



This is the author's peer reviewed, accepted manuscript. However, the online version of record will be different from this version once it has been copyedited and typeset.
PLEASE CITE THIS ARTICLE AS DOI:10.1063/5.0091405

**Gaussian Approximation of Dispersion Potentials for Efficient
Featurization and Machine-Learning Predictions of Metal-Organic Frameworks**

Sihoon Choi¹, David S. Sholl^{1,2}, and Andrew J. Medford^{1*}

¹ School of Chemical & Biomolecular Engineering, Georgia Institute of Technology, Atlanta,
Georgia 30332, USA

² Oak Ridge National Laboratory, Oak Ridge, Tennessee 37830, USA

* Author to whom correspondence should be addressed: ajm@gatech.edu

Abstract

Energy-related descriptors in machine learning (ML) are a promising strategy to predict adsorption properties of metal-organic frameworks (MOFs) in the low-pressure regime. Interactions between hosts and guests in these systems are typically expressed as a sum of dispersion and electrostatic potentials. The energy landscape of dispersion potentials plays a crucial role in defining Henry's constants for simple probe molecules in MOFs. To incorporate more information about this energy landscape, we introduce the Gaussian-approximated Lennard-Jones (GALJ) potential, which fits pairwise Lennard-Jones potentials with multiple Gaussians by varying their heights and widths. The GALJ approach is capable of replicating information that can be obtained from the original LJ potentials and enables efficient development of Gaussian integral (GI) descriptors that account for spatial correlations in the dispersion energy environment. GI descriptors would be computationally inconvenient to compute using the usual direct evaluation of the dispersion potential energy surface. We show that these new GI descriptors lead to improvement in ML predictions of Henry's constants for a diverse set of adsorbates in MOFs compared to previous approaches to this task. .

I. Introduction

Computational screening of adsorption properties of metal-organic frameworks (MOFs) has been the subject of numerous studies, either based on molecular simulations or by coupling molecular simulation data with machine learning (ML) methods¹⁻⁵. A typical goal of these studies is to predict the pressure-dependent isotherm for a given adsorbate in a well-defined set of MOFs or similar materials. Under high pressure conditions, where a MOF's pores are close to saturation, various types of descriptors based on structural features of the pores have been found to be useful⁶⁻⁹. Obtaining the same prediction accuracy at dilute loadings has remained challenging¹⁰⁻¹². This observation is important because Henry's constants for adsorption in the dilute limit play a pivotal role in determining adsorption selectivity and in applying mixing theories for multi-component adsorption like Ideal Adsorbed Solution Theory¹³⁻¹⁶.

Recently, energy-property descriptors, as opposed to structure-property descriptors, have emerged as a promising approach to accurately describing the subtleties of adsorbate-adsorbent interactions that make predicting adsorption in the dilute limit challenging. With these approaches the potential energy distribution of adsorbates, rather than (or in addition to) the atomic structure of the material, is used as the basis for descriptor generation^{13,17,18}. For example, Bucior *et al.* succeeded at explaining H₂ adsorption in MOFs using 1-dimensional histograms of an energy landscape of H₂ and a simple linear regression model¹⁷.

In molecular simulation of adsorption in MOFs and related materials, hosts and guests interact with each other via Van der Waals and electrostatic potentials¹⁹. For simplicity, we refer to the contribution of Van der Waals interactions as dispersion interactions since long-range interactions are of more interest in this study, even though a more precise terminology might distinguish between long-range dispersion terms and short-range repulsive terms. Dispersion

interactions are often calculated using pairwise Lennard-Jones (LJ) potentials²⁰. A common approach to defining dispersion potentials between unlike pairs of atoms is to use the Lorentz-Berthelot mixing rules²¹, an empirical approach that combines LJ parameters of two atoms. The 1-dimensional energy histograms mentioned above comprise information about the fully 3-dimensional potential energy surface defined by a point probe species, which is in turn made up of a summation of many two-body interatomic dispersion potentials for each probe location. The studies mentioned above focused on making predictions about the adsorption of one adsorbate species in a collection of adsorbents. The enormous number of distinct molecular species that exists means that an important generalization of this approach is to aim at making adsorption calculations for larger sets of adsorbates^{15,16,22–24}. In recent studies, potential energy information from a given probe species has been used for this purpose. Yu *et al.*¹³ used a methane (CH₄) probe in MOF pores as a descriptor in a ML-based model to predict the Henry's constants of arbitrary molecules in MOFs. Li *et al.*¹⁸ used a methyl group (CH₃) probe to predict adsorption loadings of ethane and propane molecules in MOFs. As might be expected, these approaches typically had less accuracy when applied to complex molecules that differed significantly from the probe used in the ML model. This suggests that further development in energy-property relationships would be useful to establish approaches for translating energy information from small probes to larger and more complex molecules.

To expand the use of dispersion potentials within ML descriptors of adsorption in porous materials, we describe below an efficient scheme for representing two-body dispersion potentials based on atom-centered Gaussian functions. We show applications of this concept specifically for Lennard-Jones potentials, but the concept is readily generalizable to other dispersion potential functions. The Gaussian-approximated Lennard-Jones (GALJ) potential mimics

interactions between hosts and a single adsorption probe with a linear combination of simple Gaussian functions centered on each atom. The accuracy of the GALJ approximation is evaluated by computing Henry's constants of methane in a large group of MOFs using the GALJ potential and comparing them to the true interatomic potential, where both are integrated with Monte Carlo integration, revealing numerically equivalent results at a reduced computational cost. More importantly, we show that the GALJ representation enables efficient calculation of novel Gaussian integral (GI) features that capture the potential energy of the methane probe at varying length scales. We use the GALJ potential energy surface and the GI features as inputs to a machine learning model for predicting Henry's constants of molecular adsorption in MOFs, and show that computational efficiency and accuracy are improved when compared to a state-of-the-art model.

II. Gaussian-Approximated Lennard-Jones Potentials

We consider the situation where the total dispersion interaction between a spherical probe molecule and the atoms of an adsorbent is the sum of two-body interactions between the probe and atoms of the host. We fixed our guest molecule of interest to methane and, for each element in the periodic table of which LJ parameters are accessible from the UFF^{25,26}, we obtained a single LJ potential that expresses dispersion interactions between its single atom and a methane united atom with the Lorentz-Berthelot combining rule:

$$U_{AB} = 4\epsilon_{AB} \left[\left(\frac{\sigma_{AB}}{r} \right)^{12} - \left(\frac{\sigma_{AB}}{r} \right)^6 \right] \quad (1)$$

where A and B denote two particles in a given system, r is interatomic distance, $\sigma_{AB} = (\sigma_A + \sigma_B)/2$, and $\epsilon_{AB} = \sqrt{\epsilon_A \epsilon_B}$.

Eight Gaussians were non-linearly optimized to approximate the Eq. (1) for each atom pair, with all eight Gaussians centered at $r = 0$. This approach allows the total LJ potential, U , at any position in space to be expressed as

$$U \approx \sum_j^{\text{atoms}} \sum_{k=1}^8 G_{j,k} \quad (2)$$

$$G_{j,k} = A_{j,k} \times \exp(-\gamma_{j,k} \times r^2) \quad (3)$$

where A and $\gamma = 1/2\sigma^2$ are heights and widths of Gaussian functions from the optimization step described below, respectively.

The set of Gaussians we have adopted is not necessarily the best choice if our aim is to accurately represent the LJ potential for all atomic separations, since the accuracy of the approximation varies with interatomic distance. In describing adsorption, it is more important to capture attractive interactions rather than repulsive ones since attractions contribute much more significantly to the Henry's constant than repulsion. We aimed to manage this issue by focusing the fitting within a window of $[\sqrt[6]{2}\sigma_{AB} - 1.5 \text{ \AA}, \sqrt[6]{2}\sigma_{AB} + 6.0 \text{ \AA}]$ to include the minimum of the curve as well as a long enough tail. The range of window was slightly modified for some elements to obtain better optimization results. Heights and widths of Gaussians were optimized nonlinearly via the Nelder-Mead algorithm²⁷ with the maximum iteration steps set to 50,000. See the supplementary materials for more details of this optimization.

An example of applying this approach to CH₄-N interactions is shown in Fig. 1. Fig. 1(b) shows the absolute errors of the GALJ of a N-CH₄ pair on a log scale. For interatomic distances shorter

than 2.3 Å, the deviation increases, becoming a poor approximation for small distances. For larger interatomic distances, the GALJ is at most 0.003 kJ/mol different from the true LJ potential. Crucially, this region includes all interatomic distances which would make significant contributions to the Henry's constant for an adsorbing probe molecule. Optimized heights and widths of 8 Gaussians for each element can be found in the supplementary materials.

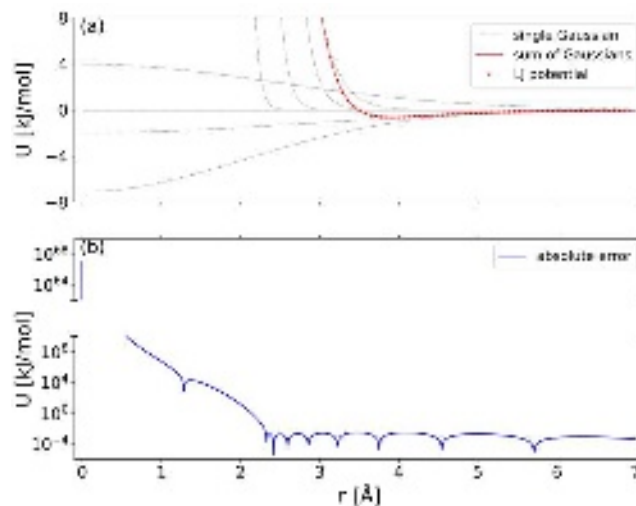


FIG. 1. (a) 8 Gaussians fitting the LJ potentials between a single N atom and a CH₄ united atom and (b) the absolute deviation between the sum of Gaussians and the true LJ potential.

(1) Precision of the Gaussian Approximation for the Dispersion PES

Given a potential energy surface for a point probe, the Henry's constant for this probe can be computed as^{28,29}

$$K_H = \frac{1}{N} \frac{1}{k_B T \rho} \sum_{i=1}^N \exp\left(-\frac{U_i}{k_B T}\right) \quad (4)$$

where N , k_B , T , ρ , and U_i are the number of Widom insertions, Boltzmann constant, temperature, density of a structure, and total probe-framework potential energy at each insertion. This

expression becomes exact in the limit of large N . Henry's constants of methane for a diverse set of 471 MOFs established by Tang *et al.*²² were calculated with 100,000 Widom insertions at 300 K by two different approaches that differ only in the way U_i is calculated (Eq. (1) vs. Eq. (2)). The fitting quality of the Gaussian approximation in Eq. (2) was assessed by mean absolute error (MAE) of log values of resulting Henry's constants. As we have described in our previous study, the MAE of $\log(K_H)$ can be expressed in a unit of kJ/mol^{13} (see S1.1).

A comparison between the methane Henry's constants computed from the full LJ-based PES (Eq. (1)) and the GALJ approximation (Eq. (2)) is shown in Fig. 2. The full LJ-based PES was computed by the Widom insertion method using RASPA¹⁹. The set of MOFs explored in these calculations have Henry's constants that vary over 20 orders of magnitude. The MAE of the GALJ calculations is 0.292 kJ/mol , which is an order of magnitude below the error corresponding to chemical accuracy. This data supports the assertion made above that the inaccuracies in the strongly repulsive core of individual LJ potentials due to the GALJ approximation do not lead to significant inaccuracies when computing adsorption-related properties.

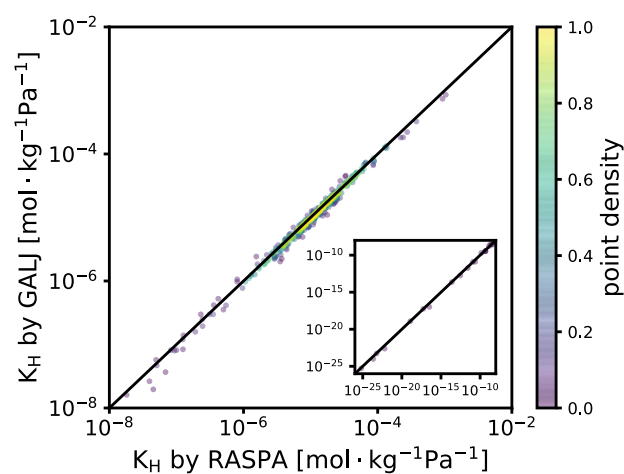


FIG. 2. Parity plot comparing Henry's constants for CH₄ in 471 MOFs computed by direct molecular simulation using RASPA and GALJ.

(2) Efficiency of GALJ in Defining Existing Descriptors

In this section, we show that using the GALJ approach leads to computational advantages in defining the descriptors used in current ML models for predicting Henry's constants in MOFs. Specifically, we consider the model defined in Section 3.2 in our previous work¹³ where Henry's constants prediction was made for 12,848 pairs of 471 MOFs and 30 molecules. Not all pairs of MOFs and molecules are present in the dataset as 1,282 pairs were dropped because they were identified to have extremely low Henry's constants ($< 10^{-15}$ mol/kgPa) by the screening algorithm defined in our earlier work¹³. Details of the data can be found at https://github.com/medford-group/predict_K_H. Along with 36 molecular descriptors that discriminate between different adsorbates¹³, three classes of MOF descriptors were used in this existing approach: textural descriptors, APRDF descriptors⁶, and 1-D histograms of the potential energy surface of a methane point probe. Here, we revisited the model by recalculating the energy descriptors using GALJ instead of RASPA. A minor advantage of using GALJ over the more usual evaluation of the potential energy with RASPA in our earlier work is that we can reduce the amount of time for generating the energy histograms (see Fig. S1). For each structure, we generated 3-D grids of the unit cell with a grid spacing of 0.2 Å. We note that the selection of grid points within a unit cell may affect the ML performance. Our initial test identified that mean absolute error may vary by 3%, and the maximum absolute error may vary by 10% at most. These deviations are within the error bars provided by training on different random training sets and may also be reduced by using a smaller grid spacing¹³. At each grid

point, dispersion potential energies of a methane probe were computed using a cutoff radius of 10 Å. In the prior study, the energy distributions of these 3-D grids were expressed as histograms of bins ranging from -26 kJ/mol to 0 kJ/mol with a bin width of 1 kJ/mol and two additional bins to count energy grid points below -26 kJ/mol and above 0 kJ/mol. In this work we also utilize percentiles of the distribution instead of histograms to more effectively focus on the low-energy region in each grid. Knowing the strong impact of low energies, we decided to consider energies lower than the first quartile. We collected the energy values at the 0.01, 0.1, 0.25, 0.5, 0.75, 1.0, 2.5, 5.0, 7.5, 10, and 25% percentiles and used these as the fingerprint instead of the histogram. As in our earlier work, a gradient boosting regressor was trained to predict $\log(K_H)$. To have a fair comparison, the same model hyperparameters and ensemble modeling with the same data splits were used as in our earlier work¹³. Specifically, a 64/16/20 training/validation/test split was applied to each data split and we compared prediction accuracy on 10 different test sets. In Table I, we provide the average and standard deviation of MAE both in unitless form and kJ/mol, coefficient of determination (r^2), and Spearman's coefficient (S) over 10 different test sets. As expected, using the GALJ replacement both in histograms and percentiles maintained the model performance in terms of all error metrics provided in Table I. From now on, the GALJ-based approach refers to the model trained with the percentiles. We also provided the maximum absolute error (MaxAE) in kJ/mol which collects the maximum of the absolute deviations between the true and predicted values. The addition of MaxAE to our analysis is to show the robustness of our model to various molecules. The MaxAEs are large because each model makes imprecise predictions for some large molecules that differ significantly from methane. This is discussed further below. A key observation from Table 1 is that the results from the original model and the model based on GALJ are equivalent to within numerical accuracy.

TABLE I. Mean Absolute Error (MAE), Maximum Absolute Error (MaxAE), Coefficient of Determination (r^2), and Spearman's Coefficient (S) of Henry's Constant Prediction

descriptors (# of descriptors)	# of MOF descriptors	MAE [unitless]	MAE [kJ/mol]	MaxAE [kJ/mol]	r^2	S
textural (3), histograms (28), and APRDF (68) ⁸	99	0.494 ± 0.016	2.836 ± 0.093	164.71 ± 55.90	0.879 ± 0.016	0.966 ± 0.004
textural (3), GALJ histograms (28), and APRDF (68)	99	0.475 ± 0.015	2.728 ± 0.085	158.57 ± 53.80	0.887 ± 0.017	0.968 ± 0.003
textural (3), GALJ percentiles (11), and ARPDF (68)	82	0.476 ± 0.011	2.735 ± 0.064	146.24 ± 62.31	0.891 ± 0.018	0.968 ± 0.003
textural (3), GALJ (11), and $\mu_{0.25,0.50}$ (22)	36	0.441 ± 0.013	2.537 ± 0.073	143.92 ± 54.88	0.910 ± 0.016	0.971 ± 0.003
textural (3), GALJ (11), and $\mu_{0.50,1.00}$ (22)	36	0.437 ± 0.011	2.508 ± 0.065	131.23 ± 58.46	0.914 ± 0.018	0.971 ± 0.003
textural (3), GALJ (11), and $\mu_{1.00,2.00}$ (22)	36	0.448 ± 0.013	2.572 ± 0.076	118.58 ± 59.69	0.912 ± 0.018	0.969 ± 0.004

III. Gaussian Integral Descriptors

A key advantage of the Gaussian decomposition of the LJ potential introduced above is to efficiently generate new descriptors for the potential energy surface of adsorbed molecules that would otherwise be computationally complex. Here we consider a Gaussian Integral (GI) descriptor scheme in which the local dispersion potential energy landscape is described by integrating Gaussian “probes” of varying widths. The GI scheme is motivated by the work of Lei and Medford, who featurized the electronic structure of chemical compounds using a combination of Maxwell-Cartesian spherical harmonics and Gaussian functions³⁰. A key advantage of transforming the LJ potential into a sum of Gaussian functions is that it enables further mathematical manipulation with well-known analytical solutions.

As an example, we consider descriptors that convolute the point-wise dispersion potential energy with a Gaussian weighting function (\hat{G}) to capture information about the local structure of the PES. Given a full PES spatially resolved in 3-D space, these descriptors have the form

This is the author's peer reviewed, accepted manuscript. However, the online version of record will be different from this version once it has been copyedited and typeset.
PLEASE CITE THIS ARTICLE AS DOI:10.1063/5.0091405

$$\begin{aligned}\mu_{\sigma}(\vec{x}) &= \iiint \hat{G}(\vec{x}) U(\vec{r}) d\vec{r}^3 \\ &= \iiint \exp\left(-\frac{\|\vec{r}-\vec{x}\|^2}{2\sigma^2}\right) U(\vec{r}) d\vec{r}^3\end{aligned}\quad (5)$$

where \hat{G} is centered at \vec{x} , the dispersion energy $U(\vec{r})$ is defined at a position \vec{r} , and $\|\cdot\|$ denotes the Euclidean norm of a given vector.

The range of interest for nearby positions that are included in this point-wise descriptor is dictated by the standard deviation (σ) of the probe Gaussian. Computing this kind of quantity directly from the full potential energy surface by direct numerical integration is computationally expensive, limiting the value of these quantities as descriptors that are to be used in large numbers of structures. Due to the Gaussian product rule, however, the analytical solution of Eq. (5) is available without any numerical integration when expressing the dispersion potential using the GALJ, making evaluation of these quantities readily feasible. The GI descriptor $\mu(\vec{x})$ can be expressed as

$$\begin{aligned}\mu_{\sigma}(\vec{x}) &= \langle \hat{G}(\vec{x}), \sum_j^{\text{atoms}} \sum_{k=1}^8 G_{j,k}(\vec{r}) \rangle \\ &= \iiint_V \hat{G}(\vec{x}) \sum_j^{\text{atoms}} \sum_{k=1}^8 G_{j,k}(\vec{r}) dV \\ &= \sum_j^{\text{atoms}} \sum_{k=1}^8 \iiint_V \hat{G}(\vec{x}) G_{j,k}(\vec{r}) dV\end{aligned}\quad (6)$$

where $\langle \rangle$ denotes an inner product between two components, and V is the volume of the system. In this case, \vec{r} is a position of atoms of adsorbents and each G_k is centered at \vec{r} . σ is the width of \hat{G} and multiple sets of GI descriptors with multiple σ 's are abbreviated by listing those widths as subscripts. The integrals in the last term in Eq. (6) can be evaluated analytically. Eq. (6) is equivalent to the simplest form in the work of Lei and Medford in which case no spherical harmonics takes part in the calculation³⁰, and thus maintains rotational and translational invariance. The similarity between the approaches hints that more complex descriptors that use spherical harmonics to capture angular variation could be explored in the future.

(1) ML Model Enhancement with the New GI Descriptors

We examined whether the addition of GI descriptors could improve the performance of the ML model predicting Henry's constants in MOFs described above. Specifically, GI descriptors were added to the GALJ-based model to incorporate information about the energy environments surrounding each grid point. Since the units of the GI descriptors are not the same as the original energy histogram descriptors, the GI descriptors were converted to percentiles as in the previous section. This created 11 additional descriptors for each σ . We generated three sets of GI descriptors, each consisting of two σ 's, resulting in the new 22 GI descriptors in each case. Our choice of σ 's was such that the weighting Gaussians cover at least a C-C bond of length 1.5 Å³¹. To prevent improvement in model simply from having more descriptors, we replaced the APRDF descriptor set in the original model with GI descriptors so that the total number of descriptors in each model decreased to 36, roughly one-third of that of the original model. The number of MOF descriptors of each model is represented in Table I. The results of this new model are compared to the original model in the last three rows of Table I. In each metric listed

in the Table I, addition of the GI descriptors improved the model performance. We reiterate that use of these descriptors is straightforward using the GALJ formalism but would be computationally impractical with a traditional representation of the potential energy surface. It is useful to look in more detail at the value of including GI descriptors to make predictions about large molecules. For extended molecules, the energy histogram from a point probe can only partially account for the interactions of the molecule with the MOF framework, since the overall interaction for an extended molecule is made up from contributions of multiple interaction centers whose positions are highly correlated. We examined five molecules that have low MAEs when using the GALJ-based model (methane, ethane, ethene, acetonitrile, and propene) and another five molecules that have high MAEs (neopentane, methyl propyl ether, methyl *tert*-butyl ether, 1,5-heptadiene and 4,4-dimethyl-1-pentene). Results for these 10 molecules are shown in Fig. 3(a). For the former group of molecules, adding the GI descriptors make only a small difference to the model's performance in all cases, with an average reduction in MAE of 0.049 kJ/mol in case of the inclusion of $\mu_{0.50,1.00}$. For the larger molecules, however, the average reduction in MAE for the same model is 1.108 kJ/mol.

Given the excellent average error of ~ 2.5 kJ/mol, it becomes increasingly important to address the more difficult cases where the maximum error can exceed 150 kJ/mol and yield qualitatively incorrect conclusions. The results in Fig. 3(b) and Table 1 also confirm the effectiveness of GI descriptors for reducing the maximum error. According to Fig. 3(b), the models without GI descriptors tends to perform poorly for many branched molecules. However, the addition of GI descriptors effectively enhances the quality of the worst predictions. Specifically, considering the effect of $\mu_{0.50,1.00}$, the average MaxAE of the extended molecules decreased by 16.2 kJ/mol, while the average MaxAE was reduced by only 0.7 kJ/mol for the smaller five molecules.

The GI scheme can be considered as scalarizing the energy environment within a sphere with a soft limit centered at each grid point, as detailed in Eq. (6). This suggests that GI descriptors should be more effective for molecules that are more spherical rather than for linear compounds. Our results are in agreement with this expectation, with the biggest improvement in both MAE and MaxAE coming from highly branched molecules such as neopentane and 4,4-dimethyl-1-pentene. On the other hand, 1,5-heptadiene, the most linear molecule among depicted on Fig. 3, shows a rise of 0.260 kJ/mol and 4.5 kJ/mol in MAE and MaxAE, respectively, after incorporating GI descriptors (see Table S2 in the supplementary materials for more details). These results support the hypothesis that the GI descriptors improve the ability of the model to describe the behavior of spherical molecules inside the pore. At the same time, the results point to the need for development of additional shape-based descriptors to fully account for the dispersion potential for arbitrary molecules.

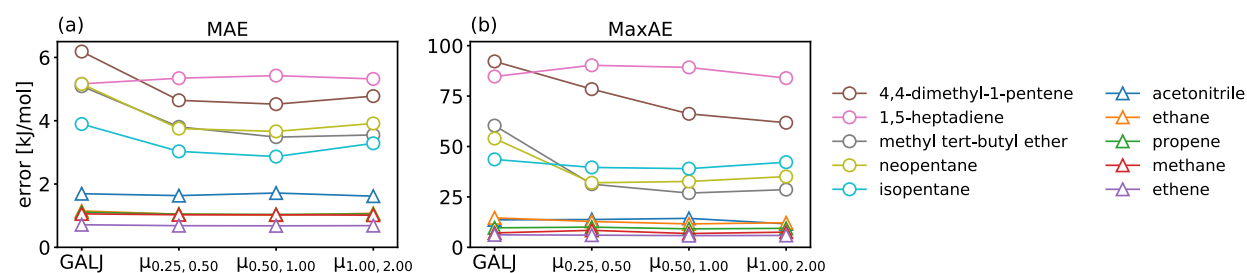


FIG. 3. Comparison in (a) mean absolute errors and (b) maximum absolute errors for 5 molecules of high MAEs (marker: ○) and 5 molecules of low MAEs (marker: △) between the GALJ-based model and the GI-augmented models.

Conclusion

We demonstrated an efficient featurization scheme which led to effective Henry's constant computation for arbitrary molecules in MOFs. Our results showed that GALJ can successfully

reproduce the conventional calculation of dispersion energies based on pair-wise potentials with faster computation but comparable accuracy. More importantly, the GALJ approach provides opportunities to incorporate information associated with local convolution of the dispersion potential via novel GI descriptors. GI descriptors were shown to lead to slightly lower average errors and significantly lower maximum errors in Henry's constants prediction while using significantly fewer descriptors than previous work, especially for extended molecules whose shape is relatively spherical. There is clearly still room for improvement in our current approach; the current model still tends to have relatively high MAEs and MaxAEs for long and linear models even after the inclusion of GI descriptors. This observation suggests that additional descriptors of energy environments are needed to better predict the adsorption of molecules with non-isotropic shapes. One possible strategy is to consider angular variations of environments using spherical harmonics^{32,33}. Exploring angular variations with spherical harmonics can be approached systematically since varying orders of spherical harmonics would lead to ample descriptions of the energy landscape. Moreover, the GALJ approach is not limited to LJ potentials and can be applied to any other two-body pairwise dispersion potentials such as the Morse potential or the Buckingham potential or can be directly parameterized based on numerically evaluated 2-body interactions. These promising advantages of the GALJ approach suggests that it will serve as the foundation for future machine-learning models that efficiently and accurately predict the adsorption properties of MOFs and other sorbent materials across multitudes of possible adsorbate molecules.

Supplementary Materials

This is the author's peer reviewed, accepted manuscript. However, the online version of record will be different from this version once it has been copyedited and typeset.
PLEASE CITE THIS ARTICLE AS DOI:10.1063/5.0091405

This is the author's peer reviewed, accepted manuscript. However, the online version of record will be different from this version once it has been copyedited and typeset.
PLEASE CITE THIS ARTICLE AS DOI:10.1063/5.0091405

See the supplementary materials for the parameters of optimized Gaussians, ML predictions of Henry's constants from models in Table I, and the data used for Figs. 2, 3, and 4.

Acknowledgements

This work was supported by the Department of Energy, Office of Science, Basic Energy Sciences, under Award #DE-SC0020306.

Data Availability

The data that support the findings of this study are openly available in GitHub at <https://github.com/medford-group/GALJ-MOF>.

References

- [1] Boyd, P. G., Chidambaram, A., García-Díez, E., Ireland, C. P., Daff, T. D., Bounds, R., Gładysiak, A., Schouwink, P., Moosavi, S. M., Maroto-Valer, M. M., Reimer, J. A., Navarro, J. A. R., Woo, T. K., Garcia, S., Stylianou, K. C. and Smit, B., “Data-driven design of metal–organic frameworks for wet flue gas CO₂ capture,” *Nature* **576**(7786), 253–256 (2019).
- [2] Chung, Y. G., Gómez-Gualdrón, D. A., Li, P., Leperi, K. T., Deria, P., Zhang, H., Vermeulen, N. A., Stoddart, J. F., You, F., Hupp, J. T., Farha, O. K. and Snurr, R. Q., “In silico discovery of metal-organic frameworks for precombustion CO₂ capture using a genetic algorithm,” *Science Advances* **2**(10), e1600909.
- [3] Lee, S., Kim, B., Cho, H., Lee, H., Lee, S. Y., Cho, E. S. and Kim, J., “Computational Screening of Trillions of Metal–Organic Frameworks for High-Performance Methane Storage,” *ACS Appl. Mater. Interfaces* **13**(20), 23647–23654 (2021).
- [4] Simon, C. M., Mercado, R., Schnell, S. K., Smit, B. and Haranczyk, M., “What Are the Best Materials To Separate a Xenon/Krypton Mixture?,” *Chem. Mater.* **27**(12), 4459–4475 (2015).
- [5] Colón, Y. J. and Snurr, R. Q., “High-throughput computational screening of metal–organic frameworks,” *Chem. Soc. Rev.* **43**(16), 5735–5749 (2014).
- [6] Fernandez, M., Trefiak, N. R. and Woo, T. K., “Atomic Property Weighted Radial Distribution Functions Descriptors of Metal–Organic Frameworks for the Prediction of Gas Uptake Capacity,” *J. Phys. Chem. C* **117**(27), 14095–14105 (2013).
- [7] Anderson, R., Biong, A. and Gómez-Gualdrón, D. A., “Adsorption Isotherm Predictions for Multiple Molecules in MOFs Using the Same Deep Learning Model,” *J. Chem. Theory Comput.* **16**(2), 1271–1283 (2020).
- [8] Fernandez, M. and Barnard, A. S., “Geometrical Properties Can Predict CO₂ and N₂ Adsorption Performance of Metal–Organic Frameworks (MOFs) at Low Pressure,” *ACS Comb. Sci.* **18**(5), 243–252 (2016).
- [9] Fernandez, M., Woo, T. K., Wilmer, C. E. and Snurr, R. Q., “Large-Scale Quantitative Structure–Property Relationship (QSPR) Analysis of Methane Storage in Metal–Organic Frameworks,” *J. Phys. Chem. C* **117**(15), 7681–7689 (2013).
- [10] Fanourgakis, G. S., Gkagkas, K., Tylianakis, E., Klontzas, E. and Froudakis, G., “A Robust Machine Learning Algorithm for the Prediction of Methane Adsorption in Nanoporous Materials,” *J. Phys. Chem. A* **123**(28), 6080–6087 (2019).
- [11] Pardakhti, M., Nanda, P. and Srivastava, R., “Impact of Chemical Features on Methane Adsorption by Porous Materials at Varying Pressures,” *J. Phys. Chem. C* **124**(8), 4534–4544 (2020).
- [12] Krishnapriyan, A. S., Haranczyk, M. and Morozov, D., “Topological Descriptors Help Predict Guest Adsorption in Nanoporous Materials,” *J. Phys. Chem. C* **124**(17), 9360–9368 (2020).
- [13] Yu, X., Choi, S., Tang, D., Medford, A. J. and Sholl, D. S., “Efficient Models for Predicting Temperature-Dependent Henry’s Constants and Adsorption Selectivities for Diverse Collections of Molecules in Metal–Organic Frameworks,” *J. Phys. Chem. C* **125**(32), 18046–18057 (2021).
- [14] Walton, K. S. and Sholl, D. S., “Predicting multicomponent adsorption: 50 years of the ideal adsorbed solution theory,” *AIChE Journal* **61**(9), 2757–2762 (2015).

- [15] Gharagheizi, F., Tang, D. and Sholl, D. S., “Selecting Adsorbents to Separate Diverse Near-Azeotropic Chemicals,” *J. Phys. Chem. C* **124**(6), 3664–3670 (2020).
- [16] Tang, D., Gharagheizi, F. and Sholl, D. S., “Adsorption-Based Separation of Near-Azeotropic Mixtures—A Challenging Example for High-Throughput Development of Adsorbents,” *J. Phys. Chem. B* **125**(3), 926–936 (2021).
- [17] Bucior, B. J., Bobbitt, N. S., Islamoglu, T., Goswami, S., Gopalan, A., Yildirim, T., Farha, O. K., Bagheri, N. and Snurr, R. Q., “Energy-based descriptors to rapidly predict hydrogen storage in metal–organic frameworks,” *Mol. Syst. Des. Eng.* **4**(1), 162–174 (2019).
- [18] Li, Z., Bucior, B. J., Chen, H., Haranczyk, M., Siepmann, J. I. and Snurr, R. Q., “Machine learning using host/guest energy histograms to predict adsorption in metal–organic frameworks: Application to short alkanes and Xe/Kr mixtures,” *J. Chem. Phys.* **155**(1), 014701 (2021).
- [19] Dubbeldam, D., Calero, S., Ellis, D. E. and Snurr, R. Q., “RASPA: molecular simulation software for adsorption and diffusion in flexible nanoporous materials,” *Molecular Simulation* **42**(2), 81–101 (2016).
- [20] Lennard-Jones, J. E., “Cohesion,” *Proc. Phys. Soc.* **43**(5), 461–482 (1931).
- [21] Lorentz, H. A., “Ueber die Anwendung des Satzes vom Virial in der kinetischen Theorie der Gase,” *Annalen der Physik* **248**(1), 127–136 (1881).
- [22] Tang, D., Wu, Y., Verploegh, R. J. and Sholl, D. S., “Efficiently Exploring Adsorption Space to Identify Privileged Adsorbents for Chemical Separations of a Diverse Set of Molecules,” *ChemSusChem* **11**(9), 1567–1575 (2018).
- [23] Tang, D., Kupgan, G., Colina, C. M. and Sholl, D. S., “Rapid Prediction of Adsorption Isotherms of a Diverse Range of Molecules in Hyper-Cross-Linked Polymers,” *J. Phys. Chem. C* **123**(29), 17884–17893 (2019).
- [24] Anstine, D. M., Tang, D., Sholl, D. S. and Colina, C. M., “Adsorption space for microporous polymers with diverse adsorbate species,” *npj Comput Mater* **7**(1), 1–9 (2021).
- [25] Rappe, A. K., Casewit, C. J., Colwell, K. S., Goddard, W. A. and Skiff, W. M., “UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations,” *J. Am. Chem. Soc.* **114**(25), 10024–10035 (1992).
- [26] Qiao, Z., Zhang, K. and Jiang, J., “In silico screening of 4764 computation-ready, experimental metal–organic frameworks for CO₂ separation,” *J. Mater. Chem. A* **4**(6), 2105–2114 (2016).
- [27] Nelder, J. A. and Mead, R., “A Simplex Method for Function Minimization,” *The Computer Journal* **7**(4), 308–313 (1965).
- [28] Maginn, E. J., Bell, A. T. and Theodorou, D. N., “Sorption Thermodynamics, Siting, and Conformation of Long n-Alkanes in Silicalite As Predicted by Configurational-Bias Monte Carlo Integration,” *J. Phys. Chem.* **99**(7), 2057–2079 (1995).
- [29] Sarkisov, L., “Toward Rational Design of Metal–Organic Frameworks for Sensing Applications: Efficient Calculation of Adsorption Characteristics in Zero Loading Regime,” *J. Phys. Chem. C* **116**(4), 3025–3033 (2012).
- [30] Lei, X. and Medford, A. J., “A Universal Framework for Featurization of Atomistic Systems,” arXiv:2102.02390 [physics] (2021).
- [31] Haynes, W. M., Lide, D. R. and Bruno, T. J., eds., [CRC handbook of chemistry and physics: a ready-reference book of chemical and physical data, 2016th–2017th, 97th Edition ed.], CRC Press, Boca Raton, Florida (2016).

This is the author's peer reviewed, accepted manuscript. However, the online version of record will be different from this version once it has been copyedited and typeset.
PLEASE CITE THIS ARTICLE AS DOI:10.1063/5.0091405

- [32] Applequist, J., "Maxwell-Cartesian spherical harmonics in multipole potentials and atomic orbitals," *Theoretical Chemistry Accounts: Theory, Computation, and Modeling (Theoretica Chimica Acta)* **107**(2), 103–115 (2002).
- [33] Lei, X. and Medford, A. J., "Design and analysis of machine learning exchange-correlation functionals via rotationally invariant convolutional descriptors," *Phys. Rev. Materials* **3**(6), 063801 (2019).

