rho_c\, exp(-\rho_c^2) + 38.7 \quad (1b)$$

$$DC_v = 6.48\, V_p S_g^{1/2} \rho_c^2\, exp(-\rho_c^2) + 38.7 \quad (1c)$$

Table 2 summarizes the performance of these equations tested against a dataset of 84,800 COFs. Notably, eqns. 1a-1c do not contain any pore geometric primary features (D_i, D_f, and D_{if}) as a variable. Also, eqns. 1b&1c depend only on three primary features (ρ_c, S_g, F_v or V_p) as variables, which are the measurable crystallographic fingerprints of any COFs. Depending on the availability of these three crystallographic features one could easily calculate deliverable methane capacity of an arbitrary COF using only paper and pencil.

We identified in total 37 single-term equations that can predict deliverable methane capacities of 84,800 COFs with R^2, AUE, RMSE, and MAPE values within 0.01, 1.0 cm^3 (STP) cm^{-3}, 1.0 cm^3 (STP) cm^{-3}, and 1%, respectively, of eqns. 1a-1c. Many of these equations either comprise 3 variables or we could transform these into 3-variable equations.

Two-term equations: We identified in total 83 top-performing unique 2-term equations, which satisfy Criterion 2 of equation selection. Among these, the first 21 2-term equations contain at least one pore geometric variables, mostly the largest included sphere (D_i). Eqn. 2a is the top-performing 2-term equation, which is strictly positive and ranked 3rd among 83 high-performing equations.

$$DC_v = 0.386\, \frac{F_v^2 (\ln(S_v))^3}{V_p^{\frac{1}{3}}} + 4160\, \frac{F_v^3 \exp(-V_p) \ln(V_p)}{\rho_c\, D_i} + 45.5 \quad (2a)$$

The first 2-term equation without any pore geometric variable is eqn. 2b, which is ranked 26th among the top performers.

$$DC_v = -0.0313\, F_v^3 S_v\, \rho_c \ln\left(\frac{\rho_c}{S_g}\right) + 0.222\, \frac{\rho_c\, exp\left(S_v^{\frac{1}{3}}\right)}{S_g \ln(F_v)} + 69.2 \quad (2b)$$

Three-term equations: We identified in total 116 high-performing 3-term equations via subsequent screening based on R^2, AUE, RMSE, and MAPE. The top-performing equation (eqn. 3a) comprises 7 out of 8 primary feature variables.

$$DC_v = 3780\, \frac{V_p F_v^2 \exp(-V_p) \ln(V_p)}{D_{if}} + 1.33 F_v S_v^{\frac{1}{2}} \exp\left(\rho_c^{\frac{1}{3}}\right) - 0.0194\, \frac{F_v \exp\left(S_v^{\frac{1}{3}}\right)}{D_i^2 D_f} + 38.1 \quad (3a)$$

The first top-performing 3-term equation with no pore geometric primary feature variable was 18th in the ranks; however, it predicted negative values for some unseen COFs. Eqn. 3b is the next top-performing 3-term equation free from pore geometric primary feature variable, which was 19th in overall ranking.

$$DC_v = 0.289\, F_v^3 S_v \rho_c \exp(-\rho_c^3) - 445\, \frac{V_p^2 S_g \exp\left(-S_g^{\frac{1}{3}}\right)}{F_v} + 0.186\, \frac{\rho_c \exp\left(S_v^{\frac{1}{3}}\right)}{S_g \ln(F_v)} + 75.8 \quad (3b)$$

Eqns. 3a&b predicted strictly positive values both for the test and unseen set COFs.

Comparison of single-term, 2-term, and 3-term equations predictability: Figures 4a-c show the correlations between actual (here GCMC calculated) data and equations (eqns. 1 to 3) calculated val-

Table 2. Accuracy of top-performing equations. Coefficient of determinant (R^2), average unsigned error (AUE), root-mean-square error (RMSE), and mean absolute percentage error (MAPE) of top-performing equations were calculated against a test set of 84,800 COFs. The unit of both AUE and RMSE is cm^3 (STP) cm^{-3}.

Equation	No. of variable	R^2	AUE	RMSE	MAPE
1a	4	0.89	7.0	10.0	5.0
1b/1c	3	0.89	7.0	10.0	5.0
2a	5	0.92	6.0	8.7	4.5
2b	4	0.91	6.3	9.1	4.8
3a	7	0.93	5.7	8.1	4.2
3b	5	0.92	6.0	8.8	4.4

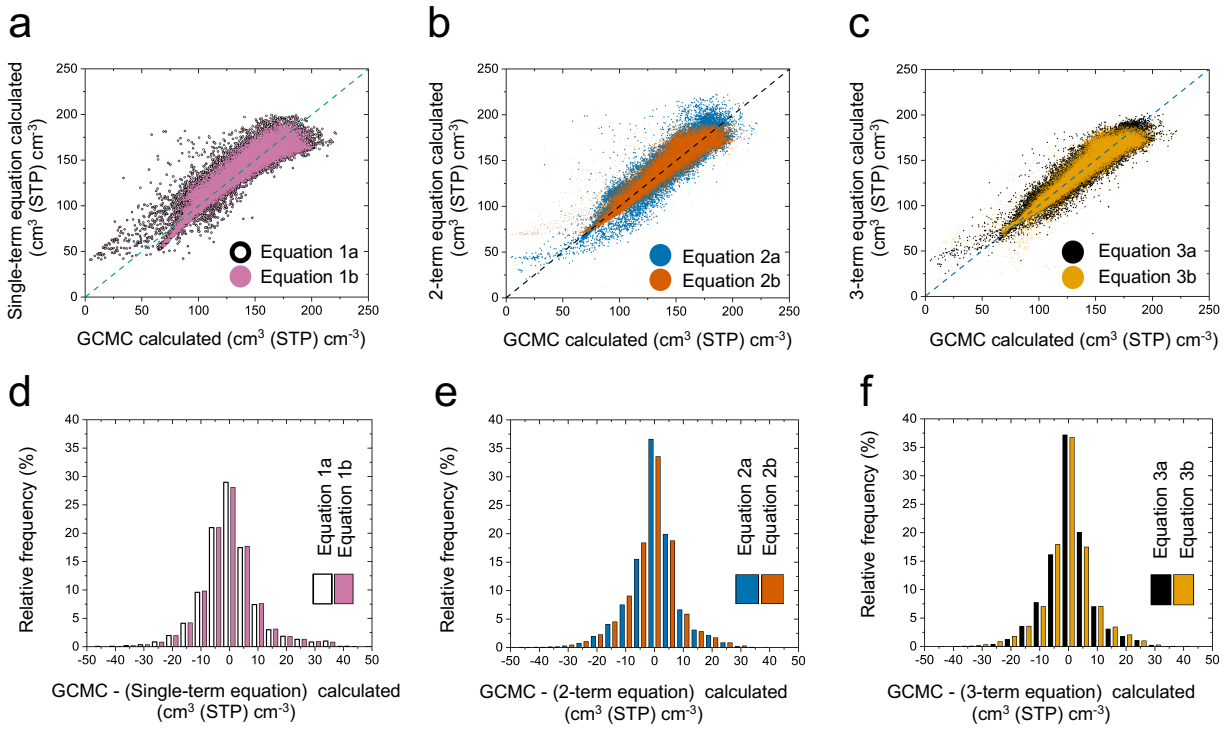

Figure 3. Predictability of high-performing equations developed here in comparison with GCMC calculated data. Correlations (a, b, c) are shown between GCMC calculated and high-performing equations that predicted deliverable methane capacities of in total 84,800 COFs. The dashed diagonal lines in a, b, c indicate perfect correlations. The distributions of differences between GCMC calculated and equations predicted deliverable capacities presented in a, b, and c are shown in d, e, and f, respectively.

ues of 84,800 COFs. Figures 4d-f show the corresponding distribution of errors (the difference between actual data and eqn. calculated values). According to Table 1 and Figs. 3a&b, the transformation of a 4-feature (eqn. 1a) to a 3-feature (eqn. 1b) single-term equation by feature substitution does not change the predictabilities of deliverable capacities. Also, Figs. 4b&c suggests that the predictabilities of equations with (eqns. 2a & 3a) and without (eqns. 2b & 3b) pore geometric features are slightly different. Most of the prediction errors shown in distribution plots (Figs. 3d-f) fall in between -10 and +10 cm^3 (STP) cm^{-3} exhibiting Gaussian distributions. The prediction accuracy of equations with a smaller number of terms slightly compromised for COFs with low deliverable methane capacities ($\lesssim$ 50 cm^3 (STP) cm^{-3}). However, the overall predictability of the equations, with a number of features (variables) between 3 and 7, for a large dataset of over 84,000 COFs is remarkable.

We can use these equations for on-demand prediction of deliverable methane capacities of COFs based on the crystallographic features at hand generating no additional features. Now, one can calculate methane capacities of COFs using only paper and pencil, a calculator, or an excel sheet with no experience of machine learning and GCMC simulations.

Calculation of not yet reported methane deliverable capacities of 449,468 COFs. We calculated deliverable volumetric methane capacities of in total 449,468 COFs (from MG-COFs and CURATED-COFs/CoRE-COFs databases) based on the high-fidelity equation developed here (eqn. 3a). To the best of our knowledge, no one has reported deliverable methane capacities of these COFs yet. Since vehicular application of methane requires a balance between volumetric and gravimetric capacities of solid-state adsorbents, we calculated deliverable gravimetric capacities (DC_g) in g g^{-1} from usable volumetric capacities (DC_v) in cm^3 (STP) cm^{-3} via eqn. 4:

$$DC_g = \frac{DC_v M_{CH4}}{22.4 \times 1000 \rho_c} \quad (4)$$

where M_{CH4} is the molar mass of methane (16.04 g mol^{-1}), and ρ_c (in g cm^{-3}) is the crystal density of a COF.

We sorted COFs based on deliverable volumetric and gravimetric capacities, subsequently, in descending order. For screening purposes, we further verified deliverable capacities of

Table 3. Screening of COFs based on deliverable methane storage capacities under the pressure swing between 65 bar and 5.8 bar at 298 K. Only COFs from MG-COFs and CURATED-COFs/CoRE COF databases were considered here since the other two databases were previously screened elsewhere. The details regarding the sources of measured data against which screenings were carried out can be found in the Supplementary Table 26.

Name	Volumetric capacity (cm^3 (STP) cm^{-3}) /gravimetric capacity (g g^{-1})	Promising COFs identified
linker91_C_linker91_C_tbd (hypo. COF)	216 /0.309 (calc.)	5
MOF-519 (real)	2010/0.157 (expt.)	9
NJU-Bai43 (real MOF)	198/0.221 (expt.)	413
UTSA-76 (real MOF)	194/0.199 (expt.)	1108
HKUST-1 (real MOF)	185/0.150 (expt.)	2,459
Al-soc-MOF-1 (real MOF)	176 /0.37 (expt.)	203
Theoretical limit	~200 (calc.)	88
AX-21 (activated carbon)	190/0.28 (expt.)	1,285

Table 4. Predicted record-setting methane storage capacities of COFs at 270 K. DC_v (cm^3 (STP) cm^{-3}) and DC_g (g g^{-1}) represent deliverable volumetric and gravimetric methane capacities, respectively. Real and hypothetical (hypoth.) are from CURATED-COFs/CoRE-COFs and MG-COFs, respectively.

Name	DC_v	DC_g
5.8-to 100-bar deliverable capacity at 298 K		
CUBE_PBB_BA2 (hypoth.)	221	0.396
CUBE_KET2_BA2 (hypoth.)	220	0.343
silicon_105-mi-noopt (hypoth.)	219	0.313
silicon_105-mi-cellopt (hypoth.)	218	0.318
MET_105-biqin (hypoth.)	217	0.299
5-to 100-bar deliverable capacity at 270 K		
3D-HNU5 (real)	285	0.643
MET_N2_BA2 (hypoth.)	280	0.501
ball_cen2-BA2_PBB_2HYD2_No1 (hypoth.)	277	0.485
cn4_fimr2-funan1_No1 (hypoth.)	276	0.507
Boro-BDC_A-irmof20_A_No65 (hypoth.)	276	0.470
100 bar total capacity at 270 K		
3D-HNU5 (real)	334	0.755
CUBE_KET2_BA2 (hypoth.)	329	0.512
silicon_105-mi-noopt (hypoth.)	326	0.467
ball_cen2-BA2_PBB_2HYD2_No1 (hypoth.)	326	0.569
bar-sibiqinfei-ket (hypoth.)	324	0.474

the top-performing 10,000 COFs via GCMC simulations. Then we screened GCMC-verified COFs against the deliverable capacities of 7 record-holding (under 65/5.8 bar pressure swing at 298K) solid-state adsorbents reported to date (see Supplementary Table 26 for the details regarding the sources and the method of curation of measured data with comments). Table 3 summarizes the number of COFs that can potentially out-perform the performance of 7 record-holding solid-state adsorbents based on usable volumetric and gravimetric capacities simultaneously. Although based on reported data (210 cm^3 (STP) cm^{-3} & 0.152) MOF-519 is currently the record-holder under 65/5.8b bar pressure swing,[2] in a later report authors have expressed some concerns regarding the reproducibility of this data (see footnote h of Table 2 in Ref. 99). However, we have identified 9 COFs that can potentially out-perform MOF-519. NJU-Bai43 (a real MOF) and UTSA-76 (a real MOF) are the next record-holding materials based on single source reported values. We identified 413 and 1,108 COFs that can potentially out-perform Bai43 and UTSA-76, respectively. Our screening identified 2,459 COFs that can potentially surpass one of the benchmarked[100–102] and well-studied[103] adsorbents HKUST-1 (a MOF with capacities 185 cm^3 (STP) cm^{-3} and 0.150 g g^{-1} under 65/5.8 bar pressure swing).

In case of COFs, Mercado et al.[25] and Martin et al.[28] previously screened Berkeley-COFs-2018 and Berkeley-COFs-2014 databases, respectively, based on GCMC calculated deliverable capacities. Mercado et al.[25] identified that the deliverable capacity of linker91_C_linker91_C_tbd (a hypothetical COF from Berkeley-COFs-2018) is the highest reported deliverable capacity to date (216.8 cm3 (STP) cm^{-3}), at 298K under the pressure swing between 65 and 5.8 bar, of any solid state-adsorbents including MOFs. Here we identified 5 promising COFs that can potentially out-perform linker91_C_linker91_C_tbd. Table 4 summarizes the deliverable capacities of these record-setting COFs including the source databases (Supplementary Table 27 compiles crystallographic properties of these COFs). To the best of our knowledge, the CUBE_PBB_BA2 (an MG-COF) sets the new record of balancing gravimetric (0.396 g g^{-1}) and volumetric (221 cm^3 (STP) cm^{-3}) deliverable methane storage capacities under the pressure swing between 65 and 5.8 bar at 298K (Fig. 4b shows a crystal structure).

Recently, Chen, Farha and co-workers reported a MOF, NU-1501-Al, with record setting total gravimetric capacity of 0.66 g g^{-1} [262 cm^3 (STP) cm^{-3}] at 100 bar/270 K and a deliverable capacity of 0.60 g g^{-1} [238 cm3 (STP) cm^{-3}] under the pressure swing between 100 and 5 bar at 270 K. To compare the performance of high-capacity COFs with NU-1501-Al, we calculated deliverable methane storage capacities of in total 2,842 COFs under pressure swing between 100 and 5 bar at 270 K using GCMC simulations. These COFs were the top-preforming candidates under the pressure swing between 65 and 5.8 bar at 298K.

We screened 2,842 COFs against the storage capacities of NU-1501-Al and discovered several COFs with record setting methane capacities. Table 4 compiles the top-performing candidates at 100 bar/270 K and under pressure swing between 5 and 100 bar at 270 K (Supplementary Tables 28 & 29 summarize the crystallographic properties of these COFs). We found that 3D-HNU5 (CURATED-COFs[104] database structure labels: 19400N3), a real COF reported by Guan et al.[105], surpasses the U.S. Department of Energy methane storage gravimetric and volumetric targets (0.5 g g^{-1} and 315 cm^3 (STP) cm^{-3}) simultaneously with uptakes of 0.755 g g^{-1} and 334 cm^3 (STP) cm^{-3} at 100 bar/270 K (Fig. 4c shows a crystal structure).

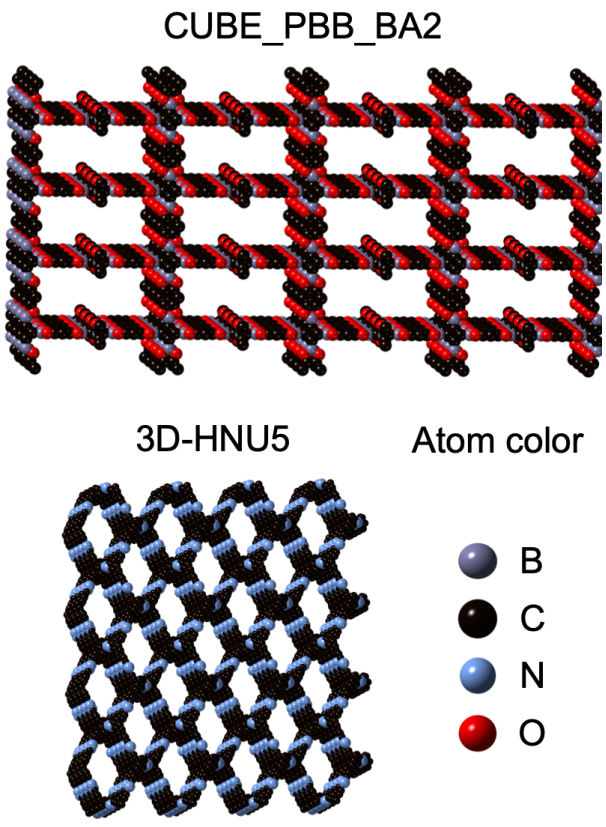

Figure 4. Crystal structure of high-capacity COFs identified in this work. Hydrogen atoms are omitted for clarity.

Also, it (285 cm^3 (STP) cm^{-3} & 0.643 g g^{-1}) outperforms the current record-holding MOF NU-1501-Al under pressure swing between 5 and 100 bar at 270 K.

CONCLUSIONS

Here we presented a systematic approach of selecting optimal feature set for the development of equations for predicting methane deliverable capacities. For this purpose, we generated addition features of COFs and grouped them based on physical, chemical, and crystallographic similarities. Also, we benchmarked 14 ML regression algorithms against 7 groups of features. Consequently, we identified a set of 8 crystallographic primary features for the development of equations. As a by-product, we developed highest - performing ML models to date for the prediction of deliverable methane capacities in COFs.

We developed a statistics-based approach for the identification of the smallest surrogate/analogous dataset (400 COFs) from a large dataset of 84,800 COFs. We carried out 636 independent SISSO calculations in a high-throughput fashion. In this way, we explored a large combinatorial space of SISSO-constructed features, to be precise 75 billon. Our method has created a potential pathway of developing equations based on large datasets for which SISSO-type symbolic regressions are currently not feasible.

We developed a set of equations with a different number of features and accuracies for the prediction of deliverable methane capacities. These variable-fidelity equations allow users for on-demand prediction of methane capacities with the features in hand, which maximize user friendliness and minimize the cost of feature generation.

Based on the high-fidelity equation (eqn. 3) developed here, we calculated and screened an extensive database of over half a million COFs for methane deliverable capacities. Based on performance evaluation of high-capacity COFs against the state-of-the-art adsorbents, we identified several record-setting COFs under different operating conditions. Among these, 3D-HNU5 (CURATED-COFs[104] database structure labels: 19400N3), a real COF reported by Guan et al.[105], is most notable since it has shown the potential of reaching DOE's volumetric and gravimetric methane capacity target simultaneously. In principle, one can apply this method of equation development for energy storage capacities of other solid-state adsorbents.

AUTHOR INFORMATION

Corresponding Author

alauddin@umich.edu; tel: 734-763-8682

Supporting Information

Supplementary Tables 1-29; Supplementary Figure 1; Metrics for ML accuracy.

ACKNOWLEDGMENT. The author acknowledges funding support from the US Department of Energy, Office of Energy Efficiency and Renewable Energy (Grant no. DE-EE0008814). This research was supported in part through computational resources and services provided by Advanced Research Computing – Technology Services (ARC-TS), a division of Information and Technology Services (ITS) at the University of Michigan, Ann Arbor.

REFERENCES

(1) Jiang, J.; Furukawa, H.; Zhang, Y.-B.; Yaghi, O. M. High Methane Storage Working Capacity in Metal–Organic Frameworks with Acrylate Links. *J. Am. Chem. Soc* **2016**, *138*, 10244–10251. https://doi.org/10.1021/jacs.6b05261.
(2) Gándara, F.; Furukawa, H.; Lee, S.; Yaghi, O. M. High Methane Storage Capacity in Aluminum Metal–Organic Frameworks. *J. Am. Chem. Soc.* **2014**, *136* (14), 5271–5274. https://doi.org/10.1021/ja501606h.
(3) Celzard, A.; Fierro, V. Preparing a Suitable Material Designed for Methane Storage: A Comprehensive Report. *Energy & Fuels* **2005**, *19*, 573–583. https://doi.org/10.1021/EF040045B.
(4) Zhang, H.; Deria, P.; Farha, O. K.; Hupp, J. T.; Snurr, R. Q. A Thermodynamic Tank Model for Studying the Effect of Higher Hydrocarbons on Natural Gas Storage in Metal–Organic Frameworks. *Energy Environ. Sci.* **2015**, *8* (5), 1501–1510. https://doi.org/10.1039/C5EE00808E.
(5) Werpy, M. R.; Santini, D.; Burnham, A.; Mintz, M. *Natural Gas Vehicles: Status, Barriers, and Opportunities Energy Systems Division*; 2010.
(6) Schroeder, A.; Fung, M. *2015 Natural Gas Vehicle Research Roadmap*; 2016.
(7) Sinor, J. E. *Comparison of CNG and LNG Technologies for Transportation Applications*; 1992.
(8) Saha, D.; Grappe, H. A.; Chakraborty, A.; Orkoulas, G. Postextraction Separation, On-Board Storage, and Catalytic Conversion of Methane in Natural Gas: A Review. *Chem. Rev.* **2016**, *116*, 11436–11499. https://doi.org/10.1021/acs.chemrev.5b00745.
(9) Sagara, T.; Klassen, J.; Ganz, E. Computational Study of Hydrogen Binding by Metal-Organic Framework-5. *J. Chem. Phys.* **2004**, *121* (24), 12543. https://doi.org/10.1063/1.1809608.
(10) Ding, S.-Y.; Wang, W. Covalent Organic Frameworks (COFs): From Design to Applications. *Chem. Soc. Rev.* **2013**, *42* (2), 548–568. https://doi.org/10.1039/C2CS35072F.
(11) Diercks, C. S.; Yaghi, O. M. The Atom, the Molecule, and the Covalent Organic Framework. *Science (80-.).* **2017**, *355* (6328), eaal1585. https://doi.org/10.1126/science.aal1585.
(12) Waller, P. J.; Gándara, F.; Yaghi, O. M. Chemistry of Covalent Organic Frameworks. *Acc. Chem. Res.* **2015**, *48* (12), 3053–3063. https://doi.org/10.1021/acs.accounts.5b00369.
(13) 2012, D. M. program. DOE MOVE program 2012,https://arpa-e.energy.gov/?q=arpa-e-programs/move. https://arpa-e.energy.gov/?q=arpa-e-programs/move.
(14) Simon, C. M.; Kim, J.; Gomez-Gualdron, D. A.; Camp, J. S.; Chung, Y. G.; Martin, R. L.; Mercado, R.; Deem, M. W.; Gunter, D.; Haranczyk, M.; et al. The Materials Genome in Action: Identifying the Performance Limits for Methane Storage. *Energy Environ. Sci.* **2015**, *8* (4), 1190–1199. https://doi.org/10.1039/C4EE03515A.
(15) Tong, M.; Lan, Y.; Yang, Q.; Zhong, C. High-Throughput Computational Screening and Design of Nanoporous Materials for Methane Storage and Carbon Dioxide Capture. *Green Energy Environ.* **2018**, *3* (2), 107–119. https://doi.org/10.1016/J.GEE.2017.09.004.
(16) Simon, C. M.; Kim, J.; Gomez-Gualdron, D. A.; Camp, J. S.; Chung, Y. G.; Martin, R. L.; Mercado, R.; Deem, M. W.; Gunter, D.; Haranczyk, M.; et al. The Materials Genome in Action: Identifying the Performance Limits for Methane Storage. *Energy Environ. Sci.* **2015**, *8* (4), 1190–1199. https://doi.org/10.1039/C4EE03515A.
(17) Boyd, P. G.; Lee, Y.; Smit, B. Computational Development of the Nanoporous Materials Genome. *Nat. Rev. Mater.* **2017**, *2* (8), 17037. https://doi.org/10.1038/natrevmats.2017.37.
(18) Materials Genome Initiative for Global Competitiveness. [Washington, D.C.]. Executive Office of the President, National Science and Technology Council. 2011.
(19) Tong, M.; Lan, Y.; Yang, Q.; Zhong, C. Exploring the Structure-Property Relationships of Covalent Organic Frameworks for Noble Gas Separations. *Chem. Eng. Sci.* **2017**, *168*, 456–464. https://doi.org/10.1016/J.CES.2017.05.004.
(20) Tong, M.; Lan, Y.; Qin, Z.; Zhong, C. Computation-Ready, Experimental Covalent Organic Framework for Methane Delivery: Screening and Material Design. *J. Phys. Chem. C 2018,* **2018**, *122*, 13009–13016.

https://doi.org/10.1021/acs.jpcc.8b04742.
(21) GitHub - core-cof/CoRE-COF-Database https://github.com/core-cof/CoRE-COF-Database (accessed Mar 28, 2020).
(22) Yan, T.; Lan, Y.; Tong, M.; Zhong, C. Screening and Design of Covalent Organic Framework Membranes for CO2/CH4 Separation. *ACS Sustain. Chem. Eng.* **2019**, *7*(1), 1220–1227. https://doi.org/10.1021/acssuschemeng.8b04858.
(23) Ongari, D.; Boyd, P. G.; Barthel, S.; Witman, M.; Haranczyk, M.; Smit, B. Accurate Characterization of the Pore Volume in Microporous Crystalline Materials. **2017**. https://doi.org/10.1021/acs.langmuir.7b01682.
(24) Côté, A. P.; Benin, A. I.; Ockwig, N. W.; O'Keeffe, M.; Matzger, A. J.; Yaghi, O. M. Porous, Crystalline, Covalent Organic Frameworks. *Science* **2005**, *310*(5751), 1166–1170. https://doi.org/10.1126/science.1120411.
(25) Mercado, R.; Fu, R.-S.; Yakutovich, A. V; Talirz, L.; Haranczyk, M.; Smit, B. In Silico Design of 2D and 3D Covalent Organic Frameworks for Methane Storage Applications. *Chem. Mater.* **2018**, *30*(15), 5069–5086. https://doi.org/10.1021/acs.chemmater.8b01425.
(26) Lan, Y.; Han, X.; Tong, M.; Huang, H.; Yang, Q.; Liu, D.; Zhao, X.; Zhong, C. Materials Genomics Methods for High-Throughput Construction of COFs and Targeted Synthesis. *Nat. Commun.* **2018**, *9*(1), 5274. https://doi.org/10.1038/s41467-018-07720-x.
(27) Mendoza-Cortes, J. L.; Pascal, T. A.; Goddard, W. A. Design of Covalent Organic Frameworks for Methane Storage. *J. Phys. Chem. A* **2011**, *115*(47), 13852–13857. https://doi.org/10.1021/jp209541e.
(28) Martin, R. L.; Simon, C. M.; Medasani, B.; Britt, D. K.; Smit, B.; Haranczyk, M. In Silico Design of Three-Dimensional Porous Covalent Organic Frameworks via Known Synthesis Routes and Commercially Available Species. *J. Phys. Chem. C* **2014**, *118*(41), 23790–23802. https://doi.org/10.1021/jp507152j.
(29) Mercado, R.; Fu, R.-S.; Yakutovich, A. V; Talirz, L.; Haranczyk, M.; Smit, B. In Silico Design of 2D and 3D Covalent Organic Frameworks for Methane Storage Applications. *Chem. Mater* **2018**, *30*, 38. https://doi.org/10.1021/acs.chemmater.8b01425.
(30) Reymond, J.-L. The Chemical Space Project. *Acc. Chem. Res.* **2015**, *48*(3), 722–730. https://doi.org/10.1021/ar500432k.
(31) Gropp, C.; Ma, T.; Hanikel, N.; Yaghi, O. M. Design of Higher Valency in Covalent Organic Frameworks. *Science (80-.).* **2020**, *370*(6515), eabd6406. https://doi.org/10.1126/science.abd6406.
(32) Sturluson, A.; Huynh, M. T.; Kaija, A. R.; Laird, C.; Yoon, S.; Hou, F.; Feng, Z.; Wilmer, C. E.; Colón, Y. J.; Chung, Y. G.; et al. The Role of Molecular Modelling and Simulation in the Discovery and Deployment of Metal-Organic Frameworks for Gas Storage and Separation*. *Mol. Simul.* **2019**, *45*(14–15), 1082–1121. https://doi.org/10.1080/08927022.2019.1648809.
(33) Fernandez, M.; Woo, T. K.; Wilmer, C. E.; Snurr, R. Q. Large-Scale Quantitative Structure–Property Relationship (QSPR) Analysis of Methane Storage in Metal–Organic Frameworks. *J. Phys. Chem. C* **2013**, *117*(15), 7681–7689. https://doi.org/10.1021/jp4006422.
(34) Li, S.; Chung, Y. G.; Simon, C. M.; Snurr, R. Q. High-Throughput Computational Screening of Multivariate Metal – Organic Frameworks (MTV-MOFs) for CO 2 Capture. **2017**. https://doi.org/10.1021/acs.jpclett.7b02700.
(35) Thornton, A. W.; Simon, C. M.; Kim, J.; Kwon, O.; Deeg, K. S.; Konstas, K.; Pas, S. J.; Hill, M. R.; Winkler, D. A.; Haranczyk, M.; et al. Materials Genome in Action: Identifying the Performance Limits of Physical Hydrogen Storage. *Chem. Mater.* **2017**, *29*(7), 2844–2854. https://doi.org/10.1021/acs.chemmater.6b04933.
(36) Gomez-Gualdron, D. A.; Gutov, O. V.; Krungleviciute, V.; Borah, B.; Mondloch, J. E.; Hupp, J. T.; Yildirim, T.; Farha, O. K.; Snurr, R. Q. Computational Design of Metal–Organic Frameworks Based on Stable Zirconium Building Units for Storage and Delivery of Methane. *Chem. Mater.* **2014**, *26*(19), 5632–5639. https://doi.org/10.1021/cm502304e.
(37) Gómez-Gualdrón, D. A.; Colón, Y. J.; Zhang, X.; Wang, T. C.; Chen, Y.-S.; Hupp, J. T.; Yildirim, T.; Farha, O. K.; Zhang, J.; Snurr, R. Q. Evaluating Topologically Diverse Metal–Organic Frameworks for Cryo-Adsorbed Hydrogen Storage. *Energy Environ. Sci.* **2016**, *9*(10), 3279–3289. https://doi.org/10.1039/C6EE02104B.
(38) Wilmer, C. E.; Leaf, M.; Lee, C. Y.; Farha, O. K.; Hauser, B. G.; Hupp, J. T.; Snurr, R. Q. Large-Scale Screening of Hypothetical Metal–Organic Frameworks. *Nat. Chem.* **2011**, *4*(2), 83–89. https://doi.org/10.1038/nchem.1192.
(39) Bucior, B. J.; Bobbitt, N. S.; Islamoglu, T.; Goswami, S.; Gopalan, A.; Yildirim, T.; Farha, O. K.; Bagheri, N.; Snurr, R. Q. Energy-Based Descriptors to Rapidly Predict Hydrogen Storage in Metal-Organic Frameworks. *Mol. Syst. Des. Eng. 2018. DOI 10.1039/c8me00050f.* https://doi.org/10.1039/c8me00050f.
(40) Toyao, T.; Suzuki, K.; Kikuchi, S.; Takakusagi, S.; Shimizu, K.; Takigawa, I. Toward Effective Utilization of Methane: Machine Learning Prediction of Adsorption Energies on Metal Alloys. *J. Phys. Chem. C* **2018**, *122*(15), 8315–8326. https://doi.org/10.1021/acs.jpcc.7b12670.
(41) Fanourgakis, G. S.; Gkagkas, K.; Tylianakis, E.; Klontzas, E.; Froudakis, G. A Robust Machine Learning Algorithm for the Prediction of Methane Adsorption in Nanoporous Materials. *J. Phys. Chem. A* **2019**, acs.jpca.9b03290. https://doi.org/10.1021/acs.jpca.9b03290.
(42) Ohno, H.; Mukae, Y. Machine Learning Approach for Prediction and Search: Application to Methane Storage in a Metal–Organic Framework. *J. Phys. Chem. C* **2016**, *120*, 23963–23968. https://doi.org/10.1021/acs.jpcc.6b07618.
(43) Pardakhti, M.; Moharreri, E.; Wanik, D.; Suib, S. L.; Srivastava, R. Machine-Learning Using Combined Structural and Chemical Descriptors for Prediction of Methane Adsorption Performance of Metal Organic Frameworks (MOFs). 1–15.
(44) Fanourgakis, G. S.; Gkagkas, K.; Tylianakis, E.; Froudakis, G. E. A Universal Machine Learning Algorithm for Large-Scale Screening of Materials. *J. Am. Chem. Soc.* **2020**, *142*(8), 3814–3822. https://doi.org/10.1021/jacs.9b11084.
(45) Pardakhti, M.; Moharreri, E.; Wanik, D.; Suib, S. L.; Srivastava, R. Machine Learning Using Combined Structural and Chemical Descriptors for Prediction of Methane Adsorption Performance of Metal Organic Frameworks (MOFs). *ACS Comb. Sci.* **2017**, *19*(10), 640–645. https://doi.org/10.1021/acscombsci.7b00056.
(46) Pardakhti, M.; Nanda, P.; Srivastava, R. Impact of Chemical Features on Methane Adsorption by Porous Materials at Varying Pressures. *J. Phys. Chem. C* **2020**, *124*(8), 4534–4544. https://doi.org/10.1021/acs.jpcc.9b09319.
(47) Himanen, L.; Jäger, M. O. J.; Morooka, E. V.; Federici Canova, F.; Ranawat, Y. S.; Gao, D. Z.; Rinke, P.; Foster, A. S. DScribe: Library of Descriptors for Machine Learning in Materials Science. *Comput. Phys. Commun.* **2020**, *247*, 106949. https://doi.org/10.1016/j.cpc.2019.106949.
(48) Broom, D. P.; Webb, C. J.; Fanourgakis, G. S.; Froudakis, G. E.; Trikalitis, P. N.; Hirscher, M. Concepts for Improving Hydrogen Storage in Nanoporous Materials. *Int. J. Hydrogen Energy* **2019**, *44*(15), 7768–7779. https://doi.org/10.1016/J.IJHYDENE.2019.01.224.
(49) Sezginel, K. B.; Uzun, A.; Keskin, S. Multivariable Linear Models of Structural Parameters to Predict Methane Uptake in Metal–Organic Frameworks. *Chem. Eng. Sci.* **2015**, *124*, 125–134. https://doi.org/10.1016/J.CES.2014.10.034.
(50) Chung, Y. G.; Haldoupis, E.; Bucior, B. J.; Haranczyk, M.; Lee, S.; Zhang, H.; Vogiatzis, K. D.; Milisavljevic, M.; Ling, S.; Camp, J. S.; et al. Advances, Updates, and Analytics for the Computation-Ready, Experimental Metal–Organic Framework Database: CoRE MOF 2019. *J. Chem. Eng. Data* **2019**, *64*(12), 5985–5998. https://doi.org/10.1021/acs.jced.9b00835.
(51) Makarov, D. E.; Metiu, H. Fitting Potential-Energy Surfaces: A Search in the Function Space by Directed Genetic Programming. *J. Chem. Phys.* **1998**, *108*(2), 590–598. https://doi.org/10.1063/1.475421.
(52) Makarov, D. E.; Metiu, H. Using Genetic Programming to Solve the Schrödinger Equation. *J. Phys. Chem. A* **2000**, *104*(37), 8540–8545. https://doi.org/10.1021/jp000695q.
(53) Schmidt, M.; Lipson, H. Distilling Free-Form Natural Laws from Experimental Data. *Science (80-.).* **2009**, *324*(5923), 81–85. https://doi.org/10.1126/science.1165893.
(54) Commercial Eureqa Software Sold by DataRobot: Https://Www.Datarobot.Com/Nutonian/.
(55) Raymond, C.; Chen, Q.; Xue, B.; Zhang, M. Genetic Programming with Rademacher Complexity for Symbolic Regression. In *2019 IEEE Congress on Evolutionary Computation, CEC 2019 - Proceedings*; Institute of Electrical and Electronics Engineers Inc., 2019; pp 2657–2664. https://doi.org/10.1109/CEC.2019.8790341.
(56) Udrescu, S. M.; Tegmark, M. AI Feynman: A Physics-Inspired Method for Symbolic Regression. *Sci. Adv.* **2020**, *6*(16), eaay2631. https://doi.org/10.1126/sciadv.aay2631.
(57) Udrescu, S.-M.; Tegmark, M. Symbolic Pregression: Discovering Physical Laws from Distorted Video. **2020**.
(58) Ouyang, R.; Curtarolo, S.; Ahmetcik, E.; Scheffler, M.; Ghiringhelli, L. M. SISSO: A Compressed-Sensing Method for Identifying the Best Low-Dimensional Descriptor in an Immensity of Offered Candidates. *Phys. Rev. Mater.* **2018**, *2*, 83802. https://doi.org/10.1103/PhysRevMaterials.2.083802.
(59) Bartel, C. J.; Millican, S. L.; Deml, A. M.; Rumptz, J. R.; Tumas, W.; Weimer, A. W.; Lany, S.; Stevanović, V.; Musgrave, C. B.; Holder, A. M.

Physical Descriptor for the Gibbs Energy of Inorganic Crystalline Solids and Prediction of Temperature-Dependent Materials Chemistry.
(60) Cao, G.; Liu, H.; Ouyang, R.; Acosta, C. M.; Ghiringhelli, L. M.; Zhou, Z.; Scheffler, M.; Carbogno, C.; Zhang, Z. High-Throughput Descriptor for Predicting Potential Topological Insulators in the Tetradymite Family. **2018**.
(61) Archetti, F.; Lanzeni, S.; Messina, E.; Vanneschi, L. Genetic Programming for Computational Pharmacokinetics in Drug Discovery and Development. *Genet. Program. Evolvable Mach.* **2007**, *8* (4), 413–432. https://doi.org/10.1007/s10710-007-9040-z.
(62) Rosenwald, A.; Wright, G.; Chan, W. C.; Connors, J. M.; Campo, E.; Fisher, R. I.; Gascoyne, R. D.; Muller-Hermelink, H. K.; Smeland, E. B.; Giltnane, J. M.; et al. The Use of Molecular Profiling to Predict Survival after Chemotherapy for Diffuse Large-B-Cell Lymphoma. *N. Engl. J. Med.* **2002**, *346* (25), 1937–1947. https://doi.org/10.1056/NEJMoa012914.
(63) M. Lichman. 2013. UCI Machine Learning Repository. (2013). Http://Archive.Ics.Uci.Edu/Ml.
(64) Willems, T. F.; Rycroft, C. H.; Kazi, M.; Meza, J. C.; Haranczyk, M. Algorithms and Tools for High-Throughput Geometry-Based Analysis of Crystalline Porous Materials. *Microporous Mesoporous Mater.* **2012**, *149* (1), 134–141. https://doi.org/10.1016/j.micromeso.2011.08.020.
(65) Ouyang, R.; Curtarolo, S.; Ahmetcik, E.; Scheffler, M.; Ghiringhelli, L. M. *SISSO: A Compressed-Sensing Method for Systematically Identifying Efficient Physical Models of Materials Properties*; 2017.
(66) Mercado, R.; Fu, R.-S.; Yakutovich, A. V; Talirz, L.; Haranczyk, M.; Smit, B. *Supporting Information for In Silico Design of 2D and 3D Covalent Organic Frameworks for Methane Storage Applications*.
(67) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12* (Oct), 2825–2830.
(68) Breiman, L.; Friedman, J. H.; Olshen, R. A.; Stone, C. J. *Classification And Regression Trees*; Routledge, 2017. https://doi.org/10.1201/9781315139470.
(69) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45* (1), 5–32. https://doi.org/10.1023/A:1010933404324.
(70) Breiman, L. Bagging Predictors. *Mach. Learn.* **1996**, *24* (2), 123–140. https://doi.org/10.1023/A:1018054314350.
(71) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
(72) Matsumoto, M.; Nishimura, T. Mersenne Twister: A 623-Dimensionally Equidistributed Uniform Pseudo-Random Number Generator. *ACM Trans. Model. Comput. Simul.* **1998**, *8* (1), 3–30. https://doi.org/10.1145/272991.272995.
(73) Marsland, S. *Machine Learning*; Chapman and Hall/CRC, 2011. https://doi.org/10.1201/9781420067194.
(74) Oliphant, T. E. *Guide to NumPy*; 2006.
(75) van der Walt, S.; Colbert, S. C.; Varoquaux, G. The NumPy Array: A Structure for Efficient Numerical Computation. *Comput. Sci. Eng.* **2011**, *13* (2), 22–30. https://doi.org/10.1109/MCSE.2011.37.
(76) Martin, M. G.; Siepmann, J. I. Transferable Potentials for Phase Equilibria. 1. United-Atom Description of n-Alkanes. *J. Phys. Chem. B* **1998**, *102* (14), 2569–2577.
(77) Mayo, S. L.; Olafson, B. D.; Goddard III, W. A. DREIDING: A Generic Force Field for Molecular Simulations. *J. Phys. Chem* **1990**, *94* (91), 8897–8909.
(78) Lorentz, H. A. Ueber Die Anwendung Des Satzes Vom Virial in Der Kinetischen Theorie Der Gase. *Ann. Phys.* **1881**, *248* (1), 127–136. https://doi.org/10.1002/andp.18812480110.
(79) D. Berthelot, "Sur le mélange des gaz," Comptes Rendus Hebdomadaires des Séances de l'Académie des Sciences, vol. 126, pp. 1703–1855, 1898. - Open Access Library http://www.oalib.com/references/13808846 (accessed Apr 26, 2020).
(80) Sadus, R. J. *Molecular Simulation of Fluids: Theory, Algorithms, and Object-Orientation.*; Elsevier: Amsterdam, 1999.
(81) Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*; Oxford University Press: New York, NY, 1989.
(82) Frenkel, D.; Smit, B. *Understanding Molecular Simulation : From Algorithms to Applications*, 2nd ed.; Academic Press, Inc.: Orlando, FL, 2001.
(83) Dubbeldam, D.; Torres-Knoop, A.; Walton, K. S. Molecular Simulation On the Inner Workings of Monte Carlo Codes On the Inner Workings of Monte Carlo Codes. *Mol. Simul.* **2013**, *39*, 14–15. https://doi.org/10.1080/08927022.2013.819102.
(84) Sandler, S. I. *An Introduction to Applied Statistical Thermodynamics*; John Wiley & Son Ltd: New York, NY, 2010.
(85) Hill, T. L. *An Introduction to Statistical Thermodynamics*; Dover Publications, 1986.
(86) Dubbeldam, D.; Calero, S.; Ellis, D. E.; Snurr, R. Q. RASPA: Molecular Simulation Software for Adsorption and Diffusion in Flexible Nanoporous Materials. *Mol. Simul.* **2016**, *42* (2), 81–101. https://doi.org/10.1080/08927022.2015.1010082.
(87) Sandler, S. I. *Chemical, Biochemical, and Engineering Thermodynamics*, 4th ed.; Wiley, 2006.
(88) Ouyang, R.; Curtarolo, S.; Ahmetcik, E.; Scheffler, M.; Ghiringhelli, L. M. *SISSO: A Compressed-Sensing Method for Systematically Identifying Efficient Physical Models of Materials Properties*; 2017.
(89) Wolpert, D. H. The Lack of a Priori Distinctions between Learning Algorithms. *Neural Comput.* **1996**, *8* (7), 1341–1390. https://doi.org/10.1162/neco.1996.8.7.1341.
(90) Wolpert, D. H.; Macready, W. G. No Free Lunch Theorems for Optimization. *IEEE Trans. Evol. Comput.* **1997**, *1* (1).
(91) Anscombe, F. J. Graphs in Statistical Analysis. *Am. Stat.* **1973**, *27* (1), 17–21.
(92) Urdan, T. C. *Statistics in Plain English* , 4th ed.; Routledge, 2017.
(93) Zwillinger, D.; Kokoska, S.; Raton, B.; New, L.; Washington, Y. *Standard Probability and Statistics Tables and Formulae CRC*; 2000.
(94) Oliphant, T. E. Python for Scientific Computing. *Comput. Sci. Eng.* **2007**, *9* (3), 10–20. https://doi.org/10.1109/MCSE.2007.58.
(95) Millman, K. J.; Aivazis, M. Python for Scientists and Engineers. *Comput. Sci. Eng.* **2011**, *13* (2), 9–12. https://doi.org/10.1109/MCSE.2011.36.
(96) Mann, H. B.; Whitney, D. R. On a Test of Whether One of Two Random Variables Is Stochastically Larger than the Other. *Ann. Math. Stat.* **1947**, *18* (1), 50–60. https://doi.org/10.1214/aoms/1177730491.
(97) Pearce, E. M.; Howell, B. A. (Bobby A.; Pethrick, R. A. (Richard A.; Zaikov, G. E. (Gennadiĭ E. *Physical Chemistry Research for Engineering and Applied Sciences. Volume 2, Polymeric Materials and Processing*.
(98) Meurer, A.; Smith, C. P.; Paprocki, M.; Čertík, O.; Kirpichev, S. B.; Rocklin, M.; Kumar, Am. T.; Ivanov, S.; Moore, J. K.; Singh, S.; et al. SymPy: Symbolic Computing in Python. *PeerJ Comput. Sci.* **2017**, *2017* (1), e103. https://doi.org/10.7717/peerj-cs.103.
(99) Jiang, J.; Furukawa, H.; Zhang, Y.-B.; Yaghi, O. M. High Methane Storage Working Capacity in Metal–Organic Frameworks with Acrylate Links. *J. Am. Chem. Soc.* **2016**, *138* (32), 10244–10251. https://doi.org/10.1021/jacs.6b05261.
(100) Mason, J. A.; Veenstra, M.; Long, J. R. Evaluating Metal–Organic Frameworks for Natural Gas Storage. *Chem. Sci.* **2014**, *5* (1), 32–51. https://doi.org/10.1039/C3SC52633J.
(101) Hulvey, Z.; Vlaisavljevich, B.; Mason, J. A.; Tsivion, E.; Dougherty, T. P.; Bloch, E. D.; Head-Gordon, M.; Smit, B.; Long, J. R.; Brown, C. M. Critical Factors Driving the High Volumetric Uptake of Methane in Cu 3 (Btc) 2. *J. Am. Chem. Soc* **2015**, *3*. https://doi.org/10.1021/jacs.5b06657.
(102) Peng, Y.; Krungleviciute, V.; Eryazici, I.; Hupp, J. T.; Farha, O. K.; Yildirim, T. Methane Storage in Metal-Organic Frameworks: Current Records, Surprise Findings, and Challenges. *J. Am. Chem. Soc.* **2013**, *135* (32), 11887–11894. https://doi.org/10.1021/ja4045289.
(103) NIST/ARPA-E Database of Novel and Emerging Adsorbent Materials. Https://Adsorption.Nist.Gov/Isodb/Index.Php#home.
(104) Ongari, D.; Yakutovich, A. V.; Talirz, L.; Smit, B. Building a Consistent and Reproducible Database for Adsorption Evaluation in Covalent-Organic Frameworks. *ACS Cent. Sci.* **2019**, *5* (10), 1663–1675. https://doi.org/10.1021/acscentsci.9b00619.
(105) Guan, P.; Qiu, J.; Zhao, Y.; Wang, H.; Li, Z.; Shi, Y.; Wang, J. A Novel Crystalline Azine-Linked Three-Dimensional Covalent Organic Framework for CO2 Capture and Conversion. *Chem. Commun.* **2019**, *55* (83), 12459–12462. https://doi.org/10.1039/c9cc05710b.