Data Mining for Predicting Gas Diffusivity in Zeolitic-imidazolate Frameworks (ZIFs)

Panagiotis Krokidas,[1,*] Stelios Karozis,[2] Salvador Moncho,[3] George Giannakopoulos,[4] Edward N. Brothers,[3] Michael E. Kainourgiakis,[2] Ioannis G. Economou[1,5*] and Theodore A. Steriotis[1]

[1]Institute of Nanoscience and Nanotechnology, National Centre for Scientific Research "Demokritos", 15341 Aghia Paraskevi Attikis, Athens, Greece

[2]Institute of Nuclear & Radiological Sciences and Technology, Energy & Safety, National Centre for Scientific Research "Demokritos", 15341 Aghia Paraskevi Attikis, Athens, Greece

[3]Science Program, Texas A&M University at Qatar, P.O. Box 23874, Education City, Doha, Qatar

[4]Institute of Informatics and Telecommunications, National Centre for Scientific Research "Demokritos", 15341 Aghia Paraskevi Attikis, Athens, Greece

[5]Chemical Engineering Program, Texas A&M University at Qatar, P.O. Box 23874, Education City, Doha, Qatar

[*]Corresponding author at p.krokidas@inn.demokritos.gr

[*]Corresponding author at ioannis.economou@qatar.tamu.edu

ZIF Building units

Table S1 shows the metals, linkers and functional groups considered. Metals were limited to those that can adopt tetrahedral co-ordination and an oxidization state of 2+. Some combinations are well-known; for instance, the imidazole (Im), combined with methyl as a functional group, provides the original linker (2-methylimidazole (mIm)) of ZIF-8.[1] The same holds for benzine-imidazole (bIm),[2,3,4] tetrazole (tz)[5] and 4,5-dichloroimidazole (dClIm).[6,7,8] The rest of the main organic parts considered here are being incorporated in ZIFs for the first time. In terms of functional groups, only the following have been used in previous experimental studies: $-CH_{3,}$ (ZIF-8's original functional group),[9] $-NH_{2,}$[10,11,12] -H,[9] -Br[13] and -Cl.[13] In total, we have designed 72 ZIF-8 analogues, by following a combinatory replacement scheme of metals, linkers, and functional groups, as reported in Table S1. Details for all the variants can be found in Table S2, while additional potential structures are summarized in Table S3. Several of the structures examined in this work have not been synthesized experimentally, yet.

Table S1. Metals, linkers and functional groups used in the ZIF-8 replacement scheme.

Replacement part	Type
Metals	Zn^{2+}, Co^{2+}, Cd^{2+}, Be^{2+}, Mg^{2+}, Mn^{2+}, Cu^{2+}
Linkers	imidazole (Im), 1-H-1,3-benzimidazole (bIm), tetrazole (tz), 4,5-dichloroimidazole, (dClIm), 4,5-dibromoimidazole (dBrIm), 4,5-difthoroimidazole (dFIm), 4,5- (dIIm)
Functional Groups	-H, $-CH_3$, -Cl, -Br, -I, -F, $-NH_2$, -CHO

Table S2. The ZIF-8 variants considered.

	Name	Metal	Linker	Functional Group
1	ZIF-8	Zn	mIm	-CH$_3$
2	ZIF-67	Co	mIm	-CH$_3$
3	CdIF-1	Cd	mIm	-CH$_3$
4	BeIF-1	Be	mIm	-CH$_3$
5	Cu-ZIF-8	Cu	mIm	-CH$_3$
6	Mg-ZIF-8	Mg	mIm	-CH$_3$
7	Mn-ZIF-8	Mn	mIm	-CH$_3$
8	ZIF-8-Br	Zn	mIm	-Br
9	Co-ZIF-8-Br	Co	mIm	-Br
10	ZIF-8-Cl	Zn	mIm	-Cl
11	ZIF-8-Im_1	Zn	mIm/mIm/Im	-CH$_3$/-CH$_3$/-H
12	ZIF-8-Im_2	Zn	mIm/Im/Im	-CH$_3$/-H/-H
13	ZIF-8-Im_3	Zn	Im/Im/Im	-H/-H/-H
14	ZIF-7-8	Zn	mIm/mIm/bIm	-CH$_3$/-CH$_3$/-H
15	Co-ZIF-7-8	Co	mIm/mIm/bIm	-CH$_3$/-CH$_3$/-H
16	Be-ZIF-7-8	Be	mIm/mIm/bIm	-CH$_3$/-CH$_3$/-H
17	Cu-ZIF-7-8	Cu	mIm/mIm/bIm	-CH$_3$/-CH$_3$/-H
18	Mg-ZIF-7-8	Mg	mIm/mIm/bIm	-CH$_3$/-CH$_3$/-H
19	Mn-ZIF-7-8	Mn	mIm/mIm/bIm	-CH$_3$/-CH$_3$/-H
20	ZIF-7-8-Cl	Zn	mIm/mIm/bIm	-CH$_3$/-CH$_3$/-Cl
21	ZIF-7-8-Br	Zn	mIm/mIm/bIm	-CH$_3$/-CH$_3$/-Br
22	ZIF-7-8-I	Zn	mIm/mIm/bIm	-CH$_3$/-CH$_3$/-F
23	ZIF-7-8-F	Zn	mIm/mIm/bIm	-CH$_3$/-CH$_3$/-Cl
24	Cd-ZIF-7-8-Cl	Cd	mIm/mIm/bIm	-CH$_3$/-CH$_3$/-Br
25	Cd-ZIF-7-8-Br	Cd	mIm/mIm/bIm	-CH$_3$/-CH$_3$/-Br
26	Co-ZIF-7-8-Br	Co	mIm/mIm/bIm	-CH$_3$/-CH$_3$/-Cl
27	Co-ZIF-7-8-Cl	Co	mIm/mIm/bIm	-CH$_3$/-CH$_3$/-Cl
28	Co-ZIF-7-8-F	Co	mIm/mIm/bIm	-CH$_3$/-CH$_3$/-F
29	Co-ZIF-7-8-I	Co	mIm/mIm/bIm	-CH$_3$/-CH$_3$/-I
30	Be-ZIF-7-8-F	Be	mIm/mIm/bIm	-CH$_3$/-CH$_3$/-F
31	Be-ZIF-7-8-I	Be	mIm/mIm/bIm	-CH$_3$/-CH$_3$/-I
32	Cu-ZIF-7-8-F	Cu	mIm/mIm/bIm	-CH$_3$/-CH$_3$/-F
33	Cu-ZIF-7-8-Cl	Cu	mIm/mIm/bIm	-CH$_3$/-CH$_3$/-Cl
34	Cu-ZIF-7-8-Br	Cu	mIm/mIm/bIm	-CH$_3$/-CH$_3$/-Br
35	Cu-ZIF-7-8-I	Cu	mIm/mIm/bIm	-CH$_3$/-CH$_3$/-I
36	Mg-ZIF-7-8-Br	Mg	mIm/mIm/bIm	-CH$_3$/-CH$_3$/-Br
37	Mg-ZIF-7-8-I	Mg	mIm/mIm/bIm	-CH$_3$/-CH$_3$/-I
38	Mn-ZIF-7-8-Br	Mn	mIm/mIm/bIm	-CH$_3$/-CH$_3$/-Br
39	Mn-ZIF-7-8-I	Mn	mIm/mIm/bIm	-CH$_3$/-CH$_3$/-I

40	Cd-ZIF-7-8-I	Cd	mIm/mIm/bIm	-CH$_3$/-CH$_3$/-I
41	Tetrz-ZIF-8	Zn	tetrz	-CH$_3$
42	Co-Tetrz-ZIF-8	Co	Tetrz	-CH$_3$
43	Be-Tetrz-ZIF-8	Be	Tetrz	-CH$_3$
44	Cu-Tetrz-ZIF-8	Cu	Tetrz	-CH$_3$
45	Tetrz-ZIF-8-NH$_2$	Zn	Tetrz	-NH$_2$
46	Be-Tetrz-ZIF-8-NH$_2$	Be	Tetrz	-NH$_2$
47	Co-Tetrz-ZIF-8-NH$_2$	Co	Tetrz	-NH$_2$
48	dClm-ZIF-8	Zn	dClm	-CH$_3$
49	Co-dClm-ZIF-8	Co	dClm	-CH$_3$
50	Be-dClm-ZIF-8	Be	dClm	-CH$_3$
51	Cd-dClm-ZIF-8	Cd	dClm	-CH$_3$
52	Mg-dClm-ZIF-8	Mg	dClm	-CH$_3$
53	Cu-dClm-ZIF-8	Cu	dClm	-CH$_3$
54	dFm-ZIF-8	Zn	dFm	-CH$_3$
55	Co-dFm-ZIF-8	Co	dFm	-CH$_3$
56	Be-dFm-ZIF-8	Be	dFm	-CH$_3$
57	Cd-dFm-ZIF-8	Cd	dFm	-CH$_3$
58	Mg-dFm-ZIF-8	Mg	dFm	-CH$_3$
59	Cu-dFm-ZIF-8	Cu	dFm	-CH$_3$
60	dIm-ZIF-8	Zn	dIm	-CH$_3$
61	Co-dlm-ZIF-8	Co	dIm	-CH$_3$
62	Be-dlm-ZIF-8	Be	dIm	-CH$_3$
63	Cu-dlm-ZIF-8	Cu	dIm	-CH$_3$
64	Cd-dlm-ZIF-8	Cd	dIm	-CH$_3$
65	Mg-dlm-ZIF-8	Mg	dIm	-CH$_3$
66	dBrm-ZIF-8	Zn	dBrm	-CH$_3$
67	Co-dBrm-ZIF-8	Co	dBrm	-CH$_3$
68	Be-dBrm-ZIF-8	Be	dBrm	-CH$_3$
69	Cd-dBrm-ZIF-8	Cd	dBrm	-CH$_3$
70	Mg-dBrm-ZIF-8	Mg	dBrm	-CH$_3$
71	Cu-dBrm-ZIF-8	Cu	dBrm	-CH$_3$
72	ZIF-8-CHO	Zn	mIm	-CHO

Table S3. Some possible linkers, metals and functional groups that can be incorporated in ZIF-8 topology and the resulting number of combinations.

Building unit	Name	Number	Combinations
Linkers	methylimidazole, benzimidazole, methyl-triazole, methyl-tetrazole, dimethyl benzimidazole, dichloroimidazole, nitroimidazole,	17	17×9×14 = **2142**

	dinitroImidazole, bromoimidazole, dibromoimidazole, fthoroImidazole, difthoroImidazole, iodoImidazole, diIodoImidazole, cyanoImidazole, dicyanoimidazole, Purinate		
Metals	Be, Cu, Mg, Co, Zn, Fe, Mn, Cd, Ni	9	
Functional Groups	$-CH_3$, -Br, Cl-, -CHO, I-, -F, -phIm, -aIm, -eIm, -SH, $-NO_2$, $-NH_2$	14	

Computational methodology

Force field development. For each ZIF-8 variant, a set of case specific force field terms was developed. The force field used consists of the following terms: bond stretching (Eq. 1), bond angle bending (Eq. 2) and torsional angle distortion (Eq. 3) for the bonded intra-molecular interactions, as well as Lennard Jones (LJ) and electrostatic terms, for the non-bonded intra- and inter-molecular interactions (Eq. 4):

$$U^{stretch}(l) = \frac{k_l}{2}(l - l_0)^2 \tag{1}$$

$$U^{bend}(\theta) = \frac{k_\theta}{2}(\theta - \theta_0)^2 \tag{2}$$

$$U^{torsion}(\varphi) = k_\varphi\left[1 + \cos\left(m\varphi - \varphi_0\right)^2\right] \tag{3}$$

$$U(r_{ij}) = 4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] + \frac{1}{4\pi\varepsilon_0}\frac{q_i q_j}{r_{ij}} \tag{4}$$

where k_l, k_θ and k_φ are constants that characterize the bond length, bond angle and torsional angle stiffness, respectively; l, θ and φ correspond to the bond length, bond angle and torsional angle, respectively, and subscript 0 refers to the equilibrium value; ε_{ij} and σ_{ij} are the Lennard-Jones energy and size parameters, r_{ij} is the distance between atoms i and j, q_i is the charge at atom i and ε_0 is the permittivity of vacuum.

More specifically, we have used the values of AMBER for the Lennard-Lones parameters, while we calculated all bonding term (bond lengths, angles, etc.) parameters and the charges with the use of DFT calculations, following the same procedure as in our recent works.[14,15,16] The hybrid density functional B3LYP with a large basis set (6-311g++(2d,2p)) was used with a dense integration grid ('ultrafine').[17,18] For ZIFs with Cd (which has 48 electrons) and/or I (which has 53 electrons), we used the Def2TZVPPD basis set which has a similar accuracy (triple-zeta-valence with two sets of polarization functions as well as diffuse functions) but includes an effective core potential (ECP) to replace the innermost 38 electrons in the metal. The clusters we used for each ZIF in our DFT calculations are shown in detail in the file that includes the force field values (ESI_3.xls).

Previously published forcefields were used to describe the interactions between the guest molecules and ZIF analogues. More specifically, O_2, CH_4, C_2H_6, C_2H_8, C_3H_8, C_3H_{10}, n-C_4H_{10} and iso-C_4H_{10}, were modeled through a united atom (UA) representation, using the TraPPE-UA.[19] H_2 and He were modeled with the force fields from the work of Velioglu and Keskin[20] and Talu and Mayers,[21] respectively. N_2 was modelled as a three-site molecule: two atoms are placed 1.1 Å apart to account for the moment of inertia of the molecule. A third fictional atom is placed between the two N atoms, which is massless but carried the appropriate charge. This arrangement facilitates an accurate quadrupole moment for the molecule.[22] The model accounts for the flexibility of the bond angles, while bond lengths are considered fixed. The TraPPE force field was adopted for the CO_2 guest molecule, which consists of a three-point charge linear molecule with fixed bond lengths of 1.16 Å.[19]

Simulations. Each modification was reconstructed on the molecular level, in the form of a super-cell, which corresponds to a box of 2×2×2 unit cells. Then, each ZIF variant underwent equilibration with Molecular Dynamics (MD) simulations, at the NPT ensemble, at 308 K and 1 bar for 1 ns, in order to allow for correct framework volume adjustment. The importance of applying an MD simulation at the NPT ensemble prior to the main simulations should be underlined because each new replacement unit affects the framework's volume, which in turn can affect the resulting aperture size. Thus, considering a common volume value across the different ZIF simulation boxes would be incorrect. Then, the structures underwent a similar thermal equilibration, at the NVT ensemble (again at 308 K) for 1 ns. By following a procedure that we reported in previous works,[15] we have processed the trajectories of the aperture's linkers and extract distributions of aperture sizes, from which we calculated the average aperture size for each new ZIF. The thermostat in all simulations was Nose-Hoover and the time step was set to 1fs. The length of the umbrella sampling simulations was 200-500 ps. More details on the MD simulation parameters can be found in our earlier studies.[23] The measurement of aperture and stretched aperture sizes (size of the aperture when a gas penetrant lies in its center), by assuming either circular or elliptical shapes have been presented previously.[23,15] Figure S1(a) shows the resulting distribution of aperture sizes in our dataset.

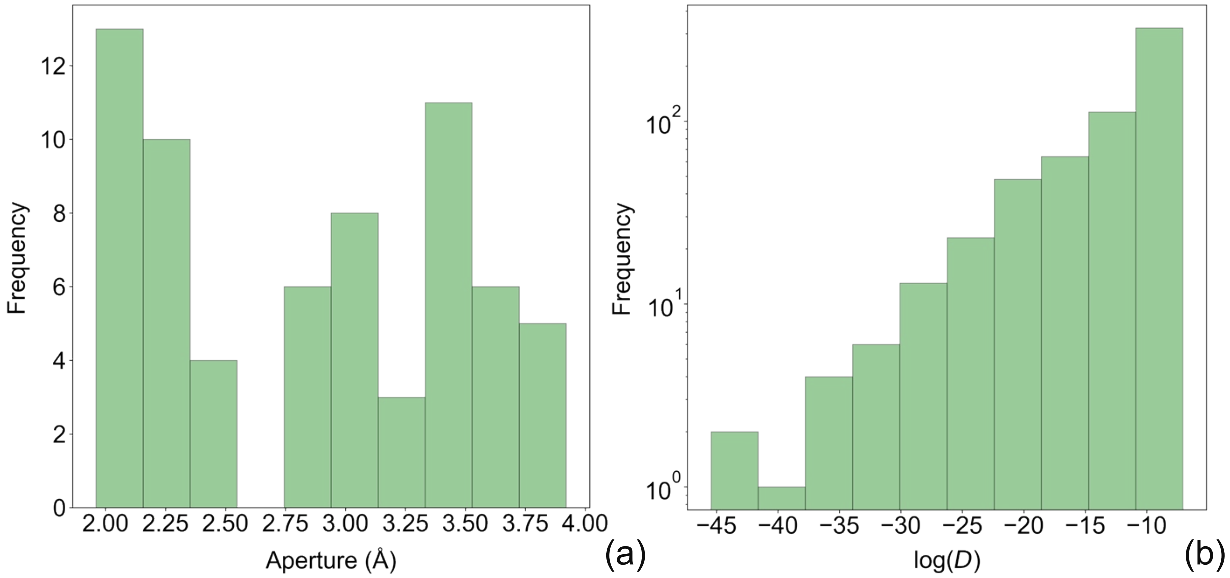

Figure S1. Bar plots showing the distribution of values of our dataset for (a) aperture and (b) the logarithm of calculated diffusivities (the units of the diffusivity in our work are in m^2/sec).

The diffusivity of several gas molecules (Table S4) was calculated in each of the 72 ZIFs. The most popular approach for estimating the diffusivity from molecular simulations, is the calculation of the mean-square displacement (MSD) from which the self-diffusivity can be extracted.[24] In a similar manner, the corrected diffusivity can be estimated by the displacement of the center of mass of a swarm of penetrants.[24,25] However, as the diffusivity decreases these approaches come with an increasing computational cost and are subject to high uncertainty. Diffusivities between 10^{-12} – 10^{-14} m^2/sec at room temperature, can be estimated by running simulations at a range of elevated temperatures and extrapolating at the desired temperature with the use of an Arrhenius-based equation.[23] Nevertheless, our experience has shown that even this approach becomes impractical for diffusivities << 10^{-14} m^2/sec. In our systems, we observed diffusivities much lower than what these approaches can reproduce, therefore we employed dynamically corrected transition state theory (dcTST),[26] with the use of umbrella sampling.[27] TST depicts very slow diffusion as a succession of prolonged periods of random collisions of a molecule with the walls of a cage, succeeded by a sudden crossing through an opening to an adjacent cage. Information about the implementation of TST in our calculations can be found in our previous work.[16]

Table S4. The gas molecules considered, along with their properties.

	Mass (g/mol)	vdW diameter[28] (Å)	Kinetic diameter[28] (Å)	Acentric factor[29]
He	4.00	2.66	2.60	-0.390
H_2	2.01	2.76	2.89	-0.217
O_2	32.00	2.94	3.46	0.022
CO_2	44.01	3.24	3.30	0.225
N_2	28.00	3.13	3.64	0.037
CH_4	16.04	3.25	3.80	0.011
C_2H_4	28.05	3.59	3.90	0.087
C_2H_6	30.07	3.72	4.00	0.099
C_3H_6	42.08	4.03	4.50	0.142
C_3H_8	44.10	4.16	4.30	0.152
n-C_4H_{10}	58.12	4.52	4.50	0.200
i-C_4H_{10}	58.12	4.80	4.42	0.183

To minimize the computational time, we assumed infinite dilution calculations (one guest molecule per ZIF). Contrary to conventional MD methods, which are subject to a high statistical uncertainty as guest molecules concentration drops, TST-based diffusivities come with exceptionally small errors for infinite dilution estimations (~10% in the calculations of this work). More specifically, for each diffusivity calculation fifty (50) evenly spaced umbrellas with a spring of 5,000 – 50,000 kJ/mol/nm^2 force constant were sampled along the axis that connects the centers of two adjacent cages and passes through the aperture's center. The complete set of the 50 umbrellas was repeated 5-10 times, with different initial random seeds. Then, the umbrellas' trajectories were analyzed with the weighted histogram analysis method (WHAM). We opted for the Bayesian Bootstrapping[30] approach of the histograms to produce the free energy curve, which allows for an accurate error estimation. In all cases, cubic periodic boundary conditions were applied to the three directions of the box and a cut-off distance of 13Å was imposed for the van der Waals interactions. The time step was set at 1.0 fs, and the integration was accomplished with the velocity-Verlet integration algorithm. The particle mesh Ewald method (PME) was used for the description of the electrostatic interactions. Figure S1(b) shows the distribution of calculated diffusivities in our dataset.

Defining the sizes for the building units

The size of the linkers and functional groups is estimated by simple length calculations as explained below. It should be mentioned that the first publication to report aperture measurements in ZIFs considered a Pauling van der Waals radius for hydrogen, which now is used in all XRD reported works.[2] Thus, for the sake of consistency, we stick to the Pauling van der Waals radii, *R*, for all outermost atoms.[31] These are shown in Table S5.

Table S5. Crystallographic van der Waals radii[31] for the outermost atoms of the aperture in the various linkers and functional groups considered in our ZIF modifications.

			R **(Å)**			
H	**F**	**Cl**	**Br**	**I**	**O**	**N**
1.20	1.35	1.80	1.95	2.15	1.40	1.50

Linkers

The length of a linker is the sum of distances and the van der Waals radii of the linker's terminal atoms (vdW), as shown in Figure S2. Im, dClm, dFm, dIm, dBrm share the same ring architecture and the measurement is shown in Figure S2(a), where the atom depicted as yellow can be hydrogen (mIm), Cl (dClm), F (dFm), Br (dBrm) and I (dIm). Tetrz and bIm linker's measurement is shown in Figure S2(b) and (c), respectively. The calculated length of all the linkers used are reported in Table S6.

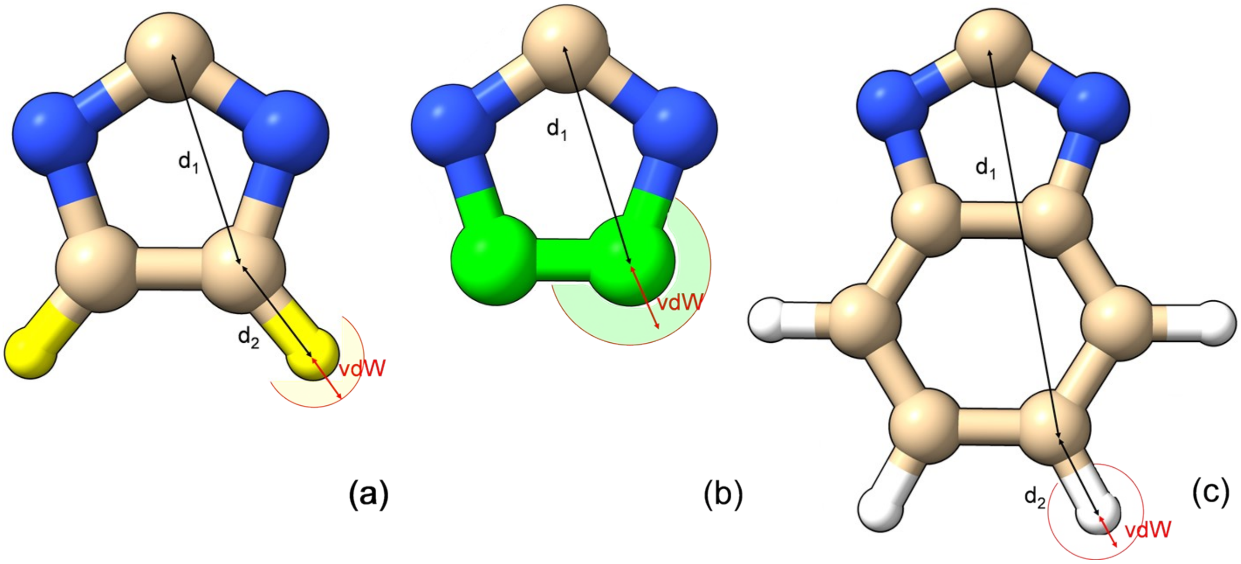

Figure S2. Geometric parameters for (a) for Im, dClm, dFm, dIm and dBrm, (b) tetrz and (c) bIm, linkers.

Table S6. Linker lengths (all values are in Å).

Linker	d_1	d_2	vdW	Length
Im	2.16	1.08	1.20	4.44
dClm	2.16	1.74	1.80	5.70
dFm	2.16	1.35	1.35	4.86
dBrm	2.16	1.90	1.95	6.01
dIm	2.16	2.10	2.15	6.41
bIm	3.72	1.08	1.20	6.00
tetrz	2.16	-	1.50	3.66

Functional groups

The functional group's length measurement starts from the upper carbon atom of the imidazole ring as shown in Figure S3. For the terminal atoms of the group, the corresponding Pauli van der Waals radii are added. Results are provided in Table S7.

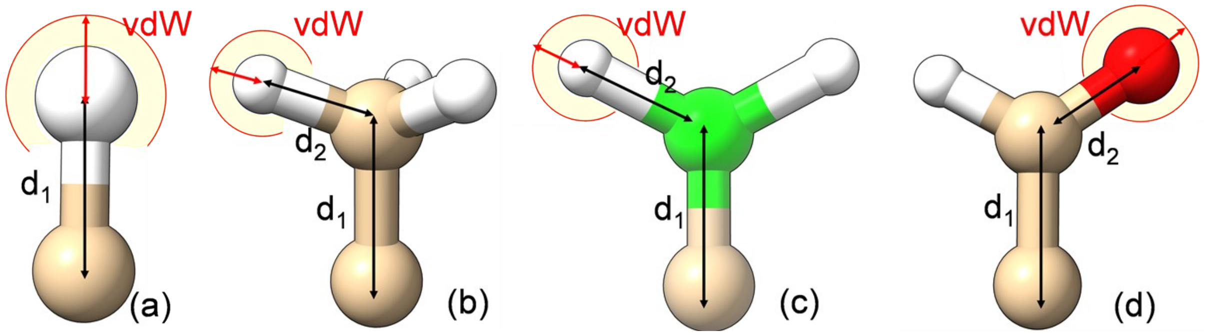

Figure S3. Geometries for (a) for -H, -Cl, -F, -I and -B, (b) -CH_3, (c) -NH_2 and (d) -CHO.

Table S7. Functional group lengths (all values are in Å).

Functional group	d_1	d_2	vdW	Length
-H	1.10		1.20	2.30
-Cl	1.74		1.80	3.54
-F	1.35		1.35	2.70
-I	2.10		2.15	4.25
-Br	1.90		1.95	3.85
-CH_3	1.48	1.10	1.20	3.78
-NH_3	1.44	1.29	1.20	3.93
-CHO	1.47	1.22	1.40	4.09

Descriptors

The aperture is formed by three organic linkers. Therefore, we have used three families of descriptors (linker 1, 2 and 3) with four features per family (i.e., mass and length of linker and functional group) adding up to twelve basic descriptors per ZIF. Additionally, the ionic radius of the metal as well as apertureAtom_σ and apertureAtom_ε were employed. The latter two stand for the Lennard-Jones σ and ε parameters, of the outermost linker atom forming the aperture, as discussed in the previous section. Finally, the aperture and stretched aperture sizes were also used to describe the ZIF variants, while penetrants were described by means of mass, acentric factor, van der Waals and kinetic diameters (Table S4). Overall, there are 17 descriptors for the framework, 4 descriptors for each gas molecule and an extra feature (stretched aperture) referring to the gas-framework interaction. Table S8 presents the descriptors in detail.

Table S8. ZIF- and gas-related descriptors employed in our ML models.

Aperture	(Å)	Diameter of the aperture – it is used as a descriptor only in the ML model of Fig. 3
Stretched Aperture	(Å)	Diameter of the aperture when a gas molecule lies in its center - it is used as a descriptor only in the ML model of Fig. 3
Predicted aperture	(Å)	Diameter of the aperture – it is predicted by M1, and it is used as a descriptor in M2 (dual-step ML model of Fig. 4)
IonicRad (Å)	(Å)	Ionic radius of ZIFs metal center
apertureAtom_sigma	(Å)	σ of the outermost atom of the linker forming the aperture
apertureAtom_e	(Å)	ε of the outermost atom of the linker forming the aperture
LinkerLength1 LinkerLength2 LinkerLength3	(Å)	Length of each of the three organic linkers of the aperture
LinkerMass1 LinkerMass2 LinkerMass3	(g/mol)	Mass of each of the three organic linkers of the aperture
Func_lenght1 Func_lenght2 Func_lenght3	(Å)	Length of the functional group of each of the three organic linkers of the aperture
Func_mass1 Func_mass2 Func_mass3	(g/mol)	Mass of the functional group of each of three organic linkers of the aperture
Mass	(g/mol)	Mass of the gas molecule
AcentricFactor		Acentric factor of the gas molecule
Size_vdW	(Å)	Van der Waals diameter of the gas molecule
Size_kinDiam	(Å)	kinetic diameter of the gas molecule

Machine Learning. The implementation of the steps, that are described below, was carried out in Python3 programming environment with the help of the scikit-learn library.

1. Data preparation: The data were subjected to a harmonization process in order to be used in the ML models. The standard scaler was used to rescale the variables in order to have zero mean unit variance, removing thus large differences of magnitude among the predictors. The latter affects the weights and parameters of many ML algorithms, such as linear regression, but not tree-based models. In Random Forest (RF) and Decision Tree (DT) algorithms, the use of non-scaled data may change the location of the data split and

thus affect the order of features. With the aforementioned data pre-treatment this type of bias is eliminated.

2. Algorithms: The main algorithm that was used in our analysis was RF regressor (general model of Figure 3 and the dual model M1 and M2). RF is suitable for avoiding overfitting, but it has the limitation that at least three descriptors are needed. Thus, for M2_simple that operates with only two descriptors (ZIF aperture and gas vdW diameter), we used the DT algorithm. Both DT and RF offer the flexibility to perform importance feature analysis that enables the extraction of physicochemical information. Moreover, the overall result of splitting and branching can be visualized, assessed, and explained. The dataset was split to train and test partitions with the *K*-fold cross-validation protocol:[32] the data are randomly split into *K* non-overlapping parts, namely folds. Then, one of the folds is selected to serve as the test set, while the rest of *K-1* folds constitute the training set. The procedure is repeated for all *K* folds. Metrics are measured for each iteration and then they are averaged over the number of *K* folds. We used *K=5*. The hyperparameters of both RF and DT are the depth and the measure of quality of a split. The best depth value was extracted from a parametric analysis, where the R^2 was calculated for different depths, and the depth value corresponding to the point where the curve reaches a plateau was kept (Figure S4).

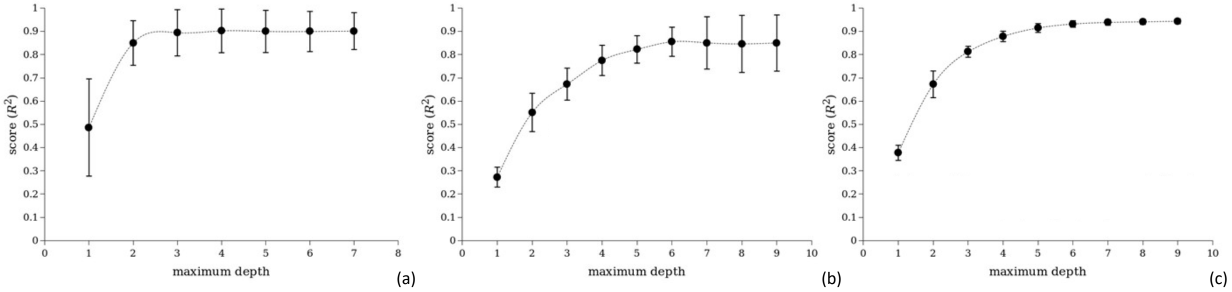

Figure S4. R^2 as a function of the depth for the models (a) M1, (b) M2_simple and (c) M2.

An additional analysis was executed to verify that we avoid over-fitting with our models, as follows: the R^2 was calculated separately for the training and test set.

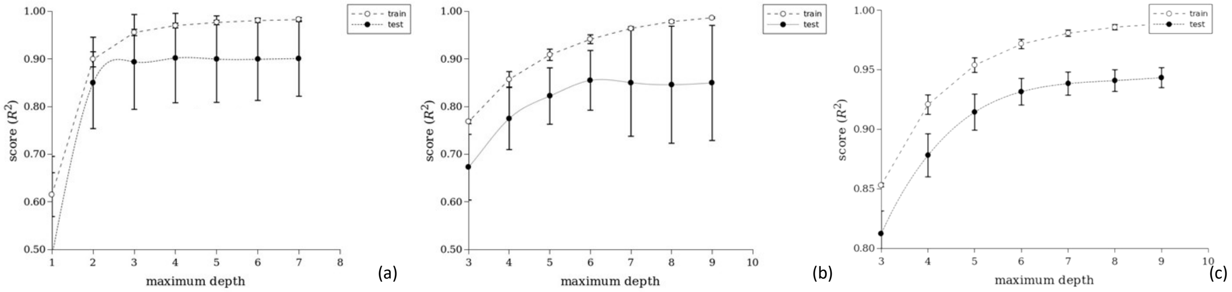

Figure S5. R^2 as a function of the depth, separately calculated for the train and test set (a) M1, for the models (b) M2_simple and (c) M2.

As shown in Figure S5, the test is always considerably below the train test in terms of performance, even after convergence has been reached. This is a direct confirmation that our ML models are not overfitted. Our analysis led us to follow a depth value of 3 for the ML model that predicts the aperture (M1 in Figure 4 of the main text) and 6 for the ML models that predict the diffusivity (M2 and M2_simple in Figure 4 of the main text).

3. Metrics: During the analysis three metrics were used in order to assess different aspect of the process:

    a. The importance feature metric was used to assess the contribution of different predictors in the diffusion process inside ZIFs. The metric is available for both RF & DT. It was used in RF and essentially revealed which predictors are used for fitting the data ("*responsible for establishing order to the dataset*").

    b. For assessing the performance per prediction value of the M1, M2 and M2_simple models, the R^2 score was chosen. It is a measure of the average squared error between predicted and true values, hence, but it does not reveal the variance.

    c. For assessing the performance of the whole distribution of predicted values of the M1, M2 and M2_simple models over actual values, the explained variance (*EV*) metric was chosen. It shows to what extend our ML model reproduces the variance of the original values distribution.

R^2 and *EV* are complementary and show not only how close the values are (R^2) but also the similarity of distributions (*EV*).

Computational Tools and Software. All DFT calculations were performed with the Gaussian 09 suite of programs.[33] The GROMACS open-source molecular simulation code (version 4.6.5)[34] was used for the energy minimization of initial structures and for all equilibrium MD simulations, umbrella sampling simulations and test particle insertions. The pull code of GROMACS was used for the umbrella sampling. The WHAM method was applied with the use of the g_wham code.[30] FORTRAN codes developed in house were used for the calculation of the correction factor from the transmission trajectories and for the aperture diameter measurement of each ZIF. The creation and handling of the massive TST-related input files were accomplished with the help of bash scripting. All the illustrations of atomistic depictions in this work were created with the use of Chimera program.[35] The ML workflow was developed by Python3[36] programming language with the help of the scikit-learn library (version 1.0.1).[37]

REFERENCES

(1) Banerjee, R.; Phan, A.; Wang, B.; Knobler, C.; Furukawa, H.; O'Keeffe, M.; Yaghi, O. M. High-Throughput Synthesis of Zeolitic Imidazolate Frameworks and Application to CO2 Capture. *Science* **2008**, *319*, 939–943.

(2) Park, K. S.; Ni, Z.; Cote, A. P.; Choi, J. Y.; Huang, R.; Uribe-Romo, F. J.; Chae, H. K.; O'Keeffe, M.; Yaghi, O. M. Exceptional Chemical and Thermal Stability of Zeolitic Imidazolate Frameworks. *Proc. Natl. Acad. Sci.* **2006**, *103*, 10186–10191.

(3) Thompson, J. A.; Blad, C. R.; Brunelli, N. A.; Lively, R. P.; Lydon, M. E.; Jones, W.; Nair, S.; Lively, R. P.; Jones, C. W.; Nair, S. Hybrid Zeolitic Imidazolate Frameworks: Controlling Framework Porosity and Functionality by Mixed-Linker Synthesis. *Chem. Mater.* **2012**, *24*, 1930–1936.

(4) Åhlén, M.; Jaworski, A.; Strømme, M.; Cheung, O. Selective Adsorption of CO2 and SF6 on Mixed-Linker ZIF-7–8s: The Effect of Linker Substitution on Uptake Capacity and Kinetics. *Chem. Eng. J.* **2021**, *422*, 130117.

(5) Li, H. Z.; Zhang, S. H.; Wang, F. Facile Syntheses of SOD-Type Tetrahedral Tetrazolate Frameworks for Acetylene Storage. *Inorg. Chem. Commun.* **2020**, *113*, 107797.

(6) Schweinefuß, M. E.; Springer, S.; Baburin, I. A.; Hikov, T.; Huber, K.; Leoni, S.; Wiebcke, M. Zeolitic Imidazolate Framework-71 Nanocrystals and a Novel SOD-Type Polymorph: Solution Mediated Phase Transformations, Phase Selection via Coordination Modulation and a Density Functional Theory Derived Energy Landscape. *Dalt. Trans.* **2014**, *43*, 3528–3536.

(7) Springer, S.; Baburin, I. A.; Heinemeyer, T.; Schiffmann, J. G.; Van Wüllen, L.; Leoni, S.; Wiebcke, M. A Zeolitic Imidazolate Framework with Conformational Variety: Conformational Polymorphs versus Frameworks with Static Conformational Disorder. *CrystEngComm* **2016**, *18*, 2477–2489.

(8) Wee, L. H.; Vandenbrande, S.; Rogge, S. M. J.; Wieme, J.; Asselman, K.; Jardim, E. O.; Silvestre-Albero, J.; Navarro, J. A. R.; Van Speybroeck, V.; Martens, J. A.; Kirschhock, C. E. A. Chlorination of a Zeolitic-Imidazolate Framework Tunes Packing and van Der Waals Interaction of Carbon Dioxide for Optimized Adsorptive Separation. *J. Am. Chem. Soc.* **2021**, *143*, 4962–4968.

(9) Hillman, F.; Hamid, M. R. A.; Krokidas, P.; Moncho, S.; Brothers, E. N.; Economou, I. G.; Jeong, H. K. Delayed Linker Addition (DLA) Synthesis for Hybrid SOD ZIFs with Unsubstituted Imidazolate Linkers for Propylene/Propane and n-Butane/i-Butane Separations. *Angew. Chemie - Int. Ed.* **2021**, *133*, 10191–10199.

(10) Kenyotha, K.; Chanapattharapol, K. C.; McCloskey, S.; Jantaharn, P. Water Based Synthesis of ZIF-8 Assisted by Hydrogen Bond Acceptors and Enhancement of CO2 Uptake by Solvent Assisted Ligand Exchange. *Crystals* **2020**, *10*, 1–23.

(11) Cho, K. Y.; An, H.; Do, X. H.; Choi, K.; Yoon, H. G.; Jeong, H. K.; Lee, J. S.; Baek, K. Y. Synthesis of Amine-Functionalized ZIF-8 with 3-Amino-1,2,4-Triazole by Postsynthetic Modification for Efficient CO2-Selective Adsorbents and Beyond. *J. Mater. Chem. A* **2018**, *6*, 18912–18919.

(12) Ding, R.; Zheng, W.; Yang, K.; Dai, Y.; Ruan, X.; Yan, X.; He, G. Amino-Functional ZIF-8 Nanocrystals by Microemulsion Based Mixed Linker Strategy and the Enhanced CO2/N2 Separation. *Sep. Purif. Technol.* **2020**, *236*, 116209.

(13) Chaplais, G.; Fraux, G.; Paillaud, J. L.; Marichal, C.; Nouali, H.; Fuchs, A. H.; Coudert, F. X.; Patarin, J. Impacts of the Imidazolate Linker Substitution (CH 3 , Cl or Br) on the Structural and Adsorptive Properties of ZIF-8. *J. Phys. Chem. C* **2018**, *122*, 26945–26955.

(14) Krokidas, P.; Moncho, S.; Brothers, E. N.; Castier, M.; Economou, I. G. Tailoring the Gas Separation Efficiency of Metal Organic Framework ZIF-8 through Metal Substitution: A Computational Study. *Phys. Chem. Chem. Phys.* **2018**, *20*, 4879–4892.

(15) Krokidas, P.; Moncho, S.; Brothers, E. N.; Castier, M.; Jeong, H. K.; Economou, I. G. On the Efficient Separation of Gas Mixtures with the Mixed-Linker Zeolitic-Imidazolate Framework-7-8. *ACS Appl. Mater. Interfaces* **2018**, *10*, 39631–39644.

(16) Krokidas, P.; Moncho, S.; Brothers, E. N.; Economou, I. G. Defining New Limits in Gas Separations Using Modified ZIF Systems. *ACS Appl. Mater. Interfaces* **2020**, *12*, 20536–20547.

(17) Becke, A. D. Density-functional Thermochemistry. III. The Role of Exact Exchange. *J. Chem. Phys.* **1993**, *98*, 5648–5652.

(18) Lee, C.; Yang, W.; Parr, R. Development of the Colle- Salvetti Correlation Energy Formula into a Functional of the Electron Density. *Phys Rev B* **1988**, *37*, 785–789.

(19) Martin, M. G.; Siepmann, J. I. Transferable Potentials for Phase Equilibria. 1. United-Atom Description of n -Alkanes. *J. Phys. Chem. B* **1998**, *102*, 2569–2577.

(20) Velioglu, S.; Keskin, S. Simulation of H2/CH4 Mixture Permeation through MOF Membranes Using Non-Equilibrium Molecular Dynamics. *J. Mater. Chem. A* **2019**, *7*, 2301–2314.

(21) Talu, O.; Myers, A. L. Molecular Simulation of Adsorption: Gibbs Dividing Surface and Comparison with Experiment. *AIChE J.* **2001**, *47*, 1160–1168.

(22) Potoff, J. J.; Siepmann, J. I. Vapor-Liquid Equilibria of Mixtures Containing Alkanes, Carbon Dioxide, and Nitrogen. *AIChE J.* **2001**, *47*, 1676–1682.

(23) Krokidas, P.; Castier, M.; Moncho, S.; Brothers, E.; Economou, I. G. Molecular Simulation Studies of the Diffusion of Methane, Ethane, Propane, and Propylene in ZIF-8. *J. Phys. Chem. C* **2015**, *119*, 27028–27037.

(24) Theodorou, D. N.; Snurr, R. Q.; Bell, A. T. Molecular Dynamics and Diffusion in Microporous Materials. In *Comprehensive Supramolecular Chemistry*; Ablerti, G., Bein, T., Eds.; Pergamon: Oxford, 1996; pp 507–548.

(25) Krokidas, P.; Castier, M.; Moncho, S.; Sredojevic, D. N.; Brothers, E. N.; Kwon, H. T.; Jeong, H. K.; Lee, J. S.; Economou, I. G. ZIF-67 Framework: A Promising New Candidate for Propylene/Propane Separation. Experimental Data and Molecular Simulations. *J. Phys. Chem. C* **2016**, *120*, 8116–8124.

(26) Verploegh, R. J.; Nair, S.; Sholl, D. S. Temperature and Loading-Dependent Diffusion of Light Hydrocarbons in ZIF-8 as Predicted Through Fully Flexible Molecular Simulations. *J. Am. Chem. Soc.* **2015**, *137*, 15760–15771.

(27) Kästner, J. Umbrella Sampling. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2011**, *1*, 932–942.

(28) Zhang, C.; Lively, R. P.; Zhang, K.; Johnson, J. R.; Karvan, O.; Koros, W. J. Unexpected Molecular Sieving Properties of Zeolitic Imidazolate Framework-8. *J. Phys. Chem. Lett.* **2012**, *3*, 2130–2134.

(29) *Perry's Chemical Engineers' Handbook*, 9th editio.; Green, D. W., Southard, M. Z., Eds.; McGraw-Hill Education: New York, 2019.

(30) Hub, J. S.; De Groot, B. L.; Van Der Spoel, D. G_wham - a Free Weighted Histogram Analysis Implementation Including Robust Error and Autocorrelation Estimates. *J. Chem. Theory Comput.* **2010**, *6*, 3713–3720.

(31) Batsanov, S. S. Van Der Waals Radii of Elements. *Inorg. Mater.* **2001**, *37*, 871–885.

(32) Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*; Springer, 2016.

(33) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery Jr., J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, Ö.; Foresman, J. B.; Ortiz, J. V;

Cioslowski, J.; Fox, D. J. Gaussian 09, Revision D.01. *Gaussian Inc.* Wallingford, CT 2009, p Wallingford CT.

(34) Hess, B.; Kutzner, C.; Van Der Spoel, D.; Lindahl, E. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *J. Chem. Theory Comput.* **2008**, *4*, 435–447.

(35) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. UCSF Chimera - A Visualization System for Exploratory Research and Analysis. *J. Comput. Chem.* **2004**, *25*, 1605–1612.

(36) Rossum, V.; Drake, F. L. Python 3 Reference Manual. Create Space: Scotts Valley, CA 2009.

(37) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, É. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.