Electronic Supplementary Material (ESI) for Journal of Materials Chemistry A.

Electronic Supplementary Information for

ReDD-COFFEE: A ready-to-use database of covalent organic framework structures and accurate force fields to enable high-throughput screenings

Juul S. De Vos,[†] Sander Borgmans,[†] Pascal Van Der Voort,[‡] Sven M. J. Rogge,[*,†] and Veronique Van Speybroeck[*,†]

[†]*Center for Molecular Modeling (CMM), Ghent University, Technologiepark-Zwijnaarde 46, 9052 Zwijnaarde, Belgium, and*

[‡]*Centre for Ordered Materials, Organometallics and Catalysis (COMOC), Ghent University, Krijgslaan 281 (S3), 9000 Ghent, Belgium*

Sven.Rogge@UGent.be; Veronique.VanSpeybroeck@UGent.be

[†]Ghent University
[‡]Ghent University

S1 Database construction

In this section of the supporting information, additional details about the construction of the ReDD-COFFEE database is given. A concise overview of the used SBUs and topologies, and how these are obtained is given in Sections S1.1 and S1.2, respectively. The structure assembly protocol in Section 2.2 of the main text is defined for a given (topology, SBUs) combination. Prior to this step, a list of possible trial combinations has to be defined, which is done in Section S1.3. During the structure assembly, close attention has to be paid to the positioning of the edges, especially when the neighboring vertices are decorated with SBUs with a different size. This is described in Section S1.4. The three criteria that are used to filter out unphysical structures use different thresholds. The values of these thresholds are discussed in Section S1.5. Finally, the structure assembly of COF-108 is discussed in detail in Section S1.6 as an example.

S1.1 Overview of secondary building units (SBUs)

The 268 687 COF structures from the ReDD-COFFEE database are assembled using 279 SBUs. Each SBU consists of a linker core, chosen from those depicted in Fig. S1, and a linkage section. The used linkages and the chemical routes to form them are defined in Fig. S2. A linkage section is uniquely characterized by the reactive group it originates from and the resulting linkage (see Fig. 1 of the main text). Therefore, each linkage section is labeled with two numbers, indicating the reactive group and resulting linkage, respectively. For example, the linkage section 02-04 originates from an aldehyde reactive group (reac02) that forms an imine linkage (link04). Labeling a linkage section only with the reactive group or the linkage would not result in a unique nomenclature, since one reactive group can form multiple linkages (*e.g.*, aldehyde can also form an (acyl)hydrazone linkage) and the synthesis of one linkage usually involves two different reactive groups (*e.g.*, an imine linkage originates from the reaction of aldehyde and amine reactive groups). To accurately mimic the environment of the SBU, a proper termination has to be introduced when deriving the cluster force fields. We have chosen to define a single termination for each linkage section, and have adopted the same nomenclature for the termination as for the linkage section. An overview of the used terminations is given in Fig. S3. Finally, the SBU names are assembled from the numbers that indicate their linker core (single number) and linkage sec-

tion (two numbers), respectively. Each SBU has only one type of linkage section. SBU 01-02-04 is therefore assembled by combining core01 with termination 02-04.

As can be observed from Figs. S1 and S3, not all linker cores can be matched with each termination. Only when the number of anchoring points are the same, a cluster can be obtained. For example, the termination 02-04 cannot be combined with an anthracene linker core (core10), since it only has one anchoring point, whereas the anthracene linker core requires two. Furthermore, the four linkages that emerge during synthesis, *i.e.*, boroxine (link02), triazine (link10), borazine (link11), and borosilicate (link03), are all terminated using a single phenyl ring. Their cores, *i.e.*, core31, core32, core33, and core34, respectively, are technically speaking no linker cores, since they have no experimental precursor. Nonetheless, they are depicted in Fig. S1 to have a concise overview of the used SBU cores. To use a similar nomenclature as the other SBUs, their labels are defined as if they would originate from an artificial reactive group (reac11). As such, the phenyl terminated boroxine ring is defined as 31-11-02, *i.e.*, SBU core 31 with termination 11-02 that forms a boroxine linkage (link02). Since the phenyl termination does not mimic the linkage environment of other linkers, it can only be adopted for core31 to core34 and can not be combined with other linker cores.

In total, there are 17 linker cores with a single anchoring point per reactive group, which can be matched with one of 14 terminations (*i.e.*, terminations 01-01, 01-02, 01-03, 02-04, 03-04, 02-05, 04-05, 02-06, 03-07, 02-08, 03-09, 05-09, 06-10, and 07-11). The remaining 13 linker cores have two anchoring points and can combine with 3 terminations (*i.e.*, terminations 08-01, 09-07, and 10-08). Together with the four trigonal linkages that emerge during synthesis, a total of 281 cluster force fields are derived using the QuickFF procedure[1,2] outlined in the main text. As the force field optimized structures of SBUs 12-05-09 and 12-06-10 largely deviate from the *ab initio* optimized structures, they are omitted from the set of used SBUs. Therefore, a final set of 279 SBUs is used as input of the structure assembly protocol.

S1.2 Extraction of topologies from the RCSR database

All 2D and 3D topologies are extracted from the Reticular Chemistry Structure Resource (RCSR) database[3] using a web scraping script. As explained in the main text, only those topologies with

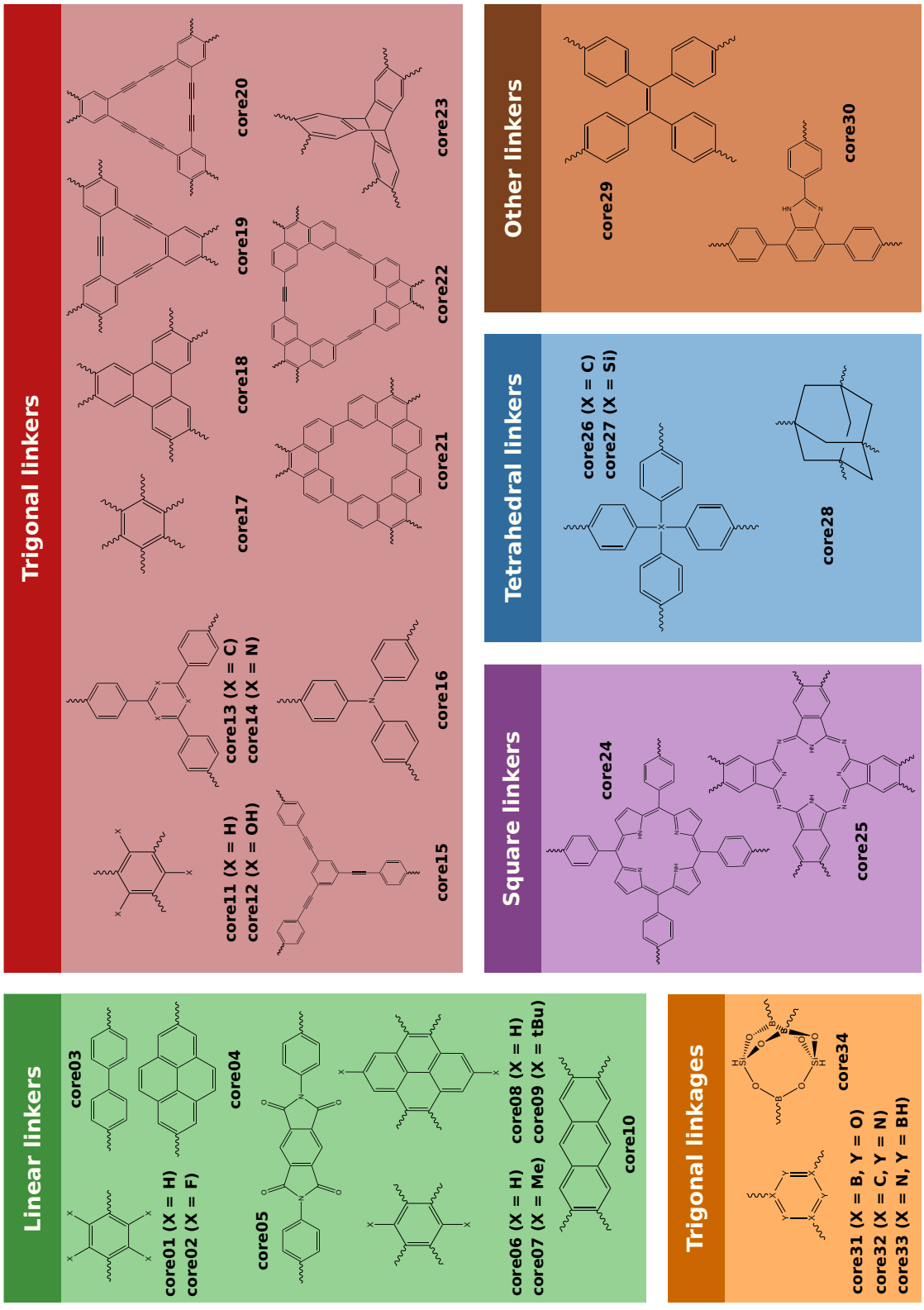


Figure S1: Classification of the linker cores included in the ReDD-COFFEE database. The linker cores are divided in linear, trigonal, square, tetrahedral, and other linkers. To give a concise overview of all SBUs, also the trigonal linkages that emerge during synthesis are visualized here. These are technically speaking no linker cores since they do not originate from an experimental precursor.

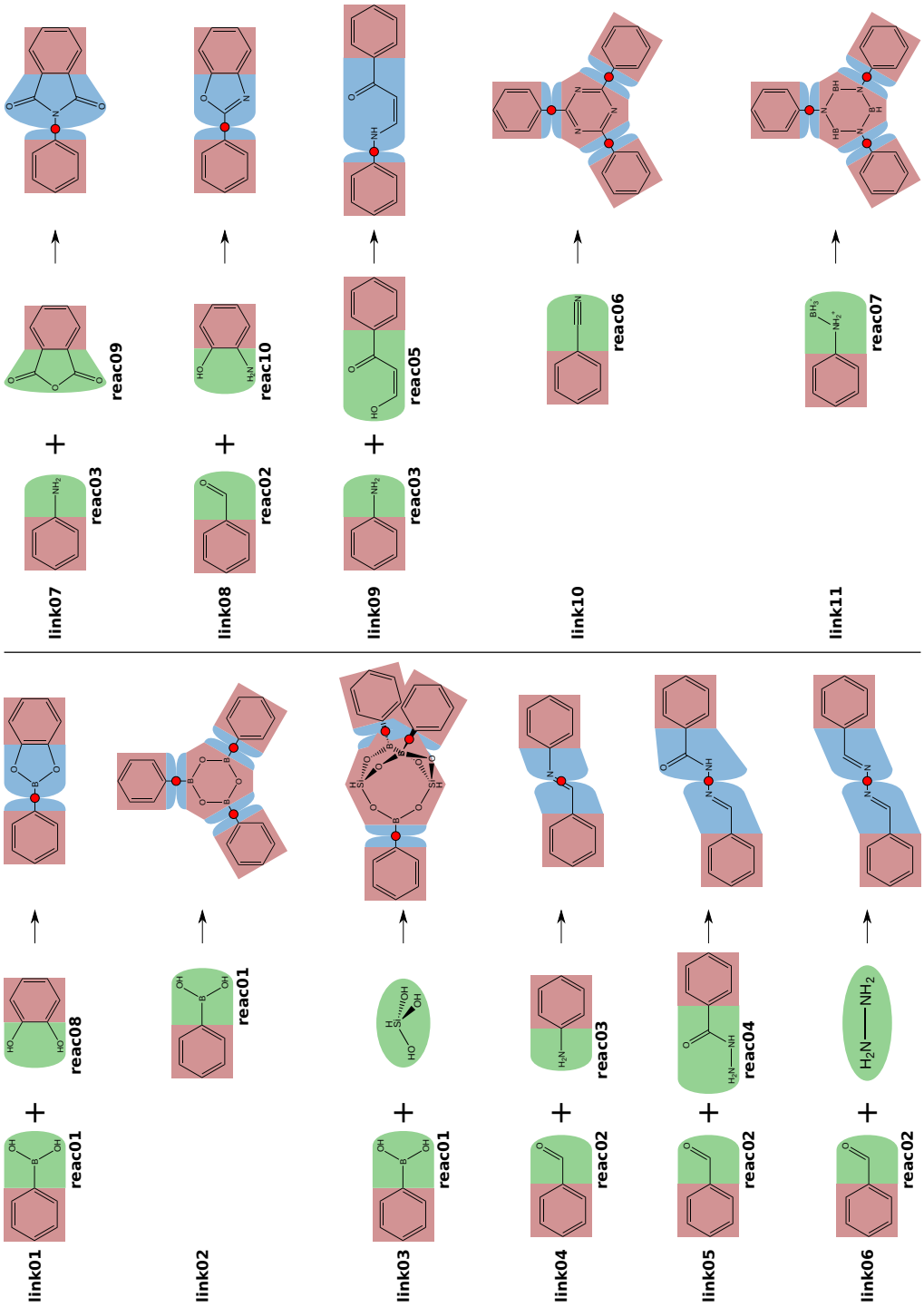


Figure S2: Classification of the 11 COF linkages included in the ReDD-COFFEE database (link01 to link11) and the 10 reactive groups from which they are formed (reac01 to reac10). The linkages are boronate ester (link01), boroxine (link02), borosilicate (link03), imine (link04), (acyl)hydrazone (link05), azine (link06), imide (link07), oxazoline (link08), (keto)enamine (link09), triazine (link10), and borazine (link11). The reactive groups are boronic acid (reac01), aldehyde (reac02), amine (reac03), hydrazide (reac04), β-ketoenol (reac05), nitrile (reac06), amine borane (reac07), catechol (reac08), anhydride (reac09), and aminophenol (reac10). The silane triol and hydrazine reagents used in the synthesis of borosilicate and azine linked COFs are not labeled, since they are not attached to a linker core.

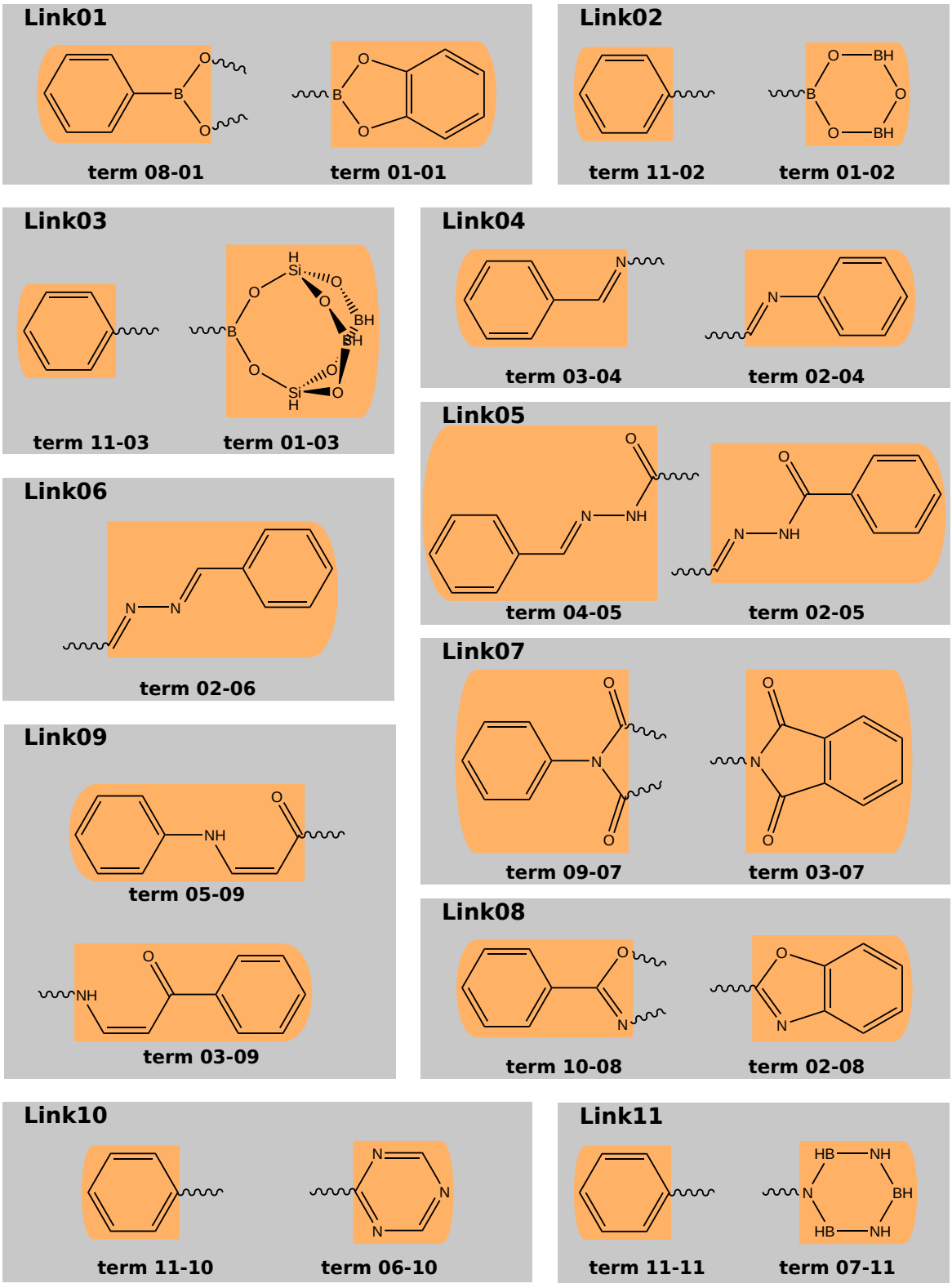


Figure S3: Classification of the 21 terminations included in the ReDD-COFFEE database, ordered per linkage type for which they can occur. All terminations are characterized using two hyphen-separated numbers, representing the reactive group of the originating precursor and the resulting linkage, respectively. Reac11 is an artificial precursor used for the linkages that emerge during synthesis.

embedding type 1 are retained. Furthermore, as all SBUs possess at most four points of extension, this is also the maximally allowed coordination number in each topology. Each topology is validated by calculating the coordination sequence and the vertex symbol of all vertices and automatically comparing it with the ones mentioned in the RCSR. If a topology could not be correctly initialized due to an invalid entry in the RCSR database, the topology was manually corrected. Following this approach, 2 495 topologies are extracted from the RCSR, among which 1 272 only contain vertices with coordination number three or four (next to edges that have coordination number two). These are the ones retained to extract our ReDD-COFFEE database.

S1.3 Initial enumeration of possible (topology, SBUs) combinations

The total number of (topology, SBUs) combinations that can be obtained for a specific topology is the product of the number of SBUs that can be placed on each individual Wyckoff set. However, as the number of Wyckoff sets in a topology can become very large, also the total number of (topology, SBUs) combinations increases rapidly. To limit the number of combinations to a computationally feasible amount, a combinatorial approach is followed for every linkage type, which is illustrated in Fig. S4 for the **nka** topology.

For each Wyckoff set, the SBUs that can be placed on it are selected. Only the SBUs that have a linkage section corresponding to the predefined linkage type and that have the same number of points of extension as the coordination number of the Wyckoff set qualify. For edges, also the option to not include a linker is allowed. In the example in Fig. S4, the **nka** topology consists of seven Wyckoff sets. Two of them are four-connected, another two are three-connected, and there are three edge Wyckoff sets. In this example, we consider two different SBUs for the four-connected Wyckoff set, two different SBUs for the three-connected Wyckoff set, and one SBU for the edge Wyckoff set. Including the possibility to leave an edge free, this results in two options per Wyckoff set.

Without any limitation, the total amount of (**nka**, SBUs) combinations would be $2^7 = 128$. These are symbolically visualized in the top left panel of Fig. S4. To limit the number of combinations, we can unify the Wyckoff sets with the same coordination number. When the edge Wyckoff sets are unified, all edges have to be decorated with the same SBU, even when they belong

to separate edge Wyckoff sets. The edge Wyckoff sets of the **nka** topology in the example of Fig. S4 can therefore be occupied in two different ways, instead of $2^3 = 8$ when all Wyckoff sets would be assigned individually. Similarly, when we unify the vertex Wyckoff sets with a given coordination number, all vertices with the same coordination number have to adopt the same SBU. As such, both the four- and three-connected Wyckoff sets can be occupied in two different ways, instead of $2^2 = 4$ each.

To make a list of (topology, SBUs) combinations that are used as input for the structure assembly procedure, four combination sets are consulted for each topology and linkage type. The edge Wyckoff sets and the vertex Wyckoff sets with the same coordination number are either all assigned individually (I.E. and I.V., respectively), or they are unified (U.E. and U.V., respectively). The four combination sets are therefore given by (I.V. + I.E.), (U.V. + I.E.), (I.V. + U.E.), and (U.V. + U.E.), and coincide with the four panels of Fig. S4. Each combination that appears in any combination set for which the total number of combinations is lower than a certain upper limit N_c, is included in the dataset. The total number of combinations in the resulting dataset can be higher than this upper limit, since separate combination sets contain different combinations. The upper limit N_c is chosen here as 10^4 to keep the number of structures within a feasible amount. Out of the listed combinations, only those are retained where the linkage segments combine to form the correct linkage type.

Once a (topology, SBUs) combination is assembled, the resulting structure is labeled as follows: top_SBU$_1$_SBU$_2$_..._SBU$_N$, with SBU$_i$ being the SBU that is placed on the i-th Wyckoff set. The order of the Wyckoff sets is the same as given in the RCSR database,[3] with the edge Wyckoff sets following the vertex Wyckoff sets. If no SBU is assigned to an edge Wyckoff set, this is indicated with “None". As an example, COF-108 is labeled bor_18-08-01_26-01-01_None. The first vertex Wyckoff set of the **bor** topology is decorated with the trigonal SBU 18-08-01 (HHTP), the second with the tetrahedral SBU 26-01-01 (TBPM), and the third Wyckoff set is a vacant edge.

S1.4 Placing the edges at an adequate location between the vertices

In the default topologies, the edges are positioned exactly in the middle between its neighboring vertices. However, the center of a two-connected SBU that has to be placed on it, only coincides

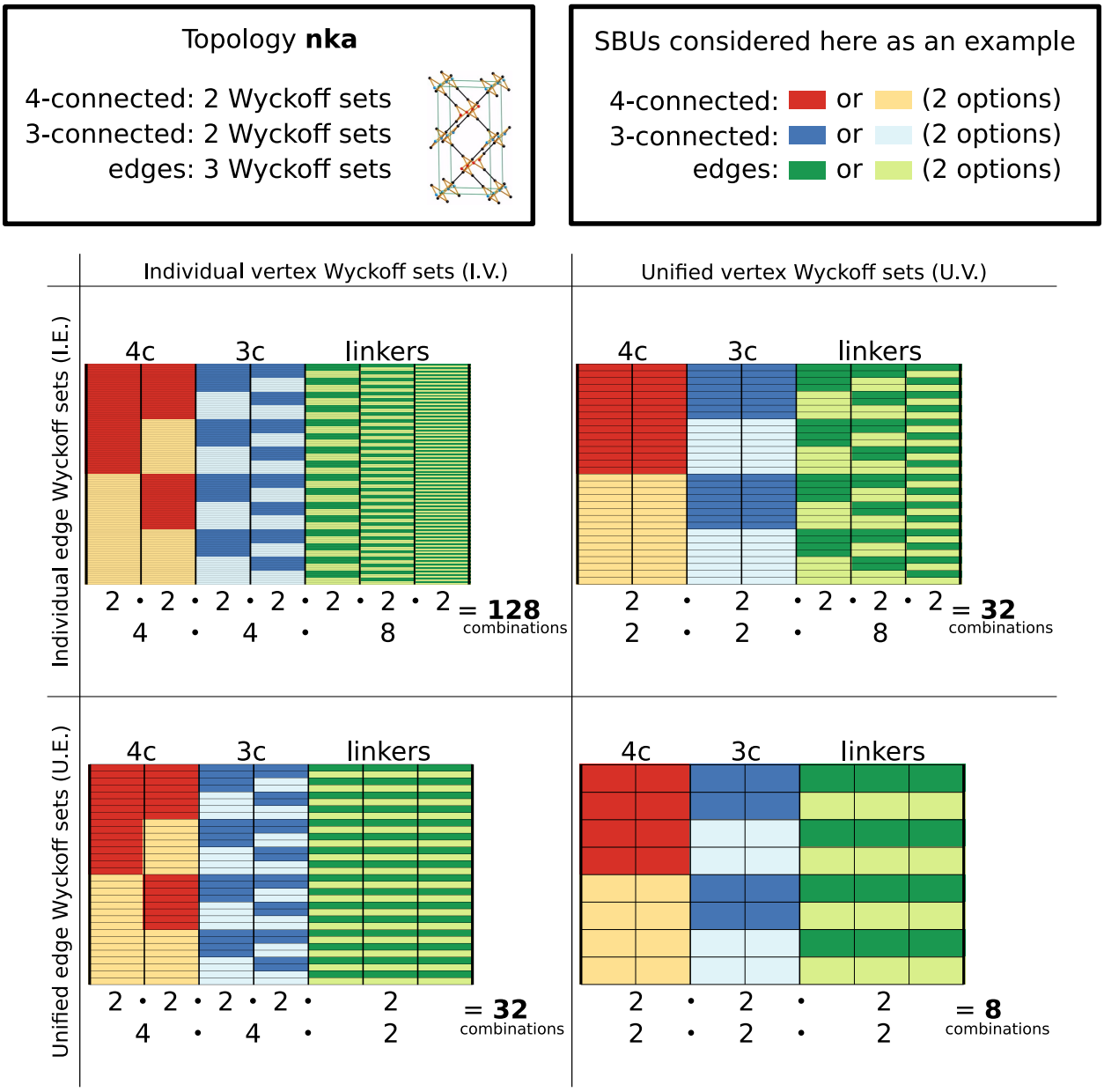


Figure S4: An example of the initial enumeration of (topology, SBUs) combinations for the nka topology. To limit the number of combinations, the vertex Wyckoff sets or edge Wyckoff sets can be unified and assigned the same SBU.

with this position when its neighboring SBUs have the same size, *i.e.*, when $d_{SBU_1} = d_{SBU_3}$ in Fig. S5. In the other cases, it should be shifted towards the vertex to which the smallest SBU is assigned. The optimal center of the linker, and thus the location of the edge in the topology is therefore given by:

$$\begin{aligned}\mathbf{P}(E) &= \mathbf{P}(V_1) + \frac{d_{12}}{d_{13}} \frac{\mathbf{P}(V_3) - \mathbf{P}(V_1)}{||\mathbf{P}(V_3) - \mathbf{P}(V_1)||} \\ &= \mathbf{P}(V_1) + \frac{d_{SBU_1} + d_{SBU_2}}{d_{SBU_1} + 2d_{SBU_2} + d_{SBU_3}} \frac{\mathbf{P}(V_3) - \mathbf{P}(V_1)}{||\mathbf{P}(V_3) - \mathbf{P}(V_1)||}\end{aligned} \tag{S1.1}$$

in which $\mathbf{P}(N)$ denotes the position of node N, which can either be the edge E or one of the vertices V_1 or V_3 to which SBU_1 and SBU_3 are assigned, respectively. All other variables are defined in Fig. S5. The positions of all edges are relocated immediately after the rescaling of the topology in Step 1 of the structure assembly procedure.

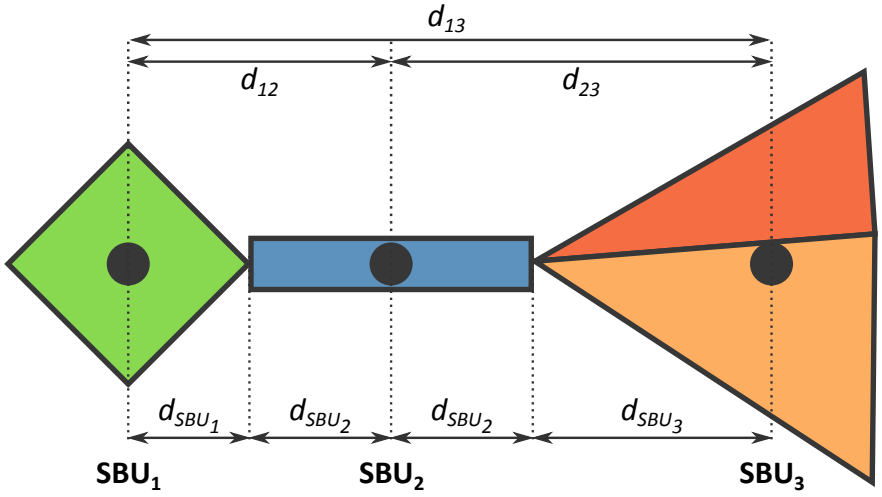


Figure S5: Example of a relocation of the edge position in order to fit SBU_2 properly between its neighboring SBUs. As SBU_1 and SBU_3 are not equally large, the edge should be translated towards the vertex on which SBU_1 is placed.

S1.5 Selection criteria

Each of the structures in the list of possible (topology, SBUs) combinations is used as input for the structure assembly procedure as described in Section 2.2 of the main text. However, not all these combinations result in a structure that is added to the ReDD-COFFEE database. The

several filters that are used to discard combinations from the database are listed below and the total number of combinations and topologies that are retained after each step are listed in Table S1. Each of the three introduced thresholds are discussed in the following paragraphs.

- **After Step 0**: An initial enumeration of matching combinations is performed as explained in Section S1.3 with the maximum number of (topology, SBUs) combinations for each (topology, linkage type) pair being 10^4.
- **After Step 1**: Only those (topology, SBUs) combinations for which an isotropic rescaling is possible are retained. This coincides with Filter I in Fig. 2 of the main text.

$$\sigma_f < \sigma_{f,\max} = 0.22 \text{ Å/l.u.} \tag{S1.2}$$

  The arbitrary length unit l.u. expresses the dimension of the unit cell vectors in the RCSR.

- **After Step 2**: Combinations for which the geometric mismatch is too high are rejected, as indicated by Filter II in Fig. 2 of the main text.

$$\text{RMSD} < \text{RMSD}_{\max} = 0.11 \text{ Å} \tag{S1.3}$$

- **After Step 4a**: As there is a large freedom in possible topologies that are used in the structure assembly protocol, there is no limit on the number of atoms or the unit cell volume of the resulting material. Therefore, we only attempted to optimize structures that satisfy

$$N_{\text{atom}} \leq 10\,000 \tag{S1.4}$$

$$V_{\text{init}} \leq 10\,000 \text{ nm}^3 \tag{S1.5}$$

- **After Step 4b**: Not for all structures a minimum in the potential energy surface was found in a reasonable time. The database only contains those structures for which the optimization was successfully converged with the default convergence criteria as implemented in Yaff.[4]
- **After Step 4c**: As implemented by Filter III in Fig. 2 of the main text, only those relaxed

materials for which the strain in the structure does not deform the SBUs to a large extent are finally added to the ReDD-COFFEE database.

$$E_{\text{def}} < E_{\text{def, max}} = 14 \text{ kJ/mol} \tag{S1.6}$$

Filter I: rescaling standard deviation σ_f

As the topology has to be rescaled isotropically in our approach, the rescaling factors f_i calculated for each edge Wyckoff set have to be close to one another. If this is not the case, SBUs can start to overlap, as is illustrated in Fig. S6. Therefore, we require the standard deviation σ_f of all calculated rescaling factors to be below the threshold $\sigma_{f,\text{max}}$. In Fig. S7, a histogram of the standard deviation of the rescaling factors over all 5 537 951 initial (topology, SBUs) combinations is shown, together with a detail in the range of 0 to 1 Å/l.u.

As can be observed, most of the combinations have a rescaling factor standard deviation close to zero. Furthermore, the SBUs quickly start to overlap when the standard deviation of the rescaling factors increases. For instance, the structure depicted in Fig. S6 already shows overlap for a rescaling standard deviation σ_f of 1.06 Å/l.u. To include only structures in which the SBUs fit almost perfectly in the topology, the threshold $\sigma_{f,\text{max}}$ is chosen to be 0.22 Å/l.u.

Filter II: largest root-mean-square deviation RMSD

To avoid combinations that result in too large a geometric mismatch in the database, a threshold is defined for the largest root-mean-square deviation (RMSD) that is allowed when introducing an SBU in a topology. However, some SBUs possess a large degree of flexibility and can tolerate such geometric mismatches. For example, COF-108 is one of the earliest synthesized COFs for which the largest RMSD is as high as 0.09 Å, as illustrated in Section S1.6. In Fig. S8, the distribution of the largest RMSD is plotted for all 749 859 (topology, SBUs) combinations that have passed the first filter. The threshold of RMSD_{max} = 0.11 Å is chosen such that the experimentally observed COF-108 and the peak observed in Fig. S8 around 0.10 Å are included in the database. An example of a structure that is rejected from the database is visualized in Fig. S9. In this

Table S1: Overview of the total number of (topology, SBUs) combinations and resulting structures that are retained after each step in the structure assembly procedure. A distinction is made between all linkages and the dimensionality of the topologies. Both the number of structures and topologies (between brackets) are reported. All structures retained after Step 4c are deposited in the ReDD-COFFEE database.

Linkage type	Dimensionality	Step 0	Step 1	Step 2	Step 4a	Step 4b	**Step 4c**
Boronate Ester	2D	37024 (124)	6779 (107)	3853 (105)	3808 (105)	1143 (88)	**767 (80)**
	3D	362846 (1148)	87892 (1147)	45589 (985)	38962 (917)	36673 (916)	**33600 (898)**
Boroxine	2D	1930 (56)	279 (36)	241 (35)	241 (35)	129 (33)	**70 (21)**
	3D	13908 (330)	2694 (285)	2477 (279)	2364 (269)	2182 (269)	**1687 (243)**
Triazine	2D	5020 (56)	360 (36)	324 (34)	324 (34)	245 (34)	**148 (22)**
	3D	34248 (330)	3589 (285)	3378 (279)	3251 (269)	3119 (269)	**2511 (236)**
Borosilicate	2D	1930 (56)	279 (36)	240 (35)	240 (35)	133 (32)	**124 (30)**
	3D	13908 (330)	2676 (285)	2459 (279)	2306 (264)	2166 (264)	**1999 (246)**
Imine	2D	32580 (124)	7373 (107)	3741 (83)	3709 (83)	1881 (82)	**531 (60)**
	3D	275994 (1148)	93806 (1147)	52974 (992)	45452 (921)	42941 (921)	**20839 (845)**
(Acyl)hydrazone	2D	32580 (124)	7386 (106)	3948 (89)	3894 (89)	1267 (87)	**229 (56)**
	3D	275994 (1148)	95602 (1147)	55118 (1073)	45332 (962)	42681 (962)	**39183 (962)**
Borazine	2D	1930 (56)	277 (36)	240 (34)	240 (34)	199 (34)	**164 (30)**
	3D	13908 (330)	2695 (285)	2478 (279)	2351 (266)	2302 (266)	**2150 (263)**
(Keto)enamine	2D	30365 (124)	6779 (106)	3437 (82)	3372 (82)	909 (79)	**830 (77)**
	3D	260492 (1148)	88949 (1147)	48650 (971)	39408 (865)	38050 (865)	**37910 (864)**
Azine	2D	337188 (124)	10118 (109)	4551 (82)	4524 (82)	1697 (81)	**952 (68)**
	3D	3006366 (1148)	142794 (1148)	70827 (1012)	62003 (973)	59354 (971)	**54356 (932)**
Imide	2D	37024 (124)	6795 (107)	3855 (105)	3800 (105)	1463 (91)	**1096 (82)**
	3D	362846 (1148)	88034 (1147)	45742 (987)	38640 (902)	36817 (902)	**35866 (900)**
Oxazoline	2D	37024 (124)	6783 (107)	3850 (105)	3805 (105)	1553 (91)	**945 (78)**
	3D	362846 (1148)	87920 (1147)	45609 (985)	39029 (917)	37005 (917)	**32730 (899)**
Total	2D	554595 (124)	53208 (109)	28280 (106)	27957 (106)	10619 (100)	**5856 (95)**
	3D	4983356 (1148)	696651 (1148)	375301 (1090)	319098 (1041)	303290 (1039)	**262831 (1019)**

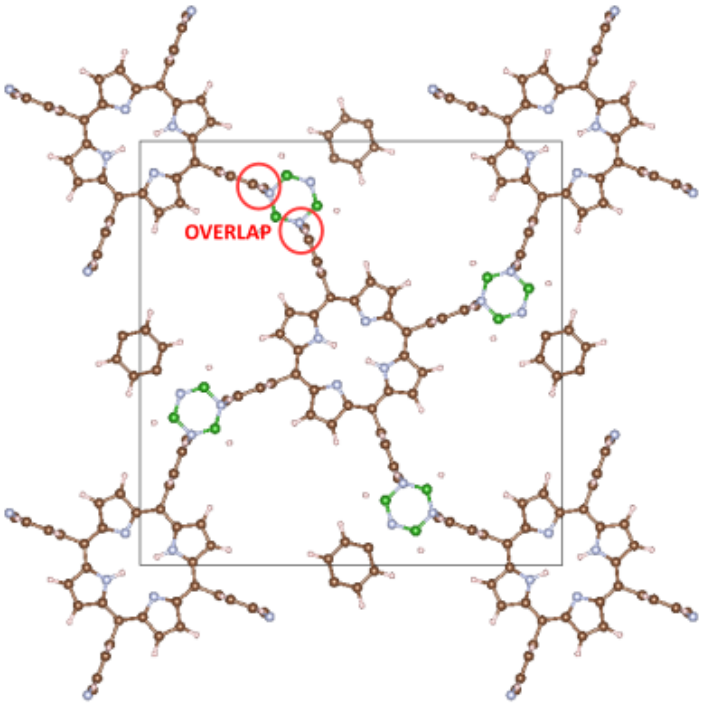


Figure S6: Atomistic structure of mcm_33-11-11_24-07-11_None_01-07-11. Due to the large rescaling standard deviation σ_f = 1.06 Å/l.u., the SBUs overlap even after rescaling using the average rescaling factor.

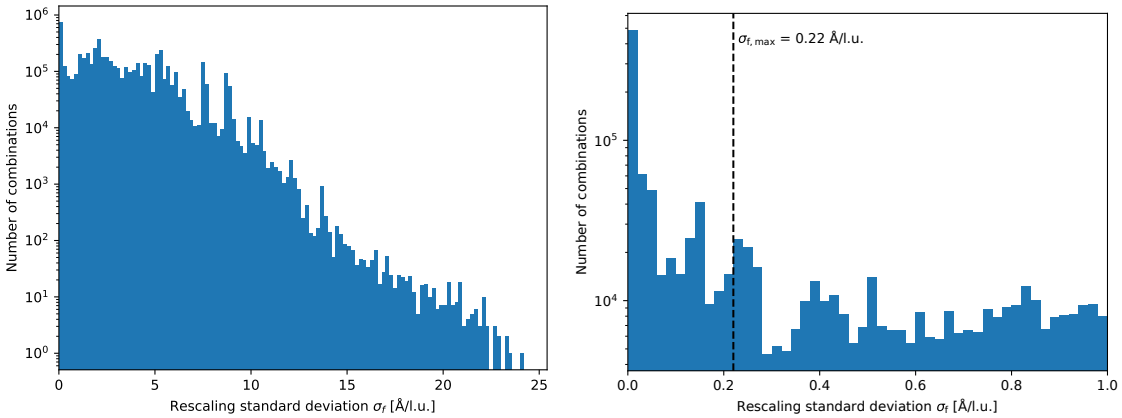


Figure S7: Distribution of rescaling factor standard deviations σ_f over all 5 537 951 combinations. Left: the full distribution. Right: detail of the distribution in the range between 0 and 1. The threshold $\sigma_{f,\max}$ is indicated with a dashed line.

example, the tetrahedral SBU 27-02-06 is placed on the square vertex of the 2D **sql** topology, which results in a RMSD of 0.16 Å. Despite the fact that the tetrahedral SBU possesses quite some flexibility, it would require too much energy to adopt a planar configuration, as required by the topology.

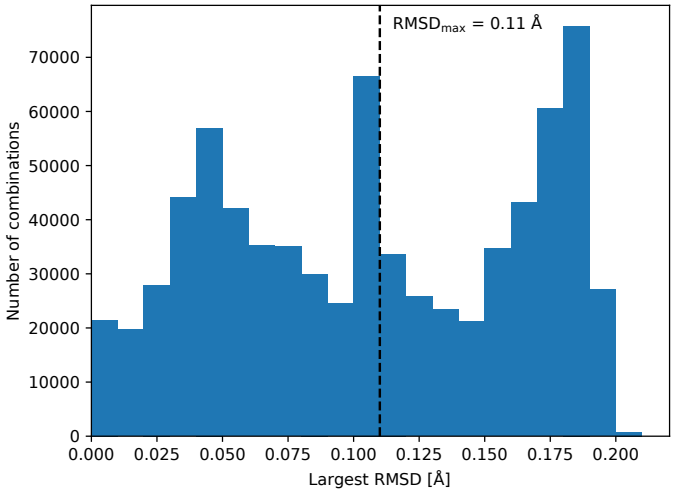


Figure S8: Distribution of the largest root-mean-square deviations of the 749 859 combinations that passed filter 1. The threshold $\mathrm{RMSD}_{\mathrm{max}}$ is indicated with a dashed line.

Filter III: Total deformation energy E_{def}

The second filter introduced in the structure assembly process allows some deviation between the SBU positions and the location of the topological nodes, as COF building blocks can have a large degree of flexibility. However, there are also SBUs that are quite rigid. To check that all SBUs were able to find a low-energy configuration during the optimization, a last filter is defined for the deformation energy E_{def}. After visually inspecting several optimized structures, we decided that SBUs in structures with a deformation energy exceeding 14 kJ/mol are largely deformed and unphysical. Therefore, these materials have a very low synthetic likelihood and are removed from the ReDD-COFFEE database. In Fig. S10, the atomistic structures of eight COFs are plotted, and their deformation energy is interpreted.

A histogram of the deformation energy of all 313 909 optimized structures is provided in Fig. S11, together with a kernel density distribution for each of the linkage types. As can be observed, only a minority of the optimized structures (14.41%) are discarded from the database. In contrast

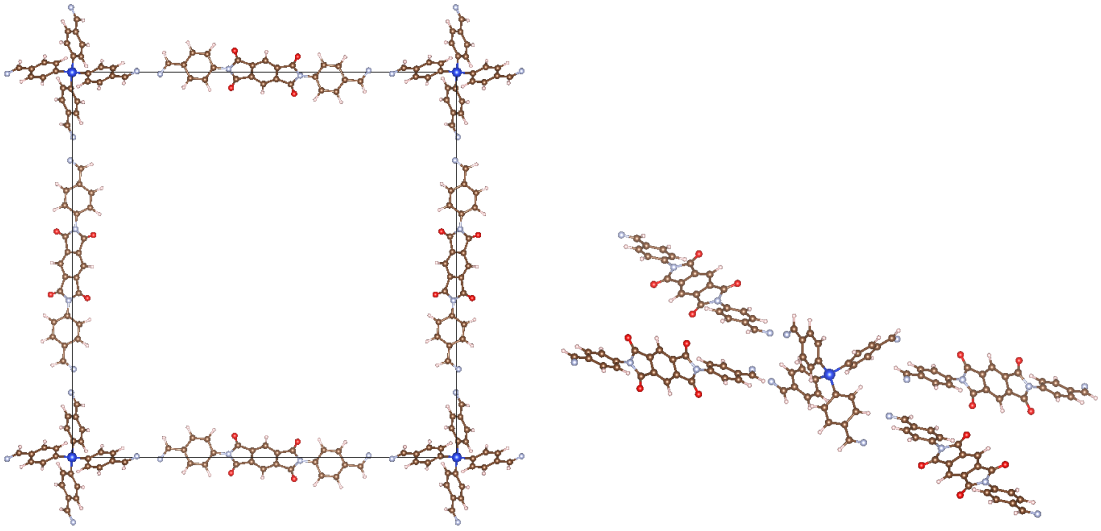
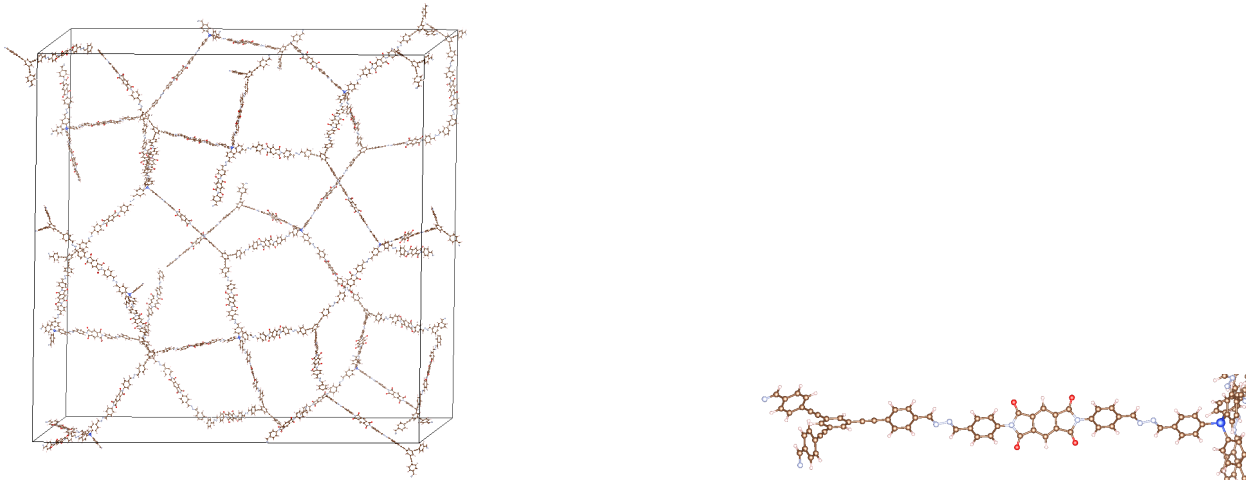


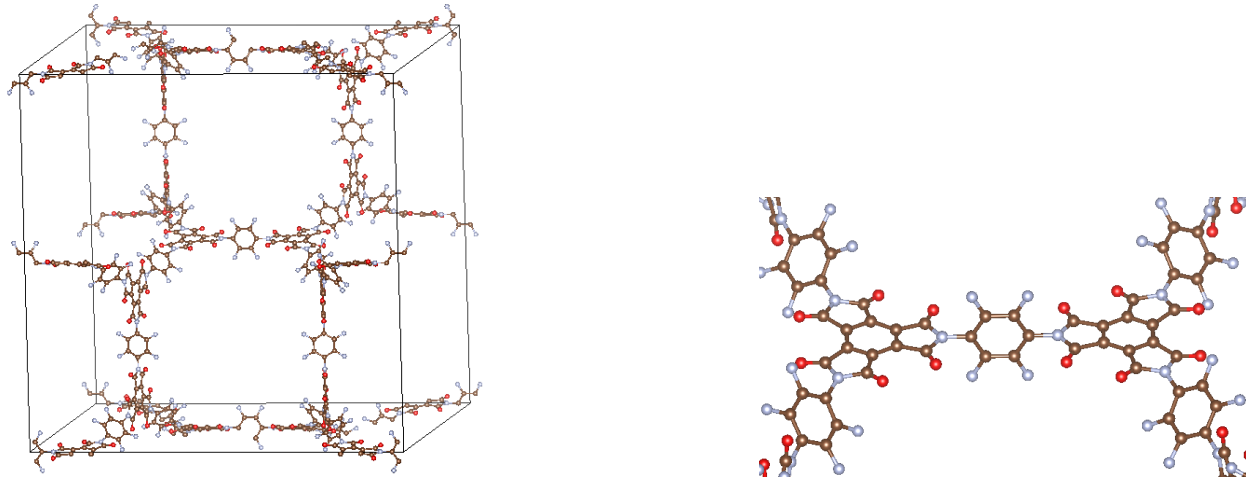
Figure S9: Atomistic structure of sql_27-02-06_05-02-06 with a maximum RMSD of 0.16 Å. This large RMSD originates from trying to place the tetrahedral SBU 27-02-06 on a square vertex in the **sql** topology. Left: unit cell of the structure viewed along the c-axis. Right: detail of the structure with the tetrahedral SBU and its direct environment.

to the other linkage types, there is a larger amount of structures with an imine linkage that are rejected.

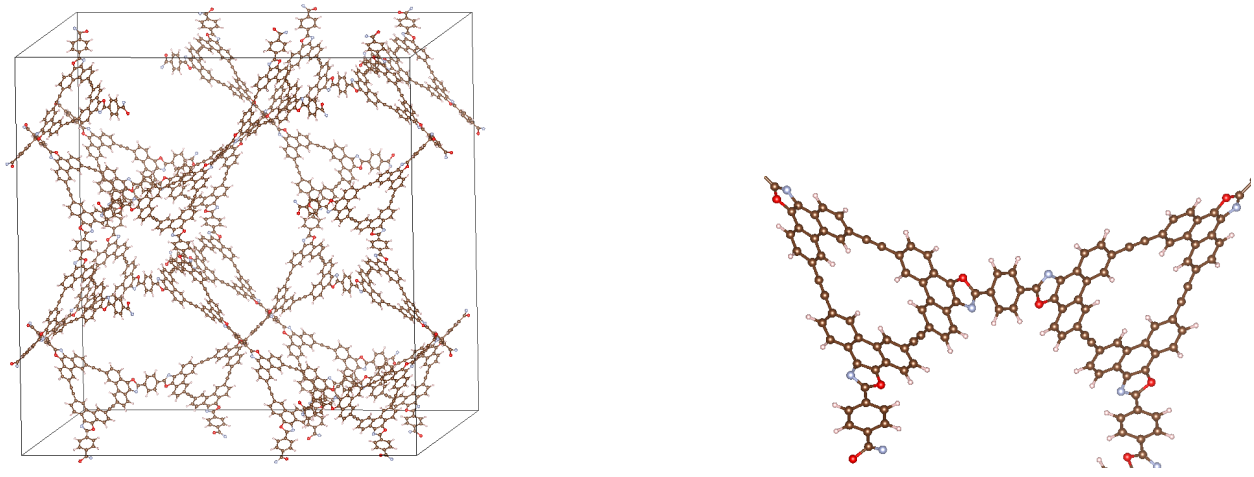
As visualized in Fig. S12, a suboptimal placement of the SBUs in an imine COF can relax towards an unphysical imine configuration, which is a local minimum on the potential energy surface. These suboptimal SBU configurations can occur when there is a geometric mismatch between the SBUs and the topological nodes. In these cases, the points of extension of neighboring SBUs do not overlap, which has a large influence for the energetic considerations of the structure assembly process. However, despite a suboptimal initial structure, most linkages optimize towards a physical configuration. Only for the imine linkage, the aldehyde hydrogen can interfere with the amine nitrogen, as these are non-bonded but closely placed together, resulting in the unphysical configuration. During molecular dynamics (MD) simulations at elevated temperatures, the system quickly leaves this suboptimal configuration and finds a more physical imine configuration. While a workaround for these suboptimal configurations would therefore be to perform a short MD run before the final optimization stage, as suggested by Ongari *et al.* while developing the CURATED database,[5] we chose not to include this MD run for computational efficiency.



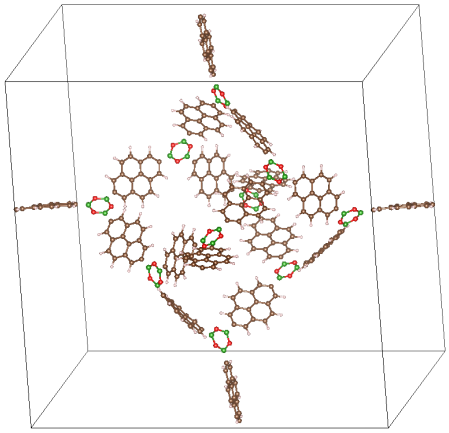
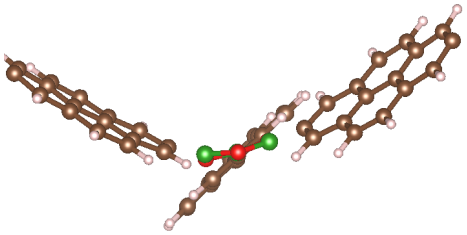
(a) ctn_15-02-06_27-02-06_05-02-06 (E_{def} = 0.14 kJ/mol): a structure without geometric mismatch between the trigonal and tetrahedral building blocks and the vertices. They are nicely connected with the linear linker.



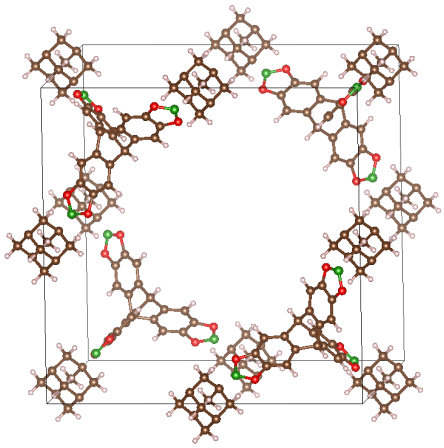
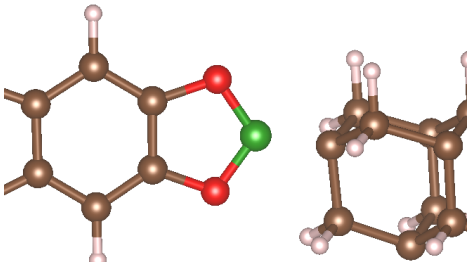
(b) lig_17-09-07_02-03-07_02-03-07 (E_{def} = 5.04 kJ/mol): no geometric mismatch between the trigonal building blocks and the vertices.



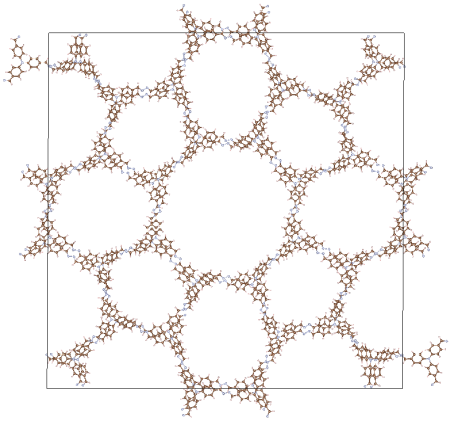
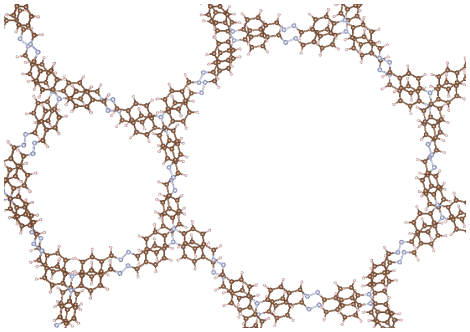
(c) srs-c3_22-10-08_01-02-08_01-02-08 (E_{def} = 10.91 kJ/mol): the trigonal SBUs fit nicely on the vertices of the topology. However, the dihedral angle between the two-connected linker and the oxazoline linkage deviates from zero, whereas it preferably relaxes to a planar configuration. Therefore, the deformation energy E_{def} increases, but is still acceptable.

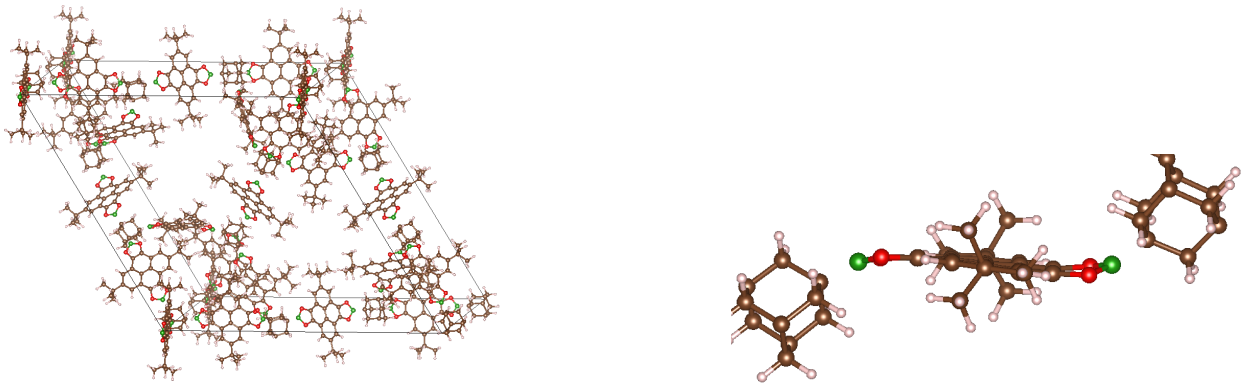
(d) bod_31-11-02_31-11-02_04-01-02_04-01-02 (E_{def} = 16.13 kJ/mol): the trigonal boroxine ring is placed on a suboptimal vertex. Therefore, the two-connected linkers pull it out of its preferred configuration. The deformation energy increases significantly due to the relatively large contortion of this small building block. Therefore, this structure is rejected.

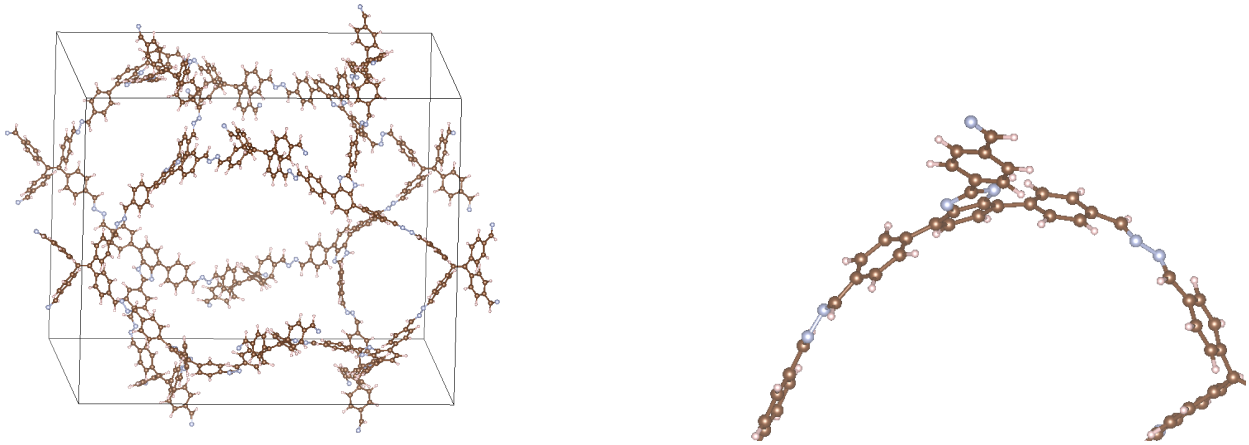
(e) bor_23-08-01_28-01-01_None (E_{def} = 18.07 kJ/mol): whereas the **bor** topology is experimentally observed, it requires flexible building blocks to decorate its vertices to compensate for the introduced geometric mismatch (*cf.* COF-108 in Section S1.6). The triptycene and adamantane cages are too rigid, which results in a high deformation energy. This material is, therefore, discarded from the ReDD-COFFEE database.

(f) ply_16-02-06_16-02-06_16-02-06_16-02-06_16-02-06_None_None_None_None_None_None_None (E_{def} = 19.23 kJ/mol): planar trigonal SBUs preferably assemble in a 2D hexagonal topology. However, in the **ply** topology, they are forced into five- and seven-membered rings, resulting in a larger deformation energy.



(g) bal_28-01-01_28-01-01_09-08-01_09-08-01_09-08-01_09-08-01 (E_{def} = 21.63 kJ/mol): again, the SBUs are placed on vertices with a geometric mismatch. In combination with the rigid adamantane cage, this results in a large deformation of the boronate ester linkages. Therefore, this structure is not allowed in the database.



(h) pto_30-02-06_29-02-06_None (E_{def} = 30.26 kJ/mol): with an increasing geometric mismatch, it becomes more difficult to accommodate for the introduced misfit. Even flexible SBUs are not able to compensate for the suboptimal configurations.

Figure S10: Atomistic configuration of several optimized structures and a detail of their structure. The materials are ordered from low to high deformation energy. On the left, the full unit cell is visualized, while the structure details are provided on the right. Color code: hydrogen (white), boron (green), carbon (brown), nitrogen (blue), oxygen (red), fluorine (blue), silicon (dark blue).

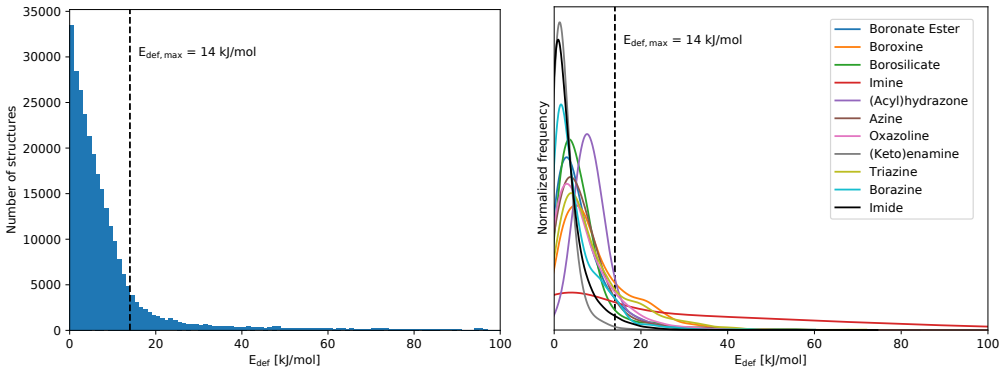


Figure S11: Distribution of deformation energies over the 313 909 optimized structures. Left: histogram of all structures. Right: kernel density of all individual linkages. The threshold deformation energy is indicated with a dashed line.

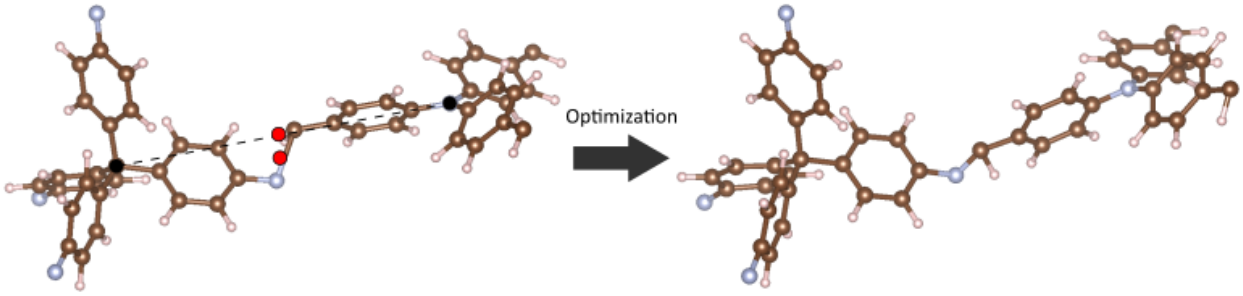
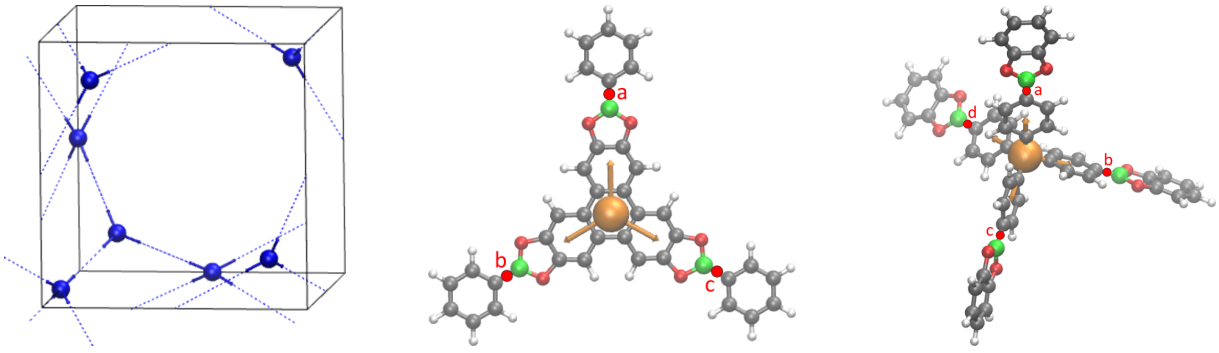


Figure S12: Some structures optimize towards a suboptimal imine configuration. Left: the initial structure as obtained after Step 3 of the structure assembly process. Black nodes are the vertices of the topology. Red nodes are the points of extension. Right: the optimized structure, including the suboptimal imine configuration.

S1.6 Case study 1: the structure assembly of COF-108

As an illustration of the technical explanation of the structure assembly procedure in the main text, we focus here on the *in silico* structure assembly of COF-108 as a case study. COF-108[6] is one of the first reported COFs in which the SBUs HHTP (18-08-01) and TBPM (26-01-01) assemble in the **bor** topology to form a boronate ester linkage. Therefore, with the nomenclature introduced in Section S1.3, this COF is labeled bor_18-08-01_26-01-01_None in the ReDD-COFFEE database.

Step 0 In Step 0 of the structure assembly process, both the **bor** topology and the SBUs 18-08-01 and 26-01-01 are initialized. As COF-108 does not contain a two-connected linker, the edge Wyckoff set of the **bor** topology is left vacant. The topology and SBUs are visualized in Fig. S13, together with the cluster termination and points of extension of the SBUs. To be able to define the different SBU configurations, their points of extension are labeled with lowercase letters. Also the unit vectors oriented from the center of the SBUs towards its points of extension are visualized. At this point, although the exact material configuration is not yet known, the system-specific force field can already be obtained and used in Step 3, since the cluster force fields are derived and it is known which SBUs are connected.



(a) bor topology **(b)** SBU 18-08-01 **(c)** SBU 26-01-01

Figure S13: Initialization of the bor topology and SBUs 18-08-01 and 26-01-01 used in the structure assembly of COF-108. The unit vectors pointing from the vertices towards its neighbors are indicated in blue in the topology. The unit vectors oriented from the center of each SBU towards its points of extension, are indicated in orange. The points of extension (red dots, labeled with lowercase letters) are consistently positioned in the middle of the bond between the boron atom and its neighboring carbon atom.

Step 1 The **bor** topology is rescaled in Step 1 of the structure assembly approach to fit the 18-08-01 and 26-01-01 SBUs. In the RCSR database, the **bor** topology is defined as such that the distance between each connected vertex is exactly 1 length unit (l.u.). After the rescaling, this should be equal to the distance between the centers of the SBUs. Since there is only one set of edge Wyckoff sets in the **bor** topology, along which the 18-08-01 and 26-01-01 SBUs are connected, this completely defines the rescaling factor. The radius of an SBU is defined as the mean of the distances from the center to the points of extension. For the 18-08-01 and 26-01-01 SBUs, this is 6.55 Å and 5.20 Å, respectively. Therefore, the rescaling factor of this edge Wyckoff set is 11.75 Å/l.u. As there is only one edge Wyckoff set, the standard deviation between the different calculated rescaling factors is zero and the (topology, SBUs) combination passes the first filter.

Step 2 and Step 3 As explained in the main text, the total number of SBU configurations is equal to the number of permutations of the points of extension, which is $N!$, with N being the number of points of extension of the SBU. In the **bor** topology, there are four three-connected vertices which can be decorated with 6 SBU configurations and three four-connected vertices that can be decorated with 24 SBU configurations. The total number of material configurations, which is the product of SBU configurations for each node, is therefore, $6 \times 6 \times 6 \times 6 \times 24 \times 24 \times 24 = 17\ 915\ 904$, which is already huge for this small topology. To decrease the number of tested configurations, our additive top-down approach is implemented. This starts at a central vertex of the topology, in this case V_{11}, and sequentially adds new SBUs one-by-one, following a breadth-first iteration. Since we only iterate over each SBU configuration once, the number of times a specific configuration is tested, reduces drastically to $6 + 6 + 6 + 6 + 24 + 24 + 24 = 96$. The final SBU configuration selected for each step in the iteration is visualized in Fig. S14.

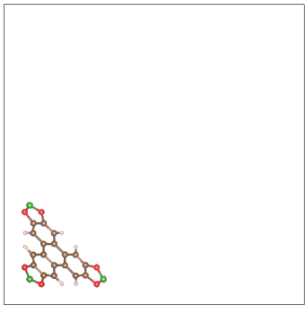
In each iteration, a favorable SBU configuration is selected based on geometric and energetic considerations in Step 2 and Step 3, respectively. This is illustrated in Fig. S15 and Table S2 for the second iteration of the structure assembly of COF-108 in which the SBU 26-01-01 is inserted on the vertex V_{21}. In Step 2, the RMSD between the unit vectors pointing (i) from the center of the SBU towards the points of extension (orange arrows in Fig. S13), and (ii) from the node

towards its neighbors in the topology (blue arrows in Figs. S13 and S15) is calculated for each SBU configuration. These values are listed in Table S2. From all SBU configurations (24 for SBU 26-01-01), only those that minimize the RMSD proceed to Step 3. In this example, the minimal RMSD is 0.086 Å, which is obtained for configurations 1, 8, 17, and 24. As can be observed in Fig. S15, this relatively large RMSD is obtained since the points of extension are not perfectly oriented towards the neighboring nodes, since the vertices V_{21}, V_{22}, and V_{23} of the **bor** topology possess the lower D_{2d} point symmetry, instead of the tetrahedral T_d point symmetry of the points of extension of the SBU.

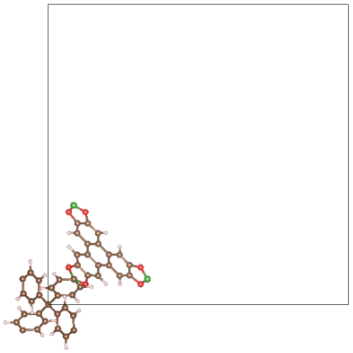
In Step 3, the deformation energy E_{def} for the four remaining configurations is calculated. Since only the neighboring vertex V_{11} is occupied with an SBU already, the deformation energy is completely defined by the linkage between those building blocks. It is not necessary to calculate the deformation energy of the configurations with a high RMSD, as the large geometric mismatch automatically results in a high deformation energy, which is illustrated in Table S2. If the internal geometry of the SBU would have the same point symmetry as its points of extension, then the SBU configurations with the same RMSD would result in the same deformation energy. However, this is not the case for the SBU 26-01-01, as the orientation of the phenyl rings is different in each configuration. Configuration 1 minimizes the deformation energy and is therefore selected to be finally placed on the **bor** topology, as is visualized in Fig. S14b.

In the following iterations, the remaining SBUs are inserted in their most likely configuration. As we follow a breadth-first iteration, the number of linkages with SBUs that are already present is higher as compared to a depth-first or a random iteration. Therefore, the deformation energy takes into account a larger fraction of the linkages present in the periodic material.

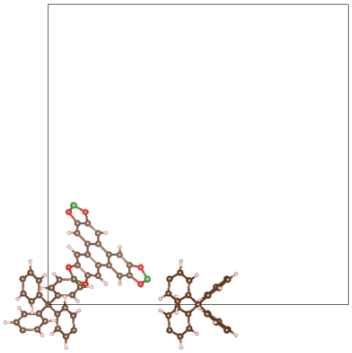
Step 4 The geometric mismatch between the tetrahderal SBU 26-01-01 and the vertices in the **bor** topology resulted in a large RMSD of 0.086 Å. Once all SBUs are inserted on the topology, the structure is relaxed using its system-specific force field in Step 4. Since the SBUs are sufficiently flexible, a relaxed structure with a small energy penalty for the geometric mismatch can be obtained, which is visualized in Fig. S14h. Despite the large RMSD, the deformation energy of the final structure is 4.85 kJ/mol, which is well below the threshold of 14 kJ/mol. Therefore,



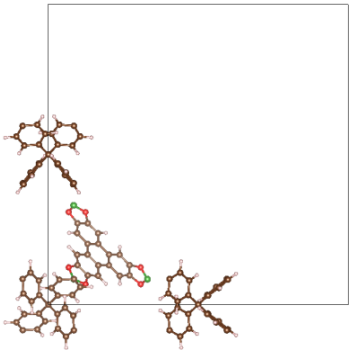
(a) Iteration 1: Decorate vertex V_{11}.



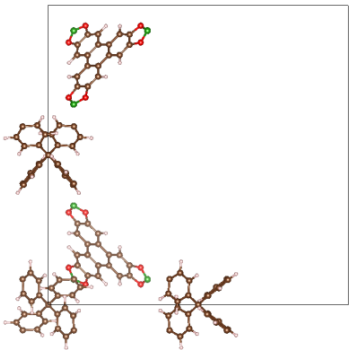
(b) Iteration 2: Decorate vertex V_{21}.



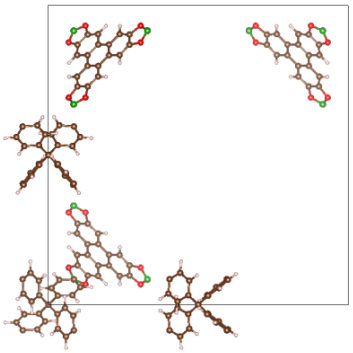
(c) Iteration 3: Decorate vertex V_{22}.



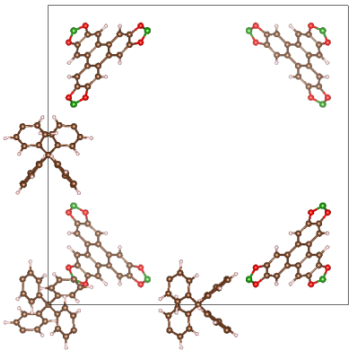
(d) Iteration 4: Decorate vertex V_{23}.



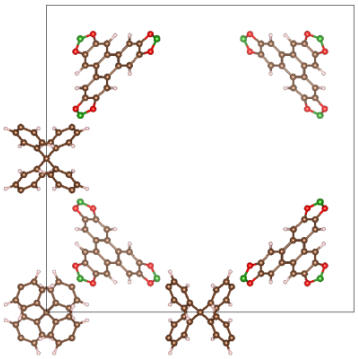
(e) Iteration 5: Decorate vertex V_{12}.



(f) Iteration 6: Decorate vertex V_{13}.

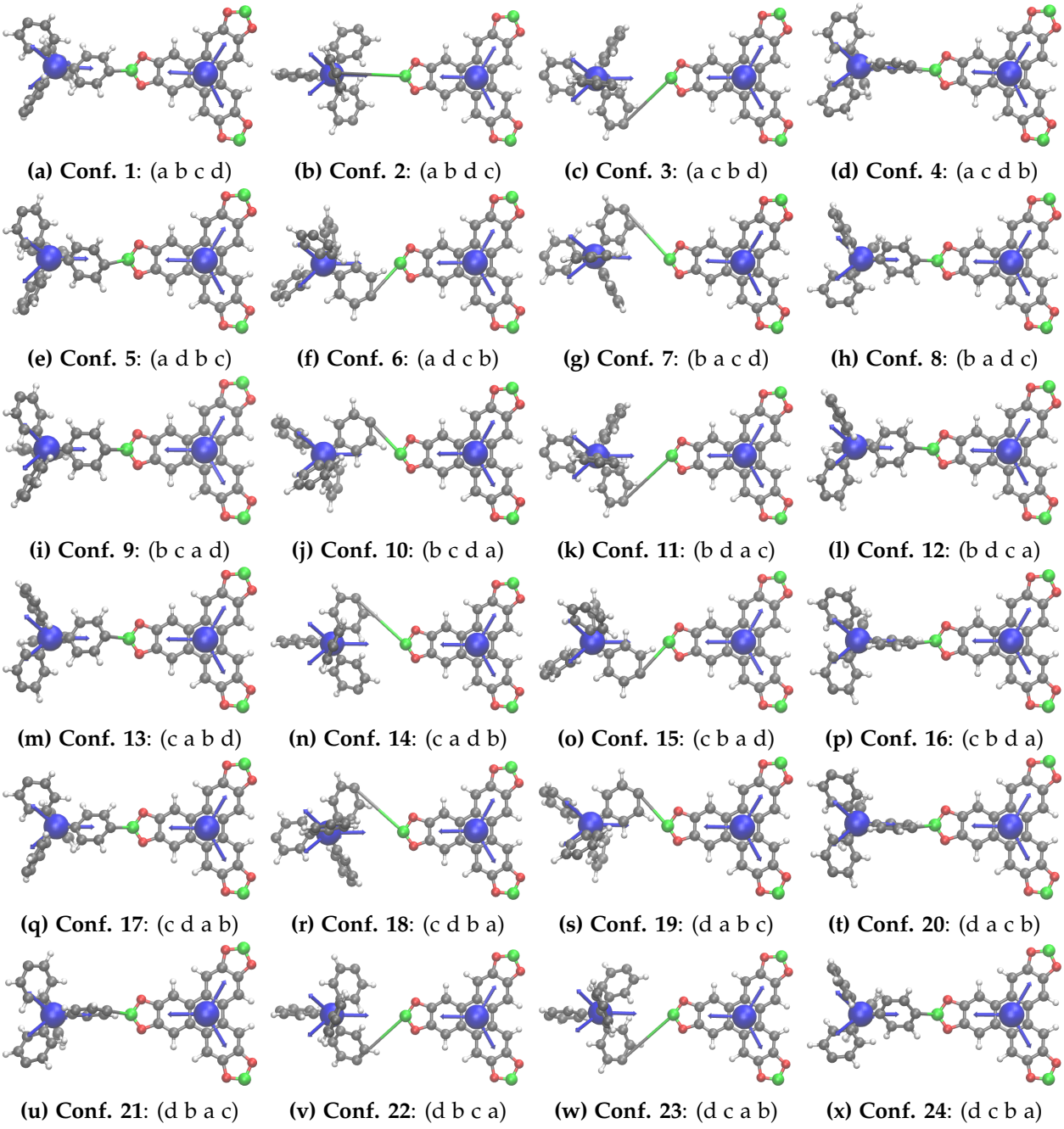


(g) Iteration 7: Decorate vertex V_{14}.



(h) Relaxed structure

Figure S14: Illustration of the additive top-down approach. The SBUs are added one-by-one, following a breadth-first iteration through the topological graph. Once all vertices are decorated, the structure is relaxed using its system-specific force field. Color code: hydrogen (white), boron (green), carbon (brown), oxygen (red).



(a) Conf. 1: (a b c d) **(b) Conf. 2**: (a b d c) **(c) Conf. 3**: (a c b d) **(d) Conf. 4**: (a c d b)

(e) Conf. 5: (a d b c) **(f) Conf. 6**: (a d c b) **(g) Conf. 7**: (b a c d) **(h) Conf. 8**: (b a d c)

(i) Conf. 9: (b c a d) **(j) Conf. 10**: (b c d a) **(k) Conf. 11**: (b d a c) **(l) Conf. 12**: (b d c a)

(m) Conf. 13: (c a b d) **(n) Conf. 14**: (c a d b) **(o) Conf. 15**: (c b a d) **(p) Conf. 16**: (c b d a)

(q) Conf. 17: (c d a b) **(r) Conf. 18**: (c d b a) **(s) Conf. 19**: (d a b c) **(t) Conf. 20**: (d a c b)

(u) Conf. 21: (d b a c) **(v) Conf. 22**: (d b c a) **(w) Conf. 23**: (d c a b) **(x) Conf. 24**: (d c b a)

Figure S15: All 24 SBU configurations for the SBU 26-01-01 that decorates vertex V_{21} in the bor topology. Also the already inserted 18-08-01 SBU of vertex V_{11} is visualized as the resulting linkage defines the selected configuration. Blue spheres are the locations of the vertices and the blue arrows indicate the position of the neighboring nodes in the topology. Color code: hydrogen (white), boron (green), carbon (gray), oxygen (red).

Table S2: The geometric and energetic parameters to select the most favorable SBU configuration for the 24 SBU configurations of SBU 26-01-01 depicted in Fig. S15. Resulting deformation energies E_{def} are reported here relative to the lowest energy configuration. The values that minimize the RMSD and E_{def} are indicated in bold. Since the configurations that do not minimize the RMSD are already discarded in Step 2, E_{def} should not be calculated for them. They are nonetheless reported here as an illustrative example, using light blue text to discriminate between the values that would actually be calculated during the procedure.

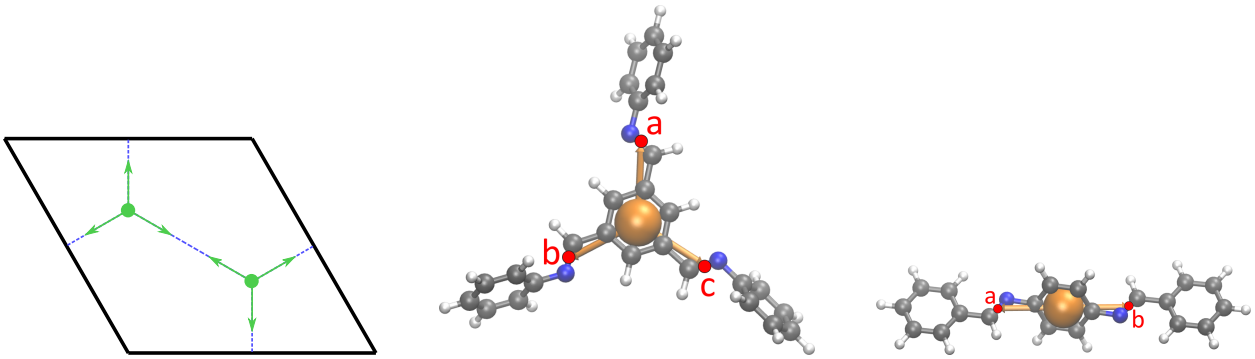
	RMSD [Å]	E_{def} [kJ/mol]		RMSD [Å]	E_{def} [kJ/mol]
Conf. 1	**0.086**	**0.0**	Conf. 13	0.113	608.61
Conf. 2	0.312	19859.09	Conf. 14	0.311	19904.97
Conf. 3	0.311	19904.93	Conf. 15	0.303	34882.95
Conf. 4	0.113	809.31	Conf. 16	0.113	820.30
Conf. 5	0.113	642.01	Conf. 17	**0.086**	6.85
Conf. 6	0.303	34735.10	Conf. 18	0.312	19497.16
Conf. 7	0.312	20048.74	Conf. 19	0.303	4385.07
Conf. 8	**0.086**	20.63	Conf. 20	0.113	800.59
Conf. 9	0.113	621.32	Conf. 21	0.113	823.14
Conf. 10	0.303	4352.64	Conf. 22	0.311	19773.86
Conf. 11	0.311	19961.69	Conf. 23	0.312	20405.04
Conf. 12	0.113	597.78	Conf. 24	**0.086**	10.37

COF-108 is added to the ReDD-COFFEE database. If the SBUs would have been more rigid, they would not have been able to accommodate for the introduced geometric mismatch. This is illustrated for the structure in Fig. S10e, where the SBUs 23-08-01 and 28-01-01 are used, which represent a triptycene linker and an adamantane cage. Since these building blocks are more rigid than the HHTP and TBPM SBUs, the deformation energy increases to 18.07 kJ/mol, and thus this structure is not included in the database.

S1.7 Case study 2: structure assembly of COF-LZU1

A second, yet shorter, case study illustrates the structure assembly of COF-LZU1, in which the SBUs TFB (11-02-04) and DAB (01-03-04) combine in a **hcb** topology. COF-LZU1[7] is the first synthesized 2D imine COF and was proposed as catalyst for the Suzuki-Miyaura coupling reaction. Its label in the ReDD-COFFEE database is hcb_11-02-04_01-03-04.

Step 0 The **hcb** topology and the SBUs adopted in the structure assembly of COF-LZU1 are visualized in Fig. S16. Again, the points of extension are labeled with lowercase letters. In this case, the location of the points of extension is chosen to be in the middle of the central N=C imine bond.



(a) hcb topology **(b)** SBU 11-02-04 **(c)** SBU 01-03-04

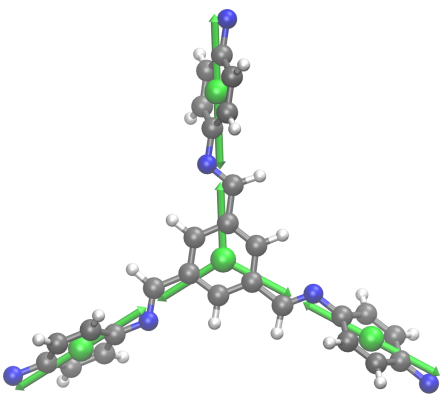
Figure S16: Initialization of the hcb topology and SBUs 11-02-04 and 01-03-04 used in the structure assembly of COF-LZU1. The unit vectors pointing from the vertices towards its neighbors are indicated in green in the topology. The unit vectors oriented from the center of each SBU towards its points of extension, are indicated in orange. The points of extension (red dots, labeled with lowercase letters) are consistently positioned in the middle of the N=C imine bond.

Step 1 Whereas the edge in COF-108 was not occupied and connected two different SBUs, this is not the case for COF-LZU1. In this case, the edge connects two 11-02-04 SBUs, with radius 3.28 Å, with a 01-03-04 linker, with radius 3.20 Å. Therefore, the distance between the two centers of the 11-02-04 SBUs is 12.96 Å, which corresponds to the initial distance of 1 l.u. in the **hcb** topology. Therefore, the rescaling factor for this (topology, SBUs) combination is 12.96 Å/l.u. Again, since there is only one edge Wyckoff set, the standard deviation of the different rescaling factors is zero, and the combination is accepted.

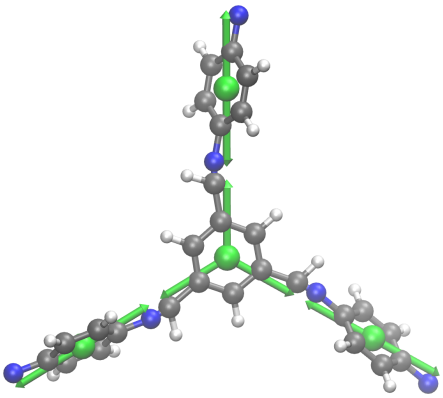
Step 2 and Step 3 A similar reasoning as in the case of COF-108 can be made. However, in this case, additional configurations are available for the SBU 01-03-04, since this is a linear linker with two points of extension that has a rotational freedom. Again, an iterative procedure is followed, where for each node, a favorable SBU configuration is selected based on geometric and energetic considerations. After the initial SBU 11-02-04 is placed on the vertex V_{11}, the three edges E_{11}, E_{12}, and E_{13} are decorated with the SBU 01-03-04. For the last iteration, in which the SBU 11-02-04 is

placed on the vertex V_{12}, six SBU configurations are available, which are visualized in Fig. S17. Since all three neighbors are already inserted in the topological unit cell, all linkages have to be taken into account when calculating the deformation energy. In this case, all configurations have the same RMSD of 0.0 Å, but a different deformation energy is observed. Configurations 2, 4, and 6 have a deformation energy E_{def} that is 1.22 kJ/mol higher than the equivalent configurations 1, 3, and 5 due to the suboptimal realization of the imine linkage. Therefore, one of the latter is adopted for this vertex.

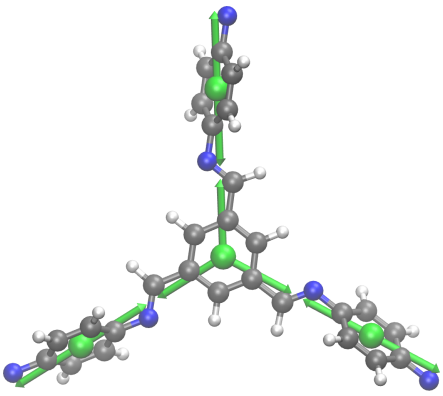
Step 4 After all SBUs are inserted in the topology, the structure is relaxed using its system-specific force field. For this material, the deformation energy of the optimized structure is well below the threshold of 14 kJ/mol, and therefore the structure of COF-LZU1 is added to the ReDD-COFFEE database.



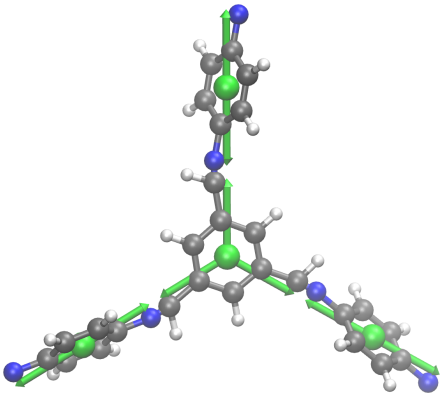
(a) Conf. 1: (a b c)



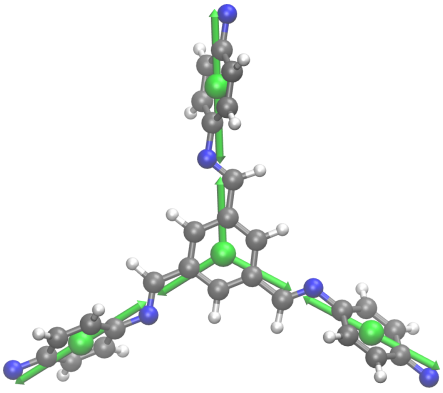
(b) Conf. 2: (a c b)



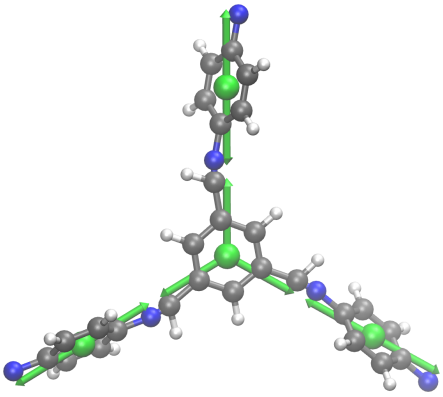
(c) Conf. 3: (b c a)



(d) Conf. 4: (b a c)



(e) Conf. 5: (c a b)



(f) Conf. 6: (c b a)

Figure S17: The six SBU configurations for the SBU 11-02-04 that decorates vertex V_{12} in the hcb topology. Also the already inserted 01-03-04 SBUs of edges E_{11}, E_{12}, and E_{13} are visualized as the resulting linkages define the selected configuration. Green spheres are the locations of the nodes and the green arrows indicate the position of the neighboring nodes in the topology. The name of each configuration determines the points of extension that are oriented towards the neighboring edges E_{11}, E_{12}, and E_{13}, respectively. Color code: hydrogen (white), carbon (gray), nitrogen (blue).

S2 Force field generation and validation

As explained in Section 2.4 of the main text, the system-specific force fields of the periodic structures are derived from the cluster force fields of its constituent SBUs. Additional details about the derivation of the cluster force fields are specified in Section S2.1, whereas the influence of the chosen cluster terminations is discussed in Section S2.2. The improved accuracy of the QuickFF cluster force fields over UFF is illustrated in Section S2.3. As discussed in the main text, the validation of the periodic force fields is done by comparing computationally derived PXRD patterns with an experimental pattern. Details about these results are given in Section S2.4. Some minor notes about the applicability of the periodic force fields are discussed in Section S2.5. We end this Section with a detailed example of the derivation of a term in the periodic force field overlapping two SBUs in Section S2.6.

S2.1 Additional details about the cluster force field generation

S2.1.1 Sixth-order polynomial to describe dihedral angles of triazine-linked SBUs

According to the QuickFF philosophy,[1,2] certain dihedral terms can be omitted during the fitting procedure due to various reasons (negative force constants, undefined multiplicities ...) and should later be replaced by more complex terms, if deemed necessary. This necessity becomes apparent when clear deviations from the *ab initio* geometry are observed during the force field optimization. This provides a distinct feature to easily recognize those force fields that should be amended. Among our SBUs, this feature was prominently present for any dihedral term connecting a triazine and phenyl ring. To this end, second-generation force fields were constructed for the clusters containing a triazine termination (*-06-10), and the triazine cluster 32-11-10. These force fields were amended by deriving an additional term, a sixth order polynomial:

$$E_{TORSCPOLYSIX} = \sum_{i=1}^{6} C_i \cos^i(\Psi) \tag{S2.7}$$

to reproduce the *ab initio* rotation behavior, captured by performing a rotational scan. This fitting procedure is visualized in Fig. S18, showcasing the large deviation of the rotational barrier for

the old force field due to a faulty valence term, disappearing for the amended force field. This is facilitated by fitting the new dihedral term on the difference between the *ab initio* rotational barrier and the force field rotational barrier, where the faulty dihedral term has been removed prior. As such, a perfect reproduction of the rotational barrier, within the limits of a polynomial expansion of sixth order, is expected. Although this approach would benefit the accuracy of any cluster force field, it was not included in the standard protocol to avoid significantly higher computational costs for limited accuracy gain.

S2.1.2 Consistent description of out-of-plane terms

QuickFF can adopt two different analytic forms to describe an out-of-plane term:[1,2]

$$E_{\mathrm{OOPDIST}} = \frac{K}{2}(D - D_0)^2 \tag{S2.8}$$

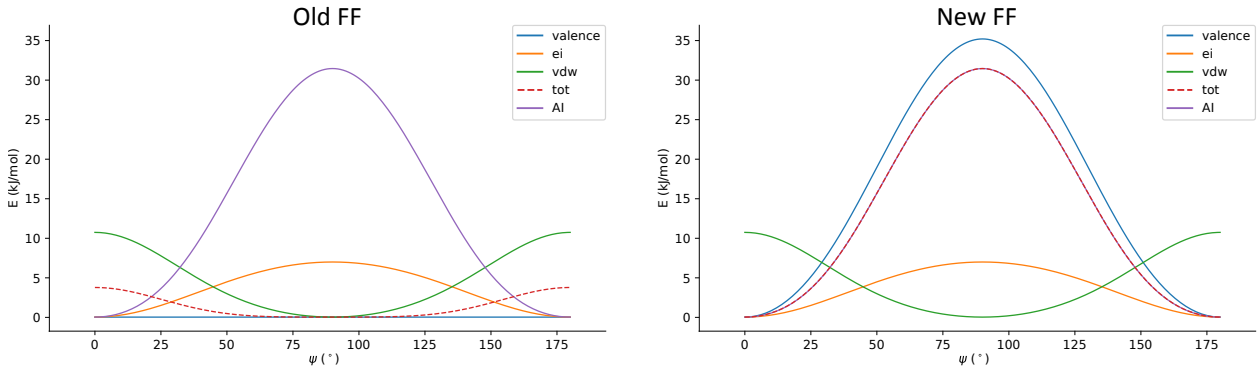
$$E_{\mathrm{SQOOPDIST}} = \frac{K}{2}(D^2 - D_0^2)^2 \tag{S2.9}$$

with D being the out-of-plane distance, D_0 the rest value of this distance, and K a force constant. When the out-of-plane distance of the *ab initio* optimized cluster is smaller than a certain threshold, which is by default 5×10^{-2} Å, the OOPDIST form of Eq. S2.8 is chosen. The SQOOPDIST form of Eq. S2.9 is adopted when the out-of-plane distance exceeds the threshold value.

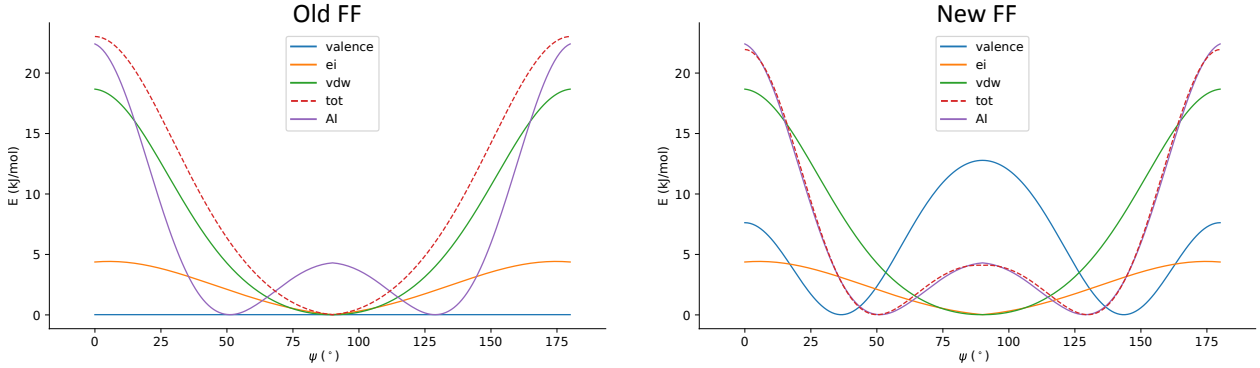
If the threshold would be left at the default value, some cluster force fields would include SQOOPDIST terms, whereas the majority only hold regular OOPDIST terms. When two SBUs are connected to form a periodic structure, the overlapping out-of-plane patterns could possibly be described with different analytic forms and would be impossible to match. To ensure that all out-of-plane patterns are consistently described by an OOPDIST term, the threshold is increased to 15×10^{-2} Å. This affects only the cluster force fields of the SBUs 12-03-09, 12-04-05, 13-04-05, 15-02-05, and 30-02-05.

S2.2 Influence of cluster termination

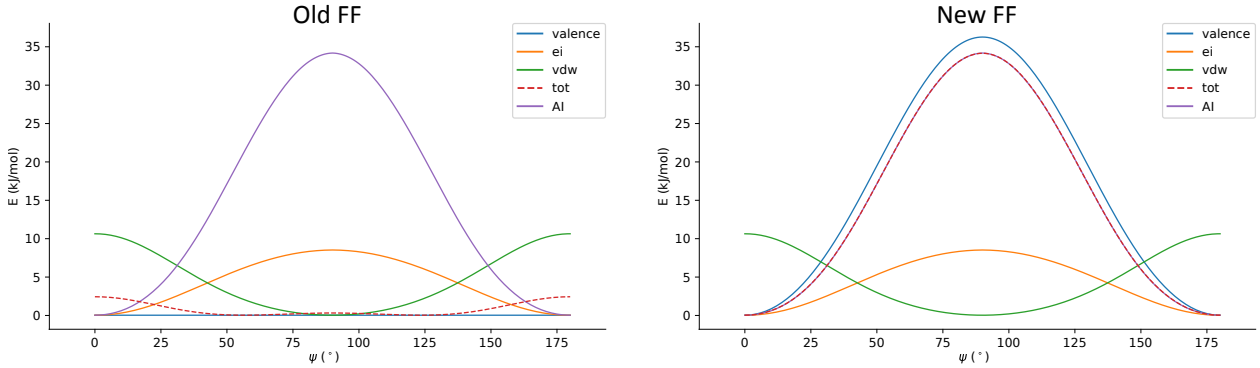
As explained in the main text, the choice of cluster termination is a trade-off between accuracy and transferability. A larger accuracy can be obtained by including the SBU environment to a



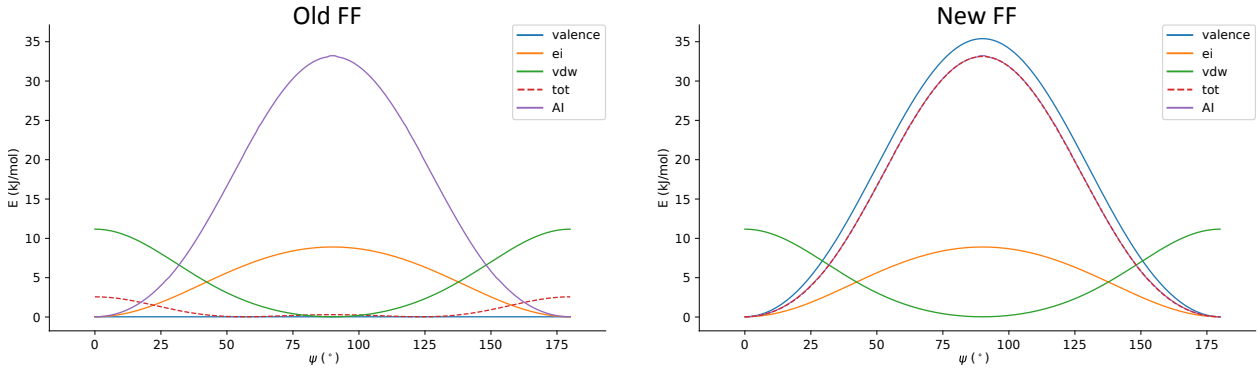
(a) SBU 01-06-10



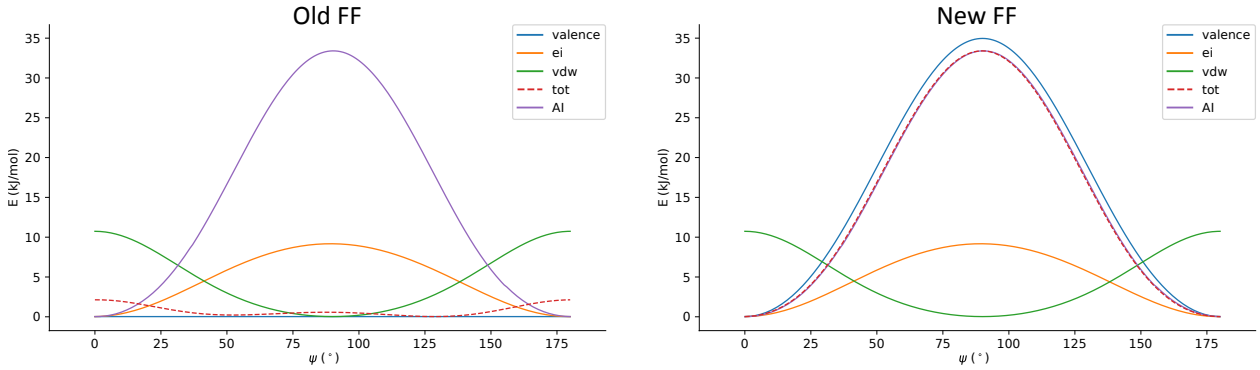
(b) SBU 02-06-10



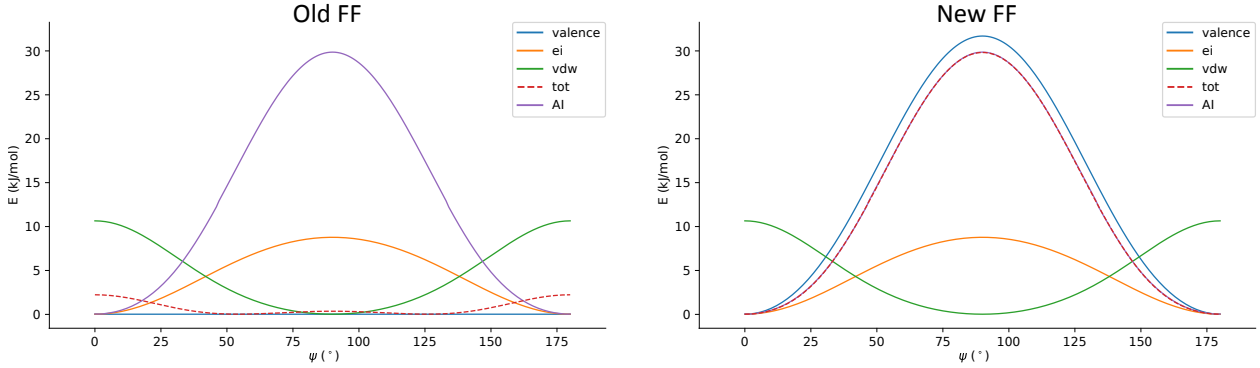
(c) SBU 03-06-10



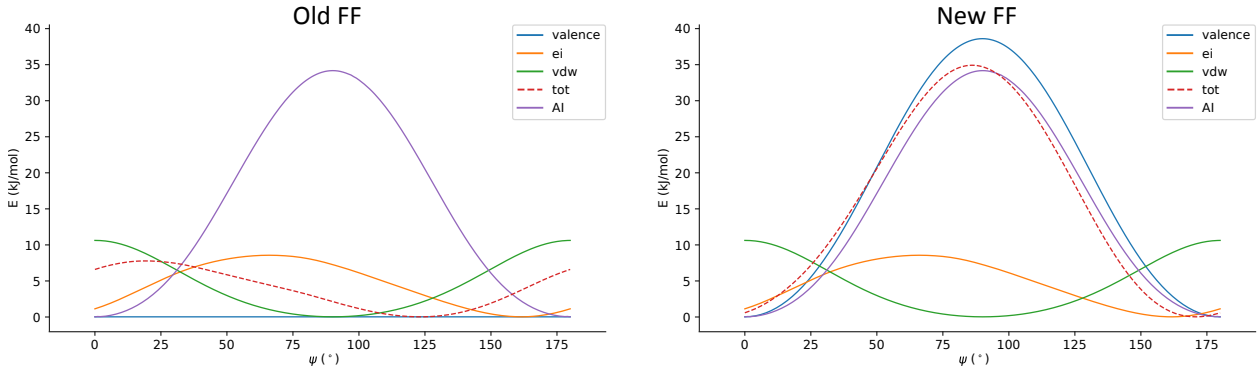
(d) SBU 04-06-10



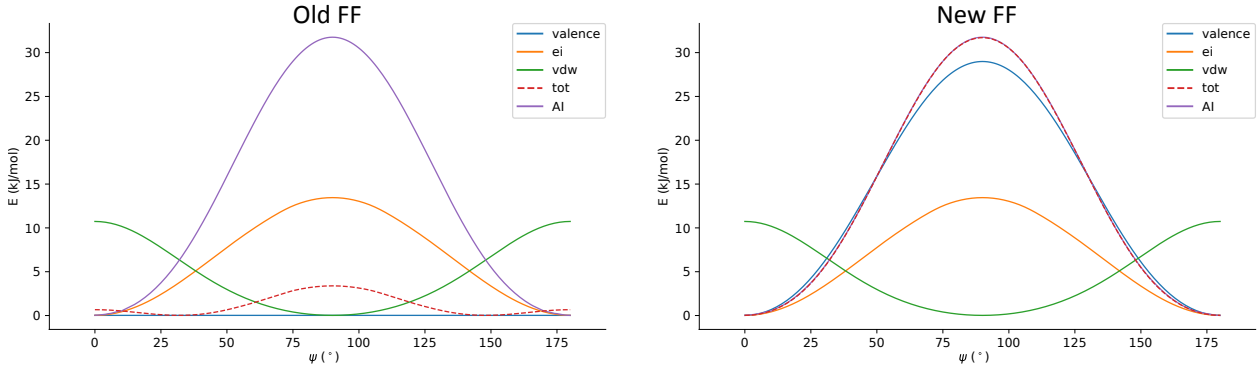
(e) SBU 05-06-10



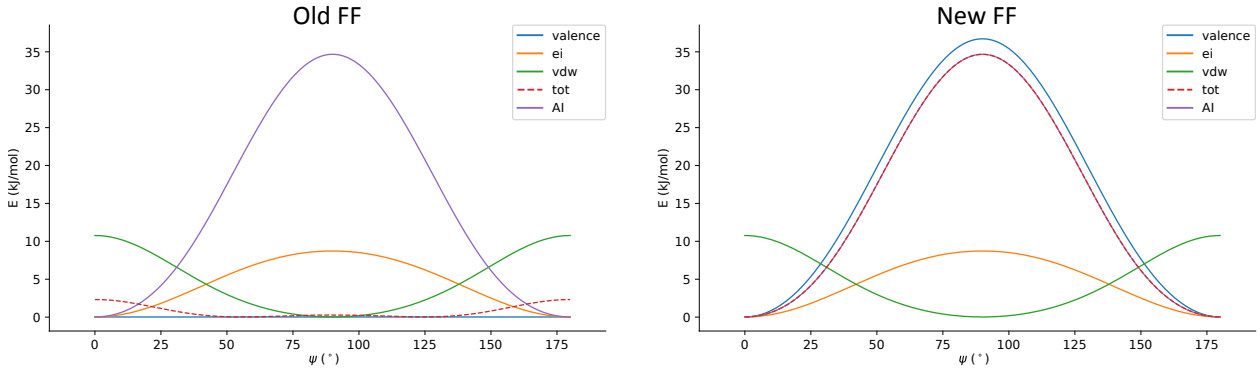
(f) SBU 11-06-10



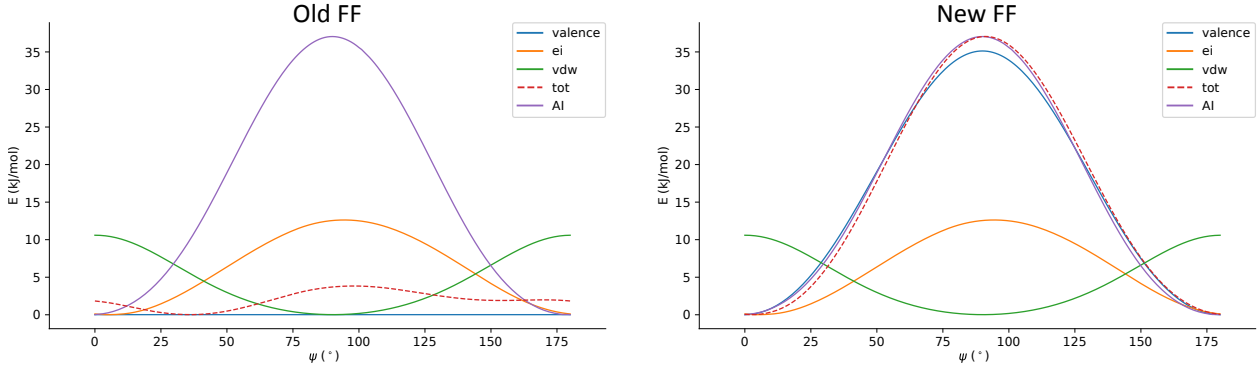
(g) SBU 13-06-10



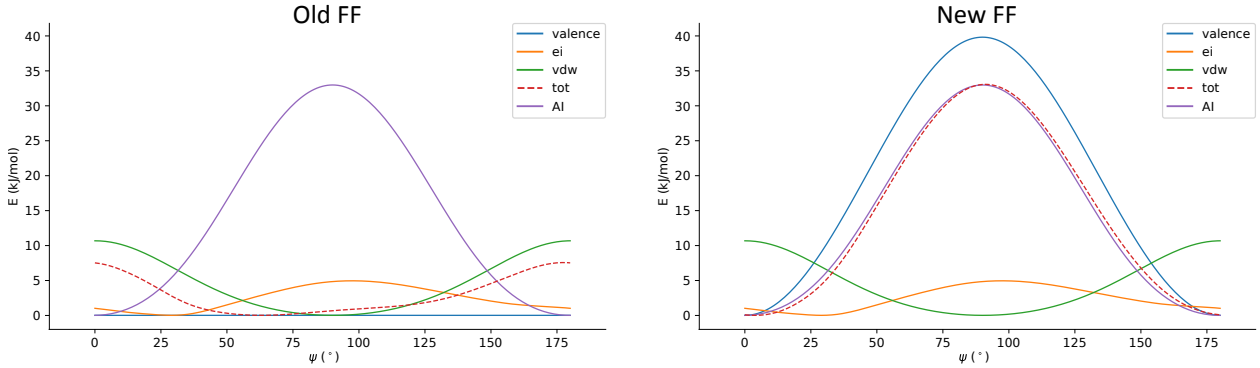
(h) SBU 14-06-10



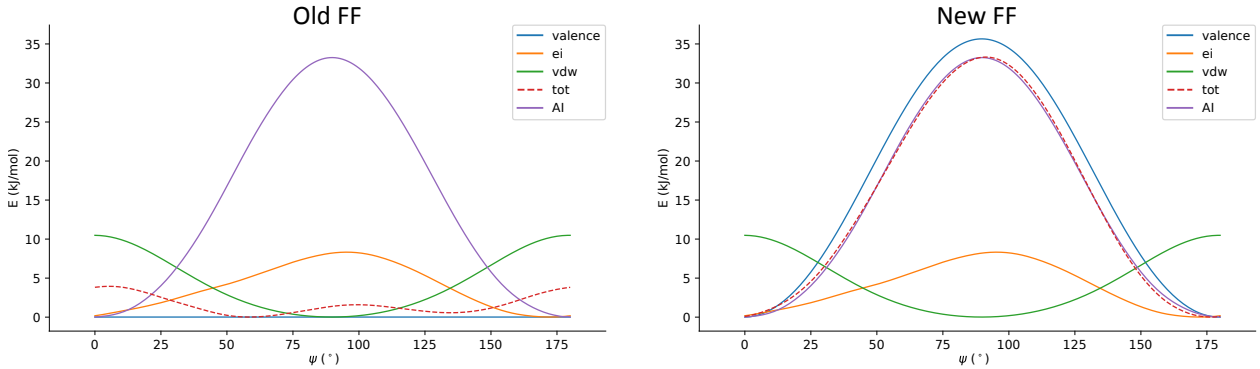
(i) SBU 15-06-10



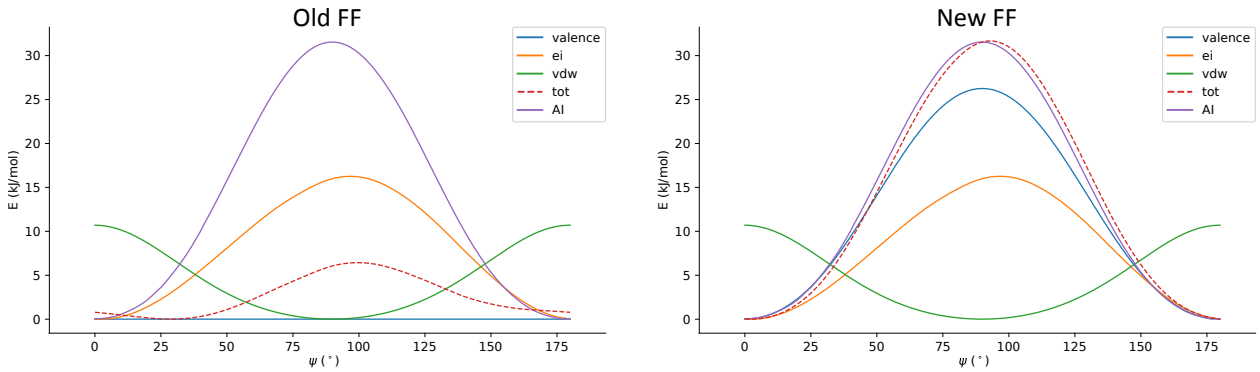
(j) SBU 16-06-10



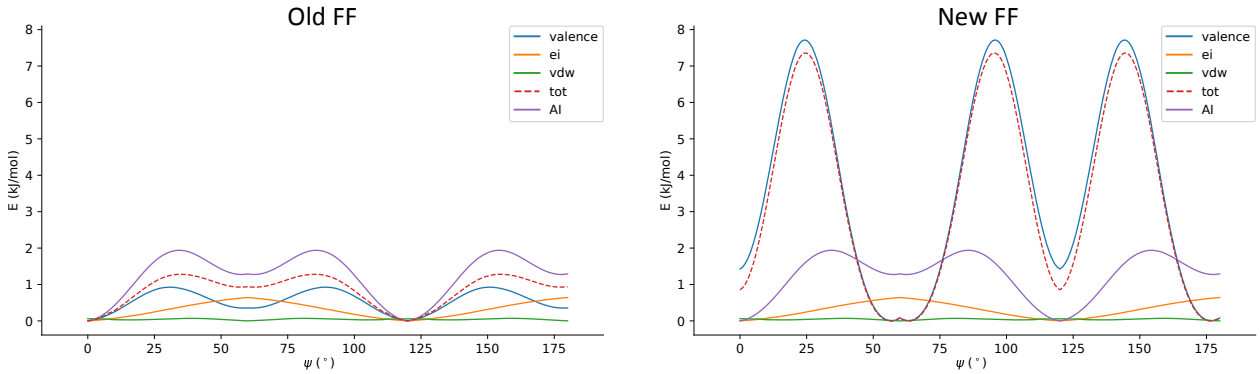
(k) SBU 24-06-10



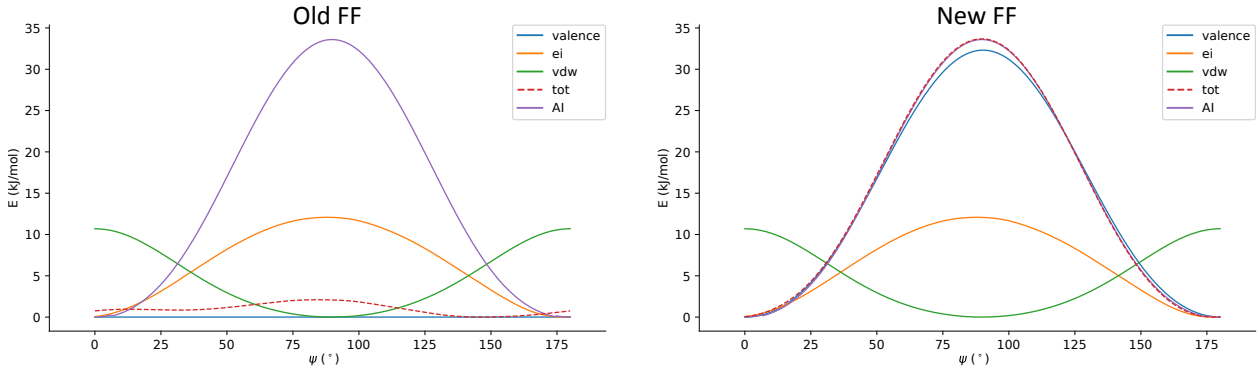
(l) SBU 26-06-10



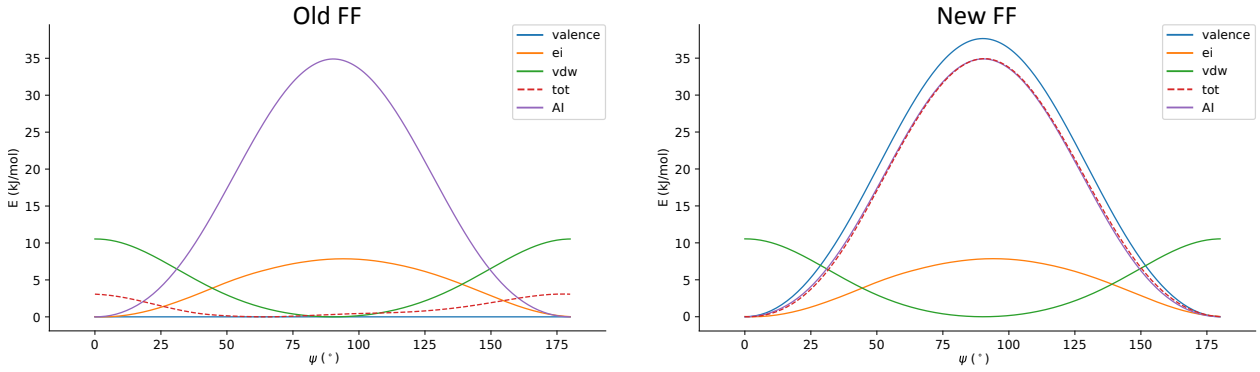
(m) SBU 27-06-10



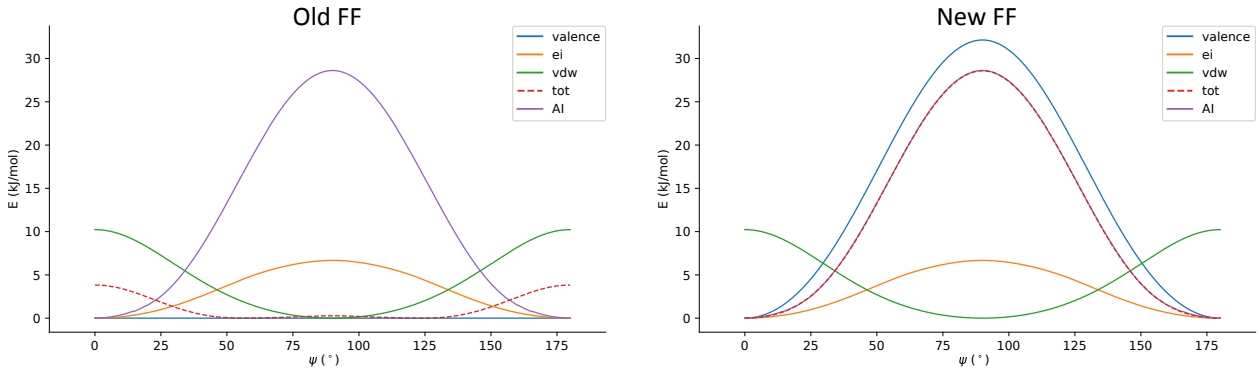
(n) SBU 28-06-10



(o) SBU 29-06-10



(p) SBU 30-06-10



(q) SBU 32-11-10

Figure S18: Reproduction of the *ab initio* rotational barrier (purple) of the triazine-phenyl dihedral angle in SBUs involved in a triazine linkage by the old and new force fields. The total force field contribution (red dashed line) is divided into a covalent (blue), electrostatic (yellow), and van der Waals (green) contribution. The old force field is directly obtained by the QuickFF algorithm. The new force field is generated by discarding the old torsional term and fitting a sixth-order polynomial instead.

larger extent. However, as this termination would not correctly mimic the SBU environment in each material, the cluster force field could only be adopted for a limited number of structures. To be able to generate the same number of periodic force fields, a larger number of clusters should be included and more *ab initio* calculations should be performed. To have a good balance between accuracy and transferability, we chose to always include the SBU environment up to the next aromatic ring and define a single termination for each linkage section.

To prove that the impact of this termination does not greatly reduce the accuracy of the cluster force fields, we compared the cluster force field parameters as derived with our default termination to the ones derived with an extended termination for four clusters, as shown in Fig. S19. As can be observed in Figs. S20-S25, there is almost no deviation in the rest values and only minor shifts in the force constants. The largest deviations are observed for the force constants of terms that overlap with the termination, for which the importance in the final periodic force field is small given the small rescaling factor. These results confirm that the adopted terminations are sufficiently accurate to derive the cluster force fields.

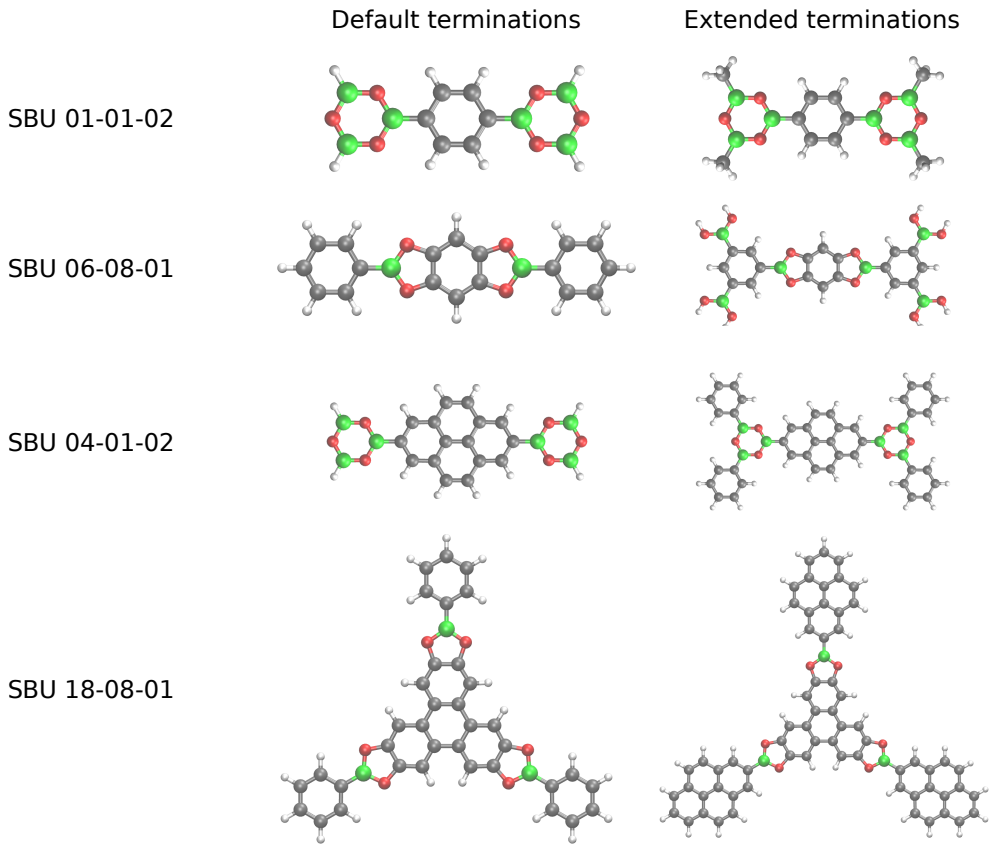
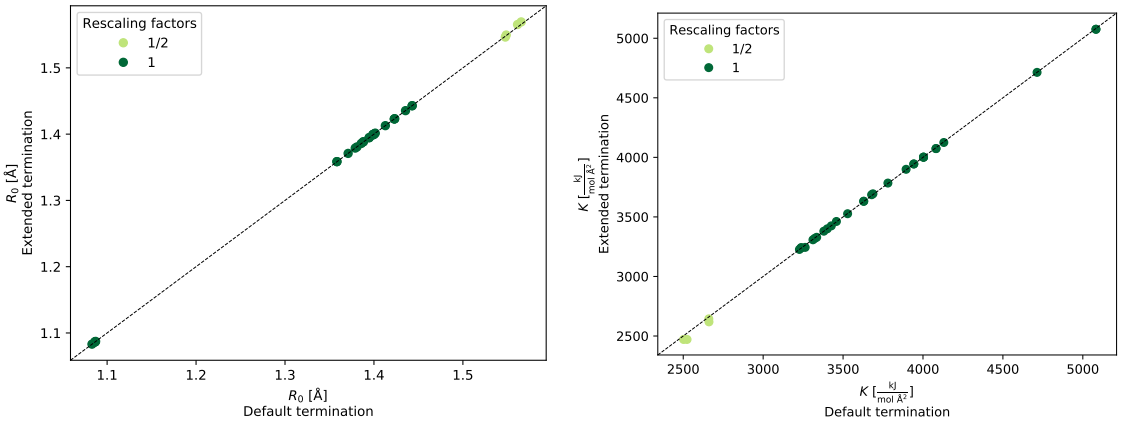
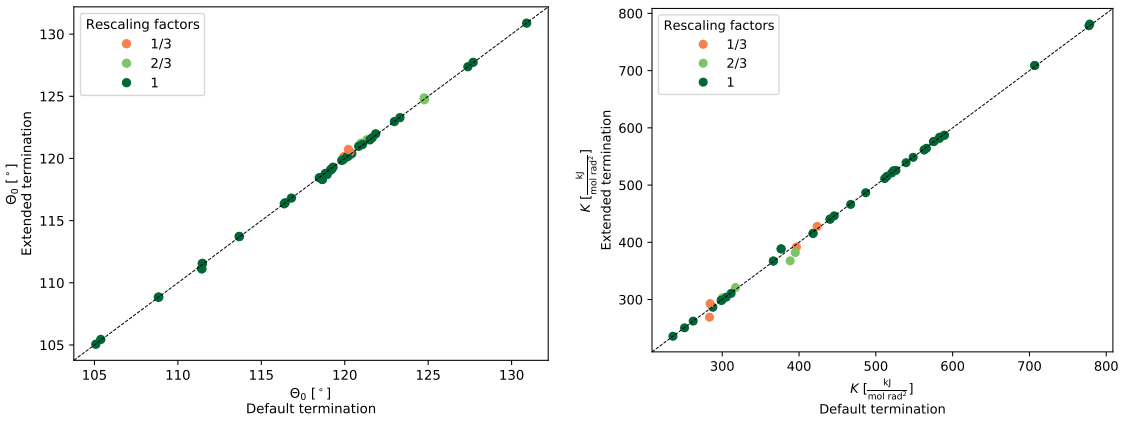


Figure S19: Different cluster terminations to validate the default termination. The extended terminations mimic the SBU environment more accurately, but can be used in a more limited number of materials.



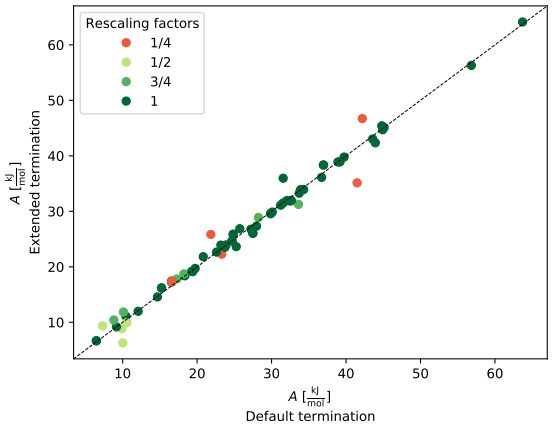
$$E_{\text{BONDHARM}} = \frac{K}{2}(R - R_0)^2 \tag{S2.10}$$

Figure S20: Force field parameters of the bond terms as compared between default and extended terminations for the four clusters depicted in Fig. S19. Left: rest bond distance R_0. Right: force constant K.



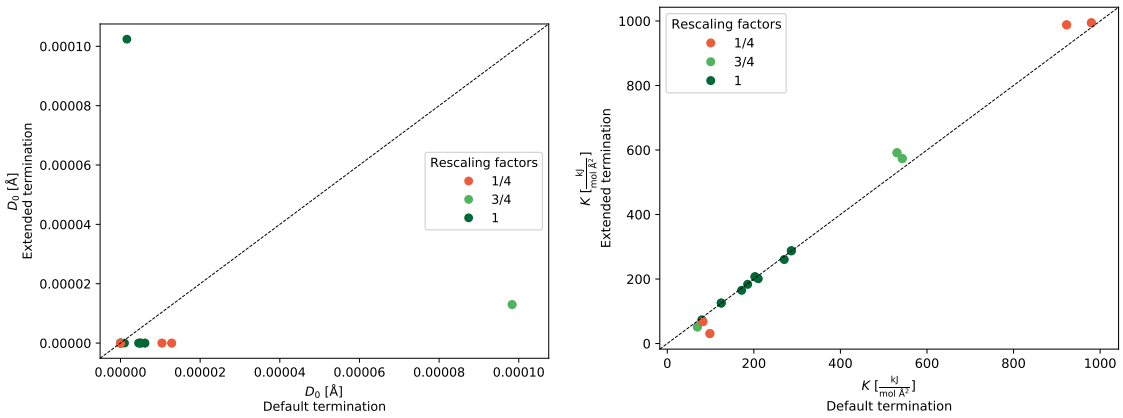
$$E_{\text{BENDAHARM}} = \frac{K}{2}(\Theta - \Theta_0)^2 \tag{S2.11}$$

Figure S21: Force field parameters of the bend terms as compared between default and extended terminations for the four clusters depicted in Fig. S19. Left: rest angle Θ_0. Right: force constant K.



$$E_{\text{TORSION}} = \frac{A}{2}\left(1 - \cos(M(\Psi - \Psi_0))\right) \tag{S2.12}$$

Figure S22: Force field parameters of the dihedral terms as compared between default and extended terminations for the four clusters depicted in Fig. S19. Only the force constant A differs between the different clusters. The multiplicity M is always exactly the same, as is the rest dihedral angle Ψ_0, which is always zero.

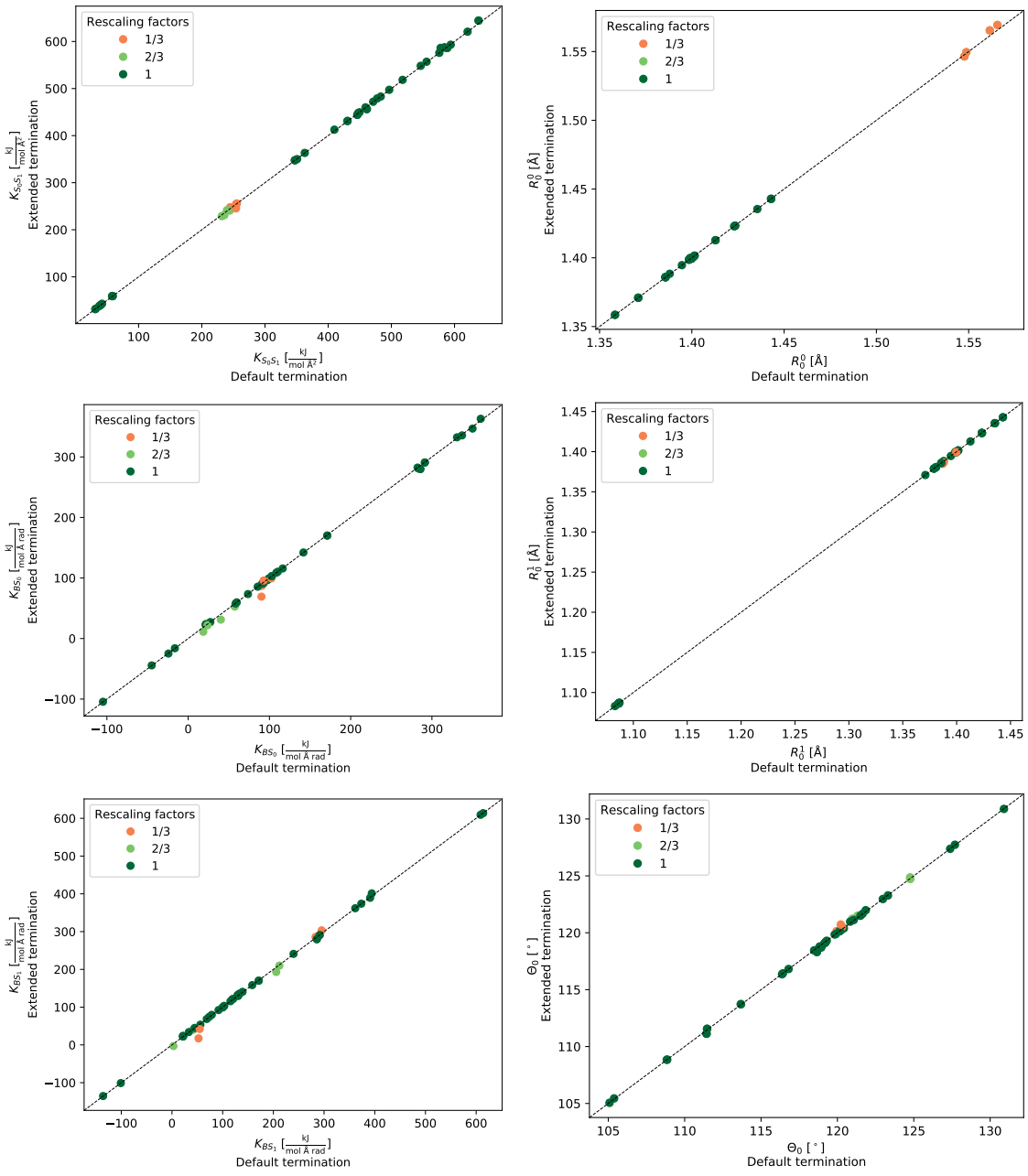


$$E_{\text{OOPDIST}} = \frac{K}{2}(D - D_0)^2 \tag{S2.13}$$

Figure S23: Force field parameters of the out-of-plane terms as compared between default and extended terminations for the four clusters depicted in Fig. S19. Left: rest out-of-plane distance D_0. Right: force constant K. The small deviations between the terminations are on the order of 10^{-4} Å and are therefore negligible.

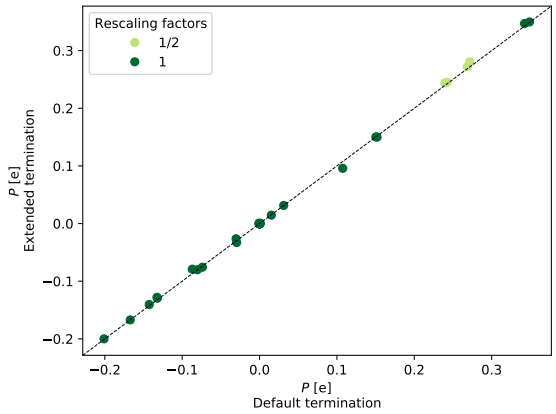
S2.3 Comparison between the UFF and QuickFF cluster force fields

To confirm that the QuickFF cluster force fields indeed reach a higher accuracy when compared to the generic UFF ones, the clusters are optimized with both force fields separately, and compared to the relaxed *ab initio* data. Both the internal coordinates (ICs), *i.e.*, bonds, bends, dihedral angles, and out-of-plane distances (oops), and the vibrational frequencies are considered. The root-mean-square deviation (RMSD) and mean deviation (MD) of all internal coordinates and frequencies are reported in Table S3. Histograms of the deviations between force field and *ab initio* derived properties are plotted in Figs. S26 and S27.



$$E_{\text{CROSS}} = K_{S_0S_1}(R^0 - R_0^0)(R^1 - R_0^1) + K_{BS_0}(\Theta - \Theta_0)(R^0 - R_0^0) + K_{BS_1}(\Theta - \Theta_0)(R^1 - R_0^1) \quad \text{(S2.14)}$$

Figure S24: Force field parameters of the cross terms as compared between default and extended terminations for the four clusters depicted in Fig. S19. Left: rest values R_0^0, R_0^1, and Θ_0. Right: force constants $K_{S_0S_1}$, K_{BS_0}, and K_{BS_1}.



$$E_{\mathrm{FIXQ}} = \frac{q_i q_j}{d_{ij}} \mathrm{erf}\left(\frac{d_{ij}}{R_{ij}}\right) \qquad \text{with} \qquad q_i = Q_{0,i} + \sum_j P_{ij} \tag{S2.15}$$

Figure S25: Force field parameters of the pairwise electrostatic constribution as compared between default and extended terminations for the four clusters depicted in Fig. S19. Only the bond charge increments P_{ij} can differ between the two clusters. Both the covalent radii R and the pre-charges $Q_{0,i}$ (put to zero) are exactly the same.

Table S3: Validation of the QuickFF cluster force field. The internal coordinates and the vibrational frequencies of the force field optimized clusters are compared with those of the *ab initio* relaxed structure. Both the root-mean-square deviation (RMSD) and the mean deviation (MD) are reported.

		QuickFF		UFF	
		RMSD	MD	RMSD	MD
ICs	bonds [Å]	4.73×10^{-3}	1.85×10^{-3}	3.56×10^{-2}	1.80×10^{-2}
	bends [°]	7.18×10^{-1}	-6.82×10^{-3}	2.87	-1.45×10^{-2}
	dihedrals [°]	9.40	-7.94×10^{-2}	22.27	-1.43×10^{-1}
	oops [Å]	4.12×10^{-2}	-3.87×10^{-3}	4.50×10^{-2}	-6.13×10^{-4}
Frequencies	all [cm^{-1}]	18.7	-2.38	3.07×10^{2}	-1.98×10^{2}
	0-100 [cm^{-1}]	7.12	-1.95	12.5	-6.84
	100-500 [cm^{-1}]	16.7	-8.43	48.8	-38.7
	500-1000 [cm^{-1}]	14.1	-8.79×10^{-1}	1.15×10^{2}	-1.08×10^{2}
	1000-3000 [cm^{-1}]	25.5	-1.79	4.40×10^{2}	-3.68×10^{2}
	>3000 [cm^{-1}]	5.99	4.58×10^{-1}	1.61×10^{2}	-6.90

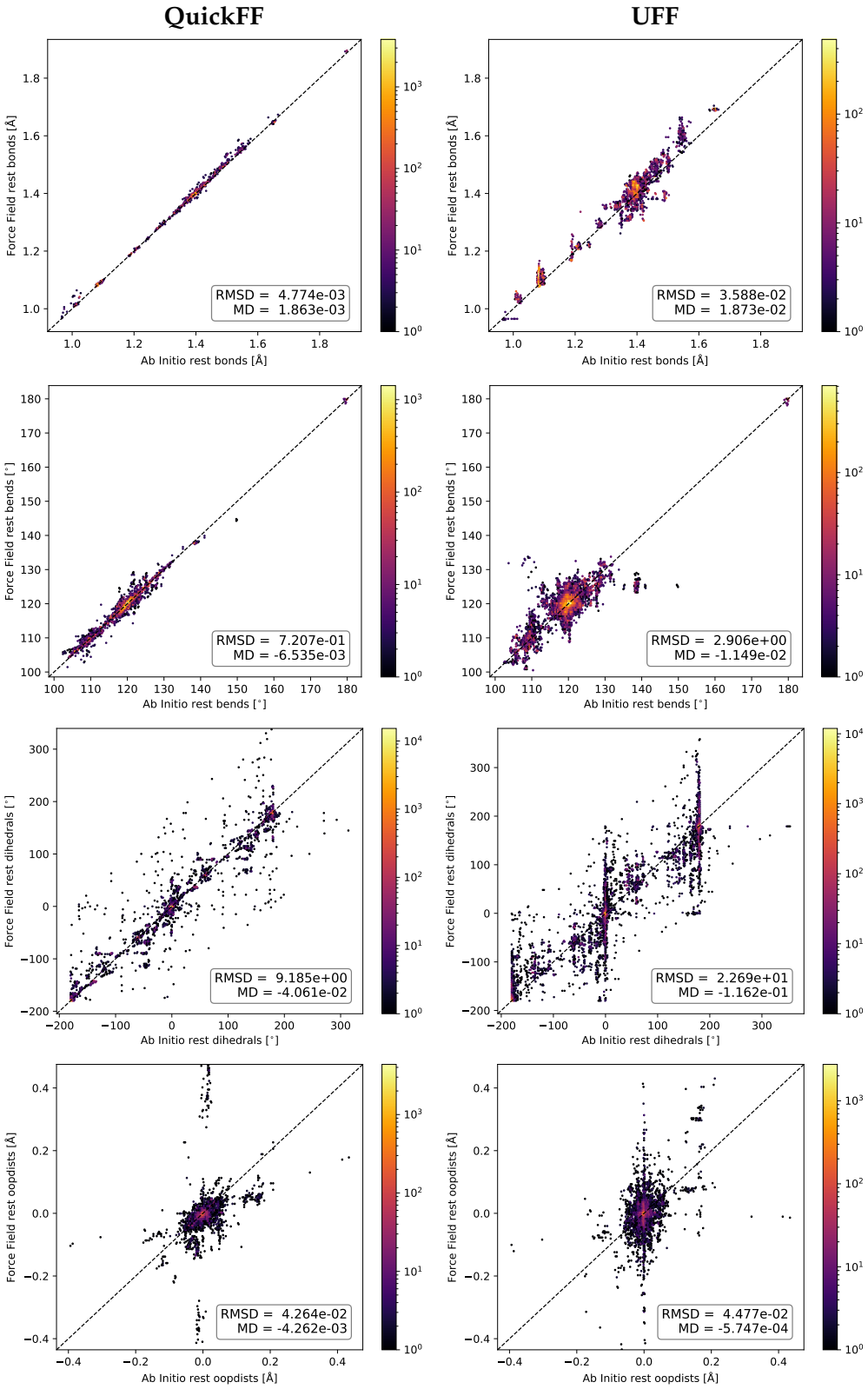


Figure S26: Internal coordinates of the QuickFF and UFF optimized clusters compared with the *ab initio* relaxed ones. From top to bottom: bonds, bends, dihedral angles and out-of-plane distances. Both the root-mean-square deviation (RMSD) and the mean deviation (MD) are reported. The colorbars indicate the density of the data points.

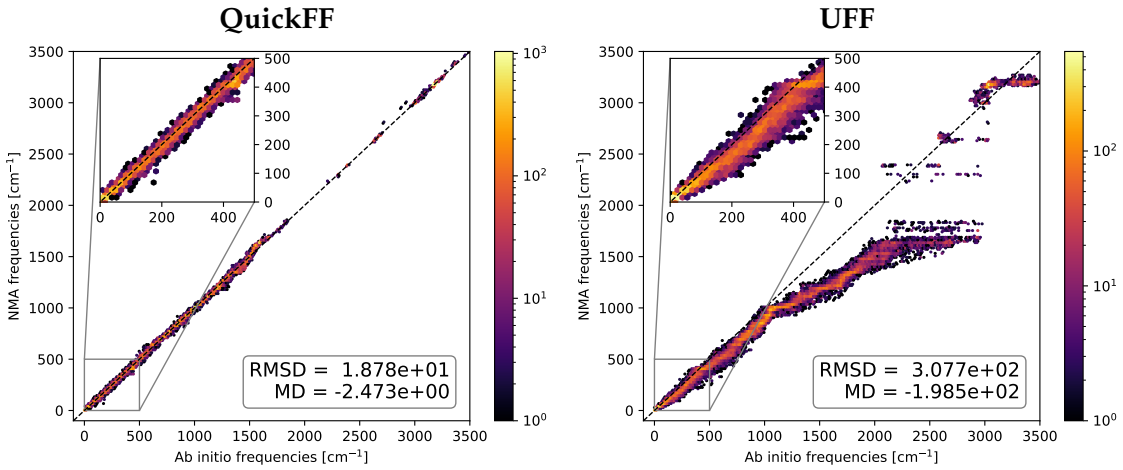


Figure S27: Vibrational frequencies of the QuickFF and UFF clusters compared with the *ab initio* derived ones. Left: QuickFF, right: UFF. Both the root-mean-square deviation (RMSD) and the mean deviation (MD) are reported. The colorbars indicate the density of the data points.

S2.4 Comparison between the UFF and QuickFF periodic force fields

S2.4.1 Calculation of the PXRD patterns

As explained in Section 4.1 of the main text, the powder X-ray diffraction (PXRD) patterns of a diverse set of seven COFs are calculated with a static and dynamic approach using both the system-specific QuickFF derived and generic UFF force fields. They are compared with the experimental pattern using two heuristic metrics, as defined in Ref. 8:

- Weighted profile residual R_{wp} ($w_i = \frac{1}{Y_{\text{ref}}(\theta_i)}$)

$$\sqrt{\frac{\sum w_i \left(Y(\theta_i) - Y_{\text{ref}}(\theta_i)\right)^2}{\sum w_i Y_{\text{ref}}(\theta_i)^2}} \qquad (\rightarrow 0) \tag{S2.16}$$

- Similarity index S_I

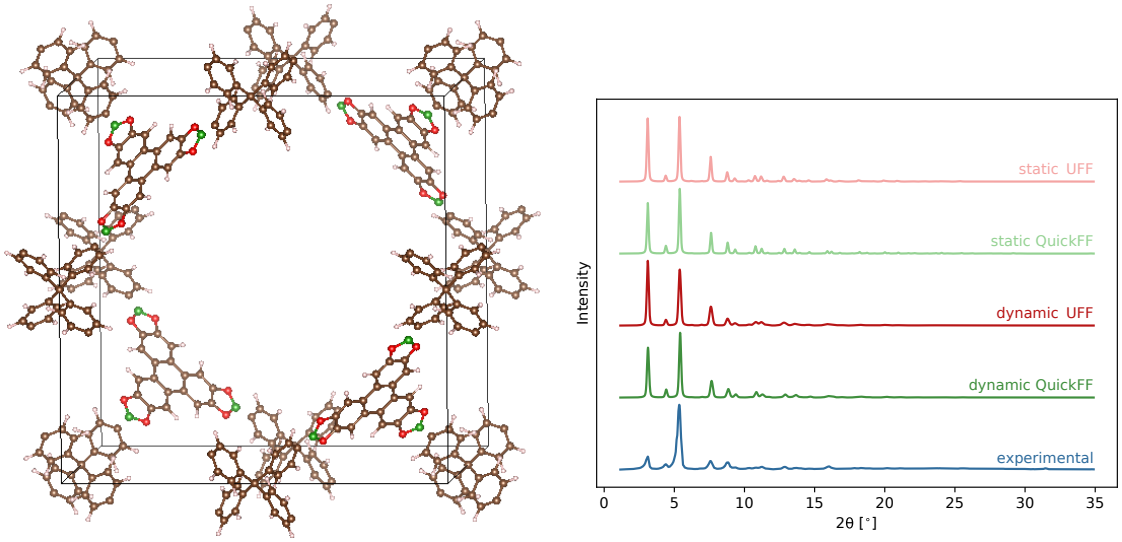
$$\frac{\sum |Y_{\text{ref}}(\theta_i)|\,|Y(\theta_i)|}{\sqrt{\sum Y_{\text{ref}}(\theta_i)^2}\sqrt{\sum Y(\theta_i)^2}} \qquad (\rightarrow 1) \tag{S2.17}$$

with each Bragg location θ_i having an intensity $Y(\theta_i)$. The reference experimental pattern is indicated with the subscript ref. The limit for exact overlap is indicated between brackets.

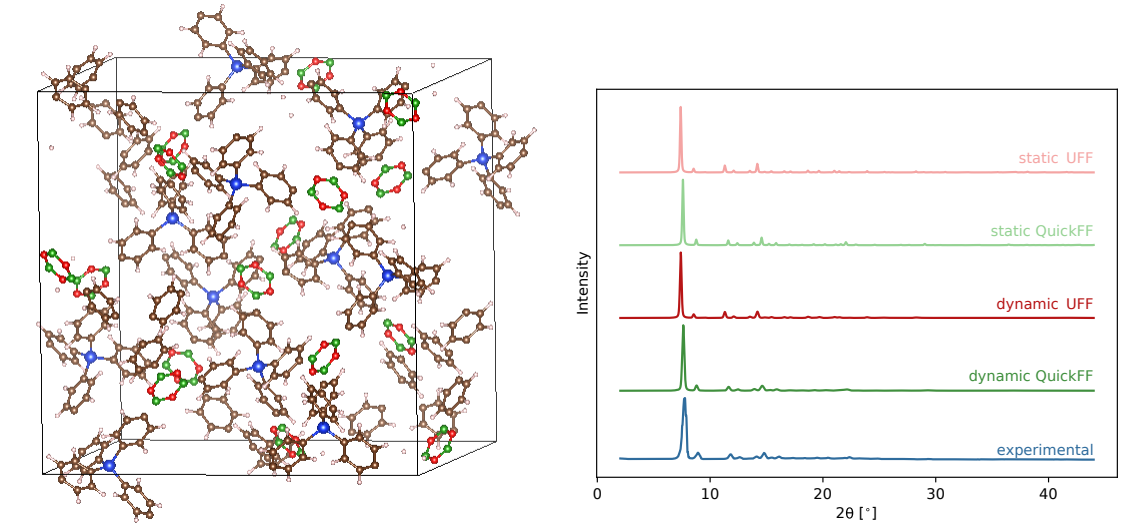
An overview of the seven COFs for which the PXRD patterns are calculated is given in Table S4. Their atomistic structure and all PXRD patterns are plotted in Fig. S28, whereas the heuristic metrics are summarized in Table S5 and Fig. S29. These confirm that the QuickFF force fields reproduce the experimental PXRD patterns better than the UFF ones.

Table S4: The seven COFs for which the PXRD patterns are calculated are diverse in terms of linkage type, topology, dimensionality, and linkers.

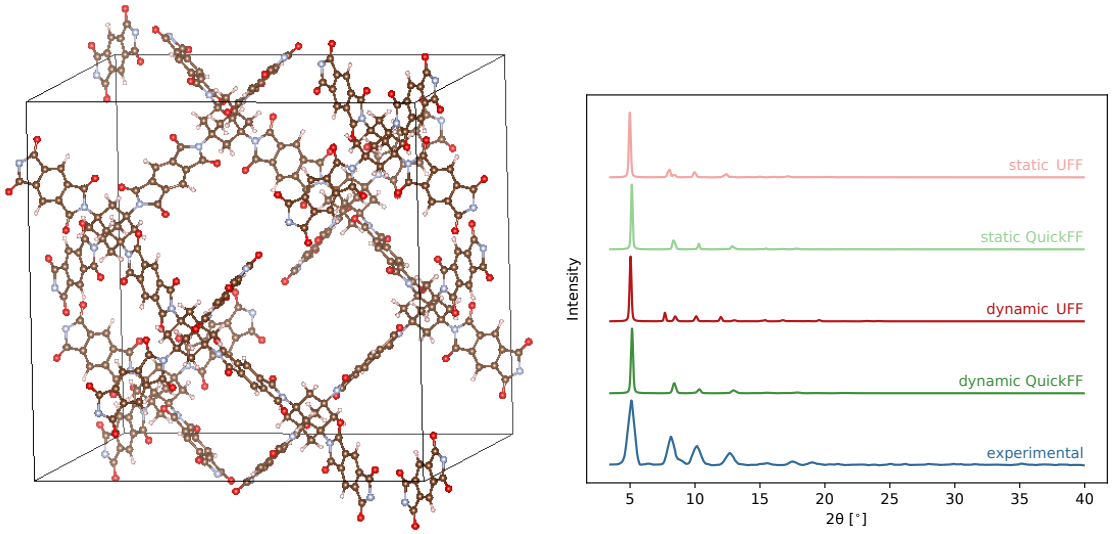
Structure	Experimental name	Linkage	Dimensionality	Ref.
bor_18-08-01_26-01-01_None	COF-108	Boronate Ester	3D	6
ctn_31-11-02_27-01-02_None	COF-103	Boroxine	3D	6
dia_28-03-07_06-09-07	PI-COF-4	Imide	3D	9
sql_24-01-01_10-08-01	COF-66	Boronate Ester	2D	10
kgm_29-03-04_01-02-04	DualPore-COF	Imine	2D	11
hcb_32-11-10_01-06-10	CTF-1	Triazine	2D	12
hcb_11-02-06_None	ACOF-1	Azine	2D	13



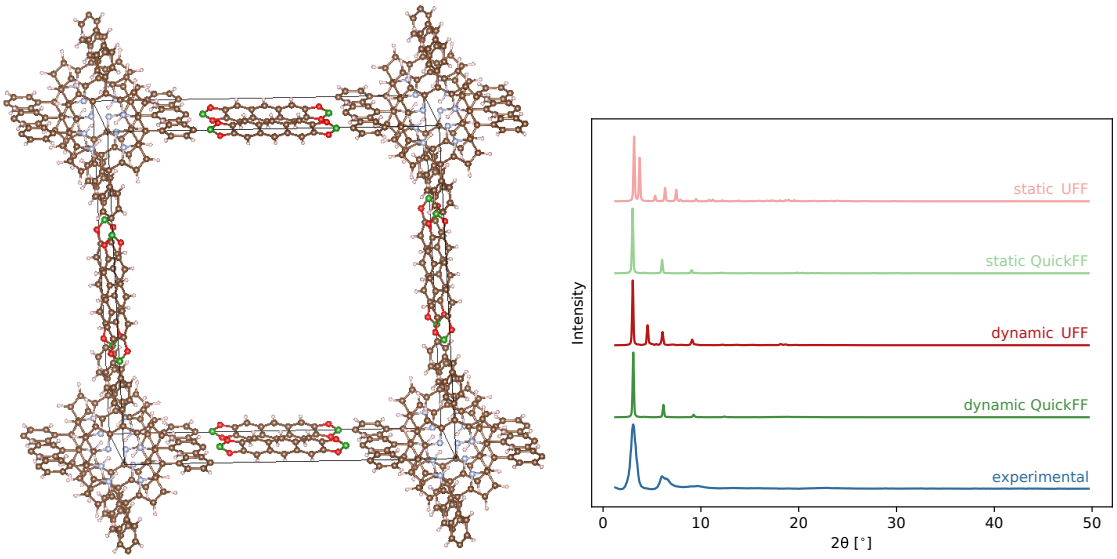
(a) COF-108 or bor_18-08-01_26-01-01_None



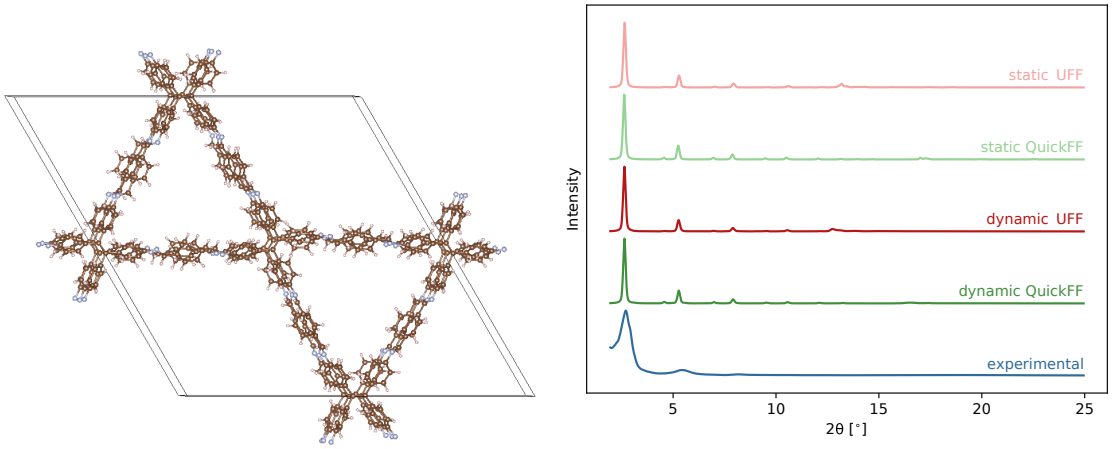
(b) COF-103 or ctn_31-11-02_27-01-02_None



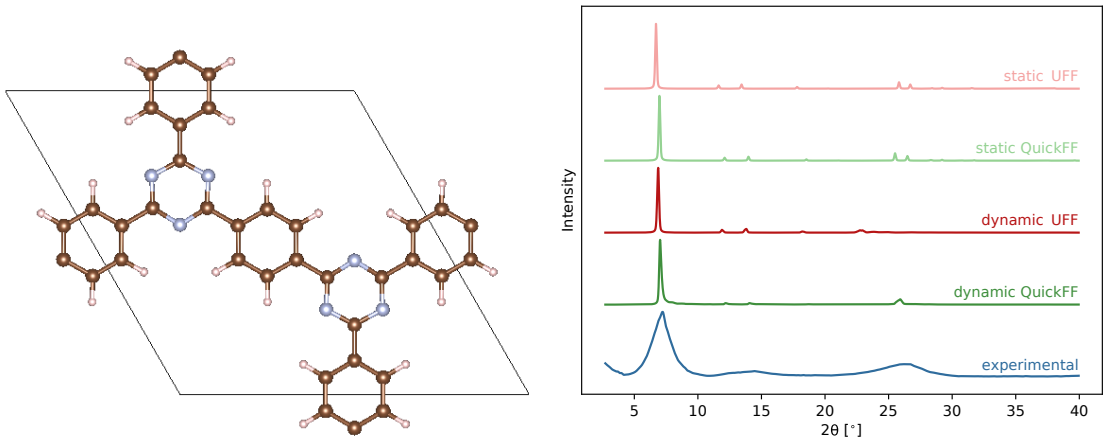
(c) PI-COF-4 or dia_28-03-07_06-09-07



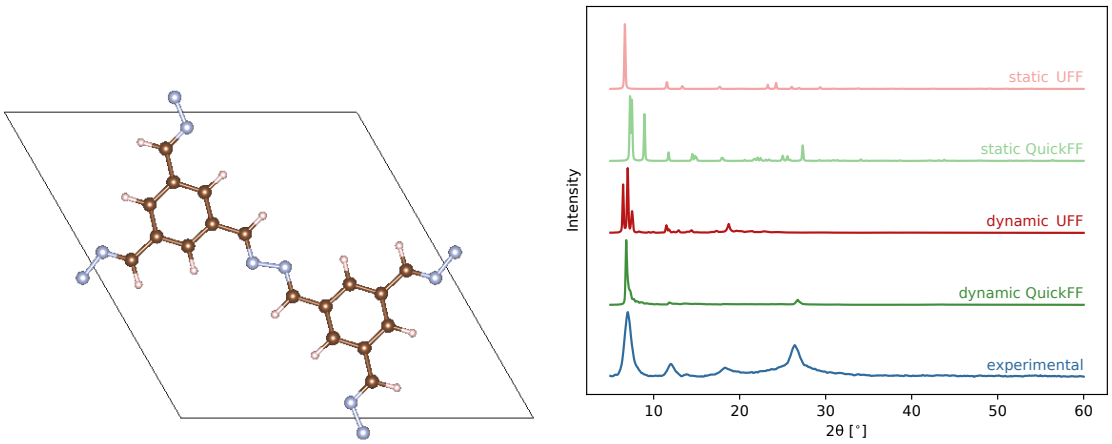
(d) COF-66 or sql_24-01-01_10-08-01



(e) DualPore-COF or kgm_29-03-04_01-02-04



(f) CTF-1 or hcb_32-11-10_01-06-10.png

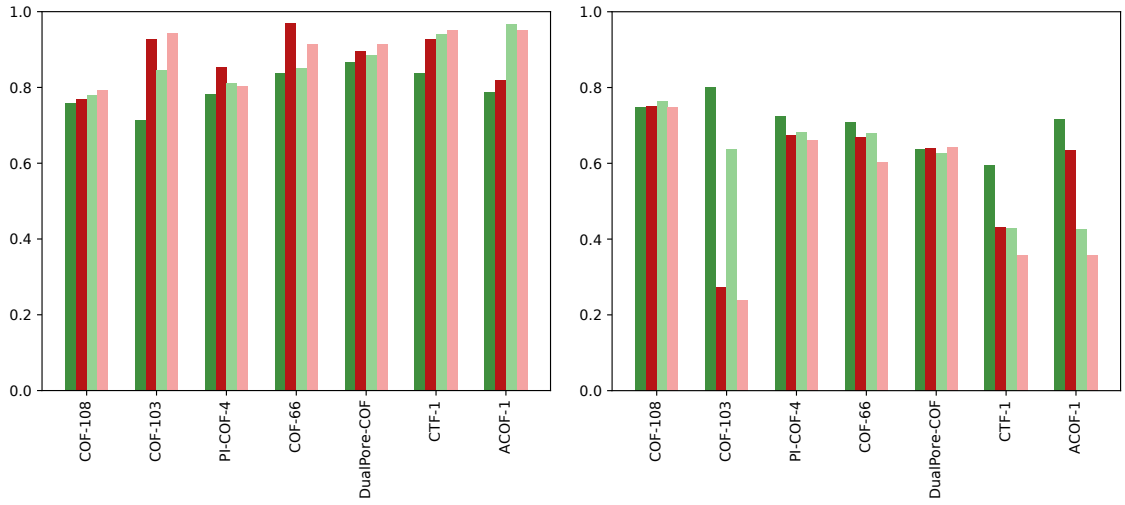


(g) ACOF-1 or hcb_11-02-06_None

Figure S28: Calculated and experimental PXRD patterns of the seven COFs together with their atomistic structure. The PXRD patterns are calculated using a static and a dynamic approach with both the QuickFF and UFF force fields.

Table S5: Comparison of the calculated patterns with the experimental pattern of the seven COFs. For each calculated pattern, the agreement with the experimental pattern is described by the weighted profile residual R_{wp} and the similarity index S_I as defined in Eqs. S2.16 and S2.17. The value that indicates the best agreement with experiment is indicated in bold.

Structure	Metric	Dynamic		Static	
		QuickFF	UFF	QuickFF	UFF
bor_18-08-01_26-01-01_None	R_{wp}	**0.76**	0.77	**0.78**	0.79
	S_I	0.75	**0.75**	**0.76**	0.75
ctn_31-11-02_27-01-02_None	R_{wp}	**0.71**	0.93	**0.85**	0.94
	S_I	**0.80**	0.27	**0.64**	0.24
dia_28-03-07_06-09-07	R_{wp}	**0.78**	0.85	0.81	**0.80**
	S_I	**0.72**	0.68	**0.68**	0.66
sql_24-01-01_10-08-01	R_{wp}	**0.84**	0.97	**0.85**	0.91
	S_I	**0.71**	0.67	**0.68**	0.60
kgm_29-03-04_01-02-04	R_{wp}	**0.87**	0.90	**0.89**	0.91
	S_I	0.64	**0.64**	0.63	**0.64**
hcb_32-11-10_01-06-10	R_{wp}	**0.84**	0.93	**0.94**	0.95
	S_I	**0.59**	0.43	**0.43**	0.36
hcb_11-02-06_None	R_{wp}	**0.79**	0.82	0.97	**0.95**
	S_I	**0.72**	0.64	**0.43**	0.36

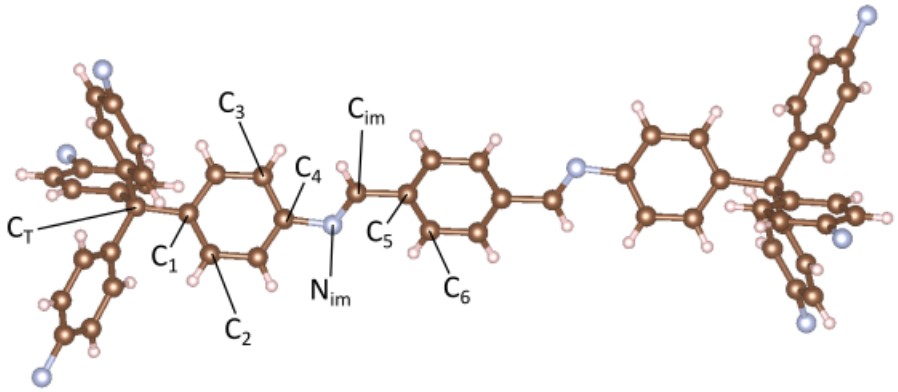


(a) Weighted profile residual R_{wp} **(b)** Similarity index S_I

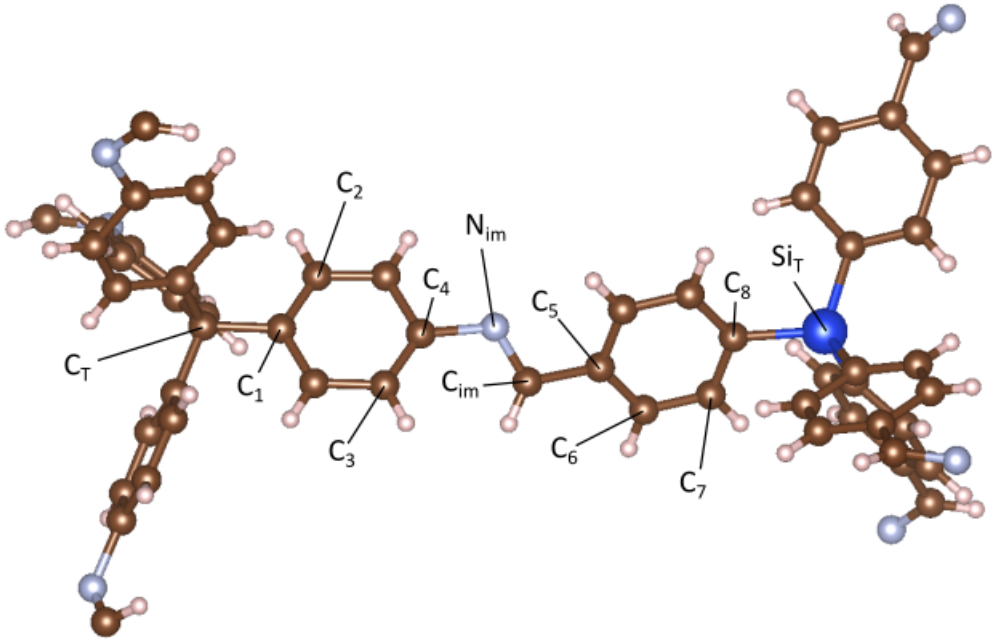
Figure S29: Overview of the heuristics to describe the agreement between the calculated patterns and the experimental ones. Color code is the same as in Fig. 3 of the main text and Fig. S28.

S2.4.2 Calculation of the single crystal structures

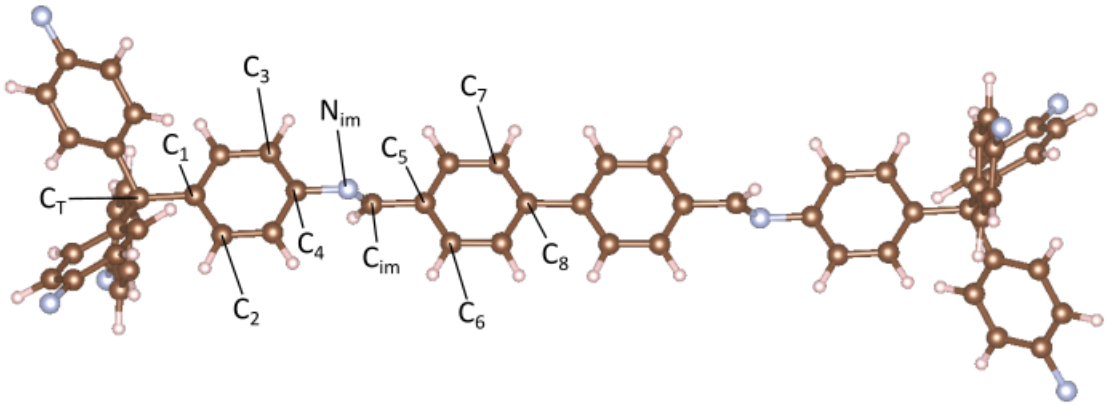
Besides the PXRD patterns, also the single crystal geometries of four COF structures is calculated. The considered materials are COF-300, LZU-111, and two different phases of COF-320. In Tables S6-S9, a selected set of internal coordinates and unit cell parameters are given and compared between the experimental and the dynamically averaged structures. Both comparisons with the system-specific QuickFF force field and the generic UFF force fields are reported. The definition of the atom labels is given in Fig. S30.



(a) COF-300



(b) LZU-111



(c) COF-320

Figure S30: Detail of the atomic structure of COF-300, LZU-111, and COF-320, and definition of the unique atom labels. Color code: hydrogen (white), carbon (brown), nitrogen (light blue), silicon (dark blue).

Table S6: Comparison between our $(N, P, \boldsymbol{\sigma}_a = \mathbf{0}, T)$ molecular dynamics simulations and single crystal X-ray diffraction (SCXRD) data[14] of a selected set of internal coordinates of COF-300 at 100 K and 1 atm.

Interatomic distance [Å]

	SCXRD[14]	QuickFF (rel. diff. [%])	UFF (rel. diff. [%])
C_T-C_1	1.547	1.574 (+1.78)	1.657 (+7.11)
C_1-C_2	1.381	1.402 (+1.55)	1.431 (+3.59)
C_2-C_3	1.377	1.396 (+1.36)	1.451 (+5.35)
C_3-C_4	1.382	1.403 (+1.53)	1.430 (+3.47)
C_4-N_{im}	1.428	1.403 (-1.80)	1.460 (+2.19)
N_{im}-C_{im}	1.250	1.278 (+2.26)	1.304 (+4.37)
C_{im}-C_5	1.487	1.477 (-0.73)	1.511 (+1.59)
C_5-C_6	1.385	1.405 (+1.45)	1.424 (+2.83)
C_6-C_6	1.391	1.389 (-0.13)	1.450 (+4.25)

Interatomic angle [°]

	SCXRD[14]	QuickFF (rel. diff. [%])	UFF (rel. diff. [%])
C_1-C_T-C_1	109.50	109.47 (-0.02)	109.52 (+0.02)
C_T-C_1-C_2	121.32	121.91 (+0.48)	120.23 (-0.90)
C_2-C_1-C_2	117.29	116.09 (-1.03)	119.47 (+1.86)
C_1-C_2-C_3	121.88	122.08 (+0.17)	119.66 (-1.82)
C_2-C_3-C_4	119.81	121.31 (+1.25)	119.84 (+0.03)
C_3-C_4-C_3	119.29	116.83 (-2.06)	119.86 (+0.48)
C_3-C_4-N_{im}	120.31	121.47 (+0.97)	119.99 (-0.26)
C_4-N_{im}-C_{im}	117.61	122.46 (+4.12)	127.89 (+8.73)
N_{im}-C_{im}-C_5	122.81	121.98 (-0.68)	119.93 (-2.35)
C_{im}-C_5-C_6	119.40	120.88 (+1.25)	119.93 (+0.44)
C_6-C_5-C_6	121.21	118.13 (-2.54)	119.94 (-1.05)
C_5-C_6-C_6	119.38	120.88 (+1.25)	119.96 (+0.49)

Dihedral angle [°]

	SCXRD[14]	QuickFF	UFF
C_3-C_4-N_{im}-C_{im}	32.83	36.83	61.42
C_4-N_{im}-C_{im}-C_5	1.63	2.65	11.84
N_{im}-C_{im}-C_5-C_6	7.20	4.51	75.35

Unit cell parameters

	SCXRD[14]	QuickFF (rel. diff. [%])	UFF (rel. diff. [%])
$\|\|\mathbf{a}\|\|$ [Å]	26.226	28.352 (+8.11)	30.359 (+15.76)
$\|\|\mathbf{b}\|\|$ [Å]	26.226	28.352 (+8.11)	30.359 (+15.76)
$\|\|\mathbf{c}\|\|$ [Å]	26.226	28.352 (+8.11)	30.359 (+15.76)
α [°]	90.00	90.02 (+0.02)	90.00 (+0.00)
β [°]	90.00	90.02 (+0.02)	90.00 (+0.00)
γ [°]	90.00	90.02 (+0.02)	90.00 (+0.00)
Volume [Å^3]	5209.63	5800.95 (+11.35)	6366.03 (+22.20)

Table S7: Comparison between our $(N, P, \boldsymbol{\sigma}_a = \mathbf{0}, T)$ molecular dynamics simulations and single crystal X-ray diffraction (SCXRD) data[14] of a selected set of internal coordinates of LZU-111 at 100 K and 1 atm.

Interatomic distance [Å]

	SCXRD[14]	QuickFF (rel. diff. [%])	UFF (rel. diff. [%])
C_T-C_1	1.529	1.574 (+2.94)	1.657 (+8.34)
C_1-C_2	1.379	1.403 (+1.74)	1.435 (+4.03)
C_2-C_3	1.389	1.396 (+0.52)	1.448 (+4.26)
C_3-C_4	1.403	1.402 (-0.06)	1.433 (+2.14)
C_4-N_{im}	1.497	1.404 (-6.20)	1.456 (-2.74)
N_{im}-C_{im}	1.254	1.278 (+1.94)	1.304 (+3.97)
C_{im}-C_5	1.478	1.478 (+0.04)	1.565 (+5.90)
C_5-C_6	1.391	1.403 (+0.87)	1.524 (+9.58)
C_6-C_7	1.389	1.393 (+0.24)	1.519 (+9.31)
C_7-C_8	1.390	1.406 (+1.13)	1.514 (+8.94)
C_8-Si_T	1.839	1.896 (+3.11)	1.986 (+7.99)

Interatomic angle [°]

	SCXRD[14]	QuickFF (rel. diff. [%])	UFF (rel. diff. [%])
C_1-C_T-C_1	109.46	109.47 (+0.01)	109.54 (+0.07)
C_T-C_1-C_2	120.50	121.86 (+1.12)	120.20 (-0.25)
C_2-C_1-C_2	118.90	116.15 (-2.31)	119.43 (+0.45)
C_1-C_2-C_3	120.86	122.01 (+0.96)	119.68 (-0.97)
C_2-C_3-C_4	121.14	121.19 (+0.04)	119.87 (-1.05)
C_3-C_4-C_3	116.99	117.02 (+0.03)	119.90 (+2.49)
C_3-C_4-N_{im}	121.50	121.40 (-0.09)	119.97 (-1.26)
C_4-N_{im}-C_{im}	117.87	121.33 (+2.93)	126.81 (+7.58)
N_{im}-C_{im}-C_5	117.42	121.94 (+3.85)	119.94 (+2.15)
C_{im}-C_5-C_6	119.97	120.93 (+0.80)	120.11 (+0.11)
C_6-C_5-C_6	119.94	118.04 (-1.59)	119.64 (-0.25)
C_5-C_6-C_7	120.02	120.88 (+0.71)	119.81 (-0.17)
C_6-C_7-C_8	120.00	121.23 (+1.02)	119.96 (-0.03)
C_7-C_8-C_7	120.02	117.49 (-2.10)	119.81 (-0.17)
C_7-C_8-Si_T	119.92	121.15 (+1.03)	119.96 (+0.03)
C_8-Si_T-C_8	109.43	109.46 (+0.02)	109.62 (+0.17)

Dihedral angle [°]

	SCXRD[14]	QuickFF	UFF
C_3-C_4-N_{im}-C_{im}	28.83	53.14	83.48
C_4-N_{im}-C_{im}-C_5	4.50	3.23	7.50
N_{im}-C_{im}-C_5-C_6	32.16	6.43	64.53

Unit cell parameters

	SCXRD[14]	QuickFF (rel. diff. [%])	UFF (rel. diff. [%])
$\|\|\mathbf{a}\|\|$ [Å]	20.396	21.216 (+4.02)	21.699 (+6.39)
$\|\|\mathbf{b}\|\|$ [Å]	20.396	21.216 (+4.02)	21.699 (+6.39)
$\|\|\mathbf{c}\|\|$ [Å]	20.396	21.216 (+4.02)	21.699 (+6.39)
α [°]	90.00	89.95 (-0.05)	90.00 (-0.00)
β [°]	90.00	89.95 (-0.05)	90.00 (-0.00)
γ [°]	90.00	89.95 (-0.05)	90.00 (-0.00)
Volume [Å^3]	12166.01	12668.09 (+4.13)	14363.32 (+18.06)

Table S8: Comparison between our $(N, P, \boldsymbol{\sigma}_a = \mathbf{0}, T)$ molecular dynamics simulations and 3D rotation electron diffraction (RED) data[15] of a selected set of internal coordinates of COF-320 at 89 K and 1 atm.

Interatomic distance [Å]

	RED[15]	QuickFF (rel. diff. [%])	UFF (rel. diff. [%])
C_T-C_1	1.530	1.574 (+2.83)	1.640 (+7.13)
C_1-C_2	1.390	1.402 (+0.86)	1.442 (+3.75)
C_2-C_3	1.390	1.396 (+0.44)	1.447 (+4.07)
C_3-C_4	1.390	1.403 (+0.92)	1.423 (+2.39)
C_4-N_{im}	1.433	1.406 (-1.88)	1.464 (+2.19)
N_{im}-C_{im}	1.290	1.279 (-0.85)	1.302 (+0.98)
C_{im}-C_5	1.440	1.476 (+2.50)	1.525 (+5.88)
C_5-C_6	1.389	1.403 (+0.98)	1.430 (+2.95)
C_6-C_7	1.389	1.392 (+0.21)	1.439 (+3.63)
C_7-C_8	1.392	1.406 (+1.01)	1.426 (+2.41)
C_8-C_8	1.442	1.492 (+3.47)	1.552 (+7.63)

Interatomic angle [°]

	RED[15]	QuickFF (rel. diff. [%])	UFF (rel. diff. [%])
C_1-C_T-C_1	109.50	109.48 (-0.02)	109.50 (-0.00)
C_T-C_1-C_2	119.96	121.84 (+1.57)	122.52 (+2.14)
C_2-C_1-C_2	120.07	116.11 (-3.30)	114.90 (-4.31)
C_1-C_2-C_3	119.96	122.08 (+1.76)	121.87 (+1.59)
C_2-C_3-C_4	119.98	121.36 (+1.15)	120.22 (+0.20)
C_3-C_4-N_{im}	119.87	121.52 (+1.38)	120.69 (+0.68)
C_4-N_{im}-C_{im}	127.06	122.18 (-3.84)	127.44 (+0.30)
N_{im}-C_{im}-C_5	124.89	122.46 (-1.94)	119.96 (-3.94)
C_{im}-C_5-C_6	119.95	121.15 (+1.00)	120.46 (+0.42)
C_6-C_5-C_6	120.10	117.57 (-2.10)	118.88 (-1.02)
C_5-C_6-C_7	119.99	121.13 (+0.95)	120.32 (+0.28)
C_6-C_7-C_8	120.06	121.44 (+1.15)	120.28 (+0.19)
C_7-C_8-C_7	119.81	117.03 (-2.33)	119.22 (-0.50)
C_7-C_8-C_8	119.84	121.44 (+1.33)	120.35 (+0.42)

Dihedral angle [°]

	RED[15]	QuickFF	UFF
C_3-C_4-N_{im}-C_{im}	19.30	31.70	60.30
C_4-N_{im}-C_{im}-C_5	4.69	4.88	11.22
N_{im}-C_{im}-C_5-C_6	1.38	6.74	70.40
C_7-C_8-C_8-C_7	81.33	37.51	68.97

Unit cell parameters

	RED[15]	QuickFF (rel. diff. [%])	UFF (rel. diff. [%])
$\|\|\mathbf{a}\|\|$ [Å]	30.170	33.377 (+10.63)	35.369 (+17.23)
$\|\|\mathbf{b}\|\|$ [Å]	30.170	33.377 (+10.63)	35.369 (+17.23)
$\|\|\mathbf{c}\|\|$ [Å]	30.170	33.377 (+10.63)	35.369 (+17.23)
α [°]	90.00	89.89 (-0.12)	90.00 (-0.00)
β [°]	90.00	89.89 (-0.12)	90.00 (-0.00)
γ [°]	90.00	89.89 (-0.12)	90.00 (-0.00)
Volume [Å^3]	6627.38	7354.46 (+10.97)	8568.62 (+29.29)

Table S9: Comparison between our $(N, P, \boldsymbol{\sigma}_a = \mathbf{0}, T)$ molecular dynamics simulations and 3D rotation electron diffraction (RED) data[15] of a selected set of internal coordinates of COF-320 at 298 K and 1 atm.

Interatomic distance [Å]

	RED[15]	QuickFF (rel. diff. [%])	UFF (rel. diff. [%])
C_T-C_1	1.551	1.577 (+1.66)	1.642 (+5.84)
C_1-C_2	1.414	1.403 (-0.75)	1.442 (+2.03)
C_2-C_3	1.397	1.397 (-0.03)	1.448 (+3.62)
C_3-C_4	1.397	1.402 (+0.35)	1.424 (+1.95)
C_4-N_{im}	1.428	1.407 (-1.48)	1.465 (+2.58)
N_{im}-C_{im}	1.320	1.279 (-3.07)	1.303 (-1.28)
C_{im}-C_5	1.463	1.477 (+0.96)	1.528 (+4.43)
C_5-C_6	1.394	1.403 (+0.65)	1.431 (+2.68)
C_6-C_7	1.398	1.392 (-0.42)	1.440 (+3.01)
C_7-C_8	1.419	1.407 (-0.83)	1.428 (+0.61)
C_8-C_8	1.519	1.494 (-1.65)	1.556 (+2.41)

Interatomic angle [°]

	RED[15]	QuickFF (rel. diff. [%])	UFF (rel. diff. [%])
C_1-C_T-C_1	109.52	109.45 (-0.07)	109.49 (-0.03)
C_T-C_1-C_2	121.12	121.83 (+0.59)	122.58 (+1.21)
C_2-C_1-C_2	117.39	115.99 (-1.19)	114.65 (-2.34)
C_1-C_2-C_3	121.18	122.02 (+0.69)	121.82 (+0.53)
C_2-C_3-C_4	120.20	121.17 (+0.80)	120.12 (-0.06)
C_3-C_4-N_{im}	120.11	121.29 (+0.98)	120.67 (+0.47)
C_4-N_{im}-C_{im}	119.28	121.20 (+1.61)	127.34 (+6.75)
N_{im}-C_{im}-C_5	121.74	122.44 (+0.57)	120.28 (-1.20)
C_{im}-C_5-C_6	120.09	121.11 (+0.85)	120.46 (+0.31)
C_6-C_5-C_6	119.52	117.42 (-1.75)	118.63 (-0.75)
C_5-C_6-C_7	120.42	121.08 (+0.55)	120.24 (-0.15)
C_6-C_7-C_8	121.42	121.46 (+0.03)	120.24 (-0.97)
C_7-C_8-C_7	116.80	116.77 (-0.03)	118.88 (+1.78)
C_7-C_8-C_8	121.60	121.47 (-0.10)	120.42 (-0.97)

Dihedral angle [°]

	RED[15]	QuickFF	UFF
C_3-C_4-N_{im}-C_{im}	87.61	48.59	63.19
C_4-N_{im}-C_{im}-C_5	0.00	6.09	10.92
N_{im}-C_{im}-C_5-C_6	86.88	9.00	68.06
C_7-C_8-C_8-C_7	0.19	34.08	66.58

Unit cell parameters

	RED[15]	QuickFF (rel. diff. [%])	UFF (rel. diff. [%])
$\|\|\mathbf{a}\|\|$ [Å]	27.930	31.525 (+12.87)	36.289 (+29.93)
$\|\|\mathbf{b}\|\|$ [Å]	27.930	31.525 (+12.87)	36.289 (+29.93)
$\|\|\mathbf{c}\|\|$ [Å]	27.930	31.525 (+12.87)	36.289 (+29.93)
α [°]	90.00	90.90 (+1.00)	90.00 (+0.00)
β [°]	90.00	90.90 (+1.00)	90.00 (+0.00)
γ [°]	90.00	90.90 (+1.00)	90.00 (+0.00)
Volume [Å^3]	6899.71	7525.91 (+9.08)	8943.10 (+29.62)

S2.5 Additional notes on the applicability of the periodic force fields

Besides a measure for the synthetic likelihood, the deformation energy can also be adopted as an indicator of the reliability of the periodic system-specific force fields. The cluster force fields are fitted to the potential energy surface in an equilibrium point. Therefore, they are most accurate when the geometry remains close to its minimum. However, by introducing the SBU in a periodic framework, topological constraints are applied that force the SBU out of its equilibrium configuration. If the deviation from the minimum is too large, the cluster force field will not be sufficiently accurate anymore, since the system has moved from its harmonic region. Such a large SBU deformation is associated with a high deformation energy. Therefore, a low deformation energy indicates that the SBUs can stay close to the equilibrium geometry at which the cluster force fields are fitted, also in the periodic framework, and thus results in a high accuracy of the periodic system-specific force field.

The long-range Coulomb and van der Waals interactions between different SBUs are not present in the *ab initio* reference data, as only cluster calculations are performed. As such, the partial charges derived from the *ab initio* cluster data might not reproduce the periodic values accurately. This limitation has the largest impact on dense materials, such as layered 2D COFs or highly interpenetrated 3D COFs, where SBUs are placed closely together without forming a covalent bond. However, COFs are built up only from organic linkers, for which long-range interactions are less dominant than for example in MOFs, where the orbitals of the metal ions have a large spatial extent. Despite these minor limitations, cluster force fields are a powerful approach to generate system-specific periodic force fields for a wide range of reticular materials and have already proven to attain a high accuracy in MOFs[16–22] and COFs.[8,23,24]

S2.6 Case study: derivation of an angle term overlapping two SBUs

The workflow for deriving periodic force fields, as already elaborated upon in Section 2.4 of the main text, is illustrated in Fig. S31, together with its application on a single force field term as an instructive example. As force field terms corresponding to atom types contained within a single SBU core domain are trivially translated to a periodic term, we consider an overlap term,

e.g., the C-B-O angle highlighted on the right hand side in Fig. S31. The force field parameters of the associated angle terms for the constituent SBUs (11-01-01 and 06-08-01) are derived using QuickFF, which assigns it a harmonic term (see Eq. S2.18), and are reported in Table S10.

$$E_{\text{BENDAHARM}} = \frac{K}{2}(\Theta - \Theta_0)^2 \tag{S2.18}$$

In accordance with the distribution of the C-B-O atoms over the two SBUs (one in 11-01-01 and two in 06-08-01), the rescaling factors α are 1/3 and 2/3, respectively. The force field parameters of the periodic force field are derived from the ones of the cluster force field using a weighted average, with weights α_i, as illustrated in Fig. S31. These are summarized in Table S10.

Table S10: Illustration of the derivation of an angle term that spans two SBUs in the periodic force field. The parameters Θ_0 and K are derived from the ones of the cluster force fields of the consitituent SBUs (06-08-01 and 11-01-01) using a weighted average with weights α_i.

	Θ_0 [°]	K [kJ/mol/rad^2]	Rescaling factor α_i []
Cluster force field 06-08-01	124.76	395.12	2/3
Cluster force field 11-01-01	124.73	371.64	1/3
Periodic force field	124.75	387.29	1

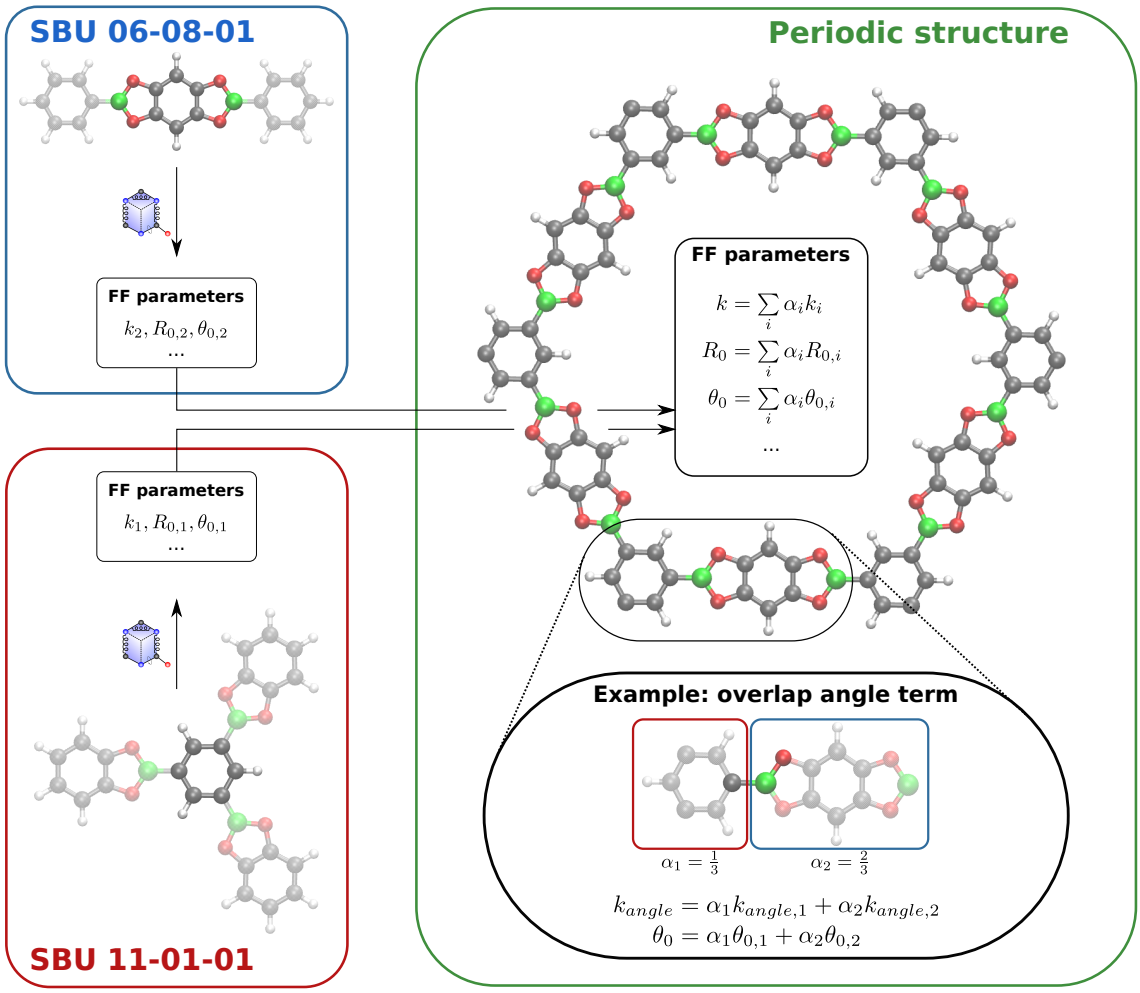


Figure S31: An illustration of the derivation of the force field for a periodic structure. The periodic structure is built up from SBUs 11-01-01 and 06-08-01, for which a cluster force field is derived with QuickFF. The termination of the clusters are semi-transparent. The parameters of the periodic structure are obtained as a weighted average of the parameters of the cluster force fields for its constituent SBUs. For the overlap angle term considered here: $\alpha_1 = 1/3$ for SBU 11-01-01 and $\alpha_2 = 2/3$ for SBU 06-08-01.

S3 Diversity metrics and subset selection

In Section 4.2 of the main text, the diversity of the ReDD-COFFEE database is compared with some other COF databases. These databases are introduced in Section S3.1. To be able to define the diversity metrics, each domain of the COF chemical space has to be featurized. For the chemical environment of the linker cores, linkages, and functional groups, revised autocorrelation functions (RACs) are adopted, which are introduced in detail in Section S3.2. A detailed definition of the three diversity metrics, *i.e.*, the variety V, the balance B, and the disparity D, is given is Section S3.3. Furthermore, a diverse subset of 10 000 COFs is extracted from the database using the derived features. The evolution in the diversity metrics upon changing the subset size is visualized in Section S3.4.

S3.1 COF databases

To assess the diversity of our ReDD-COFFEE database, its coverage of the material space is compared with that of the following four COF databases:

- **CoRE COF**:[25] Constructed with the aim to contain nearly all experimental COFs published in literature. Structure files are solvent-free and disorder-free. From v3 (Nov. 2018) on, also QEq charges are incorporated. CoRE stands for Computation-Ready, Experimental COF database. In this work, v4 from February 2020 is adopted, which contains 449 COFs.
  `https://core-cof.github.io/CoRE-COF-Database/`
- **CURATED**:[5] Contains experimental and density functional theory (DFT) optimized structures. Point charges are extracted from DFT calculations. CURATED stands for Clean, Uniform and Refined with Automatic Tracking from Experimental Database. In this work, v8 from February 2021 is adopted, which contains 632 COFs.
  `https://www.materialscloud.org/discover/curated-cofs`
- **Martin**:[26] Top-down approach (using Zeo++[27]) to generate COFs using commercially available precursors into **dia**, **bor**, or **ctn** topologies using imine, boronate ester, or borosilicate linkages. Afterwards, the structures are relaxed with DFT. 620 structures are generated and from them a total of 4 147 interpenetrated structures are derived.

http://www.nanoporousmaterials.org/databases/

- **Mercado**:[28] Top-down approach (using Zeo++[27]) to generate COFs with amide, amine, imine, or carbon-carbon linkages using 666 organic building blocks and 839 2D or 3D topologies. These topologies are selected based on symmetry considerations. Afterwards, the structures are relaxed with a transferable UFF+DREIDING force field. In this work, the latest version of this database (v3 from October 2018) is adopted, which contains 69 840 COFs.

  https://archive.materialscloud.org/record/2018.0003/v3

To be able to compare the structures in each of the databases, the preprocessing routines listed below are applied. The number of structures to which these routines apply are listed in Table S11.

1. Bonds are detected using the detect_bonds method of the molmod package.[29] Even if two hydrogen atoms are so close that they can form a bond, the bond is not formed. This step is skipped for the structures in the ReDD-COFFEE database as the bonds are already known.

2. Remove unphysical structures. A structure is defined as unphysical if it contains a carbon atom with more than four neighbors or if the framework is not connected periodically. These cases possibly occur due to incorrect periodic boundary conditions in the structure files.

3. Guests are removed from the structure. A guest is defined as a molecule that is not covalently bound to the framework.

4. Structures for which either the calculation of the pore geometry using Zeo++ or the computation of the revised autocorrelation (RAC) features failed are discarded. The pore geometry could not be calculated for nine COFs in our ReDD-COFFEE database. These materials have unit cells with very large or very small unit cell angles or lengths, or a combination of both. For eight structures in the database of Mercado, no linkage could be identified. Upon inspection, these structures were unphysical ones.

Table S11: Number of structures that are influenced by the preprocessing routines. The retained structures are obtained by removing the unphysical structures and the ones for which the geometry or RAC calculation failed from the original structures.

	CoRE	CURATED	Martin	Mercado	ReDD-COFFEE
Original structures	449	688	4 147	69 840	268 687
Structures with an H-H bond	4	0	1 234	147	0
Unphysical structures	4	2	210	305	0
Structures with guests	14	27	23	13	0
Pore geometry or RAC calculation failed	0	0	0	8	9
Retained structures	445	686	3 937	69 527	268 678

S3.2 Revised autocorrelation functions in COFs

The revised autocorrelation functions (RACs) are used in the main manuscript to describe the chemical environment of the linkers, linkages, and functional groups. The linkages that hold together the COF are identified by scanning the material graph for linkage patterns, as discussed below. Removing these linkages reveals the linker graph, which contains the separated linkers that constitute the COF. The parts of these linkers that are attached to the COF's skeleton with exactly one bond and do not exist of a single hydrogen atom are defined as the functional groups. The skeleton of a linker is the part to which the linkage is connected in the material graph. In the example of Fig. S32, the imine linkages are indicated in red in the material graph. The other atoms of the material graph are marked in black. In the linker graph, the fluor atoms are identified as functional groups, which are highlighted in green. The set of atoms in the linker graph that are no functional groups, are indicated in blue.

Once each of these environments is defined, the start and scope atom lists on which the RACs are calculated can be determined, as summarized in Fig. S32. For the linkage environment, the start atom list only contains atoms that are present in a linkage (red in Fig. S32), while the scope atom list is built up from all atoms in the material (black and red in Fig. S32). Therefore, these are calculated on the full material graph. The RACs from the linker environment are calculated by iterating over atom pairs that are both present in the same linker (blue and green in Fig. S32). While the scope list of the functional group environment again contains the linker atoms, the start atoms have to be present in a functional group (green in Fig. S32). Both the linker and

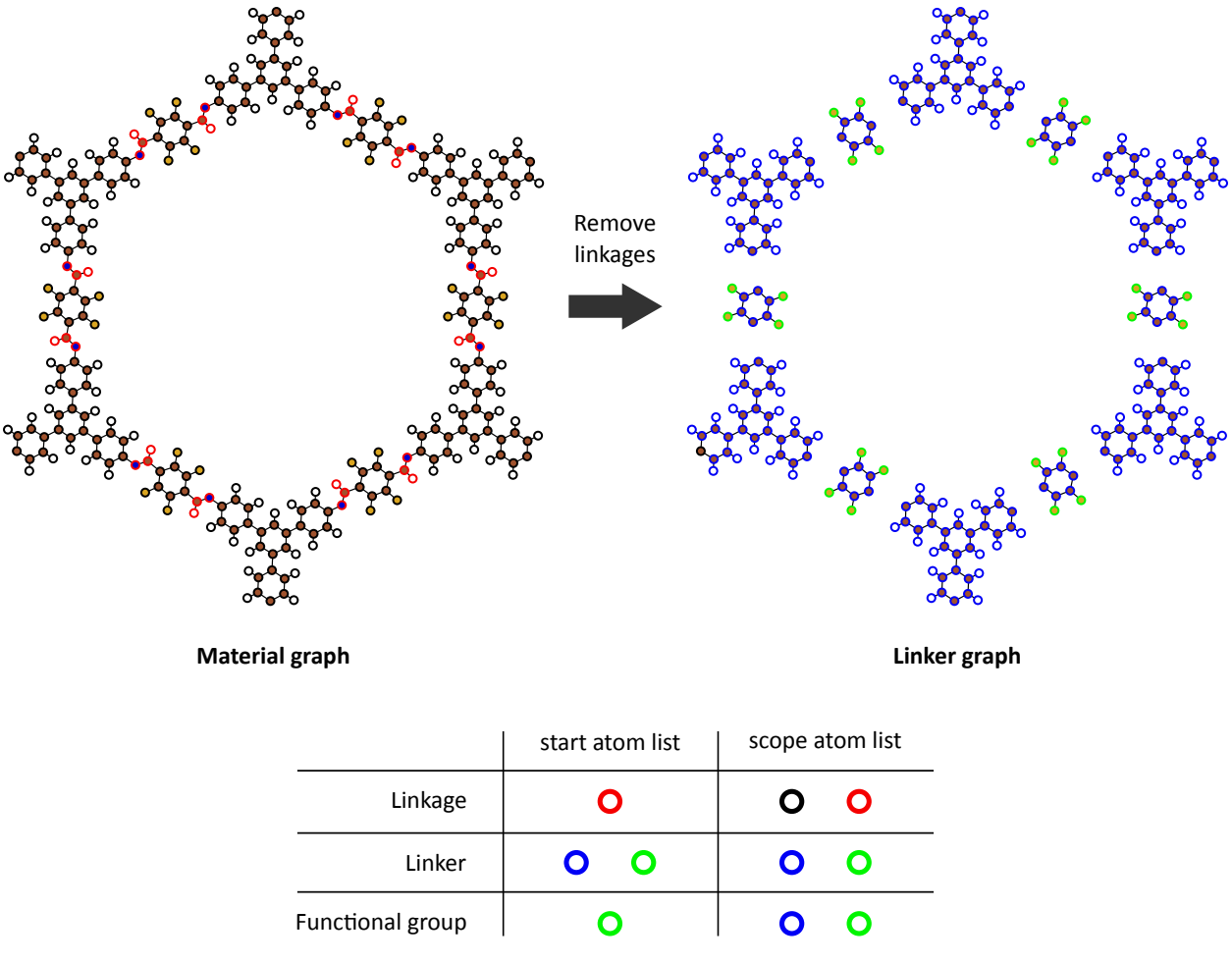


Figure S32: Illustration of the start and scope atom lists for the different chemical environments. The linkage atoms are indicated with red, whereas the other atoms in the material graph are indicated with black. By removing the linkages from the material graph, the linker graph (blue and green atoms) is obtained. Functional groups are highlighed in green. Color code for the atoms: hydrogen (white), carbon (brown), nitrogen (blue), fluorine (orange).

functional group RACs are calculated on the linker graph.

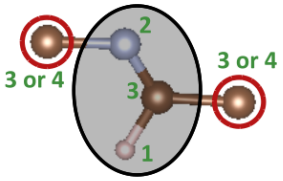
Pattern matching to identify COF linkages. The first step in determining the start and scope atom lists to calculate the RACs in COFs is identifying the linkages that hold the linkers together. This is done by scanning the material graph for linkage patterns. The linkage patterns that are used in this study are visualized in Fig. S33. To avoid that certain parts of the material graph would be wrongly recognized as a linkage, also atoms of the linker cores can be present in a linkage pattern. These atoms are only used to identify the pattern, but are not classified as linkage. Furthermore, additional restrictions are put on the connectivity of the atoms of some

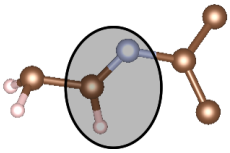
linkage patterns. Such restrictions can be either a pair of atoms in the pattern that are forbidden to be covalently bounded, or a specification of the number of covalently bounded neighbors to some atoms. These restrictions are visualized in Fig. S33. When no linkage was identified using the linkage patterns, a new scan was performed to partition the material into individual SBUs that are connected by a single carbon-carbon or carbon-nitrogen bond. To this end, SBU patterns for all carbon-carbon linked building blocks with a connectivity larger than two used in the study of Mercado *et al.* were implemented, as well as the carbon-nitrogen linked porphyrin ring used in Red-PV-COF.[30] The material graph is scanned to identify each of these SBUs, starting with the largest building blocks. SBUs are only detected if none of its atoms were already assigned to a larger SBU. In these materials, no linkage atoms are identified, but the bonds between the linker cores are broken to construct the linker graph.

The number of structures in which each linkage type is identified for each database is listed in Table S12. The fraction of structures in the ReDD-COFFEE database with two or more linkage types is relatively large. This can be explained by the presence of three linker cores containing a linkage type in its core among the set of linker cores from which the database is generated. As can be observed in Fig. S1, core05, core14, and core30 contain an imide, triazine, and benzimidazole fragment, respectively, the latter belonging to the class of "Other" linkage types. When one of these cores is adopted in a structure with another linkage type, such as imine, the structure will be identified to have a mixed linkage. The number of structures in each database that have a mixed linkage containing one of these linkage types are also listed in Table S12.

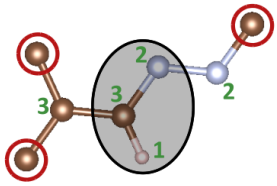
S3.3 Diversity metrics

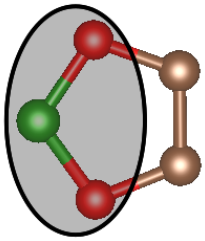
The diversity of structures present in (a subset of) the material space is determined by three diversity metrics. The full material space is defined as the union of all subsets that are used in the study, which are the four databases defined in Section S3.1 together with our ReDD-COFFEE database. Four domains are defined that describe the COF chemistry. The pore geometry is represented by a set of eight geometric properties, *i.e.*, the mass density, diameters of the largest included sphere, free sphere, and included sphere along the free path, gravimetric and volumetric accessible surface areas, gravimetric accessible volume, and pore fraction. The chemical



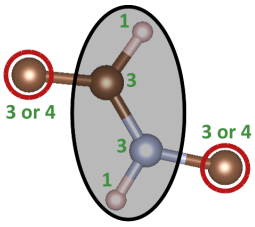


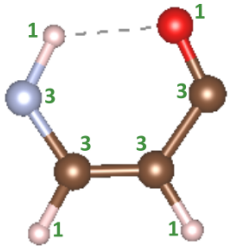
(a) Imine



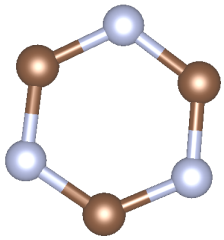


(b) Boronate Ester

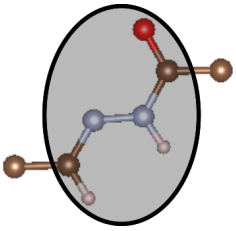


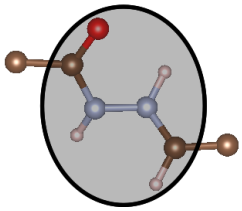


(c) (Keto)enamine

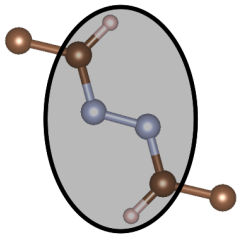


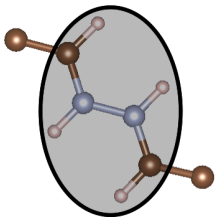
(d) Triazine



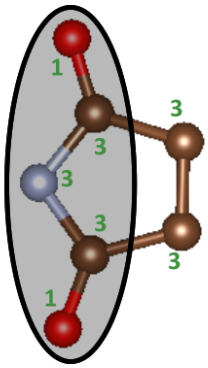


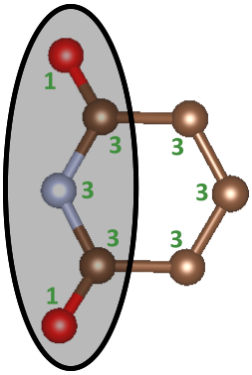
(e) (Acyl)hydrazone



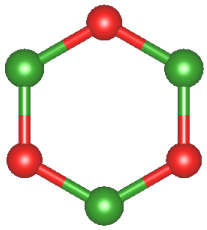


(f) Azine

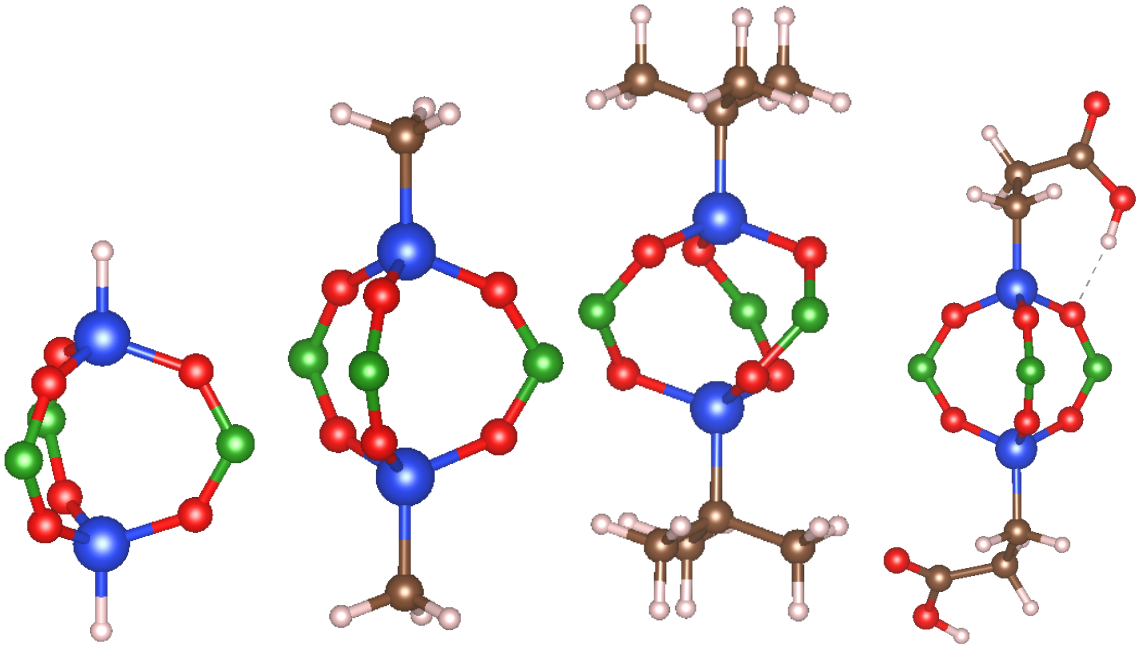




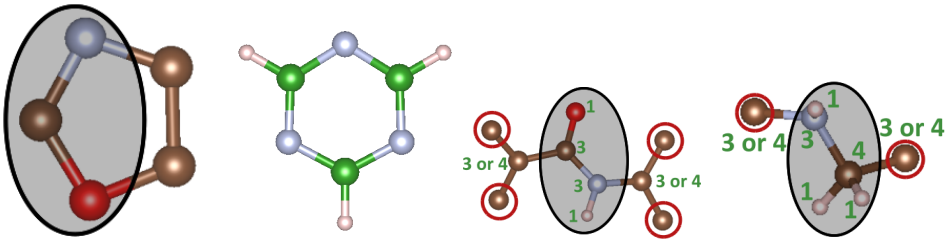
(g) Imide



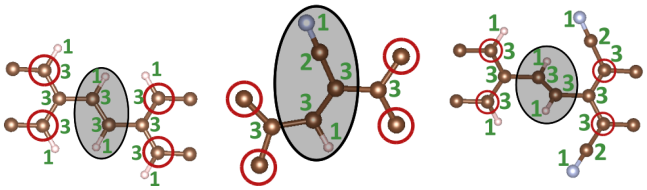
(h) Boroxine



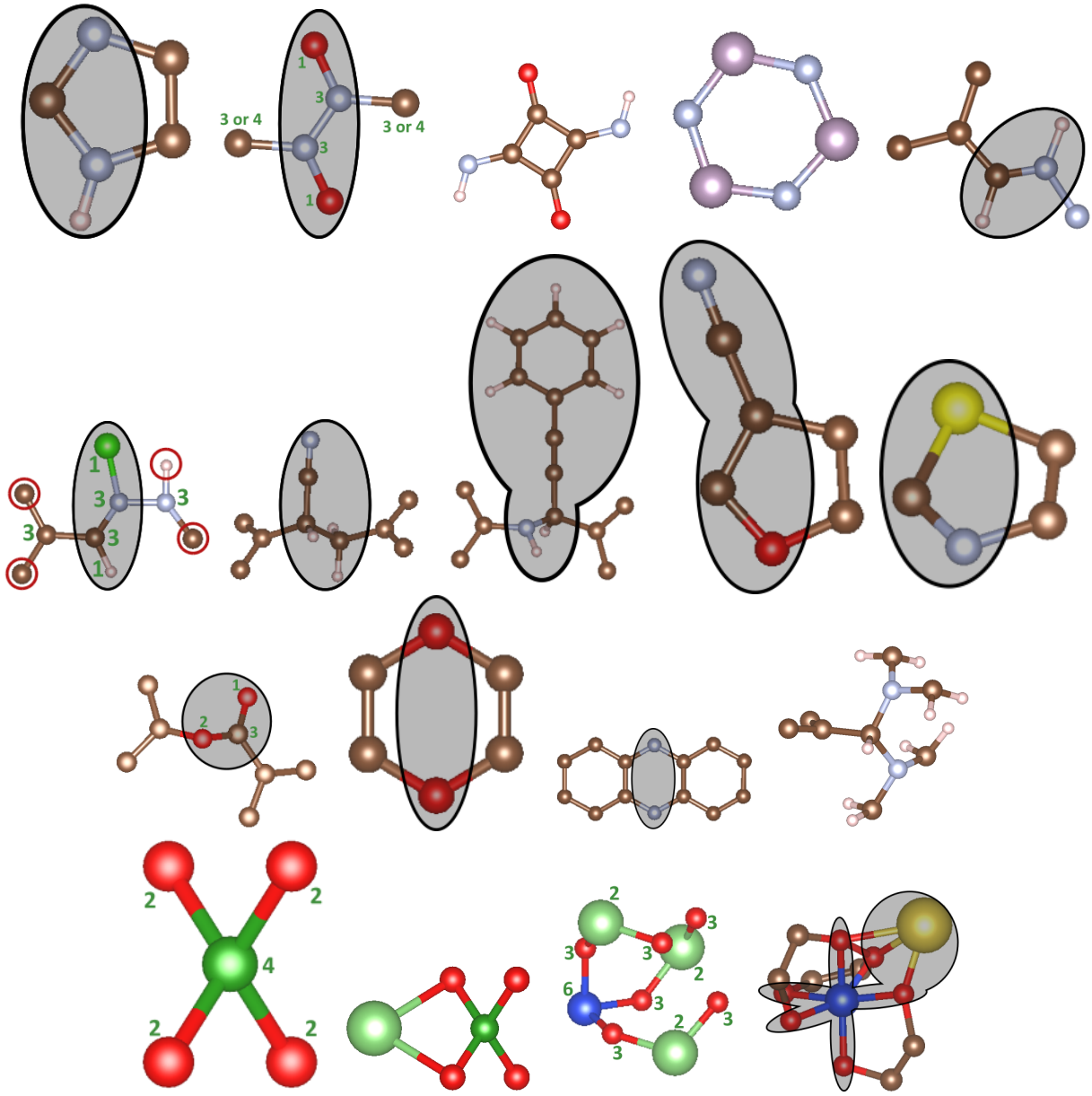
(i) Borosilicate

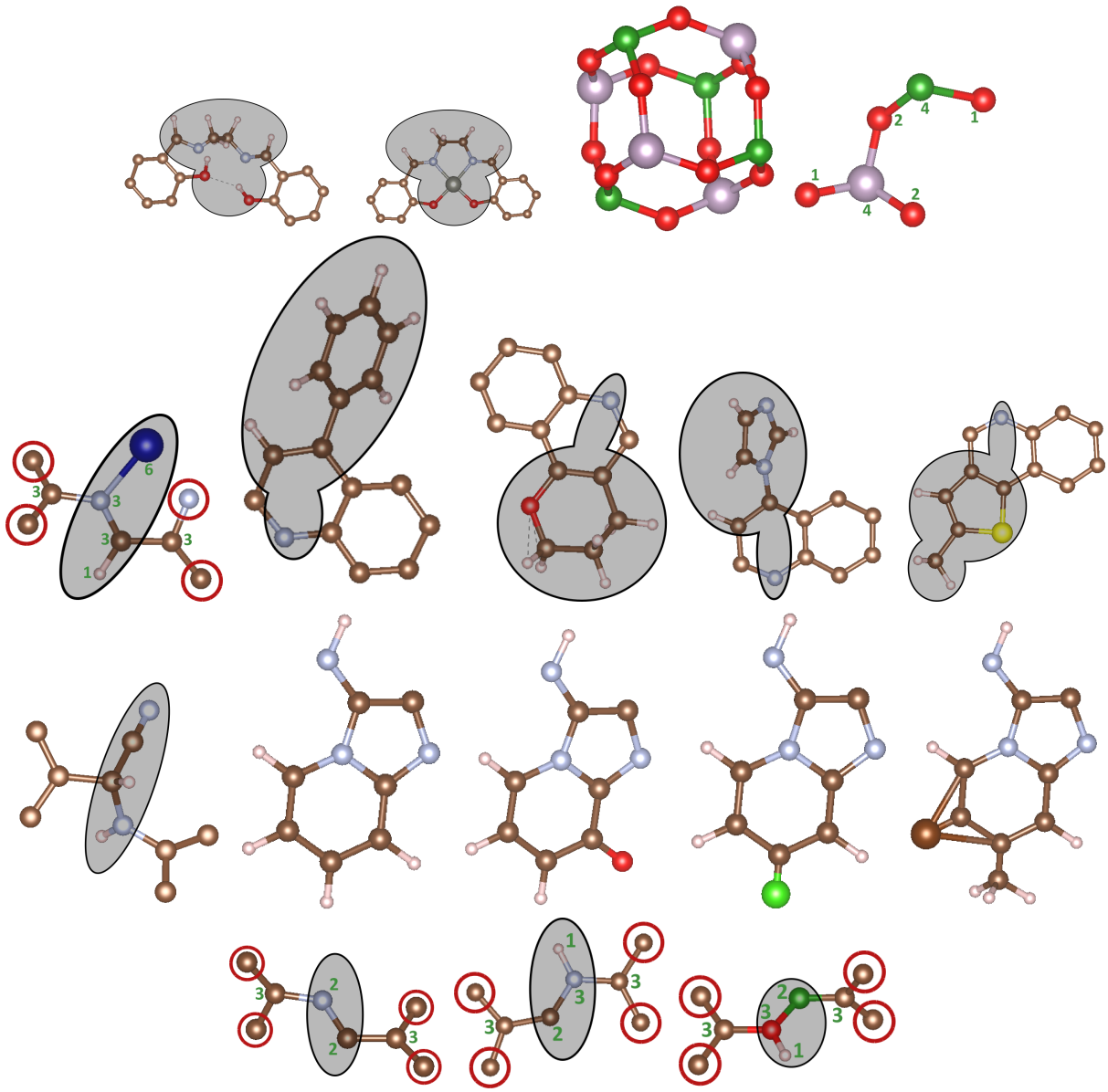


(j) Oxazoline **(k)** Borazine **(l)** Amide **(m)** Amine



(n) Olefin


3 or 4
3 or 4



(o) Other linkage types

Figure S33: Overview of the linkage patterns used to detect linkages in the material graph. The patterns are grouped according to the classes defined in Fig. 4 in the main text. If multiple patterns are given, any of them is recognized as that specific linkage type. The atoms indicated with red circles are not allowed to connect with each other. If a green number is placed beside an atom, the pattern is only detected if this number matches the coordination number. The atoms that are identified as linkage atoms are indicated in gray. When no gray area is present, all atoms in the linkage pattern are linkage atoms. Color code: hydrogen (white), lithium (pale green), boron (dark green), carbon (brown), nitrogen (light blue), oxygen (red), sodium (light yellow), silicon (dark blue), phosphorus (purple), sulfur (dark yellow), chlorine (bright green), cobalt (dark blue), zinc (gray), bromine (light brown).

Table S12: Distribution of frequently occurring linkage types in five COF databases, both in absolute value as relative compared to the full database. This is the numerical data for Fig. 4 in the main text. Also the number of structures that have a mixed linkage containing a triazine, imide, or other linkage are enumerated.

Linkage	CoRE	CURATED	Martin	Mercado	ReDD-COFFEE
Imine	164 (36.85%)	290 (42.27%)	1781 (45.24%)	32707 (47.04%)	14154 (5.27%)
Boronate Ester	63 (14.16%)	65 (9.48%)	2003 (50.88%)	0 (0.00%)	28427 (10.58%)
(Keto)enamine	49 (11.01%)	42 (6.12%)	0 (0.00%)	0 (0.00%)	26158 (9.74%)
Triazine	10 (2.25%)	19 (2.77%)	0 (0.00%)	2144 (3.08%)	2461 (0.92%)
(Acyl)hydrazone	23 (5.17%)	22 (3.21%)	0 (0.00%)	0 (0.00%)	26769 (9.96%)
Azine	12 (2.70%)	16 (2.33%)	0 (0.00%)	0 (0.00%)	42483 (15.81%)
Imide	8 (1.80%)	14 (2.04%)	0 (0.00%)	0 (0.00%)	32665 (12.16%)
Boroxine	4 (0.90%)	7 (1.02%)	0 (0.00%)	0 (0.00%)	1522 (0.57%)
Borosilicate	1 (0.22%)	1 (0.15%)	10 (0.25%)	0 (0.00%)	1881 (0.70%)
Oxazoline	4 (0.90%)	5 (0.73%)	0 (0.00%)	0 (0.00%)	27704 (10.31%)
Borazine	1 (0.22%)	2 (0.29%)	0 (0.00%)	0 (0.00%)	2069 (0.77%)
Amide	3 (0.67%)	5 (0.73%)	0 (0.00%)	6479 (9.32%)	0 (0.00%)
Amine	0 (0.00%)	0 (0.00%)	0 (0.00%)	5418 (7.79%)	0 (0.00%)
Carbon-Carbon	5 (1.12%)	5 (0.73%)	0 (0.00%)	18061 (25.97%)	0 (0.00%)
Olefin	1 (0.22%)	13 (1.90%)	0 (0.00%)	667 (0.96%)	0 (0.00%)
Other	28 (6.29%)	80 (11.66%)	0 (0.00%)	0 (0.00%)	0 (0.00%)
Mixed	69 (15.51%)	100 (14.58%)	143 (3.63%)	4051 (5.83%)	62394 (23.22%)
Mixed (including triazine)	43 (9.66%)	49 (7.14%)	0 (0.00%)	2793 (4.02%)	25793 (9.60%)
Mixed (including imide)	13 (2.92%)	13 (1.90%)	0 (0.00%)	0 (0.00%)	28855 (10.74%)
Mixed (including other)	6 (1.35%)	30 (4.37%)	40 (1.02%)	0 (0.00%)	16446 (6.12%)

environment of the linkers, linkages, and functional groups are described by a 48 RACs for each chemical environment, as defined S3.2.

For each set of descriptors, three diversity metrics are defined. For the first two, the structures are clustered in a specific number of bins, here chosen to be 1 000, which are defined using *k*-means clustering.

- **Variety** V: checks if each region in the material space is examined by measuring the number of bins that are sampled. In the ideal case ($V = 1$), all bins in the material space are occupied by at least one structure of the subset.

$$V = \frac{\text{number of bins occupied by structures in the subset}}{\text{total number of bins}} \tag{S3.19}$$

- **Balance** B: checks if each region in the material space is sampled equivalently by measuring the evenness of the distribution of the materials among the sampled bins. In the ideal case ($B = 1$), all bins are equally occupied by the structures in the subset.

  The probability that a structure from a certain database is assigned to bin x_i is given by

$$P(x_i) = \frac{\text{number of structures in bin } x_i}{\text{total number of structures in the subset}} \tag{S3.20}$$

  Shannon entropy is defined as:

$$H(X) = -\sum_i P(x_i) \ln(P(x_i)) \tag{S3.21}$$

  which obtains a maximum for an even distribution: $H_{\max} = \ln(N)$, with N the number of bins.

  The Pielou evenness of the distribution, which is used to quantify the balance, is then defined as:

$$B = PL(X) = \frac{1 - \exp(H(X))}{1 - \exp(H_{\max})} \tag{S3.22}$$

- **Disparity** D: checks how large the region is that is sampled in material space. When some bins cover a larger region in material space than others, the variety V would offer a false representation. As such, the disparity also measures the spread of the sampled bins.

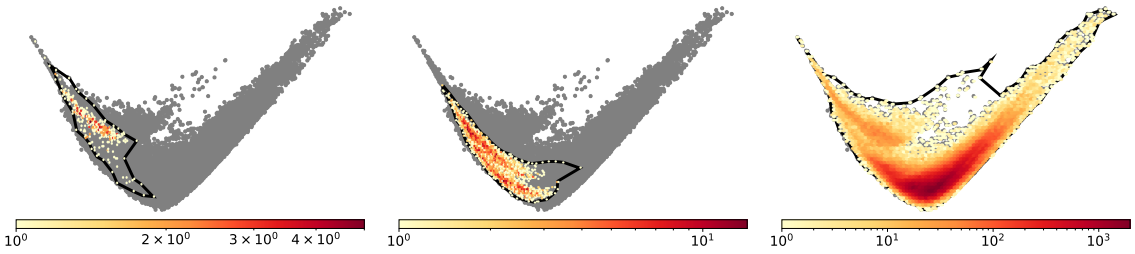
  After a principal components analysis (PCA) is performed, all structures in the material space are projected on the first two principal components and both the concave hull of the subset and of the whole material space are calculated. The disparity D is defined as the ratio of the area encompassed by both concave hulls. In the ideal case ($D = 1$), the subset and the database cover the same area of the principal components space.

$$D = \frac{\text{area of the concave hull of the subset}}{\text{area of the concave hull of the material space}} \tag{S3.23}$$

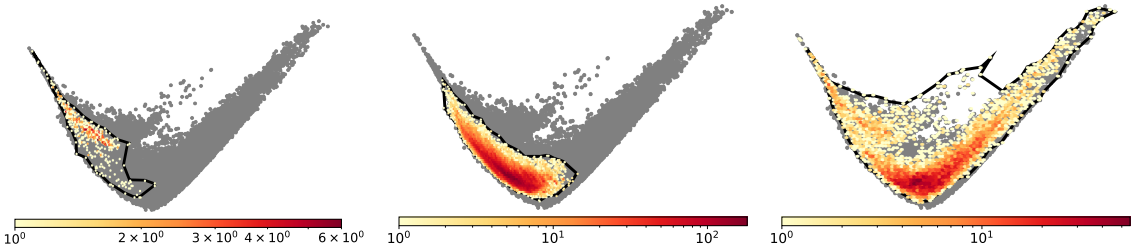
An overview of the diversity metrics computed for the five considered databases and the subset of 10 000 COFs is provided in Table S13. In Figs. S34-S41, PCA and t-SNE plots of each domain are provided. In the PCA plots, also the concave hulls of each database are indicated, which are used in the calculation of the disparities. To illustrate the balance of each database, the occupation of the sampled bins is plotted in Fig. S42.

Table S13: Diversity metrics of all domains for the five databases and the subset of 10 000 structures. This data can be used to reproduce Fig. 5 in the main text.

Domain	Database	Variety (V)	Balance (B)	Disparity (D)
Geometry	CoRE	0.092	0.544	0.167
	CURATED	0.132	0.497	0.185
	Martin	0.270	0.529	0.196
	Mercado	0.473	0.581	0.245
	ReDD-COFFE	0.997	0.685	0.986
	Subset (10 000)	0.942	0.777	0.932
Linkages	CoRE	0.104	0.360	0.289
	CURATED	0.162	0.285	0.478
	Martin	0.104	0.382	0.182
	Mercado	0.296	0.229	0.305
	ReDD-COFFEE	0.587	0.248	0.589
	Subset (10 000)	0.55	0.245	0.575
Linker cores	CoRE	0.230	0.754	0.322
	CURATED	0.332	0.723	0.406
	Martin	0.566	0.544	0.259
	Mercado	0.491	0.435	0.752
	ReDD-COFFEE	0.545	0.531	0.559
	Subset (10 000)	0.539	0.709	0.548
Functional groups	CoRE	0.083	0.106	0.213
	CURATED	0.124	0.090	0.343
	Martin	0.236	0.255	0.611
	Mercado	0.662	0.142	0.539
	ReDD-COFFEE	0.042	0.065	0.053
	Subset (10 000)	0.038	0.096	0.053

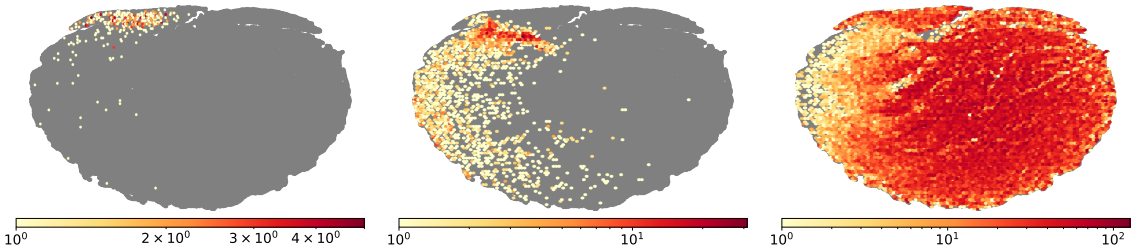


(a) CoRE **(b)** Martin **(c)** ReDD-COFFEE

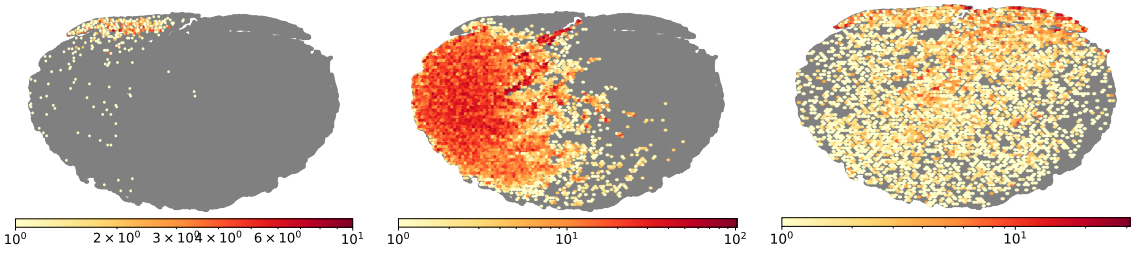


(d) CURATED **(e)** Mercado **(f)** Subset (10 000)

Figure S34: PCA plots of the structures in the five databases and the diverse subset for the pore geometry. The gray background is the full material space, which is overlaid with a histogram of the structures in the respective databases. The colorbar indicates the number of structures in each histogram bin. The boundary of the concave hull that is used to calculate the disparity is indicated with a black line.

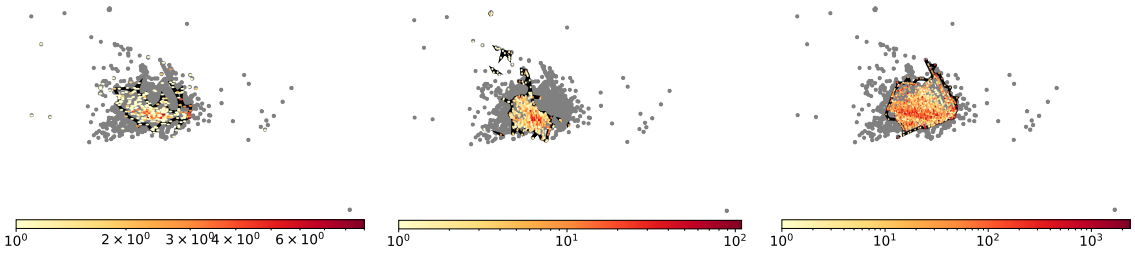


(a) CoRE **(b)** Martin **(c)** ReDD-COFFEE

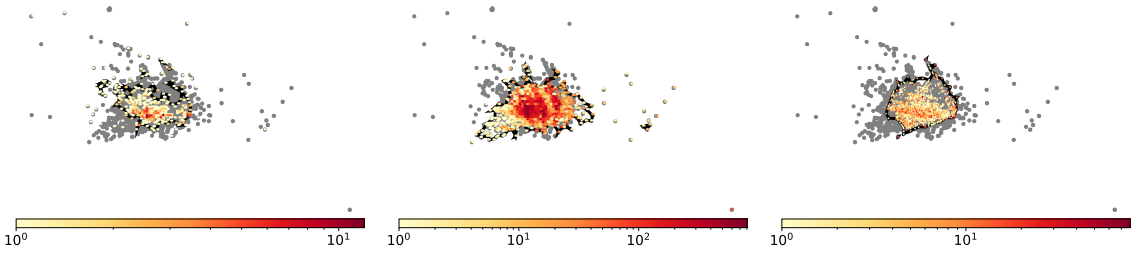


(d) CURATED **(e)** Mercado **(f)** Subset (10 000)

Figure S35: t-SNE plots of the structures in the five databases and the diverse subset for the pore geometry. The gray background is the full material space, which is overlaid with a histogram of the structures in the respective databases. The colorbar indicates the number of structures in each histogram bin.

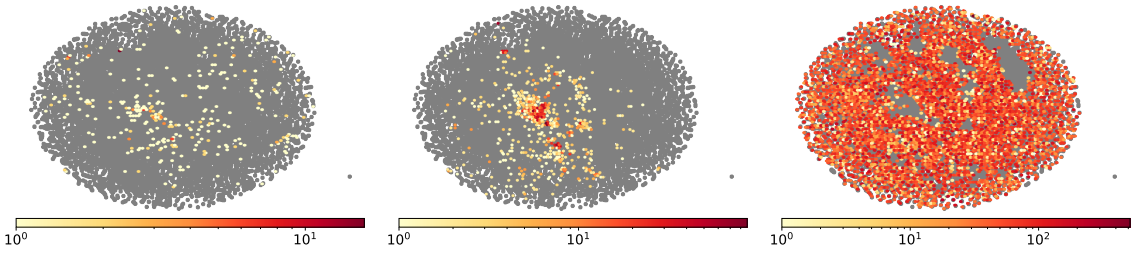


(a) CoRE **(b)** Martin **(c)** ReDD-COFFEE

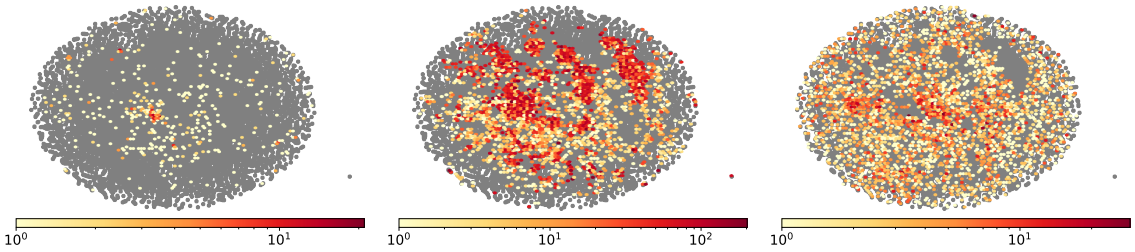


(d) CURATED **(e)** Mercado **(f)** Subset (10 000)

Figure S36: PCA plots of the structures in the five databases and the diverse subset for the linker chemical environment. The gray background is the full material space, which is overlaid with a histogram of the structures in the respective databases. The colorbar indicates the number of structures in each histogram bin. The boundary of the concave hull that is used to calculate the disparity is indicated with a black line.

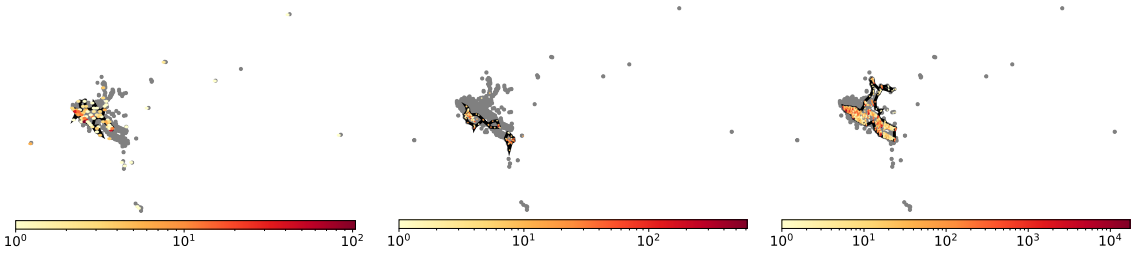


(a) CoRE **(b)** Martin **(c)** ReDD-COFFEE

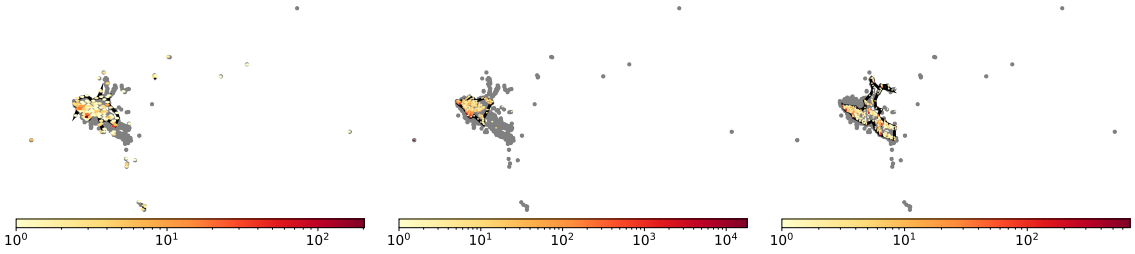


(d) CURATED **(e)** Mercado **(f)** Subset (10 000)

Figure S37: t-SNE plots of the structures in the five databases and the diverse subset for the linker chemical environment. The gray background is the full material space, which is overlaid with a histogram of the structures in the respective databases. The colorbar indicates the number of structures in each histogram bin.

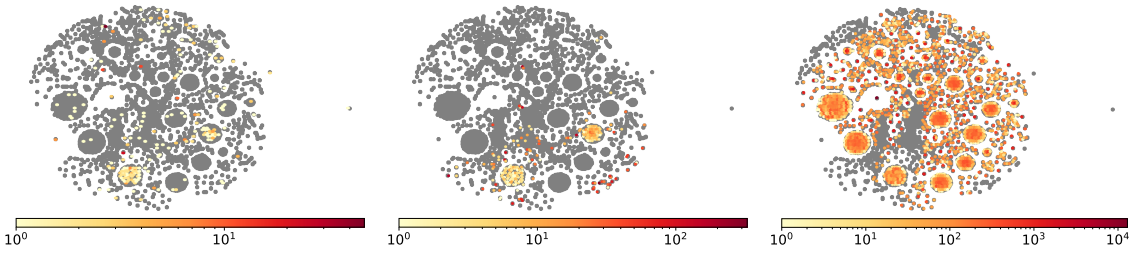


(a) CoRE **(b)** Martin **(c)** ReDD-COFFEE

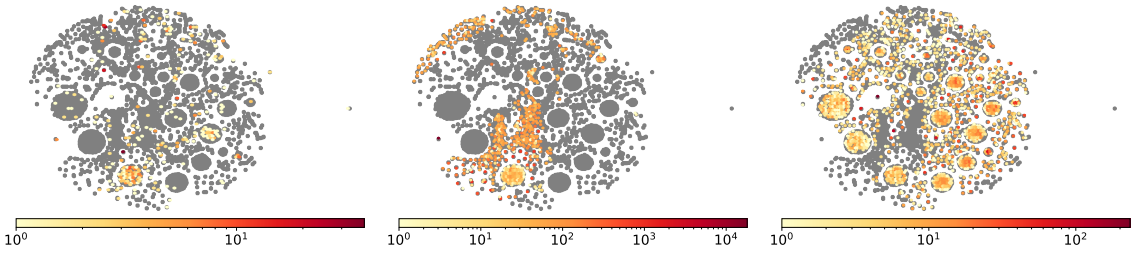


(d) CURATED **(e)** Mercado **(f)** Subset (10 000)

Figure S38: PCA plots of the structures in the five databases and the diverse subset for the linkage chemical environment. The gray background is the full material space, which is overlaid with a histogram of the structures in the respective databases. The colorbar indicates the number of structures in each histogram bin. The boundary of the concave hull that is used to calculate the disparity is indicated with a black line.

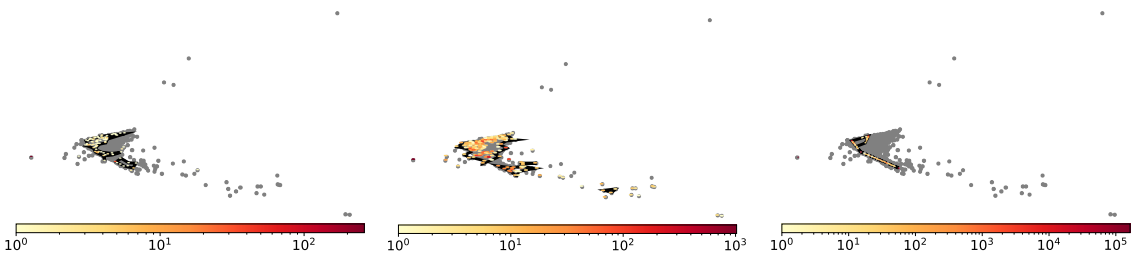


(a) CoRE **(b)** Martin **(c)** ReDD-COFFEE

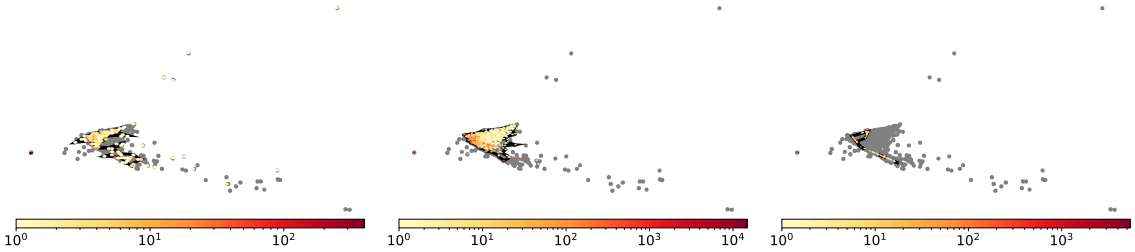


(d) CURATED **(e)** Mercado **(f)** Subset (10 000)

Figure S39: t-SNE plots of the structures in the five databases and the diverse subset for the linkage chemical environment. The gray background is the full material space, which is overlaid with a histogram of the structures in the respective databases. The colorbar indicates the number of structures in each histogram bin.

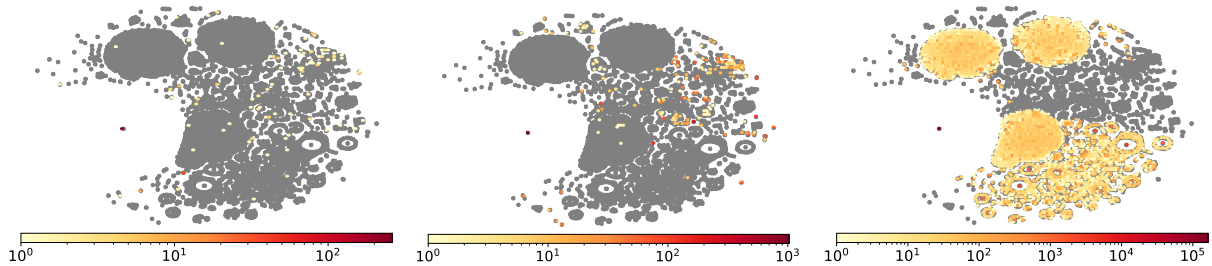


(a) CoRE **(b)** Martin **(c)** ReDD-COFFEE

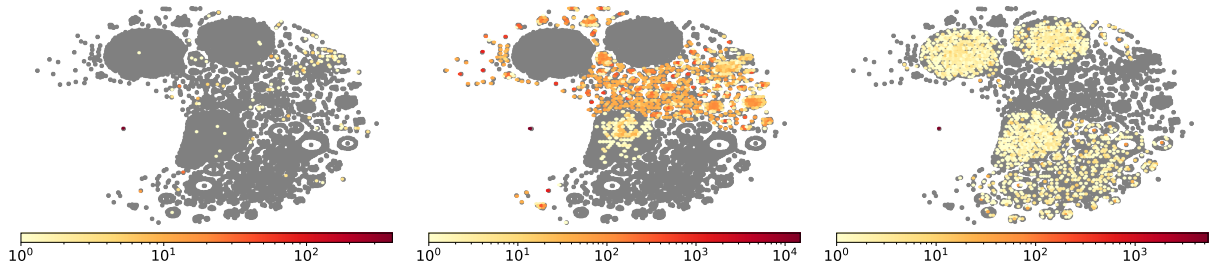


(d) CURATED **(e)** Mercado **(f)** Subset (10 000)

Figure S40: PCA plots of the structures in the five databases and the diverse subset for the functional group chemical environment. The gray background is the full material space, which is overlaid with a histogram of the structures in the respective databases. The colorbar indicates the number of structures in each histogram bin. The boundary of the concave hull that is used to calculate the disparity is indicated with a black line.

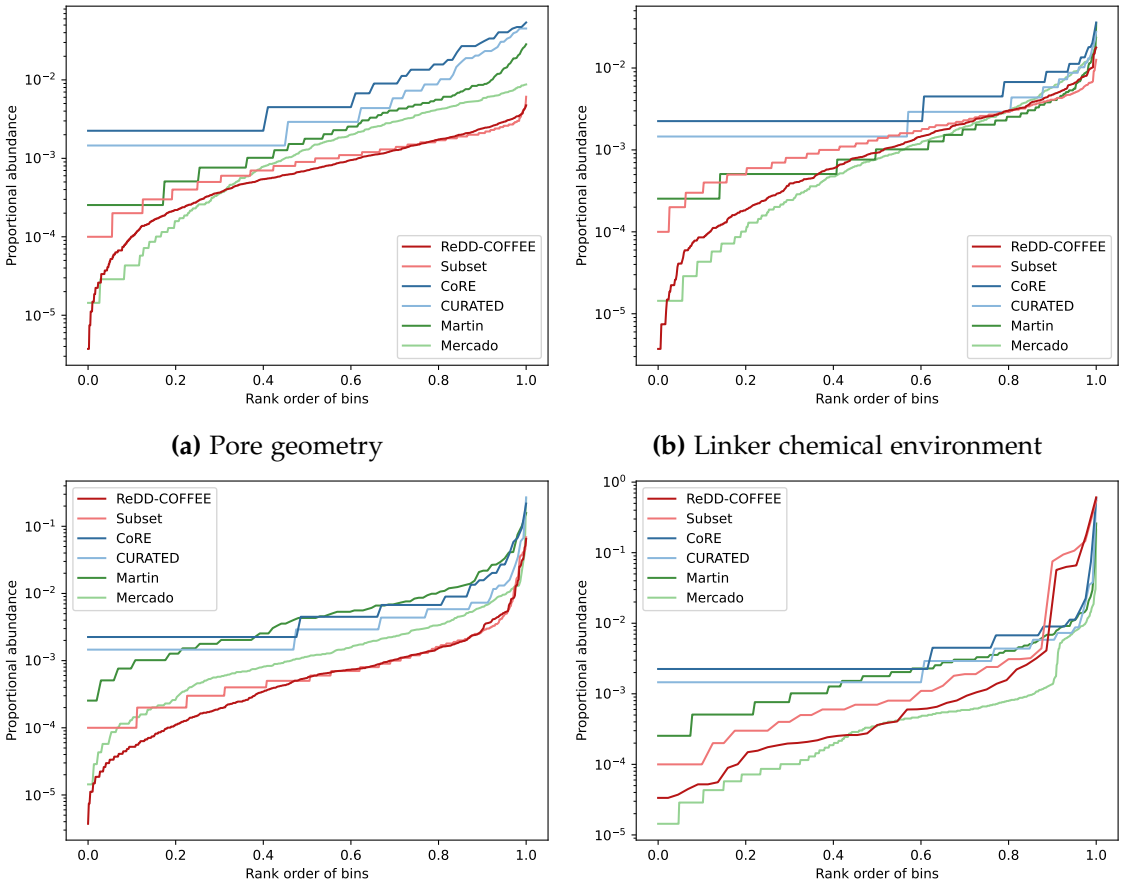


(a) CoRE **(b)** Martin **(c)** ReDD-COFFEE



(d) CURATED **(e)** Mercado **(f)** Subset (10 000)

Figure S41: t-SNE plots of the structures in the five databases and the diverse subset for the functional group chemical environment. The gray background is the full material space, which is overlaid with a histogram of the structures in the respective databases. The colorbar indicates the number of structures in each histogram bin.



(a) Pore geometry

(b) Linker chemical environment

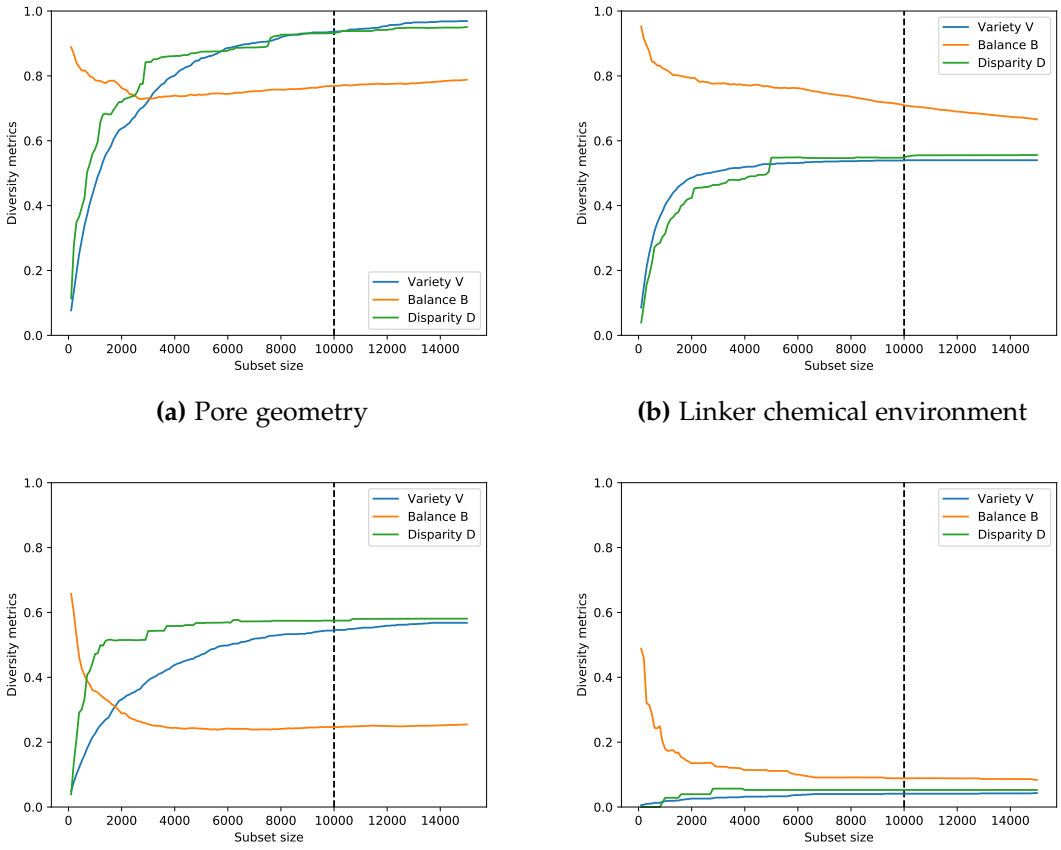
(c) Linkage chemical environment

(d) Functional group chemical environment

Figure S42: Occupation of the sampled bins for each of the four domains for the five databases and the diverse subset. The bins are ranked by increasing occupation. A flat distribution would result in a perfect evenness, and thus a high balance B.

S3.4 Subset selection

As explained in the main text, an iterative procedure is followed to select a diverse and representative subset of 10 000 structures from the complete ReDD-COFFEE database. The iterative procedure starts from a random initial structure, which is here chosen as COF-5 (hcb_18-08-01_01-01-01). In Fig. S43, the effect of an increasing subset size is plotted for each domain. Whereas the variety V and disparity D are very small as long as the subset contains a limited amount of materials, a high balance B is reached as almost empty bins are occupied. When more materials are added to the subset, the balance B drops and afterwards steadily increases. Both the variety V and disparity D increase rapidly when new materials are added to the subset.



(a) Pore geometry

(b) Linker chemical environment

(c) Linkage chemical environment

(d) Functional group chemical environment

Figure S43: Influence of the subset size on the three diversity metrics. The variety V, balance B, and disparity D are plotted in function of the subset size for the four domains, *i.e.*, the pore geometry and the chemical environment of the linkers, linkages, and functional groups.

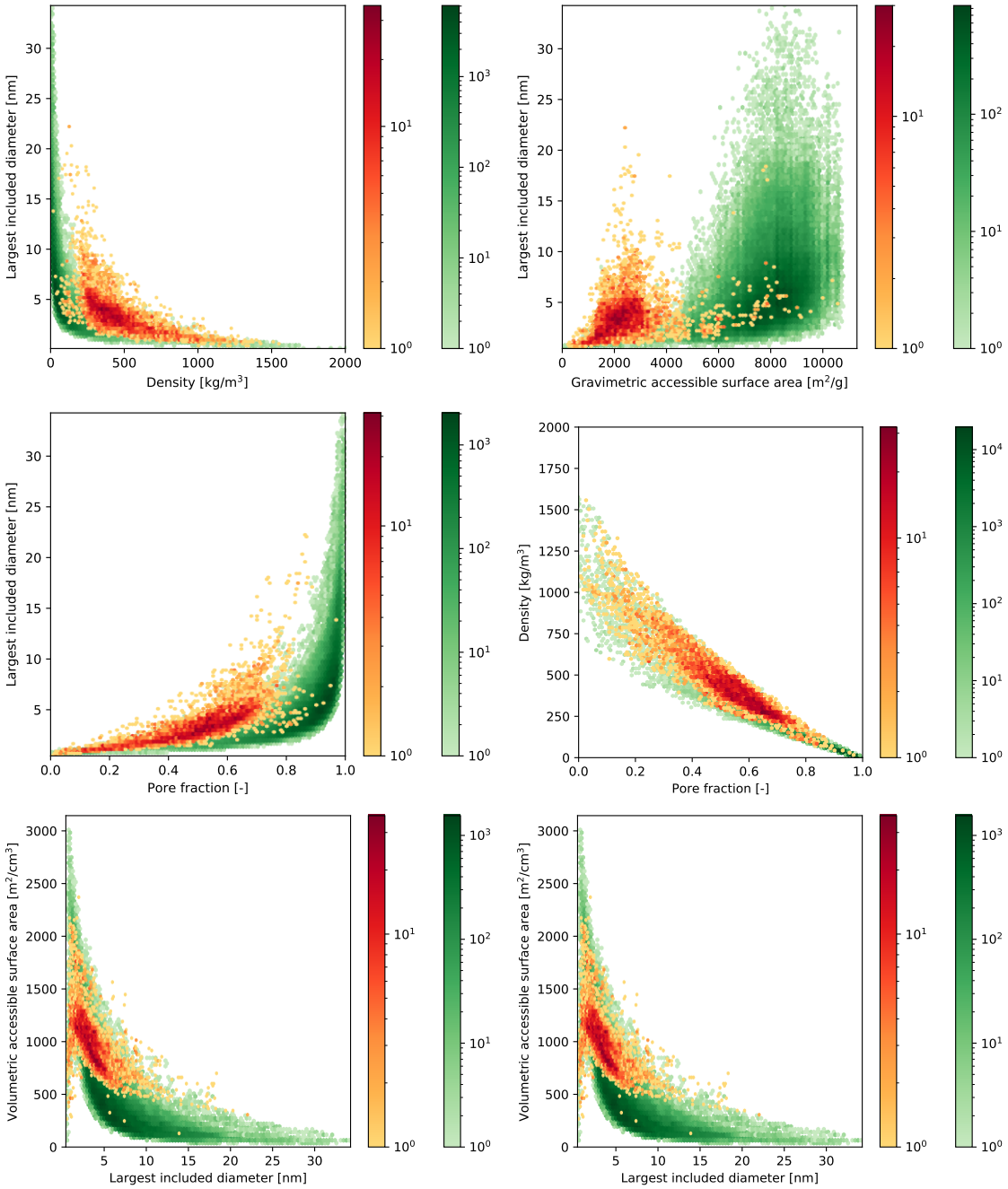
S4 Property-property relations

In Figs. 6-8 of the main text, property-property relations between textural and adsorption properties are visualized. Many more relations can be established. In this section, we have provided additional plots with property-property relations. In Section S4.1, relations solely between textural properties are discussed. As in the main text, we have included both relations solely between COFs in our ReDD-COFFEE database, and between nanoporous materials in different databases. The other material databases discussed in the manuscript are introduced here. In Section S4.2, additional relations between textural and adsorption properties are provided. Also the atomic structures of the two top-performing COFs in terms of the volumetric deliverable capacity, as discussed in the main text, are visualized.

S4.1 Textural properties

Besides the property-property relations visualized in the main text, many more relations can be explored. We have broadened the scope of reported property-property relations in Figs. S44 and S45. They support the claims that the 3D COFs in our database have a lower mass density as compared to the 2D COFs. These low density materials, therefore, possess large pores with large included diameters, high pore fractions, and high gravimetric accessible surface areas, whereas their volumetric accessible surface area decreases due to a rapidly increasing unit cell volume. 191 062 structures in the ReDD-COFFEE database (71.11%) have a pore fraction larger than 0.85 and a gravimetric accessible surface area above 7000 m^2/g, as indicated by the dashed lines in Fig. S45. High density materials only have small pore diameters and, therefore, do not accept guest molecules. Both their volumetric and gravimetric accessible surface area and volume become negligible.

In Fig. S46, two additional kernel densities are visualized, in which the structures are classified according to their linkage type. The histogram of the largest included diameter complements the one of the mass density in Fig. 6 in the main text. The largest linkage types, such as (acyl)hydrazone, (keto)enamine, and azine, lead to materials with a larger included diameter and therefore also a lower mass density as compared to smaller linkage types, such as triazine,


Largest included diameter [nm]
Density [kg/m^3]
Gravimetric accessible surface area [m^2/g]
Pore fraction [-]
Volumetric accessible surface area [m^2/cm^3]

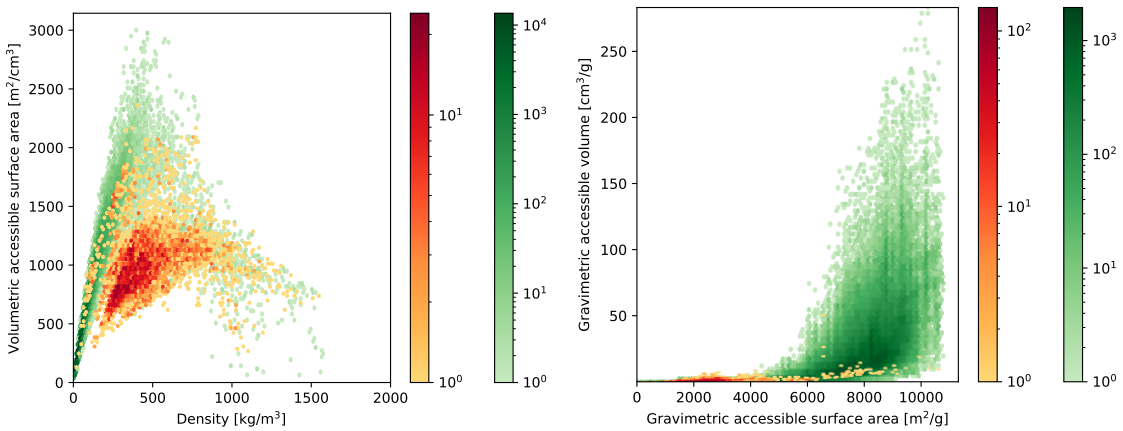


Figure S44: Additional property-property relations between textural properties for all structures in the ReDD-COFFEE database. 2D and 3D COFs are indicated with orange/red and green colorbars, similar to the figures in the main text. The colorbars indicate the number of structures in each histogram bin.

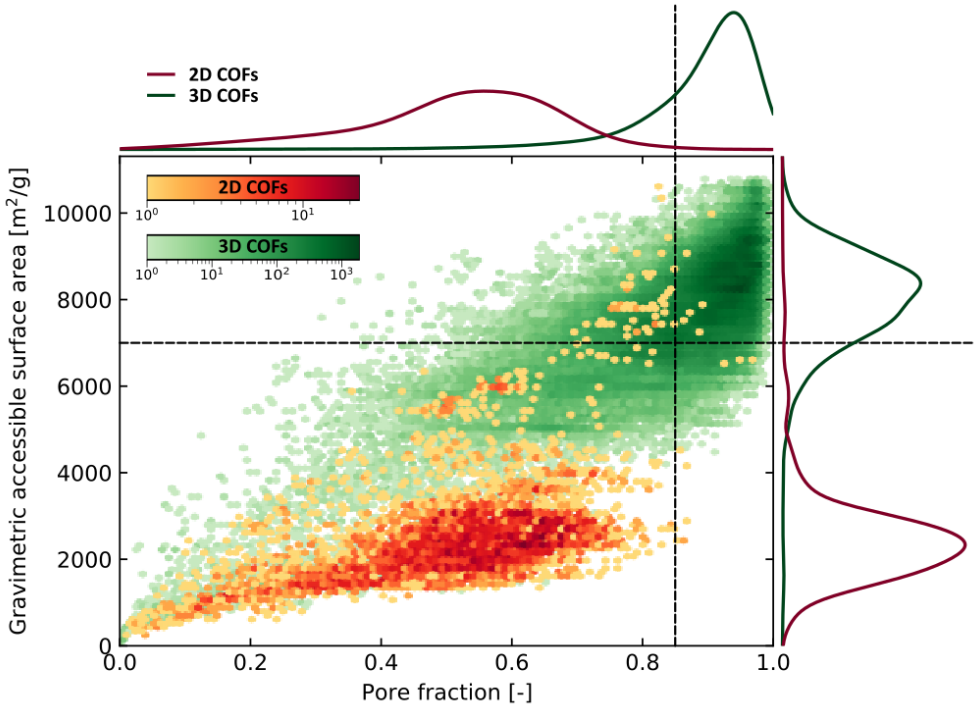


Figure S45: Property-property relation between the pore fraction and the gravimetric accessible surface area for the structures in the ReDD-COFFEE database. 1D histograms of both properties are provided at the axes. 2D and 3D COFs are indicated with orange/red and green colors, respectively. The horizontal and vertical dotted lines indicate a gravimetric accessible surface area of 7000 m^2/g and a pore fraction of 0.85.

boroxine, and borazine.

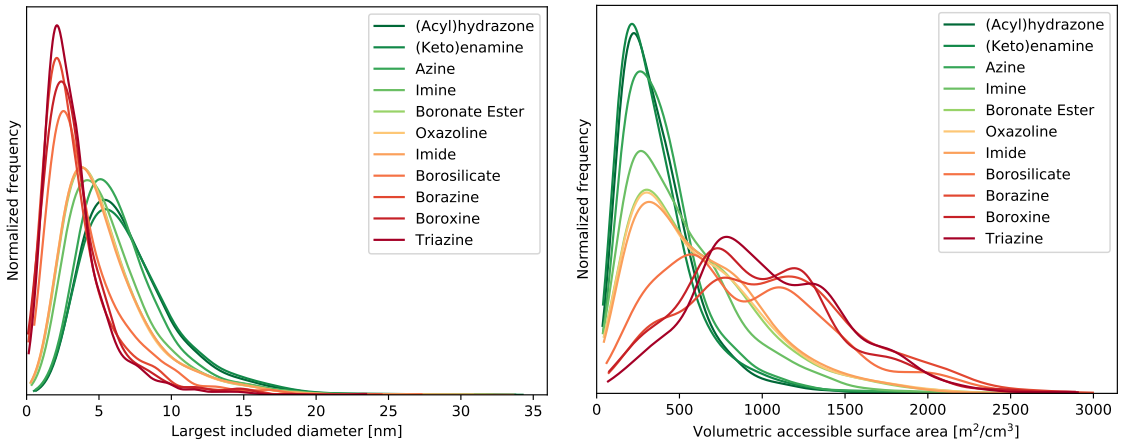


Figure S46: Distribution of the largest included diameter and the volumetric accessible surface area for each of the linkage types.

The COFs in the ReDD-COFFEE database are also compared with databases of other nanoporous materials. The following databases are included:

- **IZA**:[31] Database of 246 idealized zeolite framework types.

  `http://www.iza-structure.org/databases/`

- **QMOF**:[32] Collection of both the CSD MOF subset[33] and CoRE MOF database,[34,35] from which only the ones with a high fidelity (no overlapping atoms, missing hydrogens, ...) and a small number of atoms (<300 atoms) are retained. The reported structures are DFT-optimized using the PBE-D3(BJ) level of theory. In this work, v14 from December 2021 is adopted, which contains 20 375 MOFs.

  `https://figshare.com/articles/dataset/QMOF_Database/13147324`

- **hMOF**:[36] Bottom-up database built by snapping building blocks together based on geometric rules. From 102 building blocks, a total of 137 953 hypothetical MOFs were generated by Wilmer *et al.* in their landmark paper.

  `hmofs.northwestern.edu`

- **ToBaCCo**:[37] Top-down database of 13 512 MOFs from 41 topologies, 31 cluster building blocks, and 47 linker building blocks. In the link below, a total of 13 514 structures are provided, four of which (mof_5151.cif, mof_7874.cif, mof_8187.cif and mof_13269.cif) do

not contain an atomistic structure and were removed.

`https://github.com/tobacco-mofs/tobacco`

In these databases, many materials with no accessible surface area or volume are present. To avoid the distributions to be skewed towards these regions, they are omitted for the visualizations. Additional relations between the database are visualized in Fig. S47.

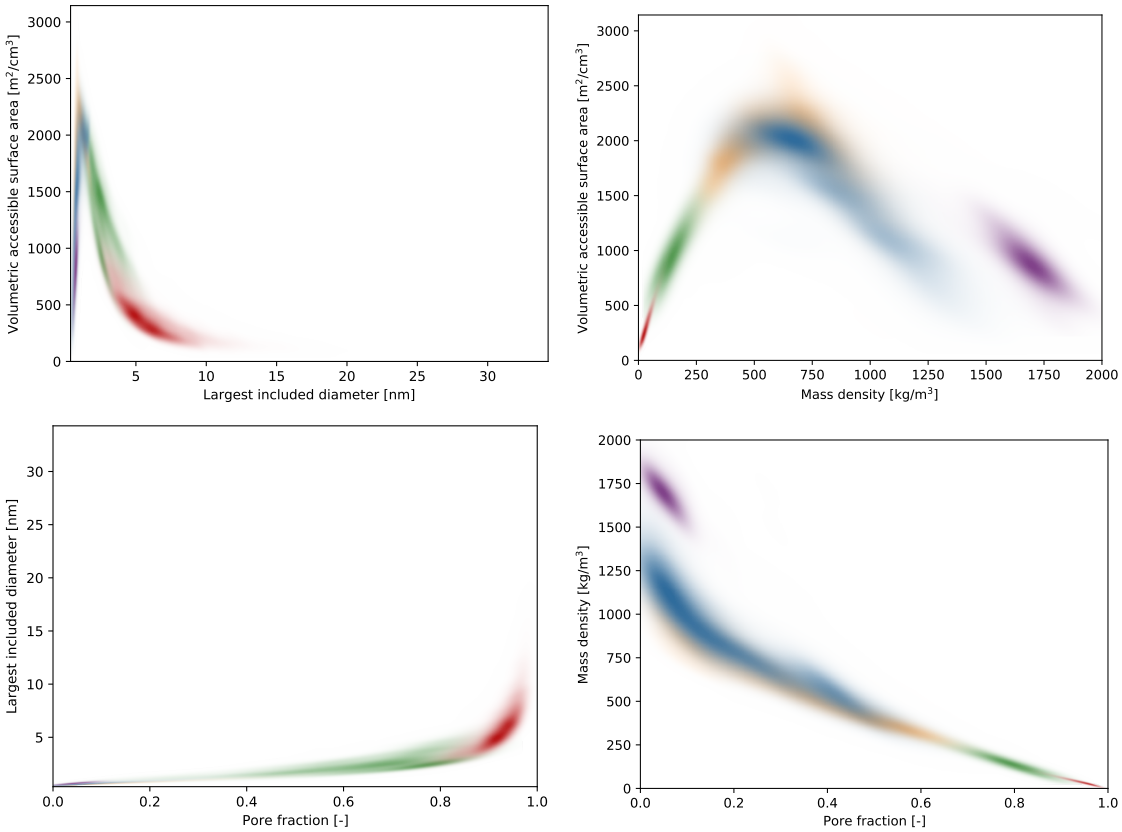


Figure S47: Additional property-property relations between textural properties between different databases of nanoporous materials. Color code is the same as in the main text: ReDD-COFFEE (red), IZA[31] (purple), QMOF[32] (blue), hMOF[36] (yellow), ToBaCCo[37] (green).

S4.2 Adsorption properties

As for the relations between textural properties, we have also included additional relations between textural and adsorption properties in this Section. In Fig. S48, the dependencies of the gravimetric and volumetric uptakes on the gravimetric accessible volume and the volumetric accessible surface area, respectively, are visualized. As can be observed, the gravimetric uptake increases linearly with increasing gravimetric accessible volume in both the low and high pres-

sure regimes. As the increase is stronger at 65 bar, the deliverable capacity also increases with increasing gravimetric accessible volume. A similar reasoning can be made for the volumetric uptake. Although the relation with the volumetric accessible surface area is less linear, the volumetric uptakes still increase with increasing volumetric accessible surface area. Again, the trend is more pronounced at high pressure and, therefore, the volumetric deliverable capacity maximizes for the largest volumetric accessible surface areas.

In Fig. S49, the dependency of the property-property relations of Fig. 8 of the main text on the dimensionality of the COF frameworks is visualized. 3D COFs have a lower mass density than 2D COFs, and, therefore, also their gravimetric accessible volume and volumetric accessible surface area are larger. This indicates that the amount of methane that can be stored in the pores of 3D COFs is larger than for 2D COFs, in terms of both the gravimetric as the volumetric deliverable capacity. However, among the structures with a large pore diameter, the 2D COFs outperform 3D COFs, since their pores stack on top of each other and form one dimensional channels that have a larger surface area than the pores of 3D COFs.

In Figs. S50 and S51, the relations between the volumetric and gravimetric deliverable capacity and some textural properties are plotted. As in the main text, histograms of the top 5% and worst 5% materials in terms of the respective deliverable capacity show the properties good or bad performing materials should possess. To achieve materials with a high volumetric deliverable capacity, a mass density between 200 and 500 kg/m^3 and a gravimetric accessible surface area between 3 000 and 7 500 m^2/g should be obtained. As discussed before, a high gravimetric accessible volume results in a large gravimetric deliverable capacity. This is accomplished by materials with a low mass density, large pore diameters, high gravimetric accessible surface areas, and large pore fractions.

Lastly, also the dependency of the heat of adsorption in the low and high pressure regimes on textural properties is investigated. These relations are depicted in Fig. S52. The methane molecules interact more strongly with the framework when small pores are available. When the pore diameter increases, the molecules are less strongly confined to the pore surface. Again, large pores result in large gravimetric accessible surface areas and high pore fractions.

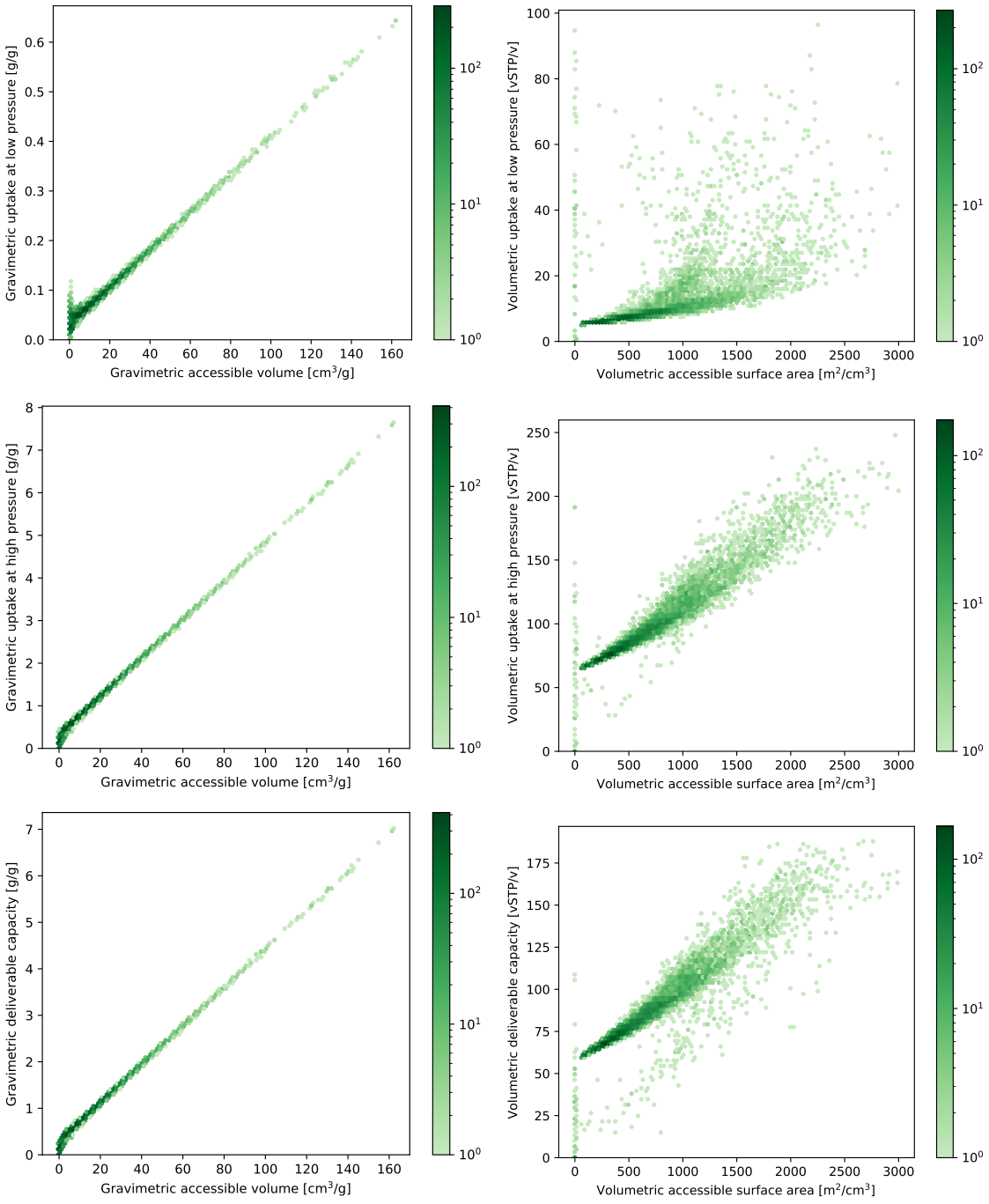


Figure S48: Additional property-property relations between textural and adsorption properties. The colorbar indicates the number of structures in each histogram bin.

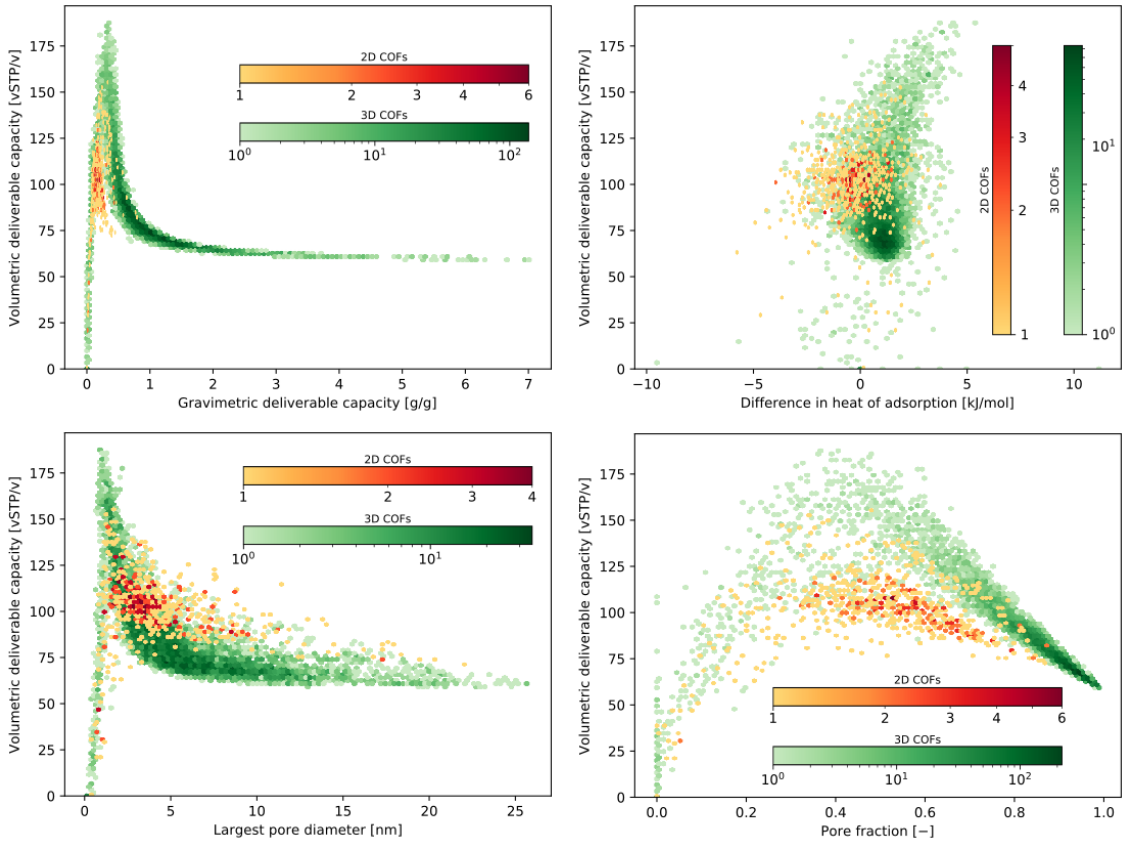


Figure S49: Breakdown of the subplots of Fig. 8 of the main text according to the dimensionality of the materials. 2D and 3D COFs are indicated with orange/red and green colorbars, respectively.

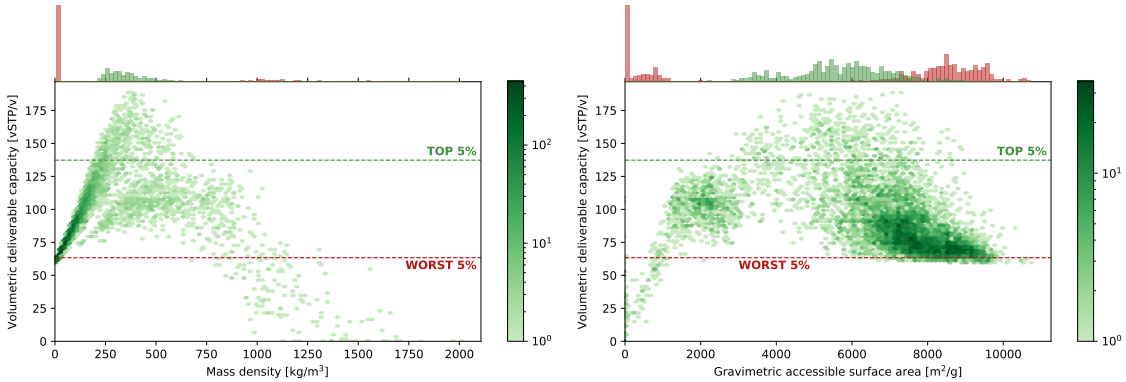


Figure S50: Dependency of the volumetric deliverable capacity on the mass density and gravimetric accessible volume. The colorbar indicates the number of structures in each histogram bin. A histogram of the top 5% and worst 5% performing structures in terms of the volumetric deliverable capacity is given on top of the plots.

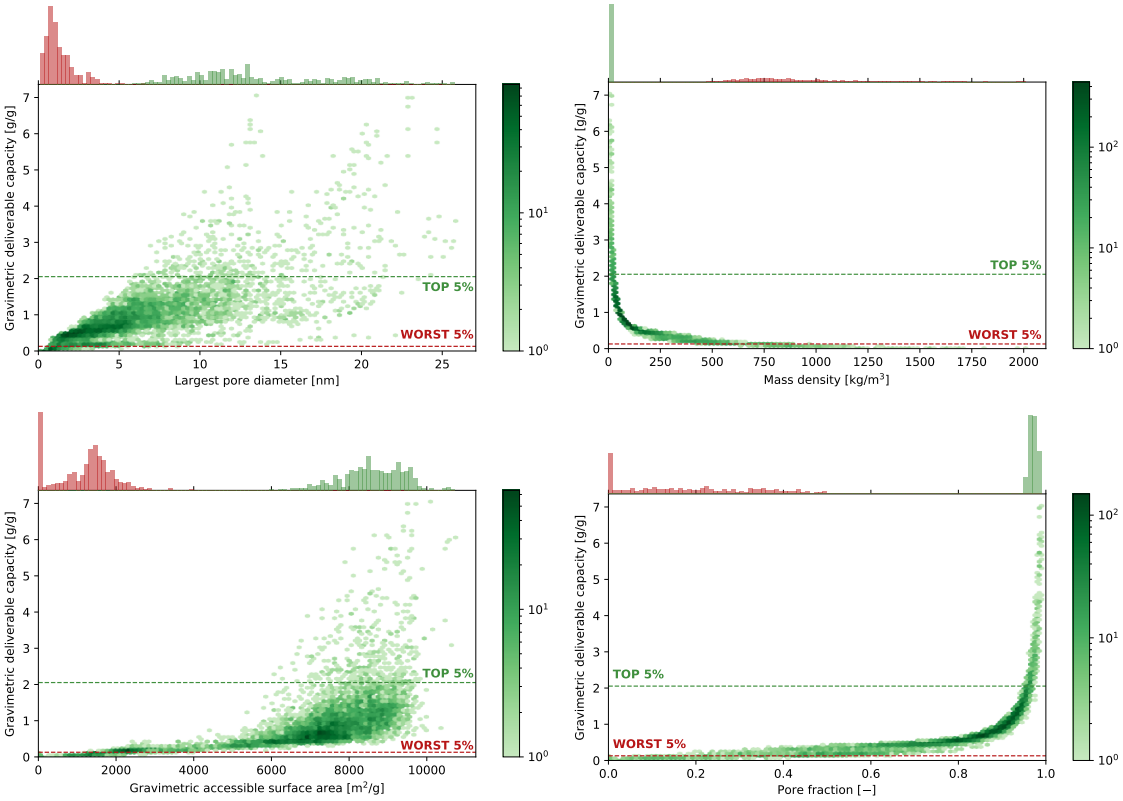


Figure S51: Dependency of the gravimetric deliverable capacity on selected textural properties. The colorbar indicates the number of structures in each histogram bin. A histogram of the top 5% and worst 5% performing structures in terms of the gravimetric deliverable capacity is given on top of the plots.

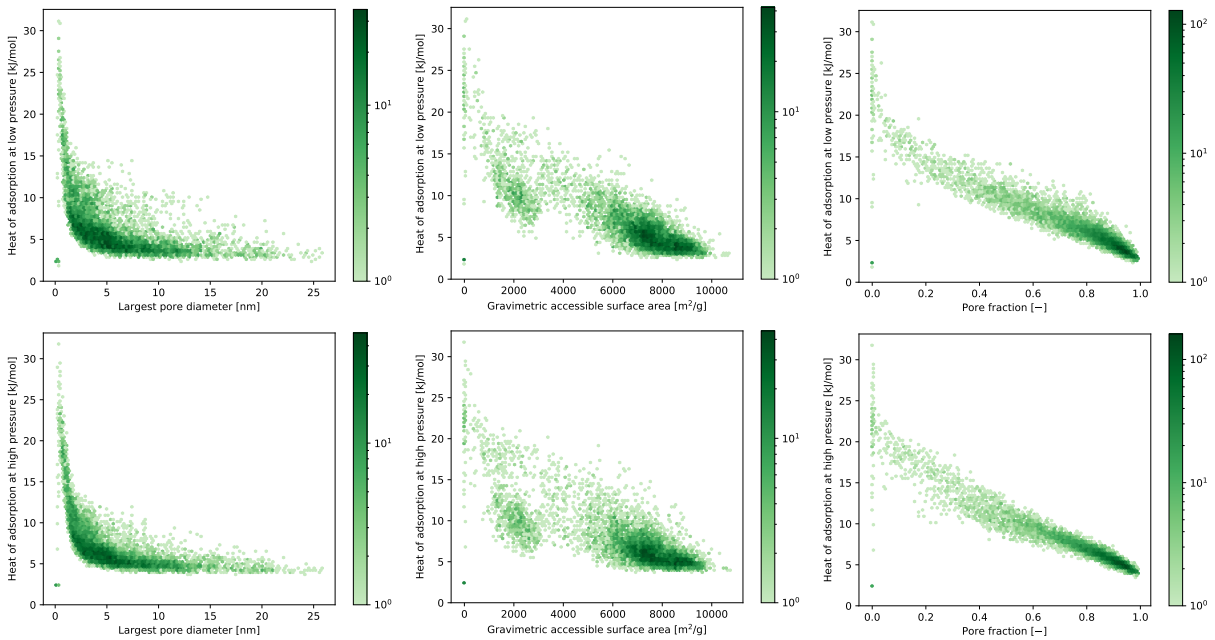


Figure S52: Dependency of the heat of adsorption at low and high pressure on selected textural properties. The colorbar indicates the number of structures in each histogram bin.

In the main text, two candidates with top-performing deliverable capacities are mentioned: ths-c3_11-01-01_06-08-01_06-08-01, with a volumetric deliverable capacity of 187.4 vSTP/v and a gravimetric deliverable capacity of 0.37 g/g, and ths-c3_11-02-04_04-03-04_04-03-04, which has a volumetric deliverable capacity of 141.1 vSTP/v and a gravimetric deliverable capacity of 0.50 g/g. These two materials are visualized in Fig. S53.

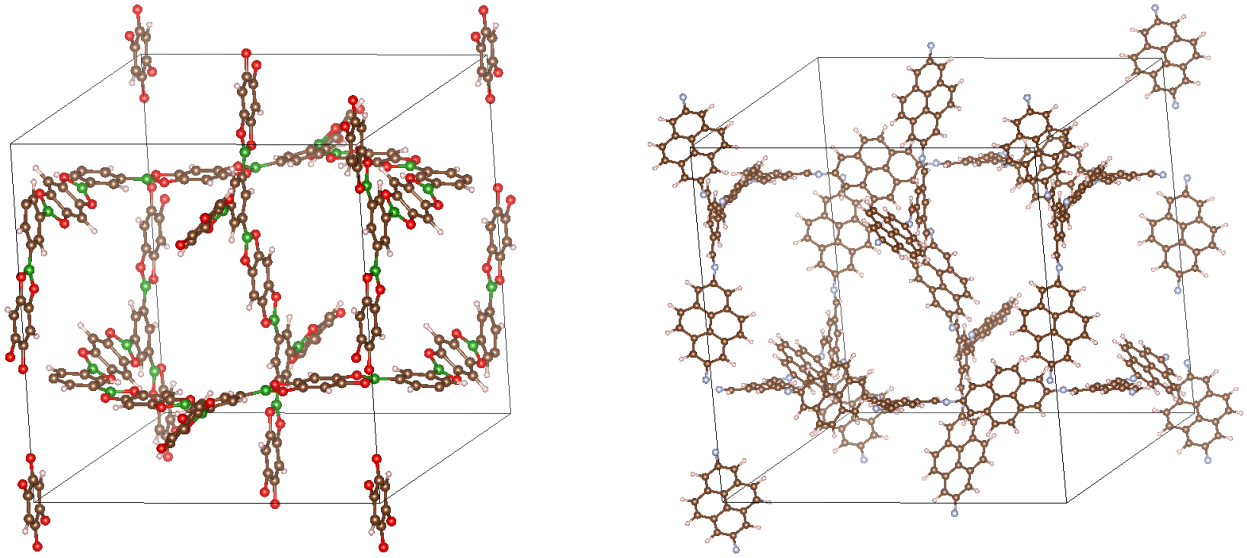


Figure S53: The two top-performing candidates for vehicular methane storage. Left: ths-c3_11-01-01_06-08-01_06-08-01. Right: ths-c3_11-02-04_04-03-04_04-03-04. Color code: hydrogen (white), boron (green), carbon (brown), nitrogen (blue), oxygen (red).

S5 Benchmark studies

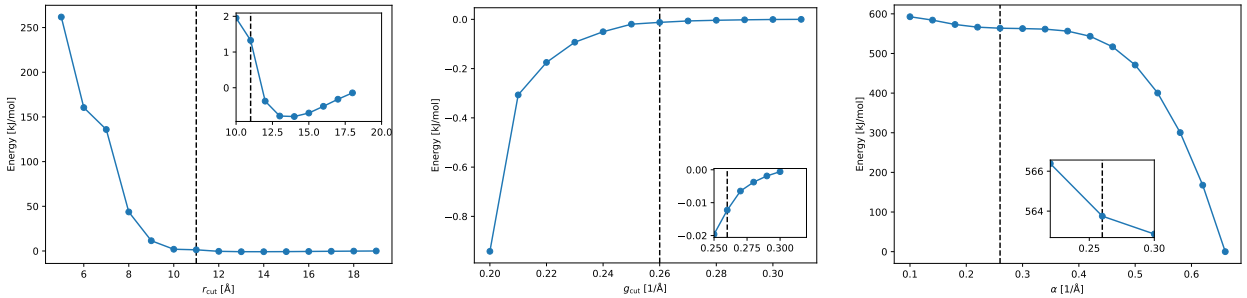
In the simulations performed in this study, many parameters and levels of theory are selected. Benchmark studies are performed to determine each of these values. In this section, the results of these studies are discussed. The force field arguments used in the structure optimizations are benchmarked in Section S5.1. For the calculation of the textural parameters with Zeo++,[38] the number of Monte Carlo samples has to be decided, which is done in Section S5.2. Finally, the level of theory used in the grand-canonical Monte Carlo (GCMC) calculations is selected in Section S5.3 based on an extensive benchmark study in which we try to reproduce the experimental isotherms of COF-1, COF-5, COF-102, and COF-103.

S5.1 Force field arguments

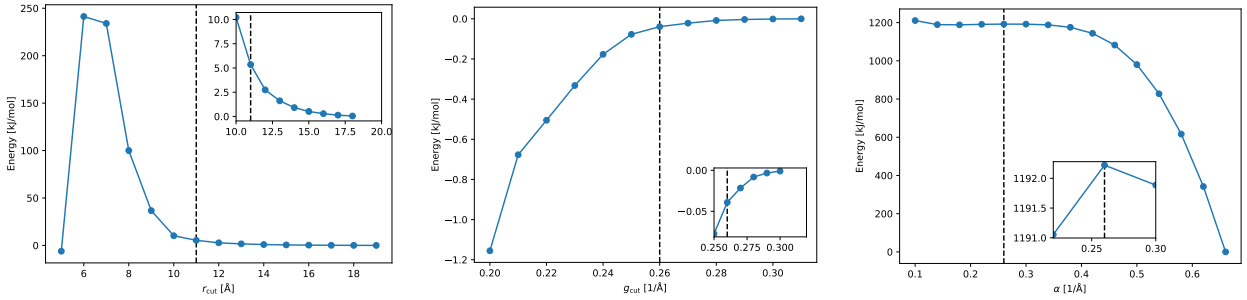
During the calculation of the force field energy in Yaff,[4] a balance is struck between computational efficiency and accuracy by choosing a finite real-space cutoff r_{cut}. However, the accuracy loss for lower values of r_{cut} can be partly compensated by introducing tail corrections, and are consequently included. Furthermore, electrostatic interactions are smoothed with a truncation model in order to avoid discontinuities in the potential energy surface. The three remaining arguments that influence the energy calculation are the real-space cutoff r_{cut}, the reciprocal space cutoff g_{cut} (or k_{max}), and the scaling factor α. These are given the following values: r_{cut} = 11 Å, g_{cut} = 0.26 Å^{-1}, α = 0.26 Å^{-1}. In Fig. S54, the energy is calculated for nine optimized structures while varying these parameters. For each plot, one parameter is changed, while the others are fixed on their default value. As can be observed, the energy does not change significantly upon increasing the values of r_{cut} and g_{cut}, whereas it does not change when slightly increasing or decreasing α. As both 2D and 3D COFs are included in the training set, this encourages us to adopt these force field arguments during our high-throughput screening.

S5.2 Zeo++

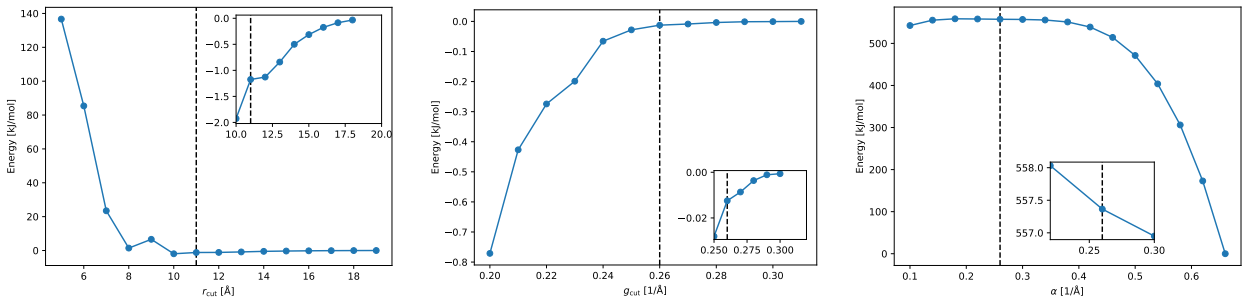
For the calculation of the accessible surface area and volume, Zeo++[38] only requires one argument: the number of Monte Carlo steps used to sample the space. In Fig. S55, the number of



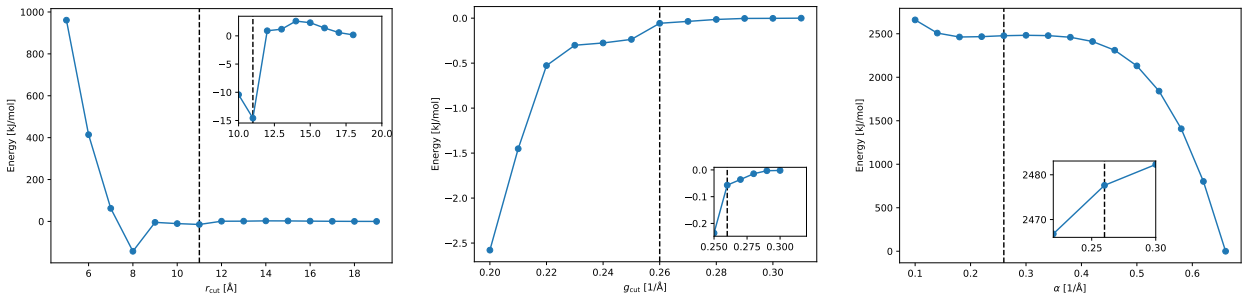
(a) dia_26-03-04_01-02-04



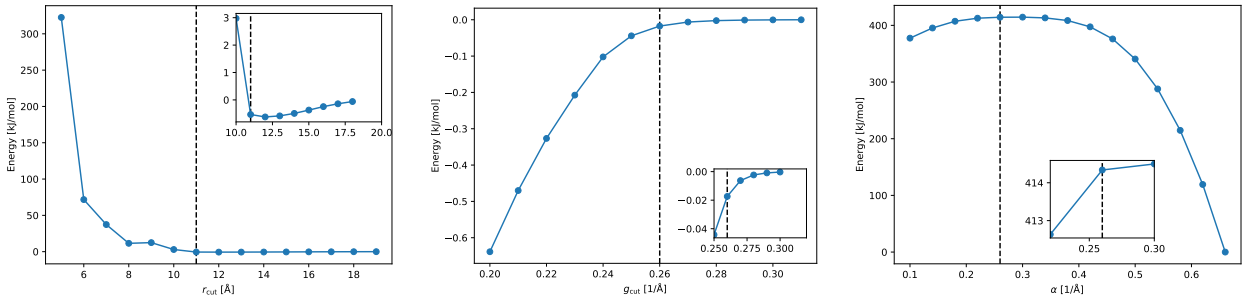
(b) dia_26-02-05_04-04-05



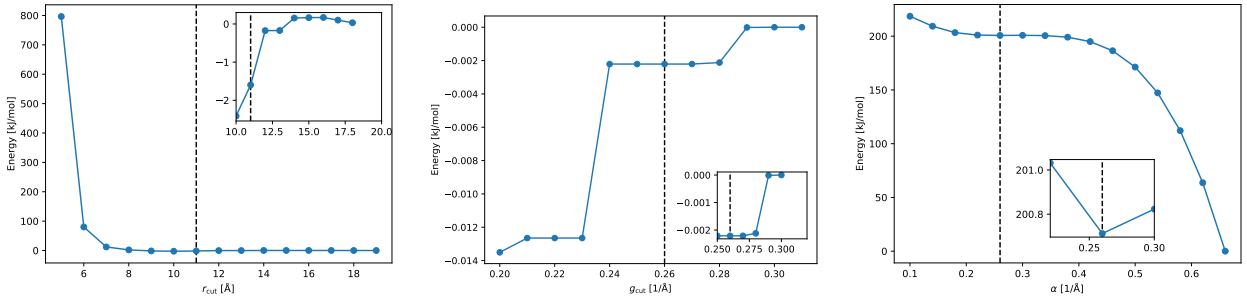
(c) bor_18-08-01_26-01-01



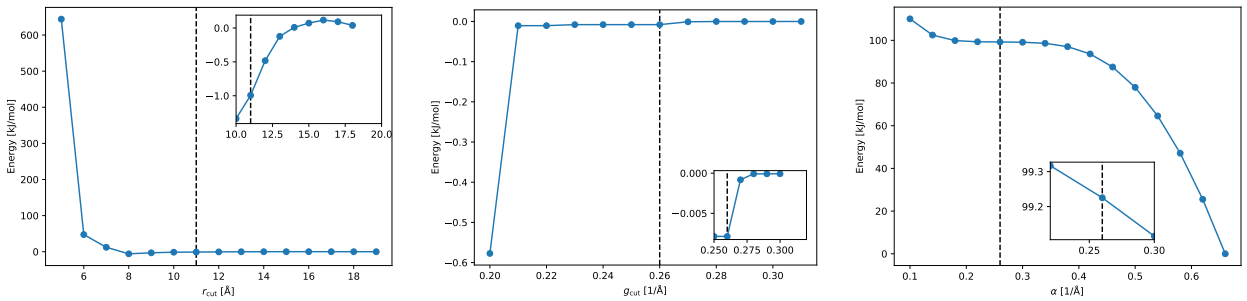
(d) ctn_31-11-02_27-01-02_None



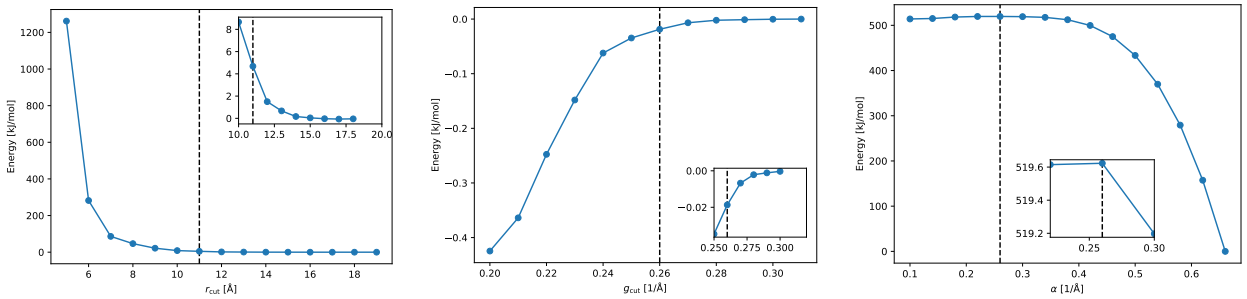
(e) pts_24-02-06_27-02-06_None



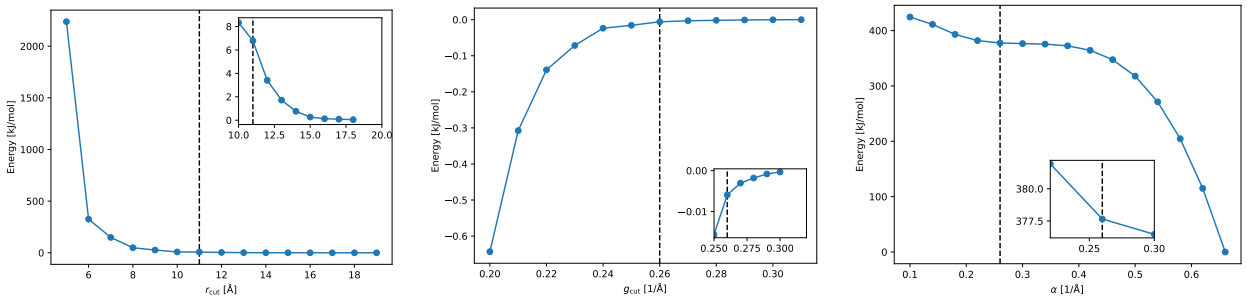
(f) hcb_32-11-10_01-06-10



(g) hcb_11-02-06_None



(h) sql_24-01-01_10-08-01



(i) kgm_29-03-04_01-02-04

Figure S54: Benchmark of the force field arguments r_{cut} (left), g_{cut} (middle) and α (right) for the QuickFF periodic force fields in nine COFs. Energy values of each plot are shifted to zero at the largest value for each parameter. The default parameters r_{cut} = 11 Å, g_{cut} = 0.26 Å^{-1}, and α = 0.26 Å^{-1} are selected, which are indicated with a dashed line. Detailed plots around the default values are provided as an inset.

Monte Carlo steps is varied for a set of 21 COFs. As can be observed in the left panels, large trends in the accessible surface area (asa) and volume (av) are immediately visible. The right plots show that the values obtained after 3 000 Monte Carlo steps do not change substantially anymore when further increasing the number of MC steps. Therefore, this value is adopted in the high-throughput calculations.

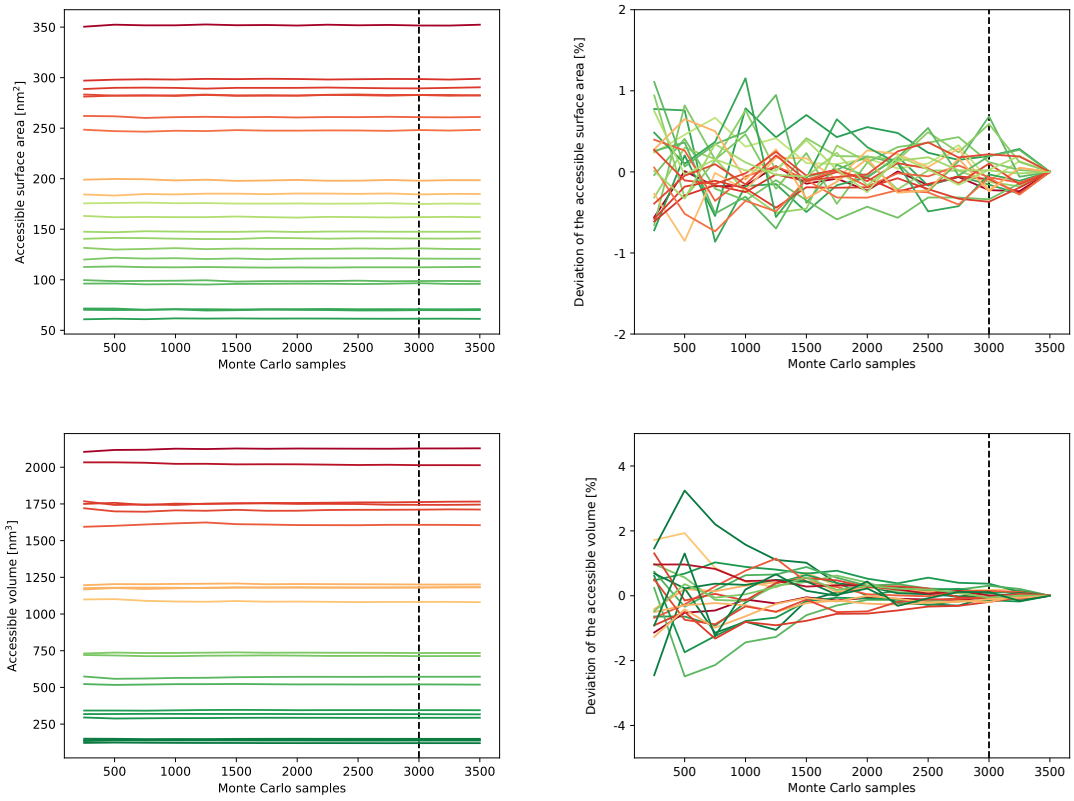


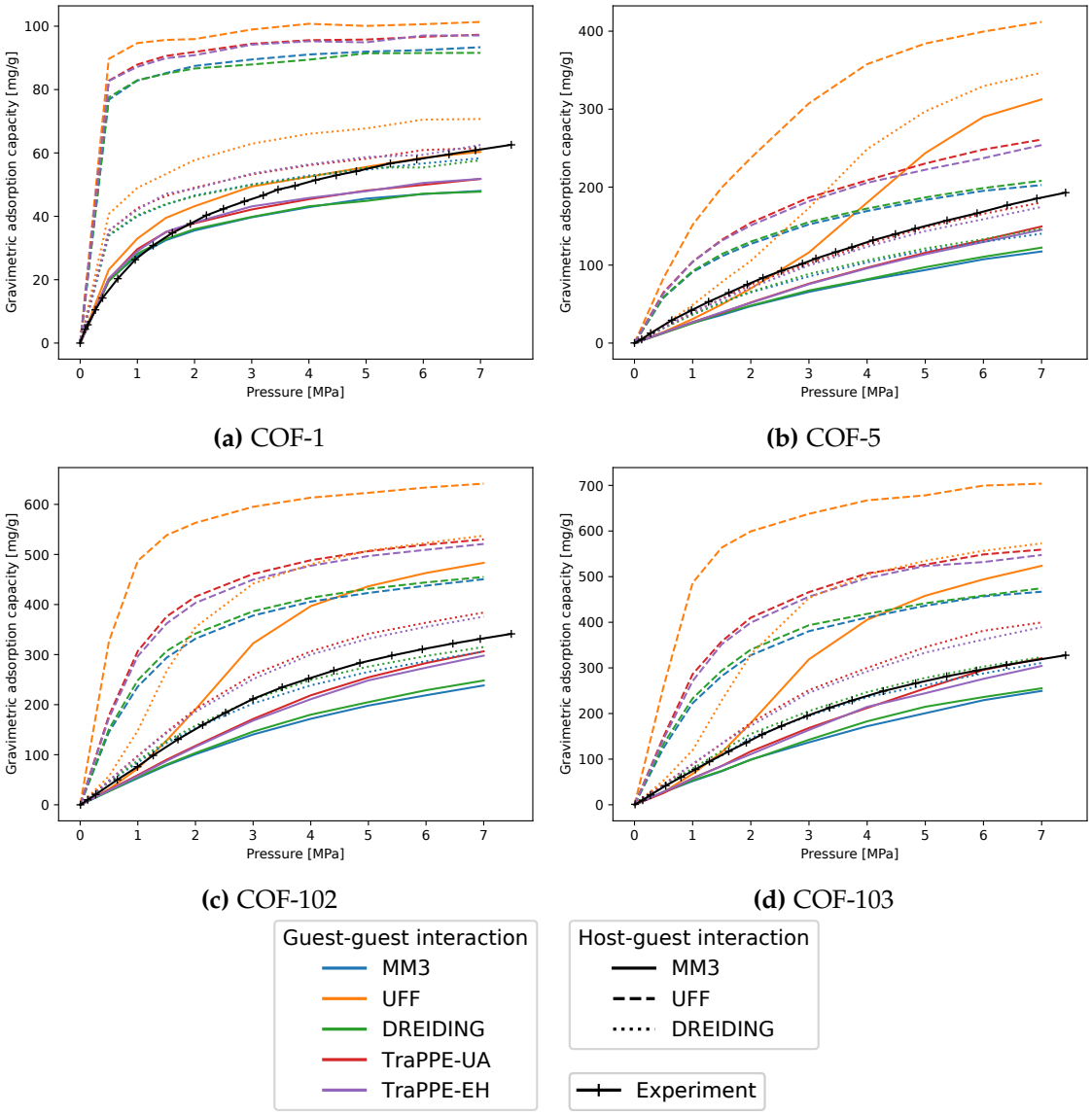
Figure S55: Benchmark of the number of Monte Carlo samples performed by Zeo++ when calculating the accessible surface area (asa, top) and volume (av, bottom) on 21 COFs. The selected number of samples is indicated with a dashed line. Left: absolute accessible surface area or volume. Right: normalized accessible area or volume.

S5.3 GCMC calculations

To benchmark the level of theory used in the grand-canonical Monte Carlo (GCMC) calculations performed with RASPA,[39] we calculated the methane isotherms of COF-1, COF-5, COF-102, and COF-103 at 298 K and compared them with the experimental isotherms.[40] A total of fifteen levels of theory is used. The guest-guest interactions can be described with either the MM3,[41]

UFF,[42] DREIDING,[43] TraPPE-UA,[44] or TraPPE-EH[45] models, whereas the MM3,[41] UFF,[42] or DREIDING[43] models can be used to describe the host-guest interactions. The MBIS partitioning scheme[46] is used to derive partial charges for the electrostatic interactions.

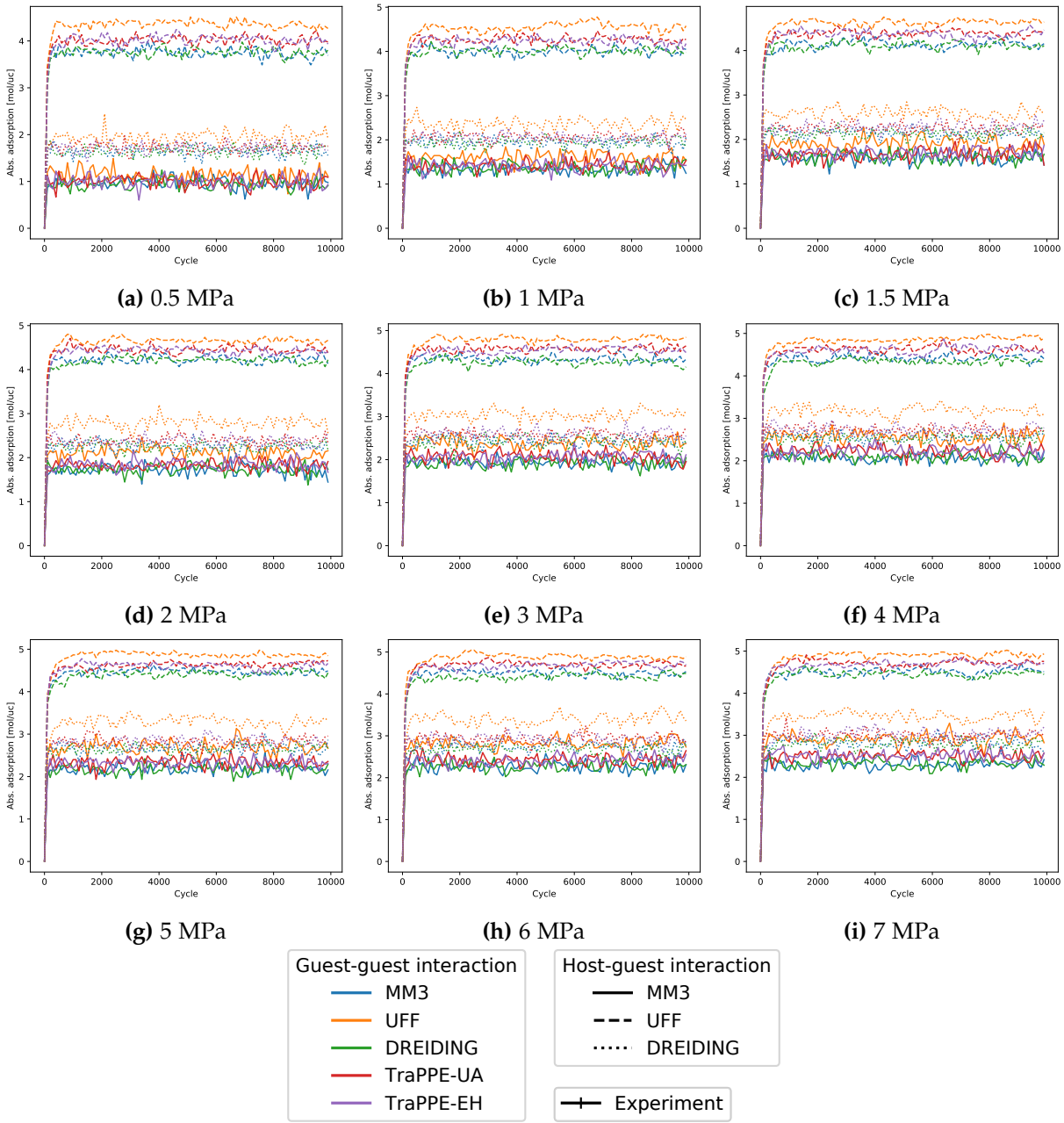
As can be observed from the isotherms in Fig. S56, the UFF host-guest model largely overestimates the methane uptake, while MM3 van der Waals host-guest interactions underestimate it. As a compromise between accuracy and time-efficiency, the TraPPE-UA model is used to describe the guest-guest interactions, whereas the DREIDING model defines the host-guest interactions. The uptakes for all runs of the four selected materials are visualized in Figs. S57-S60. It can be observed that equilibration is obtained after 5 000 cycles. Therefore, each calculation in the high-throughput screening starts with an equilibration run of 5 000 cycles, after which a production run of 10 000 cycles is performed.



(a) COF-1 **(b)** COF-5

(c) COF-102 **(d)** COF-103

Figure S56: Isotherm of four COFs calculated with different levels of theory at 298K. The experimental isotherm is indicated in black.

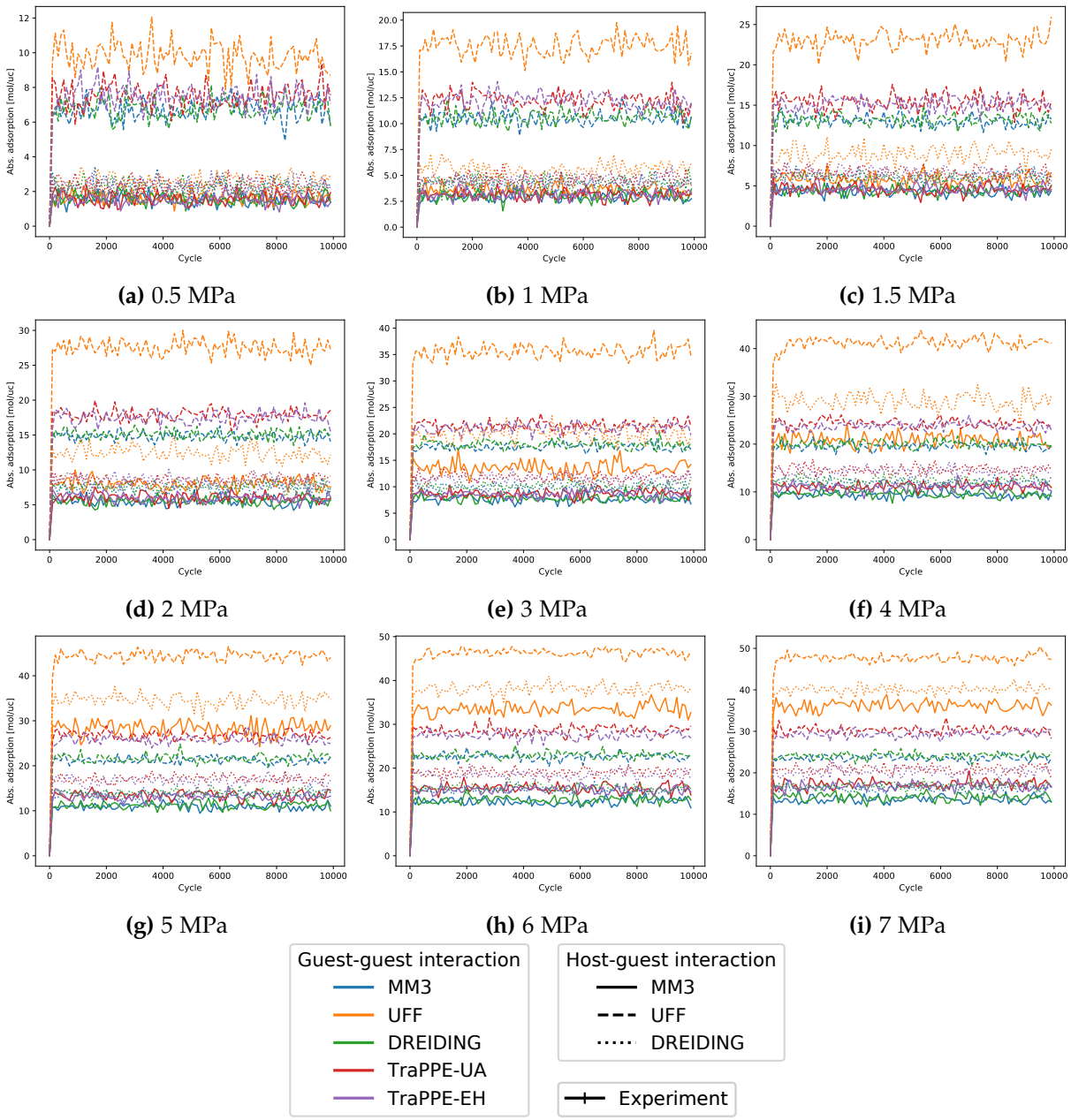


(a) 0.5 MPa **(b)** 1 MPa **(c)** 1.5 MPa

(d) 2 MPa **(e)** 3 MPa **(f)** 4 MPa

(g) 5 MPa **(h)** 6 MPa **(i)** 7 MPa

Figure S57: Equilibration of all GCMC calculations on COF-1. At every pressure, 15 GCMC calculations are performed with different levels of theory. The system is equilibrated after 5000 cycles.

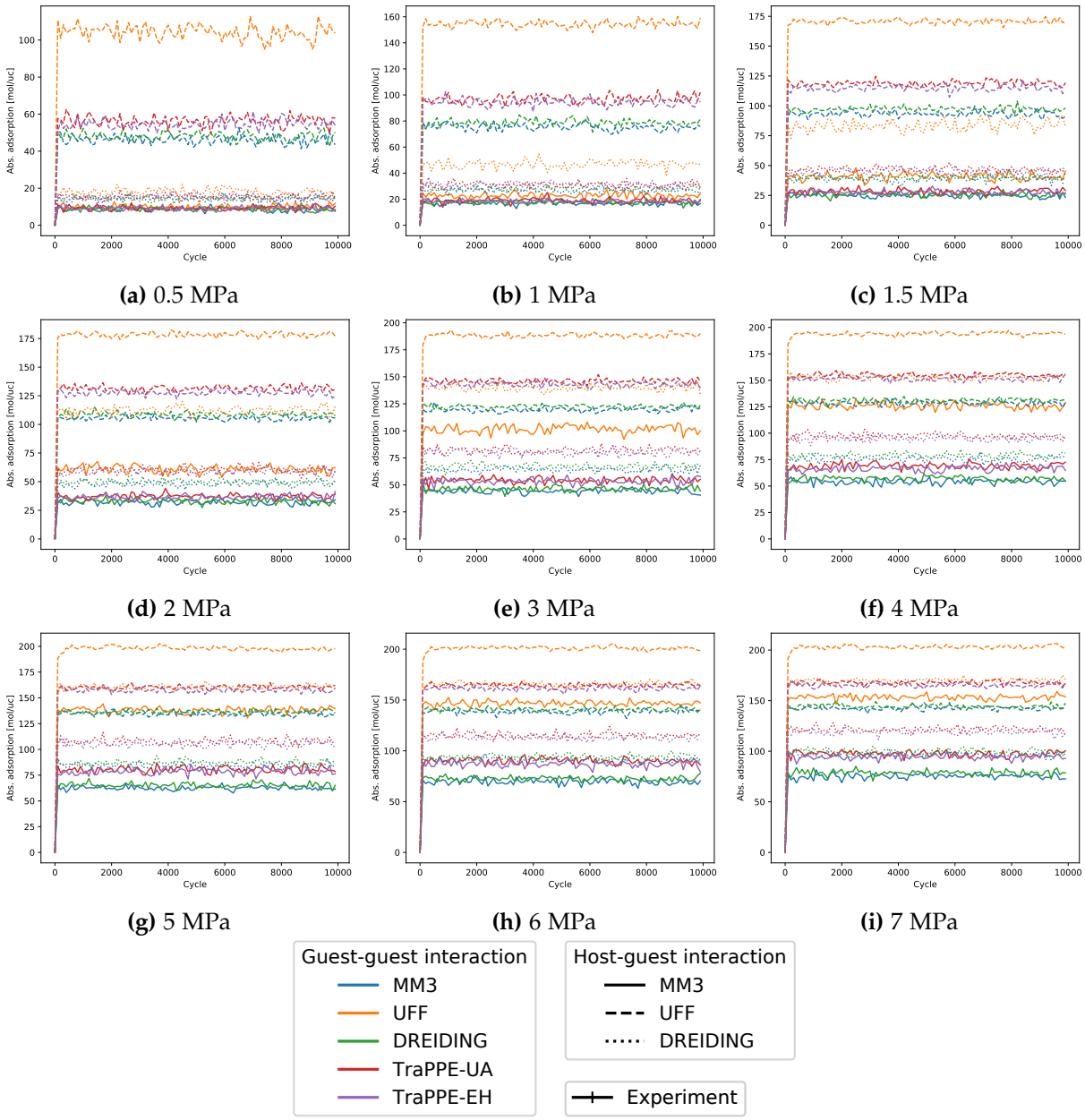


(a) 0.5 MPa **(b)** 1 MPa **(c)** 1.5 MPa

(d) 2 MPa **(e)** 3 MPa **(f)** 4 MPa

(g) 5 MPa **(h)** 6 MPa **(i)** 7 MPa

Figure S58: Equilibration of all GCMC calculations on COF-5. At every pressure, 15 GCMC calculations are performed with different levels of theory. The system is equilibrated after 5000 cycles.

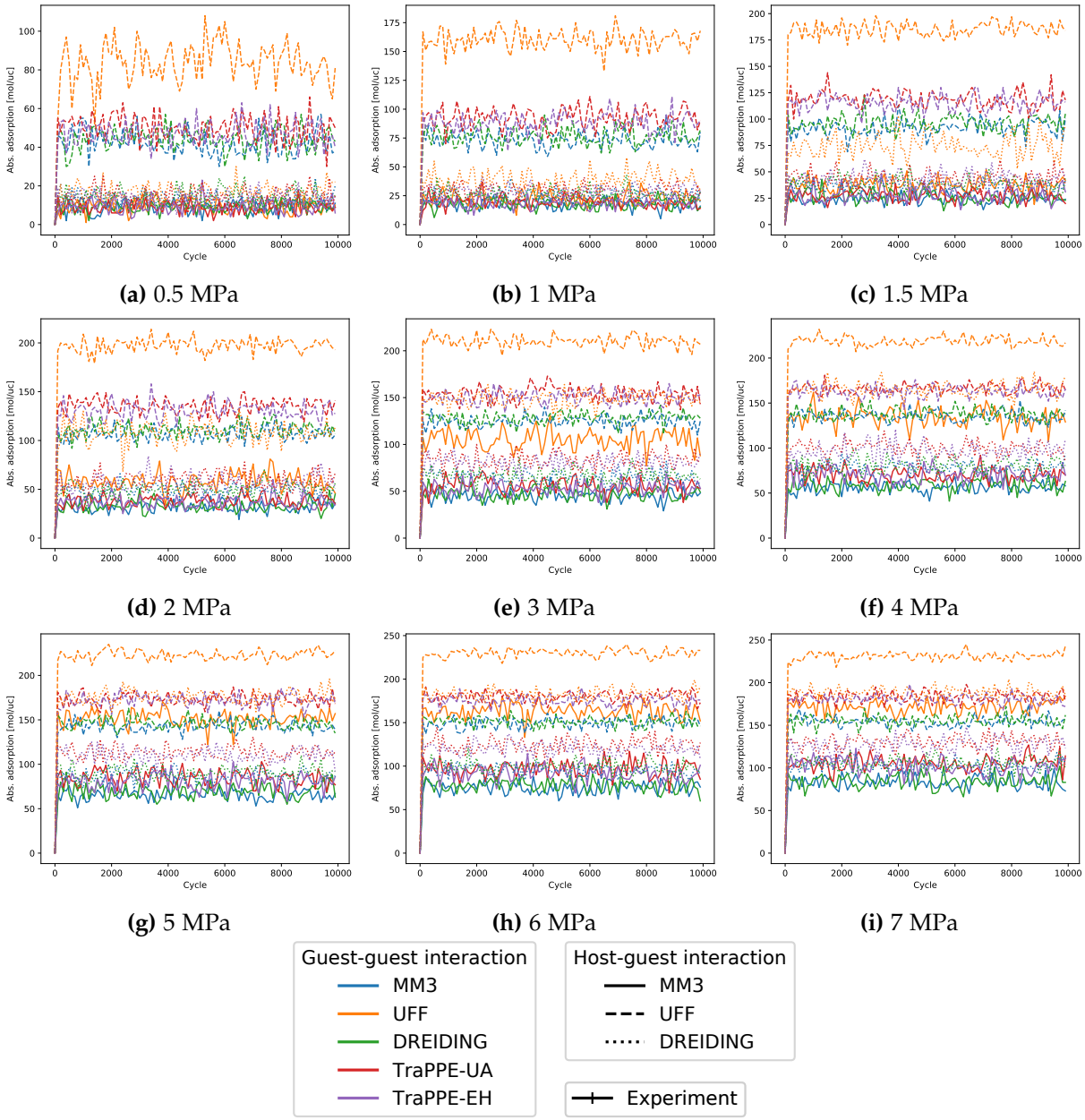


(a) 0.5 MPa **(b)** 1 MPa **(c)** 1.5 MPa

(d) 2 MPa **(e)** 3 MPa **(f)** 4 MPa

(g) 5 MPa **(h)** 6 MPa **(i)** 7 MPa

Figure S59: Equilibration of all GCMC calculations on COF-102. At every pressure, 15 GCMC calculations are performed with different levels of theory. The system is equilibrated after 5000 cycles.



(a) 0.5 MPa **(b)** 1 MPa **(c)** 1.5 MPa

(d) 2 MPa **(e)** 3 MPa **(f)** 4 MPa

(g) 5 MPa **(h)** 6 MPa **(i)** 7 MPa

Figure S60: Equilibration of all GCMC calculations on COF-103. At every pressure, 15 GCMC calculations are performed with different levels of theory. The system is equilibrated after 5000 cycles.

References

(1) Vanduyfhuys, L.; Vandenbrande, S.; Verstraelen, T.; Schmid, R.; Waroquier, M.; Van Speybroeck, V. QuickFF: a program for a quick and easy derivation of force fields for metal-organic frameworks from ab initio input. *J. Comput. Chem.* **2015**, *36*, 1015–1027.

(2) Vanduyfhuys, L.; Vandenbrande, S.; Wieme, J.; Waroquier, M.; Verstraelen, T.; Van Speybroeck, V. Extension of the QuickFF force field protocol for an improved accuracy of structural, vibrational, mechanical and thermal properties of metal-organic frameworks. *J. Comput. Chem.* **2018**, *39*, 999–1011.

(3) O'Keeffe, M.; Peskov, M. A.; Ramsden, S. J.; Yaghi, O. M. The reticular chemistry structure resource (RCSR) database of, and symbols for, crystal nets. *Acc. Chem. Res.* **2008**, *41*, 1782–1789, Accessed: 2020-02-06.

(4) Verstraelen, T.; Vanduyfhuys, L.; Vandenbrande, S.; Rogge, S. M. J. Yaff, yet another force field (v1.6.0). `http://molmod.ugent.be/software/`.

(5) Ongari, D.; Yakutovich, A. V.; Talirz, L.; Smit, B. Building a consistent and reproducible database for adsorption evaluation in covalent–organic frameworks. *ACS Cent. Sci.* **2019**, *5*, 1663–1675.

(6) El-Kaderi, H. M.; Hunt, J. R.; Mendoza-Cortés, J. L.; Côté, A. P.; Taylor, R. E.; O'Keeffe, M.; Yaghi, O. M. Designed synthesis of 3D covalent organic frameworks. *Science* **2007**, *316*, 268–272.

(7) Ding, S.-Y.; Gao, J.; Wang, Q.; Zhang, Y.; Song, W.-G.; Su, C.-Y.; Wang, W. Construction of covalent organic framework for catalysis: Pd/COF-LZU1 in Suzuki–Miyaura coupling reaction. *J. Am. Chem. Soc.* **2011**, *133*, 19816–19822.

(8) Borgmans, S.; Rogge, S. M. J.; De Vos, J. S.; Stevens, C. V.; Van Der Voort, P.; Van Speybroeck, V. Quantifying the likelihood of structural models through a dynamically enhanced powder X-ray diffraction protocol. *Angew. Chem. Int. Ed.* **2021**, *60*, 8913–8922.

(9) Fang, Q.; Wang, J.; Gu, S.; Kaspar, R. B.; Zhuang, Z.; Zheng, J.; Guo, H.; Qiu, S.; Yan, Y. 3D porous crystalline polyimide covalent organic frameworks for drug delivery. *J. Am. Chem. Soc.* **2015**, *137*, 8352–8355.

(10) Wan, S.; Gándara, F.; Asano, A.; Furukawa, H.; Saeki, A.; Dey, S. K.; Liao, L.; Ambrogio, M. W.; Botros, Y. Y.; Duan, X.; Seki, S.; Stoddart, J. F.; Yaghi, O. M. Covalent organic frameworks with high charge carrier mobility. *Chem. Mater.* **2011**, *23*, 4094–4097.

(11) Zhou, T.-Y.; Xu, S.-Q.; Wen, Q.; Pang, Z.-F.; Zhao, X. One-step construction of two different kinds of pores in a 2D covalent organic framework. *J. Am. Chem. Soc.* **2014**, *136*, 15885–15888.

(12) Yang, Z.; Chen, H.; Wang, S.; Guo, W.; Wang, T.; Suo, X.; Jiang, D.-e.; Zhu, X.; Popovs, I.; Dai, S. Transformation strategy for highly crystalline covalent triazine frameworks: from staggered AB to eclipsed AA stacking. *J. Am. Chem. Soc.* **2020**, *142*, 6856–6860.

(13) Li, Z.; Feng, X.; Zou, Y.; Zhang, Y.; Xia, H.; Liu, X.; Mu, Y. A 2D azine-linked covalent organic framework for gas storage applications. *Chem. Commun.* **2014**, *50*, 13825–13828.

(14) Ma, T.; Kapustin, E. A.; Yin, S. X.; Liang, L.; Zhou, Z.; Niu, J.; Li, L.-H.; Wang, Y.; Su, J.; Li, J.; Wang, X.; Wang, W. D.; Wang, W.; Sun, J.; Yaghi, O. M. Single-crystal x-ray diffraction structures of covalent organic frameworks. *Science* **2018**, *361*, 48–52.

(15) Zhang, Y.-B.; Su, J.; Furukawa, H.; Yun, Y.; Gándara, F.; Duong, A.; Zou, X.; Yaghi, O. M. Single-Crystal Structure of a Covalent Organic Framework. *J. Am. Chem. Soc.* **2013**, *135*, 16336–16339.

(16) Vanduyfhuys, L.; Verstraelen, T.; Vandichel, M.; Waroquier, M.; Van Speybroeck, V. Ab initio parametrized force field for the flexible metal-organic framework MIL-53(Al). *J. Chem. Theory Comput.* **2012**, *8*, 3217–3231.

(17) Bureekaew, S.; Amirjalayer, S.; Tafipolsky, M.; Spickermann, C.; Roy, T. K.; Schmid, R. MOF-FF - A flexible first-principles derived force field for metal-organic frameworks. *Phys. Status Solidi B* **2013**, *250*, 1128–1141.

(18) Tafipolsky, M.; Amirjalayer, S.; Schmid, R. Ab initio parametrized MM3 force field for the metal-organic framework MOF-5. *J. Comput. Chem.* **2007**, *28*, 1169–1176.

(19) Tafipolsky, M.; Schmid, R. Systematic first principles parameterization of force fields for metal-organic frameworks using a genetic algorithm approach. *J. Phys. Chem. B* **2009**, *113*, 1341–1352.

(20) Tafipolsky, M.; Amirjalayer, S.; Schmid, R. First-principles-derived force field for copper paddle-wheel-based metal-organic frameworks. *J. Phys. Chem. C* **2010**, *114*, 14402–14409.

(21) Wieme, J.; Vanduyfhuys, L.; Rogge, S. M. J.; Waroquier, M.; Van Speybroeck, V. Exploring the flexibility of MIL-47(V)-type materials using force field molecular dynamics simulations. *J. Phys. Chem. C* **2016**, *120*, 14934–14947.

(22) Rogge, S. M. J.; Bavykina, A.; Hajek, J.; Garcia, H.; Olivos-Suarez, A. I.; Sepúlveda-Escribano, A.; Vimont, A.; Clet, G.; Bazin, P.; Kapteijn, F.; Daturi, M.; Ramos-Fernandez, E. V.; Llabrés i Xamena, F. X.; Van Speybroeck, V.; Gascon, J. Metal-organic and covalent organic frameworks as single-site catalysts. *Chem. Soc. Rev.* **2017**, *46*, 3134–3184.

(23) Schmid, R.; Tafipolsky, M. An accurate force field model for the strain energy analysis of the covalent organic framework COF-102. *J. Am. Chem. Soc.* **2008**, *130*, 12600–12601.

(24) Amirjalayer, S.; Snurr, R. Q.; Schmid, R. Prediction of structure and properties of boron-based covalent organic frameworks by a first-principles derived force field. *J. Phys. Chem. C* **2012**, *116*, 4921–4929.

(25) Tong, M.; Lan, Y.; Yang, Q.; Zhong, C. Exploring the structure-property relationships of covalent organic frameworks for noble gas separations. *Chem. Eng. Sci.* **2017**, *168*, 456–464.

(26) Martin, R. L.; Simon, C. M.; Medasani, B.; Britt, D. K.; Smit, B.; Haranczyk, M. In silico design of three-dimensional porous covalent organic frameworks via known synthesis routes and commercially available species. *J. Phys. Chem. C* **2014**, *118*, 23790–23802.

(27) Martin, R. L.; Haranczyk, M. Construction and characterization of structure models of crystalline porous polymers. *Cryst. Growth Des.* **2014**, *14*, 2431–2440.

(28) Mercado, R.; Fu, R.-S.; Yakutovich, A. V.; Talirz, L.; Haranczyk, M.; Smit, B. In silico design of 2D and 3D covalent organic frameworks for methane storage applications. *Chem. Mater.* **2018**, *30*, 5069–5086.

(29) Verstraelen, T. MolMod Software Library. http://molmod.ugent.be/software.

(30) Skorjanc, T.; Shetty, D.; Gándara, F.; Ali, L.; Raya, J.; Das, G.; Olson, M. A.; Trabolsi, A. Remarkably efficient removal of toxic bromate from drinking water with a porphyrin–viologen covalent organic framework. *Chem. Sci.* **2020**, *11*, 845–850.

(31) Database of zeolite structures. http://www.iza-structure.org/databases/, Accessed: 2022-08-01.

(32) Rosen, A. S.; Iyer, S. M.; Ray, D.; Yao, Z.; Aspuru-Guzik, A.; Gagliardi, L.; Notestein, J. M.; Snurr, R. Q. Machine learning the quantum-chemical properties of metal-organic frameworks for accelerated materials discovery. *Matter* **2021**, *4*, 1578–1597.

(33) Moghadam, P. Z.; Li, A.; Wiggin, S. B.; Tao, A.; Maloney, A. G. P.; Wood, P. A.; Ward, S. C.; Fairen-Jimenez, D. Development of a Cambridge structural database subset: a collection of metal-organic frameworks for past, present, and future. *Chem. Mater.* **2017**, *29*, 2618–2625.

(34) 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.

(35) Chung, Y. G.; Haldoupis, E.; Bucior, B. J.; Haranczyk, M.; Lee, S.; Zhang, H.; Vogiatzis, K. D.; Milisavljevic, M.; Ling, S.; Camp, J. S.; Slater, B.; Siepmann, J. I.; Sholl, D. S.; Snurr, R. Q. Advances, updates, and analytics for the computation-ready, experimental metal-organic framework database: CoRE MOF 2019. *J. Chem. Eng. Data* **2019**, *64*, 5985–5998.

(36) Wilmer, C. E.; Leaf, M.; Lee, C. Y.; Farha, O. K.; Hauser, B. G.; Hupp, J. T.; Snurr, R. Q. Large-scale screening of hypothetical metal–organic frameworks. *Nat. Chem.* **2012**, *4*, 83–89.

(37) Colón, Y. J.; Gómez-Gualdrón, D. A.; Snurr, R. Q. Topologically guided, automated construction of metal-organic frameworks and their evaluation for energy-related applications. *Cryst. Growth Des.* **2017**, *17*, 5801–5810.

(38) 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.

(39) Dubbeldam, D.; Calero, S.; Ellis, D. E.; Snurr, R. Q. RASPA: molecular simulation software for adsorption and diffusion in flexible nanoporous materials. *Mol. Simul.* **2016**, *42*, 81–101.

(40) Furukawa, H.; Yaghi, O. M. Storage of hydrogen, methane, and carbon dioxide in highly porous covalent organic frameworks for clean energy applications. *J. Am. Chem. Soc.* **2009**, *131*, 8875–8883.

(41) Allinger, N. L.; Yuh, Y. H.; Lii, J. H. Molecular mechanics. The MM3 force field for hydrocarbons. 1. *J. Am. Chem. Soc.* **1989**, *111*, 8551–8566.

(42) Rappe, A. K.; Casewit, C. J.; Colwell, K. S.; Goddard, 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.

(43) Mayo, S. L.; Olafson, B. D.; Goddard, W. A. DREIDING: a generic force field for molecular simulations. *J. Phys. Chem.* **1990**, *94*, 8897–8909.

(44) 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.

(45) Chen, B.; Siepmann, J. I. Transferable potentials for phase equilibria. 3. Explicit-hydrogen description of normal alkanes. *J. Phys. Chem. B* **1999**, *103*, 5370–5379.

(46) Verstraelen, T.; Vandenbrande, S.; Heidar-Zadeh, F.; Vanduyfhuys, L.; Van Speybroeck, V.; Waroquier, M.; Ayers, P. W. Minimal basis iterative stockholder: atoms in molecules for force-field development. *J. Chem. Theory Comput.* **2016**, *12*, 3894–3912.