Prediction of Diffusion Coefficient Through Machine Learning Based on Transition State Theory Descriptors

Emmanuel Ren, François-Xavier Coudert

Prediction of Diffusion Coefficient Through Machine Learning Based on Transition State Theory Descriptors

Emmanuel Ren[†,‡] and François-Xavier Coudert[*,‡]

†*CEA, DES, ISEC, DMRC, Univ. Montpellier, Marcoule, 30207 Bagnols-sur-Cèze, France*

‡*Chimie ParisTech, PSL University, CNRS, Institut de Recherche de Chimie Paris, 75005 Paris, France*

E-mail: fx.coudert@chimieparistech.psl.eu

Abstract

Nanoporous materials serve as very effective media for storing or separating small molecules. To design the best materials for a given application based on adsorption, one usually assesses the equilibrium performance by using key thermodynamic quantities such as Henry constants or adsorption loading values. To go beyond standard methodologies, we probe here the transport effects occurring in the material by studying the self-diffusion coefficients of xenon inside the nanopores of framework materials. We find good correlations between the diffusion coefficients and the pore aperture size, as well as other geometrical and energetic descriptors. We used extensive molecular dynamics simulations to calculate the diffusion coefficient of xenon in 4 873 MOFs from the CoRE-MOF 2019 database, the first large-scale database of transport properties published at this scale. Based on this data, we present a tool to quickly evaluate the diffusion energy barrier that proved to be very correlated to the diffusion rate. This descriptor, alongside other geometrical characterizations, were then used to build a machine learning model that can predict the xenon diffusion coefficients in MOFs. The final trained model is quite accurate and shows a root mean square error (RMSE) on the $\log_{10}$ of the diffusion coefficient equal to 0.25.

1 Introduction

Separation processes are omnipresent in the industry in many areas, including energy, environment and health, to separate chemical mixtures into pure components through operations such as distillation, molecular sieving, etc.[1,2] These processes are account for a significant fraction of the world's energy consumption, and therefore the design of more energy efficient separation methods could help lower global energy use, carbon dioxide emissions and pollution.[3] While distillation is the chemical engineering process most commonly associated with chemical separation and purification, other options exist such as crystallization, adsorption and membranes. Microporous materials, such as porous amorphous polymers[4] or nanoporous crystalline materials,[5] are among the materials frequently used for both adsorption and the design of membranes.

In molecular separation processes based on nanoporous materials, microscopic transport properties are key to the kinetics of the adsorption process at the macroscopic scale. Two distinct use cases for nanoporous materials in separation processes exist: adsorption-based separation, which is primarily a thermodynamic process, and nanoporous separation membranes, which rely on both kinetic and thermodynamic properties. Depending on the targeted application, diffusion is either the main performance metric (for membranes) or a secondary parameter (for adsorption) that is often overlooked. In processes based on molecular sieving, for instance, gases are passed through a membrane material that selectively blocks certain atoms or molecules on the basis of their size, while allowing other particles to diffuse freely.[6] In the steady state, the performance of the separation is, in part, related to the ratio of diffusion coefficients for the species involved. On the other hand, the thermodynamic selectivity is the primary performance metric in adsorption-based separation processes commonly performed at the industrial scale, such as pressure and/or temperature swing adsorption (PSA, TSA or PTSA).[7,8] However, even in those cases, it is worth considering that the kinetic performance can enhance the overall industrial process.[9] For instance, in breakthrough experiments used to characterize the comparative adsorption performances of a gas mixture

— and somewhat akin, at the lab scale, of a pressure swing adsorption — the shape of the breakthrough curve can be explained by diffusion processes.

Despite the importance of guest transport properties, most high-throughput computational studies based on atomistic simulations of nanoporous materials so far have focused heavily on the thermodynamics of adsorption,[10–12] studying adsorption parameters such as adsorption enthalpy, adsorption entropy, Henry constant, uptake isotherms at various temperatures, loading at specific values of pressure, working capacity, and other such thermodynamic quantities.[13–15] Here, we want to explore this frequently overlooked aspect through a high-throughput screening approach, and we therefore focus on a single quantity: the guest diffusion coefficient in the low loading regime (or low pressure limit). While transport phenomena in nanoporous media are very complex, we want to start addressing the issue with a manageable quantity, and therefore have chosen the low-loading diffusion coefficient as a first target.

The diffusion coefficient of a guest molecule inside nanopores is relatively straightforward to calculation from molecular dynamics simulations.[16] However, this approach has a very high computational cost, explaining in part why it has not been routinely adopted in high-throughput studies — although some examples exist in the literature: Altintas et al. studied the diffusion of H_2 and CH_4 in 4240 MOFs;[17] Zhou and Wu calculated minimum energy paths for the diffusion of polyatomic molecules;[18] Bukowski and Snurr characterized the diffusion of a chemical warfare agent simulant in a diverse set of 776 Zr-based MOFs.[19] More recently, alternative approaches have been proposed and developed to compute the diffusion coefficient of species directly from the characteristics of the potential energy field of the diffusing particles.[20,21] These approaches are very efficient, because they do not require the explicit molecular simulation of the dynamics of the diffusion itself, but instead are built on physics-informed approximations linking the underlying potential energy surface to the transport properties.

In this article, we implemented and tested a third approach to this problem, based on the

generation of a large database of self-diffusion coefficients and the application of statistical learning. We first performed explicit molecular dynamics simulations on a subset of the CoREMOF 2019 database in order to compute diffusion coefficients for a specific guest molecule — we have chosen xenon to illustrate our algorithm. We then introduced a grid-based algorithm to calculate the diffusion energy barrier values for the guest migration inside the nanoporous material. We then trained a machine learning model on our database, using as features this energy barrier alongside key geometrical descriptors (like the pore limiting diameter). We found that the final model is sufficiently accurate for the purpose of a fast estimation of the self-diffusion coefficient, showing an RMSE on the $\log_{10}$ of the diffusion coefficient equal to 0.25. We then discuss the perspectives opened by this new methodology.

2 Methods

The several computational methods that have been proposed to calculation diffusion coefficient values in nanoporous materials can be broken down into two main categories: methods based on molecular dynamics (MD) simulations and those relying on the application of transition state theory (TST).[22] The first approach is probably the most "natural" way, leveraging physically meaningful MD trajectories of adsorbates inside the nanoporous material, but it is computationally intensive. The second approach is much faster, relying only on the potential energy surface instead of computing long trajectories, but represents an important approximation. In this work, we use statistical learning to predict MD-based diffusion coefficients using approximate values coming from a TST-based method, augmented with other descriptors of the nanoporous geometry.

2.1 Screening of transport properties

Using systematic MD simulation, we calculated xenon self-diffusion coefficient values at infinite dilution (no guest–guest interactions) for 6 525 non-disordered materials from the CoRE-

MOF (Computation-Ready Experimental MOFs) 2019 database,[23,24] chosen to be the most thermodynamically selective for Xe/Kr separation based on a previous screening study.[25] From this set, we removed a small number of structures (291) that had several, inequivalent channels — since we can only probe one channel at a time in an infinite dilution MD simulation, that would cause inaccurate sampling in materials with several channels.

For each material, MD simulations were performed using the RASPA2 software[16,26] on the calculation machines (Intel Xeon Platinum 8168 cores at 2.7 GHz). Simulations were performed in the (N, V, T) ensemble with a Nosé–Hoover thermostat at $T = 298$ K. The simulations were set up with a maximum of 500 million MD steps, but we also set a CPU time limit of 60 hours. Within this time constraint, 3 899 structures completed all 500 million steps.

We then used the MSD, corresponding to the average of the squared displacement of the xenon atom $\langle r(t)^2 \rangle$ in the nanoporous material at time t, to determine the diffusion coefficients through the Einstein equation:

$$\langle r(t)^2 \rangle = 6 D_{\text{diff}}\, t \tag{1}$$

The RASPA2 software uses a multiple-window algorithm developed by Dubbeldam et al.[27] to probe different timescales of the MSD data when running MD simulations. We used this software for our MD calculations, with the default value of 25 for sampling rate ("SampleMSDEvery" parameter), 500 million time steps and a relatively large value of 5 fs for each step, in order to obtain sufficiently long dynamics — which we confirmed was acceptable in the absence of dynamics for light elements in the system. For larger structures, which did not complete the full MD trajectory, we set up a criterion for convergence of the mean squared displacement (MSD) calculation: we included trajectories that had a good linear fit, with a correlation coefficient $R^2 > 0.9$.

The final methodological choice in our exploration was: what characteristic time range do

we use to fit the MSD profiles and obtain diffusion coefficients? Based on manual examination of the MSD data on several materials, it is clear that there was no one-size-fits-all answer, but we typically found that most materials exhibited a clear diffusion regime in one of two time windows: 2–47 ns and 50–950 ns. we decided to avoid adding physical insight or intuition into the process, and in a data-based approach, chose for each material the time window that corresponded to the fit with the highest determination coefficient (ranging from 0 to 1) — as highlighted in the case of structure `KAXQIL_clean` in Figure S9. After this fitting step, structures with a determination coefficient R^2 below 0.9 were removed, leaving 5 125 structures reported in the following and used for drawing structure–diffusivity relationships, correlation analyses and prediction model development.

2.2 Energy barrier

A few years ago, Mace et al. developed an algorithm to calculate self-diffusion coefficients from the potential energy surface of an adsorbate molecule, based on the transition state theory.[20] In their work, they used a clustering algorithm to identify the adsorption sites, transition states, and connecting tunnels within the material for a specific adsorbate, which are then leveraged to run a lattice kinetic Monte Carlo simulation from which the mean square displacements are deduced. We were inspired by this approach, but are more interested in the present work in the use of the energy barrier itself: instead of using it to determine transition probabilities, we have developed a simplified version of the algorithm that only focuses on the energy at which the diffusion tunnels have percolated (i.e., become all connected).

Recently, we developed the GrAED algorithm that efficiently computes energy grids in nanoporous materials: we divide the unit cell into small voxels, whose size is typically between 0.1 and 0.2 Å. For each voxel, we compute the host–guest interaction energy that would have a guest molecule placed in the center of that voxel. While this concept is not novel, we recently made strong improvements in its computational efficiency, by using the symmetry of the host framework and excluding from calculations the space occupied by

framework atoms.[28] This grid is a discrete representation of the potential energy surface (PES) of a guest molecular inside the framework. From this PES surface representation, the energy minima are determined. Then, in order to determine the values of energy barrier for diffusion, we then needed to develop an algorithm that detects all-connected clusters within the energy grid. A breadth-first search algorithm is employed to label different connected components within a given channel between $E_{\min}$ and $E_{\min} + i\delta E$ (at the i^{th} iteration). By monitoring changes in the number of connected components between two energy values, the code automatically detects the energy E_{TS} at which components reconnect and form a channel (allowing diffusion from one boundary to another). The activation energy E_{a} is then calculated as the difference between the calculated transition state energy E_{TS} and the minimal energy $E_{\min}$ within the channel: $E_{\text{a}} = E_{\text{TS}} - E_{\min}$.

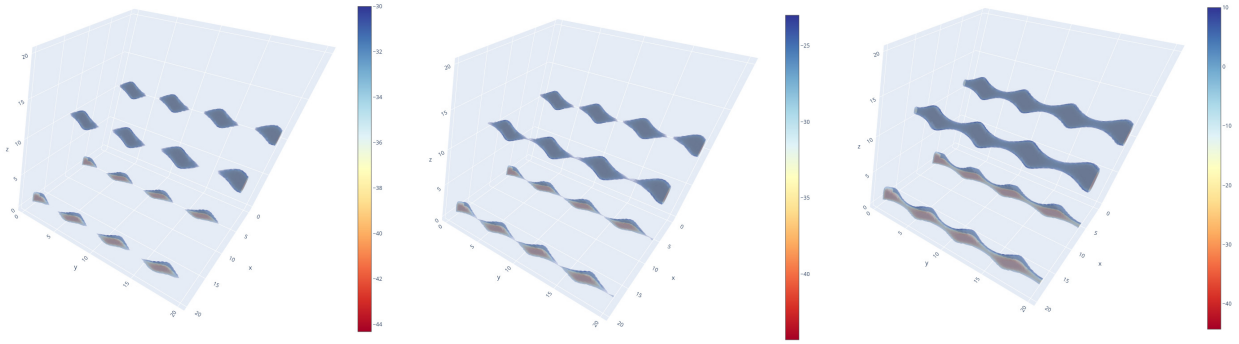
To illustrate the approach, here we show the case of `KAXQIL`,[29] where the barrier detection was performed using an energy step δE of $0.3\,\text{kJ}\,\text{mol}^{-1}$. A single symmetrically unique type of channel was identified in `KAXQIL`, with a minimal energy of -44.3 $\text{kJ}\,\text{mol}^{-1}$ — the various channels shown in Figure 1c are all symmetrically equivalent. The code detected a single merge that resulted in a fully connected component within the channel. This merging occurred at an energy of -25.7 $\text{kJ}\,\text{mol}^{-1}$ (as depicted in Figure 1b), indicating that the estimated activation energy is 18.6 $\text{kJ}\,\text{mol}^{-1}$ with an uncertainty of $0.3\,\text{kJ}\,\text{mol}^{-1}$ (due to the energy step used).[1]

In the simplest case of one unique merge of a unidimensional channel, the method demonstrates strong performance, and it becomes possible to associate the activation energy with a diffusion rate k_{diff} using the Arrhenius equation:

$$k_{\text{diff}} = A \exp\left(-\frac{E_{\text{a}}}{k_{\text{B}}T}\right) \tag{2}$$

where A is a prefactor that depends on the temperature and system (adsorbate, adsorbent). This is a simplified version of the transition probability used in TST-based methods. In the

[1]Code available at: `https://github.com/coudertlab/FaEB`



(a) $E_{\max} = -30\ \mathrm{kJ\,mol^{-1}}$ (b) $E_{\max} = -23\ \mathrm{kJ\,mol^{-1}}$ (c) $E_{\max} = 10\ \mathrm{kJ\,mol^{-1}}$

Figure 1: 3D visualization of channels within `KAXQIL` using different energy thresholds $E_{\max}$. Depending on the maximum value of energy allowed, the channel is either composed of unconnected basins (a), or they are fully connected (b) and (c). This illustrates the principle of the energy barrier detection.

case of a unidimensional channel with a single possible transition, the diffusion coefficient is directly associated with the diffusion rate. The problem can be reduced to a unidimensional random walk with a given transition probability, and the diffusion coefficient is given by $D = k_{\mathrm{diff}} L^2/2$, where L is the distance between two basins. In this special case, there exists a direct relationship between the diffusion coefficient and the activation energy, such that $\log(D) \propto E_{\mathrm{a}}$. For more complex systems, the relationship may be more complex, hence the need for our statistical learning approach.

We calculated the xenon diffusion barrier energy for all 5 125 structures previously selected for screening of diffusion coefficient. An energy step of $\delta E = 0.1\ \mathrm{kJ\,mol^{-1}}$ was employed during the energy loop to determine the minimal energy barrier for each unique channel in the material. Then, to avoid any potential noise arising from the MD simulation initialization problem, materials with significantly different energy barrier values from one channel to another (standard deviation of energy barrier values higher than $1\ \mathrm{kJ\,mol^{-1}}$) were excluded, reducing the number of structures considered to 4 873.

2.3 Energetic and geometrical descriptions

The approach chosen in this work is to rely on a fast energy evaluation to boost the performance of the ML model, instead of simply adding a maximum information on the chemical, geometrical and adsorptive properties of the material as typically done in previous studies.[30] The total set of descriptors for our model, detailed in Table 1, is relatively restricted.

The activation energy E_a and the adsorption enthalpy $\Delta_{\text{ads}}H_0^{\text{Xe}}$ within a channel were both obtained by the energy barrier algorithm described above. The adsorption enthalpy does not cost much more computational resources to compute and can be seen as a cheap descriptor that comes with the more useful barrier descriptor E_a.

We also added information on the pore size distribution inside the nanoporous material, characterized by geometrical methods with the `Zeo++` code.[31,32] As is typical, we used the largest cavity diameter (LCD) and the pore limiting diameter (PLD), as well as their difference. The pore limiting diameter is found to be a very important feature for the description of diffusion, as shown by an in-depth correlation analysis of the PLD and energy barrier.

We then added general (and easily accessible) data on the structure such as the framework mass M_f and density ρ_f. These descriptors are useful as a primary description of the structures, but are rarely what makes a difference in the ML model. Other standard geometrical descriptors such as the surface area (SA) and the void fraction or porosity (VF) were also considered. Finally, information about the dimensionality of the nanoporous channels of the materials were also found to be beneficial for the accuracy of the model the diffusion process and are therefore were also included in the final list of descriptors.

All descriptors are described in detail in Table 1, and were used to train the supervised ML model presented further in this article. The model architecture chosen is an XGBoost[33] framework, similar to that previously used in our study of the thermodynamics of Xe/Kr selectivity in nanoporous materials,[28] and the hyperparameters of the XGBoost model were determined using a random search. The SHAP interpretation tools are applied to better understand the underlying reasons behind the performance of the final ML model.[34]

Table 1: Features used in the ML model for diffusion coefficient prediction.

Feature name	Symbol	Description
`barrier`	E_{a}	Energy barrier: difference between transition state energy E_{TS} and the minimal energy E_{a} within a channel (in $\mathrm{kJ\,mol^{-1}}$)
`adsorption_enthalpy`	$\Delta_{\mathrm{ads}}H_0^{\mathrm{Xe}}(\mathrm{channel})$	Xenon adsorption enthalpy within a channel calculated using the barrier algorithm (in $\mathrm{kJ\,mol^{-1}}$)
`framework mass`	M_f	Molar mass of the framework material considered (in $\mathrm{g\,mol^{-1}}$)
`framework density`	ρ_f	Mass density of the framework material considered (in $\mathrm{kg\,m^{-3}}$)
`ASA`	SA	Surface area accessible to a 1.2 Å radius probe (in $\mathrm{m^2\,cm^{-3}}$)
`PO_VF_2.0`	$VF = \frac{V_{\mathrm{pore}}}{V_{\mathrm{tot}}}$	Void fraction or the ratio of the pore volume occupied by a 2 Å radius probe over the total material volume
`D_f_vdw_uff298`	PLD or D_f	Pore limiting diameter of the largest free sphere diameter calculated using the UFF dependent definition (in Å)
`D_if_vdw_uff298`	LCD or D_{if}	The largest included free sphere diameter in a free diffusion path calculated using the UFF dependent definition (in Å)
`delta_LCD_PLD`	LCD−PLD	Difference between the LCD and PLD values (in Å)
`1D_chan`	$\mathbb{1}_{1\mathrm{D}}$	Binary feature: 1 if there is a unidimensional channel, 0 else
`2D_chan`	$\mathbb{1}_{2\mathrm{D}}$	Binary feature: 1 if there is a bidimensional channel, 0 else
`3D_chan`	$\mathbb{1}_{3\mathrm{D}}$	Binary feature: 1 if there is a tridimensional channel, 0 else

2.4 Interaction energy calculation details

This study focuses on xenon and its interaction with materials of the MOF family. The interatomic interactions were modeled using Lennard-Jones potentials. The MOF atoms are described using the UFF forcefield,[35] whereas the Lennard-Jones parameters of xenon, taken from Ref. 36, are $\varepsilon_{\mathrm{Xe}} = 221.0$ K and $\sigma_{\mathrm{Xe}} = 4.100$ Å. To determine cross interaction parameters between xenon and all MOF atoms, we used the Lorentz–Berthelot combination rules.[37] No Coulombic interactions are considered in this force field.

3 Results and Discussion

3.1 Analysis of the diffusion coefficient values

We first analyzed the values of diffusion coefficient computed for all 5 125 structures with converged MSD data and a satisfactory linear fit. Because of the large range of variation of the diffusion coefficient D, we focus in the following on its base-10 logarithm, $\log_{10}(D)$. In order to get a good physical understanding of the process at the microscopic scale, we studied the correlation between $\log_{10}(D)$ and the geometric descriptors of the materials. After carrying out a thorough analysis for all the 12 descriptors listed in Table 1, whose results are reported in the Supporting Information, we identified the two descriptors with the strongest correlation to the diffusion coefficient: the pore limiting diameter (PLD) and the energy barrier (or diffusion activation energy). We also show in Figure S1 the distribution of values of the PLD and energy barrier across the structure retained for the model training.

As shown in Figure 2a, the activation energy is correlated with the diffusion coefficient for xenon in MOF nanopores. A stronger correlation is observed for points with a PLD around 4.5 Å, while for PLD values exceeding 6 Å, the correlation appears to be weaker compared to smaller PLD values, as illustrated in Figure 2b. This correlation between the energy barrier and the diffusion coefficient is confirmed in a different visualization in Figure 3. The points

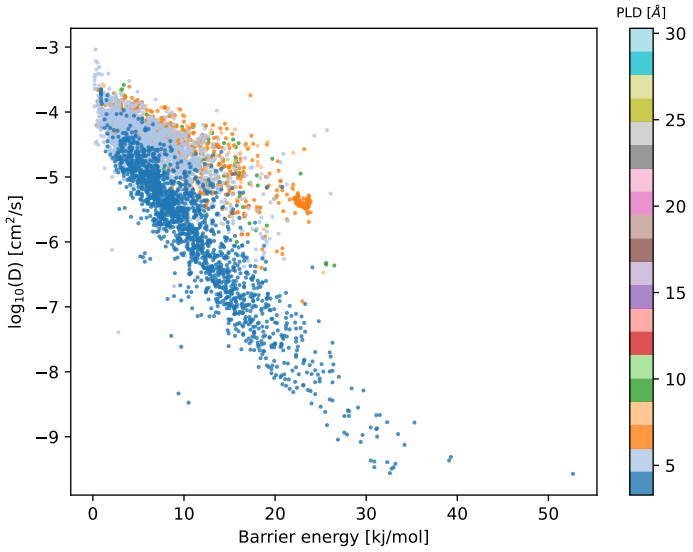
are labeled according to the energy barrier value in a given material, and the highest energy barrier points tend to be concentrated among lower diffusion coefficient values. However, a few points with very high energy barriers are also observed for diffusion coefficients that are quite low.

We see from the different representations that while both PLD and barrier energy play an important role in the diffusion coefficient, they are also complementary: the information obtained through energy calculations strengthens our comprehension of the correlation, compared to a purely geometric description. Indeed, PLD values cannot distinguish between structures over 6 Å in the "plateau" region of the diffusion/PLD graph. While a geometric analysis would interpret the different values of diffusion coefficient as statistical noise, Figure 3 reveals that higher values of barrier energies are typically associated with lower diffusion coefficients, thereby explaining the variations in diffusion coefficient across materials within this plateau based on the activation barrier values. Although the correlation is not perfect, this barrier descriptor provides better insights into this uncharted area of PLD values above 6 Å, which cannot be explained by simple geometric considerations. The barrier activation energy value sheds light on the chemical nature of the diffusion barrier that needs to be overcome.

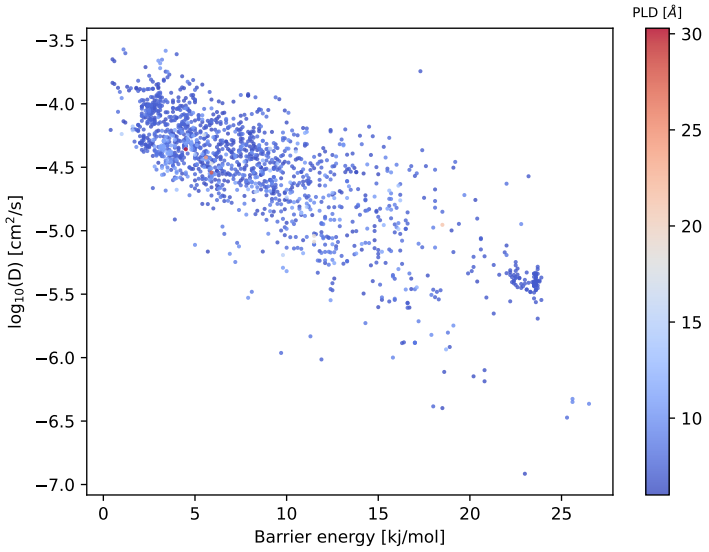
Based on this microscopic understanding, we then decided to combine standard geometrical descriptors with energy barrier values to train a machine learning model, following the approach we had previously used for high-throughput screening of thermodynamics of Xe/Kr separation[28]. We describe and analyse this ML model in the next section, and show how it can be used to evaluate the diffusion coefficient of xenon in nanoporous materials, offering a significantly faster alternative to MD simulations.

3.2 Machine learning model

The calculation of diffusion coefficients through explicit molecular simulation is an extremely time-consuming process, and is further complicated for high-throughput purposes by various



(a) All structures



(b) Subset of structures with PLD greater than 6 Å

Figure 2: Scatterplots of the $\log_{10}$ of the diffusion coefficient (in $\mathrm{cm^2\,s^{-1}}$) as a function of the diffusion activation energy E_a in $\mathrm{kJ\,mol^{-1}}$. Panel (a): for all structures; panel (b): for structures with a PLD larger than 6 Å. For all structures, the Pearson correlation coefficient is equal to $r = -0.77$, whereas for the restriction to structures with a PLD below 6 Å this correlation is stronger with a Pearson coefficient of $r = -0.85$. For structures with a PLD above 6 Å, this coefficient decreases to reach $r = -0.74$.

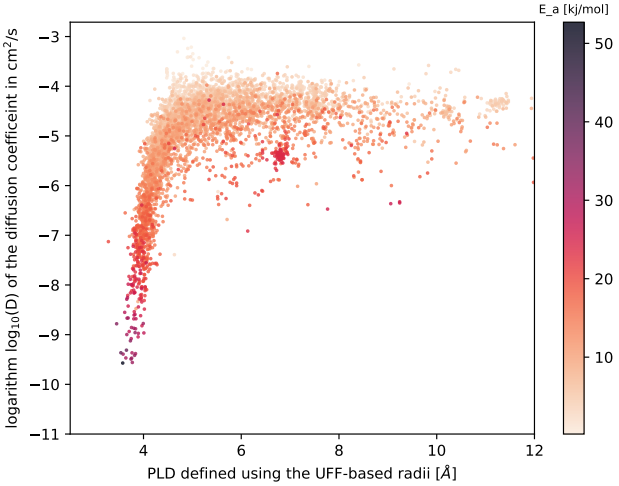


Figure 3: Scatterplot of the $\log_{10}$ of the diffusion coefficient (in $\mathrm{cm^2\,s^{-1}}$) as a function of the PLD values and labeled by the barrier activation energy. The higher barriers seem to correspond to lower diffusion coefficients, thus echoing the correlation observed in the previous Figure 2.

challenges in the fitting of the trajectory (i.e., the MSD data). Out of the 6,525 structures initially considered for this work, over one thousand were not completely evaluated through MD simulations, resulting in a success rate of approximately 75% for the direct simulation approach — and this is mainly due to either insufficient time for obtaining a usable MSD, or MSD data corresponding to non-Brownian regimes. It could be possible to use a larger time step than conventional for the MD simulation, in order to reduce the computational cost to attain the diffusion regime, but at the expense of accuracy: however, such simulations still require a typical simulation time of a couple of days per structure. On the other hand, the calculation of energy barriers with an energy step of $0.1\,\mathrm{kJ\,mol^{-1}}$ has an average time of 12 seconds, and the determination of geometric descriptors through the `Zeo++` software typically takes a few minutes at most. The MD method is therefore several orders of magnitude slower, even under highly optimistic hypotheses for MD simulation parameters.

However, the relationships between energy barrier, PLD, and diffusion coefficient remain unclear — the relatively weak correlation demonstrated in Figure 2a highlights the important limitations of the Arrhenius law as a general and direct relationship between diffusion

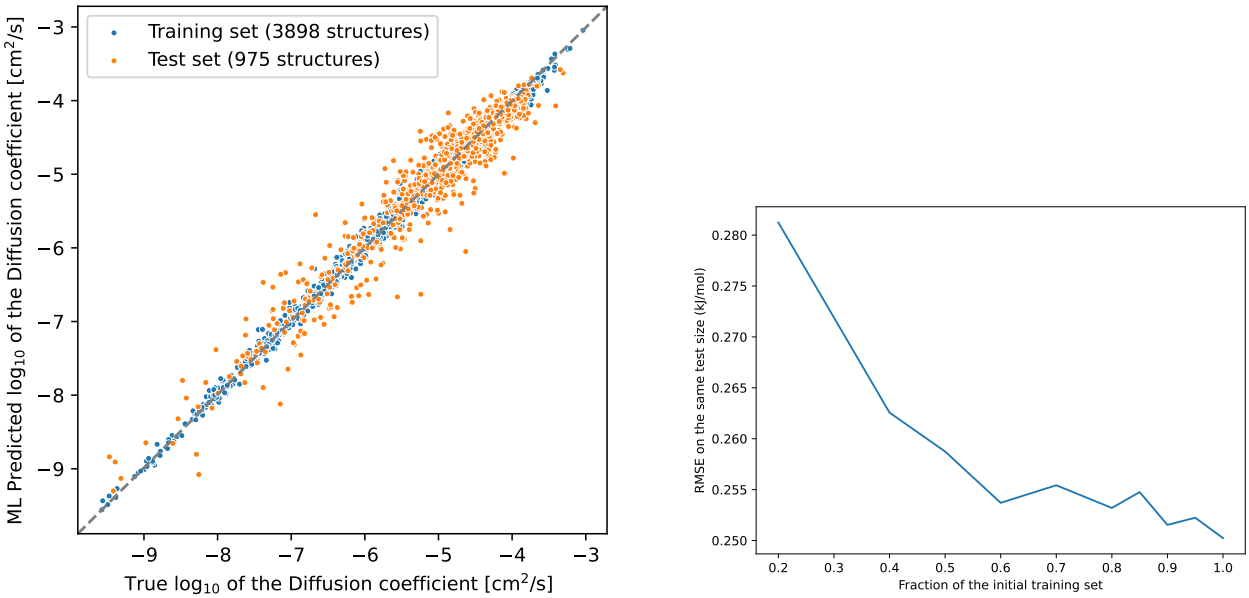
coefficient and energy barrier. The aim of the ML model is to build upon this observed correlation by introducing additional geometrical descriptors, and achieve accurate calculation of diffusion coefficients while significantly reducing the time required for predicting the diffusion coefficient of future selective materials. The ML model was trained using 80% of the 4 873 structures that survived all the different filters imposed. We employed a total of 12 descriptors listed in Table 1 to build the model. The selected hyperparameters for the XGBoost model are detailed in Table 2.

Table 2: Hyperparameters of the XGBoost model trained in this work.

parameter	value
`objective`	`reg:squarederror`
`n_estimators`	1500
`max_depth`	4
`colsample_bytree`	1
`colsample_bylevel`	0.75
`subsample`	0.75
`alpha`	0.6
`lambda`	1
`learning_rate`	0.04

With this parameterization, the ML model predicts the $\log_{10}$ of the diffusion coefficient (in units of $\mathrm{cm^2\,s^{-1}}$) with a root mean square error (RMSE) of 0.26 on the test set and a mean absolute error (MAE) of 0.18. This implies that the exponent α is known with an accuracy of approximately ± 0.2, when expressing the diffusion coefficient as $D = 10^{\alpha}$. For comparison, the previous ML model for thermodynamic selectivity predicts the $\log_{10}$ of selectivity with an error of about 0.07. It is important to note that the goal here is not to predict the exact values of the diffusion coefficient due to the inherent noise in the values generated by MD simulation (about 20% relative error for `KAXQIL`). Instead, the objective is to determine the order of magnitude of the diffusion coefficient. The proposed model achieves this objective effectively, as illustrated in Figure 4a, where the predicted diffusion coefficient aligns closely with the true values when represented on a log scale.

The training curve (Figure 4b) was examined to assess whether the model had sufficient



(a) Prediction result

(b) Training curve

Figure 4: (a) Comparison of the $\log_{10}$ of the diffusion coefficient predicted by an ML model and the true values. (b) Root mean squared errors on the same test set (20% of all data) as a function of the fraction of the training set used to train smaller models. The error decreases as the amount of data increases and seems to stabilize near 0.25.

training data or required additional data. As the amount of training data increased, the error converged to 0.25, indicating that no further data was necessary for training the model, given the descriptors available and the complexity of the model being fixed. However, it is conceivable to train a similar model using fewer data (50% instead of 80% of the total data could probably suffice to train a similar model). Furthermore, to prove that this good accuracy does not correspond to a fortuitous random train/test split, a cross-validation evaluation was performed on the whole dataset using a 5-fold cross-validation scheme. The average error (RMSE) on the five validation sets equals 0.26 with a standard deviation of 0.01, which is very similar to the one obtained with the specific train/test split we obtained here.

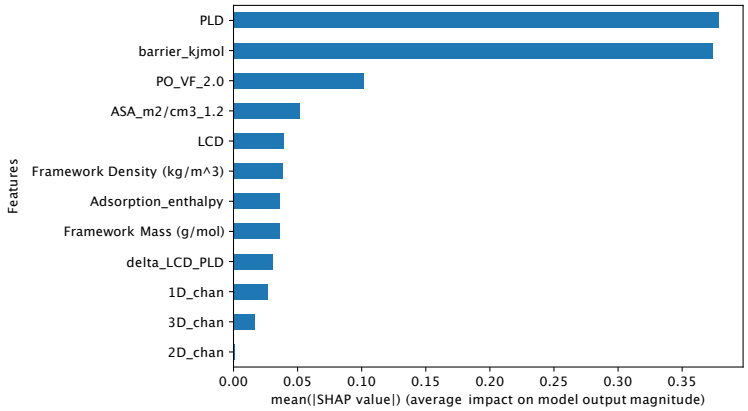


Figure 5: Feature importance determined using the average of the absolute Shapley values for each feature based on every training data. An influential feature would have a very high average absolute SHAP value. The features are detailed in Table 1.

3.3 Interpretation of the ML model

In order to get more insight into the impact of the materials' features on the diffusion coefficient, we then interpreted the ML model using the SHAP algorithms. The values of feature importance, determined using the average of the absolute Shapley values for each feature, are plotted in Figure 5. As expected, the most important features are found to be the PLD and the barrier activation energy, as demonstrated in the previous section. The void fraction also appeared to play a non-negligible role.

To unravel the relationship between these features and the target diffusion coefficient, partial dependence plots (PDPs) were examined for these features shown in Figure 6. The PLD has a contribution similar to that described in Figure 3. A linear contribution was observed when the PLD values were below 6 Å, followed by a constant contribution for PLD values above this threshold. The activation energy showed a negative correlation with the log of the diffusion coefficient, which explained the linear contribution observed in the dependence plot.

The analysis of the model also reveals less obvious contributions. Figure S5 indicates that no clear relationships can be inferred between surface areas or void fractions and the diffusion coefficient. These factors played a more secondary role, slightly adjusting the obtained values

with contributions of the order of 0.2 as shown on Figure 6. For instance, the model identifies a positive relation between the void fraction and the contribution to the diffusion coefficient, which aligns with the physical understanding that lower void fractions correspond to lower diffusion rates within the material, assuming other parameters are equal. Conversely, larger surface areas imply more interaction with the pore walls, which slows down the diffusion of particles. Regarding the LCD, the LCD – PLD difference, xenon adsorption enthalpy, framework's mass, and density, no clear contribution patterns were observed. This may be attributed to the fact that the previous features account for a substantial portion of the contribution due to the correlation between all these features.

Finally, we note that the final predicted values are only marginally influenced by the channel dimension, despite its association with a clear physical phenomenon. This could be in part explained by the fact that no clear statistical distinctions can be drawn between the materials with different channel dimensions (see Figure S3 and S4). For the model, the behavior of diffusion coefficients varies slightly depending on the dimensionality of the channel. Figure 6 illustrates that a 1D channel has a lower diffusion coefficient when all other features are similar. On the other hand, a 2D channel demonstrates a higher contribution, which is further confirmed by the partial dependence plots. A tridimensional channel exhibits an even higher diffusion coefficient. The model can distinguish between different material types based on their channel dimensionality.

4 Conclusions and perspectives

In this article, we have introduced different methods for evaluating transport properties of an adsorbate inside a nanoporous material, through the guest species' diffusion coefficient. The most accurate method is the direct molecular simulation through molecular dynamics: it requires considerable computational time and "meticulous attention" to achieve optimal accuracy. In particular, careful selection of parameters in MD simulations is essential to ob-

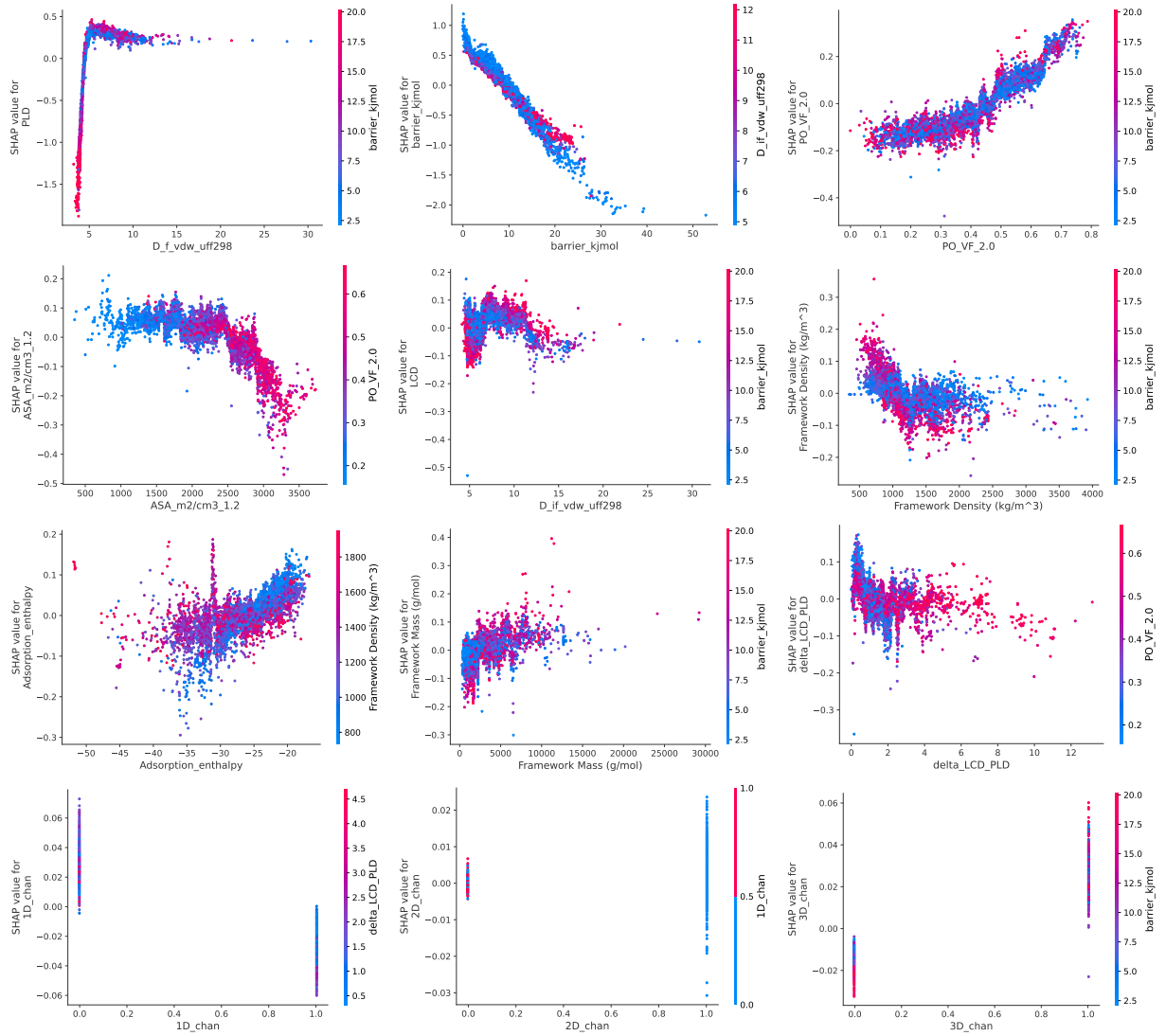


Figure 6: A SHAP dependence plot corresponds to the Shapley values as a function of the feature values for every structure. These SHAP plots show the contribution of the features to the prediction given by the ML model. Each Shapley value depends not only on the value of the feature itself but also on the other features. For this reason, the plots are labeled based on a relevant second feature. The partial dependence plots of every feature in the diffusion prediction model are presented here.

tain relevant mean square displacement data and allow the accurate calculation of a diffusion coefficient through fitting. We performed a high-throughput screening of diffusion coefficient values for xenon in 4 873 nanoporous materials from the CoreMOF 2019 database, allowing us to identify materials with notable thermodynamic and kinetic separation performance. This published database is a first of its kinds, as we are not aware of any published database of similar size for molecular diffusion through nanoporous frameworks.

We have used this data as a baseline for testing other methods, such as the calculation of diffusion activation energy, and the training of a machine learning (ML) model based on structural and energetic descriptors. The final ML model demonstrates promising performance, achieving a root mean squared error (RMSE) of only 0.25 on the base-10 logarithm of the diffusion coefficient. This indicates the ability to accurately assess the order of magnitude of diffusion properties. Such assessment can help identify potential diffusion limitations in promising materials and optimize this property to improve the performance of materials for adsorption-based separations. Furthermore, the techniques developed in this study, as well as future developments, can also be applied to membrane separation processes.

The results obtained here provide the foundation for possible future work. For instance, the effect of tortuosity on diffusion coefficient values and relevant definitions for tortuosity remain open questions. Unidimensional channels can be particularly examined, where the frequency and magnitude of changes in direction can be analyzed to quantify their occurrence[38]. Another challenge could consist in measuring different diffusion regimes, such as single-file diffusion characterized by a square root time relation in the mean square displacement (MSD)[39]. In this study, materials with MSD relations other than linear were excluded since only materials with high determination coefficients in the linear fit were considered.

To expand beyond conventional studies, the diffusion coefficient could be used to model breakthrough experiments, which is the closest a lab experiment can get from the industrial adsorption process. The recent development of the RUPTURA software[40] opens new perspectives in modeling. For instance, the axial dispersion coefficient used in a breakthrough

model can be calculated using transport properties, combined with thermodynamic data on the adsorption process of xenon and krypton. This presents an opportunity for experiment-theory comparison, fostering a virtuous feedback loop to improve modeling and facilitate the discovery of best-performing materials.

The diffusion coefficients calculated using the aforementioned methodologies solely describe self-diffusion in an infinitely diluted environment. To better describe transport properties in industrial conditions, it will be necessary, in the future, to study diffusion coefficients in a higher loading environment to account for host–host interactions. Furthermore, mixture simulations can be directly conducted to obtain the so-called Onsager diffusion coefficients, which are based on the Maxwell–Stefan diffusion equation rather than Fick's equation.[41] The calculation of such quantities requires significant computational resources, as MD simulations on mixtures at relatively high loading must be run for a sufficiently long duration to capture the diffusion regime. Therefore, applying this approach to large-scale screening is impractical, but some interesting materials can be tested to study the effects of mixtures and loading on transport properties.

Finally, we note that the work performed here relies on the "rigid host" approximation, which is an important limitation of our current methodology. Indeed, many MOFs are known to exhibit dynamic behavior, whether by local flexibility of their organic linkers or through large-scale structure changes upon external stimulation. Both types of flexibility can have a significant impact on adsorption, on its thermodynamics as well as on the transport properties of the guest molecules.[42,43] While flexibility and its impact on adsorption can be studied on a case-by-case basis, we do not believe there is any reliable methodology today that can systematically determine what structures are flexible or rigid, at the scale of thousands of structures. Moreover, classical simulations of flexible frameworks would require the use of "universal" or transferable force fields for intramolecular interactions, whose accuracy will be rather limited. We see this, for now, as a wide open and challenging question — on which perhaps we can draw inspiration from the recent methodologies to address the question of

high-throughput screening of thermal conductivity.[44]

Conflicts of interest

There are no conflicts to declare.

Acknowledgement

We thank Philippe Guilbaud and Isabelle Hablot for many discussions on the topic of adsorption-based separation. This work was supported by the Agence Nationale de la Recherche (project MATAREB, ANR-18-CE29-0009-01), and access to HPC platforms was provided by a GENCI grant (A0150807069). This work was financially supported by Orano.

Supporting Information Available

Additional discussion and results on the exploration of relevant descriptors, detailed list of features and their selection, details on ML model training. Raw data are available online at `https://github.com/fxcoudert/citable-data`, and the Grid Adsorption Energy Sampling code is available at `https://github.com/coudertlab/GrAED` and `https://github.com/eren125/xe_kr_selectivity_xgb`

References

(1) Humphrey, J.; Keller, G. *Separation Process Technology*; Builders' Guides; McGraw-Hill, 1997.

(2) Seader, J.; Henley, E.; Roper, D. *Separation Process Principles, 3rd Edition*; John Wiley Incorporated, 2010.

(3) Sholl, D. S.; Lively, R. P. Seven chemical separations to change the world. *Nature* **2016**, *532*, 435–437.

(4) Schüth, F.; Sing, K.; Weitkamp, J. *Handbook of Porous Solids*; Handbook of Porous Solids vol. 5; Wiley-VCH, 2002.

(5) Auerbach, S.; Carrado, K.; Dutta, P. *Handbook of Zeolite Science and Technology*; CRC Press, 2003.

(6) Pabby, A.; Rizvi, S.; Requena, A. *Handbook of Membrane Separations: Chemical, Pharmaceutical, Food, and Biotechnological Applications*; Taylor & Francis, 2008.

(7) Ruthven, D. *Principles of Adsorption and Adsorption Processes*; Wiley-Interscience publication; Wiley, 1984.

(8) Ruthven, D.; Farooq, S.; Knaebel, K. *Pressure Swing Adsorption*; Wiley, 1996.

(9) Kumar, R. Pressure Swing Adsorption Process: Performance Optimum and Adsorbent Selection. *Ind. Eng. Chem. Res.* **1994**, *33*, 1600–1605.

(10) Colón, Y. J.; Snurr, R. Q. High-throughput computational screening of metal–organic frameworks. *Chem. Soc. Rev.* **2014**, *43*, 5735–5749.

(11) Daglar, H.; Keskin, S. Recent advances, opportunities, and challenges in high-throughput computational screening of MOFs for gas separations. *Coord. Chem. Rev.* **2020**, *422*, 213470.

(12) Ren, E.; Guilbaud, P.; Coudert, F.-X. High-throughput computational screening of nanoporous materials in targeted applications. *Digital Discovery* **2022**, *1*, 355–374.

(13) Martin, R. L.; Simon, C. M.; Smit, B.; Haranczyk, M. In silico Design of Porous Polymer Networks: High-Throughput Screening for Methane Storage Materials. *J. Am. Chem. Soc.* **2014**, *136*, 5006–5022.

(14) Yeo, B. C.; Kim, D.; Kim, H.; Han, S. S. High-Throughput Screening to Investigate the Relationship between the Selectivity and Working Capacity of Porous Materials for Propylene/Propane Adsorptive Separation. *J. Phys. Chem. C* **2016**, *120*, 24224–24230.

(15) Boyd, P. G.; Lee, Y.; Smit, B. Computational development of the nanoporous materials genome. *Nat Rev Mater* **2017**, *2*, 1.

(16) Dubbeldam, D.; Snurr, R. Q. Recent developments in the molecular modeling of diffusion in nanoporous materials. *Mol. Simul.* **2007**, *33*, 305–325.

(17) Altintas, C.; Avci, G.; Daglar, H.; Gulcay-Ozcan, E.; Erucar, I.; Keskin, S. Computer simulations of 4240 MOF membranes for H_2/CH_4 separations: insights into structure–performance relations. *J. Mater. Chem. A* **2018**, *6*, 5836–5847.

(18) Zhou, M.; Wu, J. Massively Parallel GPU-Accelerated String Method for Fast and Accurate Prediction of Molecular Diffusivity in Nanoporous Materials. *ACS Appl. Nano Mater.* **2021**, *4*, 5394–5403.

(19) Bukowski, B. C.; Snurr, R. Q. Insights and Heuristics for Predicting Diffusion Rates of Chemical Warfare Agents in Zirconium Metal-Organic Frameworks. *ACS Appl. Mater. Interfaces* **2022**, *14*, 55608–55615.

(20) Mace, A.; Barthel, S.; Smit, B. Automated Multiscale Approach To Predict Self-Diffusion from a Potential Energy Field. *J. Chem. Theory Comput.* **2019**, *15*, 2127–2141.

(21) Gustafsson, H.; Kozdra, M.; Smit, B.; Barthel, S.; Mace, A. Predicting Ion Diffusion from the Shape of Potential Energy Landscapes. *J. Chem. Theory Comput.* **2024**, *20*, 18–29.

(22) Bukowski, B. C.; Keil, F. J.; Ravikovitch, P. I.; Sastre, G.; Snurr, R. Q.; Coppens, M.-

O. Connecting theory and simulation with experiment for the study of diffusion in nanoporous solids. *Adsorption* **2021**, *27*, 683–760.

(23) Chung, Y. G.; Camp, J.; Haranczyk, M.; Sikora, B. J.; Bury, W.; Krungleviciute, V.; Yildirim, T.; Farha, O. K.; Sholl, D. S.; Snurr, R. Q. Computation-Ready, Experimental Metal–Organic Frameworks: A Tool To Enable High-Throughput Screening of Nanoporous Crystals. *Chem. Mater.* **2014**, *26*, 6185–6192.

(24) 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*, 5985–5998.

(25) Ren, E.; Coudert, F.-X. Thermodynamic exploration of xenon/krypton separation based on a high-throughput screening. *Faraday Discuss.* **2021**, *231*, 201–223.

(26) Dubbeldam, D.; Calero, S.; Ellis, D. E.; Snurr, R. Q. RASPA: molecular simulation software for adsorption and diffusion in flexible nanoporous materials. *Mol. Simul.* **2015**, *42*, 81–101.

(27) Dubbeldam, D.; Ford, D. C.; Ellis, D. E.; Snurr, R. Q. A new perspective on the order-n algorithm for computing correlation functions. *Mol. Simul.* **2009**, *35*, 1084–1097.

(28) Ren, E.; Coudert, F.-X. Enhancing Gas Separation Selectivity Prediction through Geometrical and Chemical Descriptors. *Chem. Mater.* **2023**, *35*, 6771–6781.

(29) Banerjee, D.; Zhang, Z.; Plonka, A. M.; Li, J.; Parise, J. B. A Calcium Coordination Framework Having Permanent Porosity and High CO_2/N_2 Selectivity. *Cryst. Growth Des.* **2012**, *12*, 2162–2165.

(30) Daglar, H.; Keskin, S. Combining Machine Learning and Molecular Simulations to

Unlock Gas Separation Potentials of MOF Membranes and MOF/Polymer MMMs. *ACS Appl. Mater. Interfaces* **2022**, *14*, 32134–32148.

(31) 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*, 134–141.

(32) Pinheiro, M.; Martin, R. L.; Rycroft, C. H.; Jones, A.; Iglesia, E.; Haranczyk, M. Characterization and comparison of pore landscapes in crystalline porous materials. *J. Mol. Graph. Model.* **2013**, *44*, 208–219.

(33) Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA, 2016; pp 785–794.

(34) Lundberg, S.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. 2017; `https://arxiv.org/abs/1705.07874`, Version v2 submitted on 2017-11-25. Accessed on 2023-07-05.

(35) Rappé, A. K.; Casewit, C. J.; Colwell, K.; Goddard III, W. A.; Skiff, W. M. UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *J. Am. Chem. Soc.* **1992**, *114*, 10024–10035.

(36) Ryan, P.; Farha, O. K.; Broadbelt, L. J.; Snurr, R. Q. Computational screening of metal-organic frameworks for xenon/krypton separation. *AIChE J.* **2010**, *57*, 1759–1766.

(37) Lorentz, H. A. Ueber die Anwendung des Satzes vom Virial in der kinetischen Theorie der Gase. *Ann. Phys.* **1881**, *248*, 127–136.

(38) Bullitt, E.; Gerig, G.; Pizer, S.; Lin, W.; Aylward, S. Measuring tortuosity of the intracerebral vasculature from MRA images. *IEEE Trans. Med. Imaging* **2003**, *22*, 1163–1171.

(39) Lin, B.; Meron, M.; Cui, B.; Rice, S. A.; Diamant, H. From Random Walk to Single-File Diffusion. *Phys. Rev. Lett.* **2005**, *94*, 216001.

(40) Sharma, S.; Balestra, S. R. G.; Baur, R.; Agarwal, U.; Zuidema, E.; Rigutto, M. S.; Calero, S.; Vlugt, T. J. H.; Dubbeldam, D. RUPTURA: simulation code for breakthrough, ideal adsorption solution theory computations, and fitting of isotherm models. *Mol. Simul.* **2023**, *49*, 893–953.

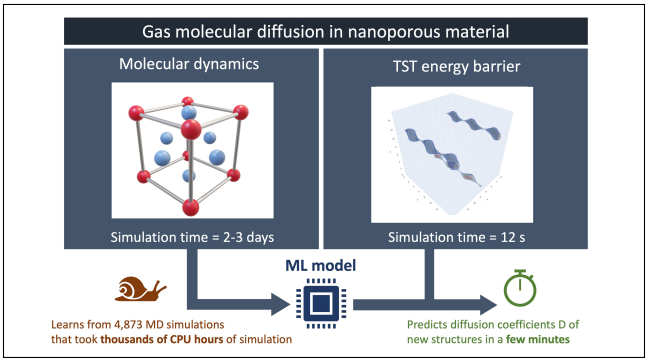
(41) Krishna, R.; van Baten, J. Onsager coefficients for binary mixture diffusion in nanopores. *Chem. Eng. Sci.* **2008**, *63*, 3120–3140.

(42) Haldoupis, E.; Watanabe, T.; Nair, S.; Sholl, D. S. Quantifying Large Effects of Framework Flexibility on Diffusion in MOFs: CH_4 and CO_2 in ZIF-8. *ChemPhysChem* **2012**, *13*, 3449–3452.

(43) Fraux, G.; Boutin, A.; Fuchs, A. H.; Coudert, F.-X. Structure, Dynamics, and Thermodynamics of Intruded Electrolytes in ZIF-8. *J. Phys. Chem. C* **2019**, *123*, 15589–15598.

(44) Islamov, M.; Babaei, H.; Anderson, R.; Sezginel, K. B.; Long, J. R.; McGaughey, A. J. H.; Gomez-Gualdron, D. A.; Wilmer, C. E. High-throughput screening of hypothetical metal-organic frameworks for thermal conductivity. *npj Comput. Mater.* **2023**, *9*.

TOC Graphic


Gas molecular diffusion in nanoporous material
Molecular dynamics
TST energy barrier
Simulation time = 2-3 days
Simulation time = 12 s
ML model
Learns from 4,873 MD simulations
that took thousands of CPU hours of simulation
Predicts diffusion coefficients D of
new structures in a few minutes