Journal of Materials Chemistry A

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Check for updates

Cite this: *J. Mater. Chem. A*, 2023, **11**, 7468

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, [a] Sander Borgmans, [a] Pascal Van Der Voort, [b] Sven M. J. Rogge *[a] and Veronique Van Speybroeck *[a]

Received 27th January 2023
Accepted 15th March 2023

DOI: 10.1039/d3ta00470h

rsc.li/materials-a

Covalent organic frameworks (COFs) are a versatile class of building block materials with outstanding properties thanks to their strong covalent bonds and low density. Given the sheer number of hypothetical COFs envisioned *via* reticular synthesis, only a fraction of all COFs have been synthesized so far. Computational high-throughput screenings offer a valuable alternative to speed-up such materials discovery. Yet, such screenings vitally depend on the availability of diverse databases and accurate interatomic potentials to efficiently predict each hypothetical COF's macroscopic behavior, which is currently lacking. Therefore, we herein present ReDD-COFFEE, the Ready-to-use and Diverse Database of Covalent Organic Frameworks with Force field based Energy Evaluation, containing 268 687 COFs and accompanying *ab initio* derived force fields that are shown to outperform generic ones. Our structure assembly approach results in a huge amount of computer-ready structures with a high diversity in terms of geometric properties, linker cores, and linkage types. Furthermore, the textural properties of the database are analyzed and the most promising COFs for vehicular methane storage are identified. By making the database freely accessible, we hope it may also inspire others to further explore the potential of these intriguing functional materials.

1 Introduction

Covalent organic frameworks (COFs)[1–6] form a class of nanoporous materials that show great potential for diverse applications, including the storage and separation of gases[7–12] or from solution,[13–15] catalysis,[16–19] energy storage,[20,21] optoelectronics,[22–24] sensing,[25] and drug delivery.[26] They are built from organic linkers, which form secondary building units (SBUs) that are held together by strong, covalently bound linkages. These covalent bonds typically result in COFs with a high mechanical, thermal, and chemical stability, and their building block structure makes them highly tuneable, similar to metal–organic frameworks (MOFs).[27,28] Substituting these building blocks, or the material's topology, can greatly influence the properties of the synthesized structure, such as the pore size,[29] adsorption capacity,[30] electronic properties,[31] or flexibility.[32] This fascinating feature makes materials for which the reticular principle holds the ideal engineering materials. Yet, the almost unlimited number of COFs that can be synthesized using reticular synthesis makes an experimental exploration of the whole COF material space for a given application unfeasible.[33,34] Therefore, computational high-throughput screenings offer an efficient and cheap alternative to accelerate materials discovery.[35,36] Vital for the predictive power of any high-throughput screening is the material database on which the computational screening is performed, which should be both diverse and uniformly distributed over the material space.[37] In addition, a cheap yet accurate description of the interatomic interactions is crucial to ensure viable property predictions. In this work, we therefore present the ReDD-COFFEE database containing 268 687 COF structures, showcasing a great diversity in terms of geometric properties, linker cores, and linkages, and

[a]*Center for Molecular Modeling (CMM), Ghent University, Technologiepark-Zwijnaarde 46, 9052 Zwijnaarde, Belgium. E-mail: Sven.Rogge@UGent.be; Veronique.VanSpeybroeck@UGent.be*

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

† Electronic supplementary information (ESI) available: Additional details about the database construction and force field generation and validation, together with extensive analysis and illustrative case studies. Further information and analysis of the diversity metrics and subset selection. Supplementary property–property relations for textural and adsorption properties. Benchmark of the adopted parameters in the force field optimizations, Zeo++ calculations, and GCMC calculations. All 268 687 COF structures and *ab initio* derived force fields of the ReDD-COFFEE database are publicly available at **https://doi.org/10.24435/materialscloud:nw-3j**. The relevant input files and computational data which generated the results of this work are available at **https://doi.org/10.5281/zenodo.7697262**. See DOI: **https://doi.org/10.1039/d3ta00470h**

accompany each material with an *ab initio* derived force field. Whereas such system-specific force fields have been developed earlier for a limited set of materials,[38–50] we have largely extended the scale of materials for which they are derived. Our ready-to-use database ensures molecular simulations can be directly started from the provided structures and force fields. This is explicitly demonstrated by performing a high-throughput screening on our database with the goal to unveil property–property relationships for textural and adsorption properties and discover attractive vehicular methane storage materials.

Computational high-throughput screenings offer a valuable alternative to accelerate materials discovery.[35,36,51,52] Such screening studies select a database of material geometries and perform several calculations on each material to predict their macroscopic behavior. Especially for MOFs, high-throughput screenings are abundant, especially in the fields of gas adsorption[53–60] and separation[61–66] processes. Recently, also the discovery of MOFs with targeted electronic and catalytic properties has gained attention.[67–70] A limited number of screening studies has been performed to shed light on their mechanical stability.[40,56] All these high-throughput screening studies start from one of the many constructed MOF databases. These databases can be divided into two major categories, depending on the origin of the structures they contain. On the one hand, experimental databases are built from already synthesized materials, containing either the synthesized structure or the structure as obtained after computational structure optimization and/or guest removal. The CoRE MOF database,[54] for instance, contains a subset of 5109 MOF structures identified from the Cambridge Structural Database (CSD)[71] and was later expanded to include 14 142 structures.[72] Moghadam *et al.* implemented an automated screening algorithm in the CSD Python API to instantly identify a MOF when it is added to the CSD.[73] At the time of its first publication, this subset contained 69 666 MOF materials.[73] Recently, the QMOF database was established, containing a subset of 15 713 materials from the previously mentioned databases for which DFT calculations can be carried out efficiently.[68] On the other hand, hypothetical databases contain *in silico* generated structures. They broaden the material space and provide a large and diverse set of structures. Although energy minimization approaches exist to generate plausible geometries, such as the Automated Assembly of Secondary Building Units (AASBU) method,[74] hypothetical databases generally rely on geometric procedures that connect SBUs with one another to form a periodic material without any optimization. These geometric procedures can be divided into two classes: bottom-up and top-down methods, depending on how the SBUs are assembled. In the bottom-up approach, SBUs are naturally grown until a periodic framework is formed. Wilmer *et al.* applied this method to a set of 102 SBUs to generate a database of 137 953 MOFs,[53] whereas a database of 324 500 MOFs was generated from 66 SBUs and 19 functional groups by Fernandez *et al.*[75] However, later research showed that the topological diversity of the structures constructed using this bottom-up approach is limited.[76] Top-down approaches typically result in a larger variety of topologies. These methods initially define the topological net, after which the SBUs are deliberately placed on the net's nodes. Multiple software packages, such as Zeo++,[77] AuToGraFS,[78] Weaver,[79] TOBASCCO,[80] and ToBaCCo,[55,56] use such top-down approach to generate hypothetical frameworks. The latter has been used to construct a database of 13 512 MOFs starting from 78 SBUs and 41 topologies.[55]

Also for COFs, several high-throughput screenings have emerged, although they mainly focus on gas adsorption and separation processes, such as methane storage,[81–83] hydrogen storage,[84,85] and carbon capture,[86–88] among others.[89–91] These high-throughput COF screenings all rely on one of the four large COF databases constructed to date. The two experimental databases, the CoRE[89] and CURATED[86] databases, focus on the boronate ester and imine COFs, which are abundantly present in literature. They naturally underrepresent other linkage types that are less frequently observed experimentally, although these linkage types may result in materials with unique features. Therefore, Martin *et al.*[82] and Mercado *et al.*[83] created two databases of hypothetical COFs to explore different regions of material space. However, because they focused on a large diversity of linker cores, they included only a limited number of linkage types, some of which are rarely realized experimentally. Moreover, when considering those linkage types that are experimentally relevant, these hypothetical databases are significantly lacking. As each linkage type provides unique properties, a suboptimal representation of these subclasses in a database will result in a largely untapped potential for COF materials. For example, boronate ester COFs have outstanding crystallinity,[6] and, whereas imine linked COFs already possess improved stability,[6] enamine COFs have an even higher stability.[92] Furthermore, triazine and hydrazone COFs provide coordination centers for transition metals that can be adopted in catalysis.[93,94] Yet, both enamine and hydrazone COFs are absent in the current hypothetical databases and are underrepresented with respect to imine COFs in the experimental ones.

To tap into the potential of these and other COF linkage types, we present in this work a hypothetical database describing a diverse set of linkage types, including both frequently observed linkages and linkages that are not often synthesized experimentally. Furthermore, an unprecedented feature of our database is that a system-specific force field is generated for each material, starting from the cluster force fields of the underlying building blocks. These are derived using our in-house developed QuickFF routine.[38,39] As such, each structure of the database can directly be used in high-throughput studies.

Besides the material's geometry and partial atomic charges derived for the COFs in the aforementioned experimental databases, the four databases lack the necessary information to model the interatomic interactions. To model these interactions, one can resort to computationally expensive *ab initio* methods, such as density functional theory (DFT), which largely limits the length and time scales that are feasible to simulate,[67,68,70] or a less expensive generic force field with reduced accuracy.[56,95] To obtain a description of the interatomic

interactions with a higher accuracy than generic force fields but a lower computational cost than *ab initio* techniques, some studies derived system-specific force fields for a smaller set of materials.[40] Such force fields are attractive to perform high-throughput screenings and to remove structures with a low synthetic likelihood from the database. For instance, many structures in the hypothetical databases of Martin *et al.*[82] and Mercado *et al.*[83] contain largely deformed SBUs and are likely unphysical. To identify the structures that are experimentally the most feasible, the synthetic likelihood of a large database of structures has been predicted with DFT energies[96] and force field free energy calculations.[95] It has been shown that while free energy calculations are necessary to predict the most favorable configuration out of a set of polymorphs, energy metrics are sufficient to predict the synthetic likelihood of hypothetical materials.[95] However, this requires one to augment the versatile COF database with a relatively inexpensive yet sufficiently accurate method to describe the interatomic interactions.

Therefore, we present in this work the ReDD-COFFEE database: a ready-to-use database of 268 687 COFs. ReDD-COFFEE is an acronym for Ready-to-use and Diverse Database of Covalent Organic Frameworks with Force field based Energy Evaluation. In our database, an optimized atomic structure and an *ab initio* derived force field is provided for each material. Essential to generate system-specific force fields for the huge amount of structures within this database, is the building block approach. In this approach, cluster force fields are fitted to quantum mechanical reference data for a limited number of smaller building blocks and then assembled to derive force fields suitable for the periodic structures. This procedure was introduced earlier through our QuickFF procedure,[38,39] but is now adopted for a much larger number of materials. By deriving cluster force fields for the underlying building blocks, the number of required *ab initio* calculations is greatly reduced, while a high accuracy of the interatomic potential is maintained.[40–50] All COFs are assembled starting from a limited set of 279 SBUs, which are selected to result in a diverse database, as elaborated in Section 4.2. In contrast to other hypothetical databases, we have explicitly included a large number of linkage types in our database. The SBUs are combined taking physical constraints into account and the structures that have the lowest synthetic likelihood are removed with a deformation energy filter. We firstly demonstrate that our system-specific force fields achieve a higher accuracy than generic force fields. Next, we show that our database has a great diversity in terms of geometric properties, linker cores, and linkages compared with the already existing COF databases. Furthermore, we establish property–property relations in between textural properties and compare them with other nanoporous materials, *i.e.*, MOFs[53,55,68] and zeolites.[97] Finally, the applicability of the database to identify COFs for vehicular methane storage is demonstrated by identifying promising candidates and determining property–property relations for adsorption properties on a diverse subset of the database containing 10 000 structures. All 268 687 optimized structures and force fields of the ReDD-COFFEE database are publicly available at **https://doi.org/10.24435/materialscloud:nw-3j** and are ready-to-use for further high-throughput screenings.

2 Methodology to construct database

2.1 Terminology to describe a COF's structure

The nomenclature of the COF building blocks used throughout this paper mimics the experimental synthesis of the material. Experimentally, a COF is synthesized by mixing several precursors, which react to form a covalently bound linkage that holds the structure together. As illustrated in Fig. 1, we herein distinguish two subregions for each precursor: a linker core, which remains unaltered during linkage formation, and reactive groups, which react with the reactive groups of other

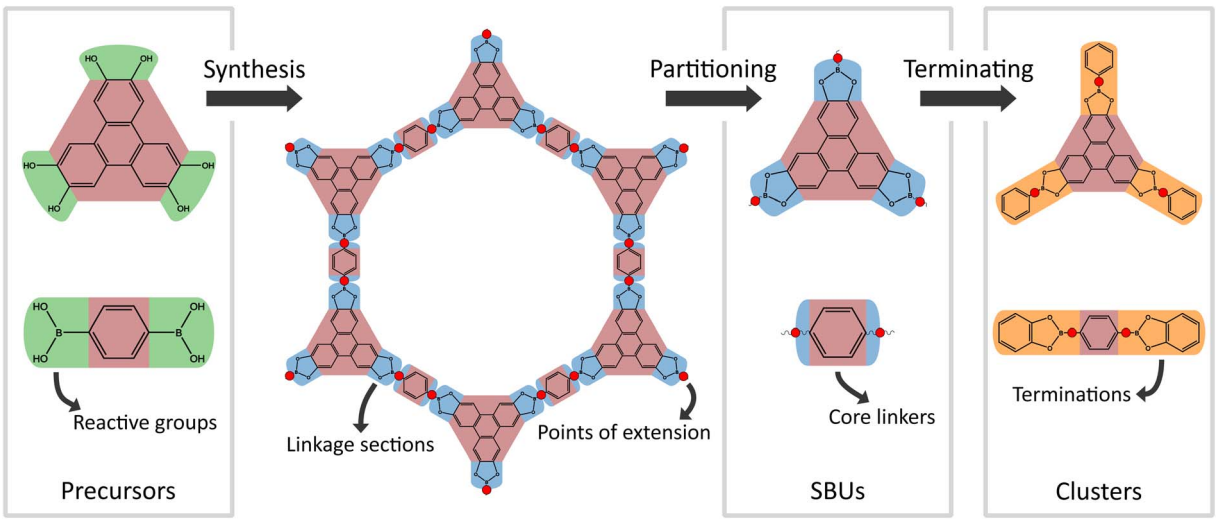

Fig. 1 Illustration of the terminology used in this paper for COF-5. Reactive groups (green) that are attached to linker cores (red) react to form linkages, consisting of linkage sections (blue) that are assigned to an SBU. Two linkage sections connect at a point of extension (red dots). Choosing an appropriate termination (orange) for the SBU defines the accuracy of the cluster model.

precursors to form the linkage. The COF itself can be partitioned into secondary building units (SBUs),[33,34] each containing the linker core and part of the linkage, which we will call the SBU's linkage section. The linkage sections of two adjacent SBUs connect at a point of extension to form a linkage. In this work, the points of extension are always located in the geometric center of the central bond of the linkage. As will be explained in Section 2.4, this partitioning of the material in SBUs enables us to derive a cluster force field for each building block, which can be combined to accurately describe the whole COF structure. To obtain an accurate cluster force field for these SBUs, their environment has to be mimicked as detailedly as possible by choosing an appropriate termination for the linkage section. The combination of the linker core with its termination is defined as the cluster for this SBU.

2.2 Structure assembly and database construction

Our top-down approach to generating a versatile COF database starts from a topological blueprint and a selection of SBUs that can be placed on its nodes (vertices and edges). A topology is a periodic graph in which the connection between vertices, *i.e.*, the connection between the SBUs of the material, is represented by edges. This is illustrated in Fig. 2 for the **pts** topology. The collection of both vertices and edges is defined as the nodes of the topology. The process of determining the topology is straightforward once the structure of the material is known and the building blocks or the linkages between them are defined. In this case, the topology directly follows by reducing each building block into a zero-dimensional vertex and storing the connectivity between building blocks as edges that connect these vertices. However, in the inverse approach, *i.e.*, when only the topology and its SBUs are known, it is more difficult to generate the material's structure at the atomic level, because it is not clear how the three-dimensional SBUs should be inserted on the nodes.[79] Since this is a vital part of our database generation, we here implement an additive top-down approach. This takes into account both geometric and energetic parameters to introduce the rigid building blocks into the topology to obtain a physical structure. As illustrated in Fig. 2, the procedure consists of a four-step process, which is explained in more detail below. Three filters check whether the structure is physical or whether it should be rejected. If the optimized material passes each filter, it is added to the final database. Additional details to construct the database are provided in Section S1 of the ESI.† In Section S1.6,† the structure assembly procedure is illustrated with a detailed case study of COF-108.

Step 0. In step 0 of Fig. 2, the topology and SBUs, *i.e.*, a so-called (topology, SBUs) combination, are selected as input. The Reticular Chemistry Structure Resource (RCSR) contains a large number of experimentally observed two- and three-

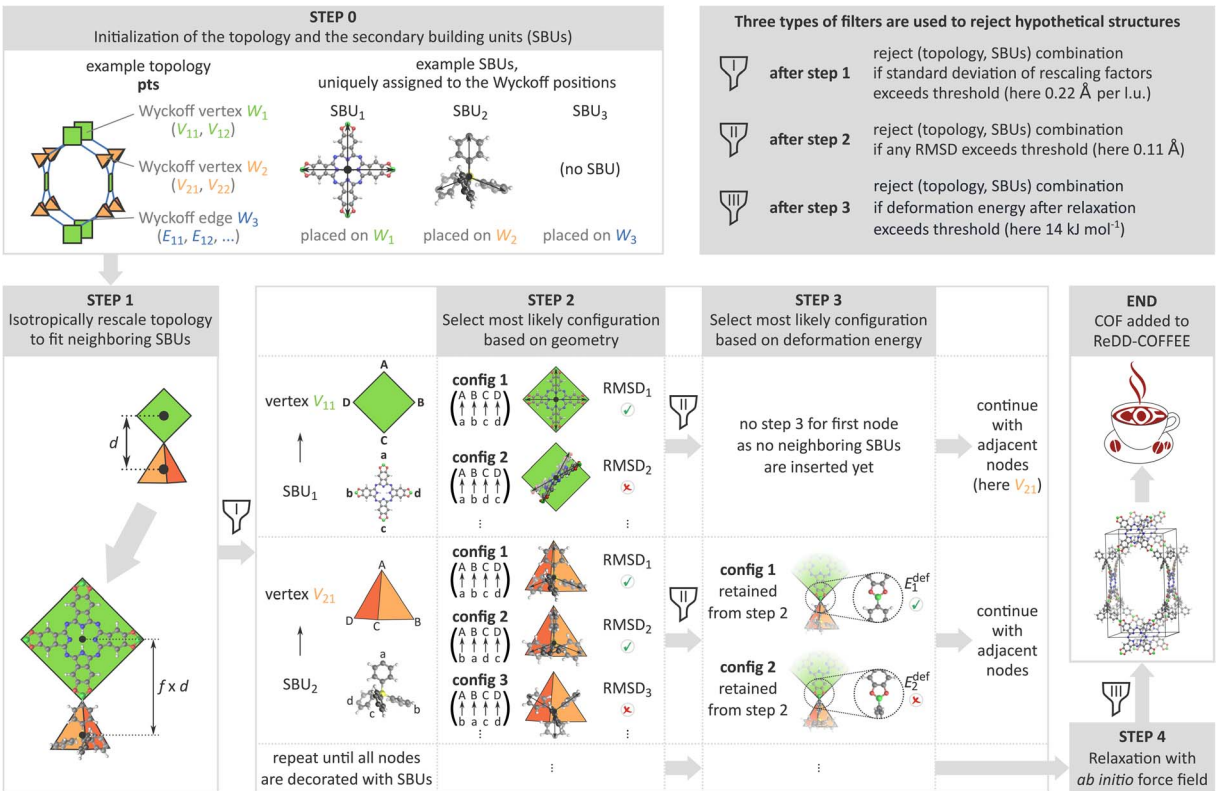

Fig. 2 The four-step process followed to assemble a COF structure. This procedure starts from a topology and a selection of SBUs for each set of equivalent nodes. In step 1, the topology is rescaled to fit the SBUs. In step 2 and step 3, the most favorable SBU configurations are chosen based on geometric and energetic considerations. By using an additive top-down approach, a low-energy structure is quickly generated. Three filters are implemented to ensure that only physical structures are added to the database.

dimensional topologies, embedded in a Euclidean representation.[98] We extracted these embeddings using a web scraping script and only retained the 2495 topologies with embedding type 1, *i.e.*, where all nearest-neighbor vertices at the same distance are connected by an edge. These topologies are most frequently observed experimentally.[99] Subsequently, an SBU is assigned to each set of equivalent nodes, *i.e.*, each Wyckoff set of the topology, based on two restrictions. Firstly, the number of points of extension of the SBU has to be the same as the coordination number of the node, as it has to be able to connect with each neighbor. Secondly, the two linkage sections of neighboring SBUs have to form one of the linkage types considered in this manuscript. For example, it is not possible to combine SBUs belonging to an imine and a boronate ester linkage. Since each Wyckoff set can be assigned a different SBU, the theoretical upper limit for the number of distinct SBUs in a material is given by the total number of Wyckoff sets of the material's topology. To also include materials without a linear linker in the database, an edge Wyckoff set may also remain vacant, in which case the SBUs that decorate the neighboring vertices are directly connected. In this work, the SBUs are chosen from a set of 279 building blocks. Four of them are SBUs that have no experimental precursor but emerge during COF synthesis as product of the reaction groups: boroxine, triazine, borazine, and borosilicate. The other 275 are generated by combining 30 experimentally observed linker cores with eleven frequently observed linkage types that correspond to eleven reactive groups. Our database contains imine,[100] boronate ester,[1] (keto)enamine,[21,92] triazine,[101] (acyl)hydrazone,[102] azine,[103] imide,[104] boroxine,[1] borosilicate,[105] oxazoline,[106] and borazine[107] COFs (see Fig. S1 and S2 of the ESI†). In this way, a total of 5 537 951 (topology, SBUs) combinations in 1272 topologies are initialized. Each of these combinations is given the unique label top_SBU_1_SBU_2_…_SBU_N, in which "top" indicates the topology and SBU_i is the SBU placed on the *i*-th Wyckoff set (see Section S1.3 of the ESI†).

Step 1. After choosing the (topology, SBUs) combination in step 0, the topology is isotropically rescaled to be able to accommodate the SBUs in step 1 of Fig. 2. To this end, a rescaling factor f_i is calculated for each edge Wyckoff set so that the distance between its connecting vertices is the same as the distance between the centers of the SBUs occupying those vertices. We require the unit cell to be rescaled isotropically, *i.e.*, with the same rescaling factor in every direction. The single rescaling factor is defined as the mean of the rescaling factors averaged over all Wyckoff sets. To prevent the rescaling factors from differing too much and avoid overlapping SBUs, the standard deviation of the set of rescaling factors is not allowed to surpass a threshold of 0.22 Å per l.u., with l.u. being the length unit of the topology (see Section S1.5 of the ESI†). This constraint corresponds to the first filter visualized in Fig. 2. After this first filter, 749 859 (topology, SBUs) combinations are passed on to step 2. Most discarded combinations contain mixed-linker topologies in which the lengths of the linear linkers are too different.

Step 2. Once the topology is rescaled to accommodate the SBUs, they can be inserted onto their assigned nodes. This is accomplished by orienting the points of extension of the SBU towards the neighboring nodes in the topology while keeping the SBU's internal geometry fixed. Once a point of extension is chosen for each neighboring node, the optimal transformation is found by Kabsch algorithm.[108] However, it is not *a priori* known which point of extension has to be oriented towards which neighboring node. Every permutation of the points of extension results in a different SBU configuration, with possibly a different internal geometry. This is illustrated in Fig. 2 by configurations 1 and 2 of the square SBU_1 that is placed on vertex V_{11}. The points of extension of SBU_1 are indicated with the lower case letters a, b, c, and d, whereas the neighboring nodes of vertex V_{11} are specified with the upper case letters A, B, C, and D. When points of extension a and b are oriented towards nodes A and B, respectively, the least strain configuration is found when assigning the points of extension c and d to the nodes C and D, as in the first configuration in Fig. 2. However, in the second configuration, these points are assigned to the nodes D and C, respectively, for which the Kabsch algorithm defines a transformation that results in a less favorable configuration. Furthermore, for some SBUs, the point symmetry of their internal geometry is lower than the point symmetry of their points of extension. In these cases, information about the internal geometry of the building block is lost by abstracting the SBU to its points of extension.[79] Therefore, even for two configurations that orient the points of extension towards the neighboring nodes equally well, different internal geometries may be found. This is illustrated for configurations 1 and 2 of the tetrahedral SBU_2 that is placed on vertex V_{21} in Fig. 2. Whereas the SBU perfectly fits the vertex in both configurations, two different internal geometries are observed, which differ in the orientation of the phenyl rings. To only retain the most favorable SBU configurations, geometric and energetic criteria are taken into account in step 2 and step 3 of the structure assembly process, respectively.

In step 2 of Fig. 2, a first selection of reasonable SBU configurations is made based on the geometric consideration whether the points of extension can be oriented towards the neighboring nodes correctly. This is quantified by the root-mean-square deviation (RMSD) between the unit vectors pointing (i) from the center of the SBU towards the points of extension, and (ii) from the node towards its neighbors in the topology. In the ideal case, these two sets of vectors would be overlapping once the SBU is positioned on the node using the transformation obtained from Kabsch algorithm.[108] However, for some configurations, such as configuration 2 of SBU_1 and configuration 3 of SBU_2 in Fig. 2, the transformation results in a less optimal configuration and a geometric mismatch between the SBU and the topological node is introduced. To avoid that these SBU configurations would be inserted in the topology, only those configurations that minimize the RMSD are selected and passed on to step 3. Furthermore, to avoid structures with too large a mismatch, a second filter checks if the minimal RMSD for each node is lower than the threshold $RMSD_{max} = 0.11$ Å (see Section S1.5 of the ESI† for the rationalization of this value). If this is not the case, for instance when trying to insert a tetrahedral SBU on the square vertex V_{11} in Fig. 2, the

(topology, SBUs) combination is rejected. After this step, 403 581 combinations in 1196 topologies are retained.

Step 3. As explained before, an identical RMSD does not imply the same internal geometry as the representation of the SBU by its points of extension does not fully capture its three-dimensional structure. Depending on the geometry of the neighboring SBUs, the linkage formed between the SBUs in one configuration can be energetically more favorable than in another configuration, although they have the same geometric mismatch, such as configurations 1 and 2 of the tetrahedral SBU that is placed on vertex V_{21} in Fig. 2. In step 3, the deformation energy E_{def}, formally introduced in Section 2.4, is used to discriminate between such configurations and identify the one in which the topological constraints have the least influence on the SBUs. This energy depends on the selected configuration of the neighboring SBUs. Therefore, it should in principle be calculated for each possible material configuration, which is a product of all SBU configurations. However, as the number of these material configurations increases exponentially with the number of nodes in the topology, a brute-force iteration over all material configurations is unfeasible, definitely when constructing an extensive database.

This challenge motivated us to introduce our *additive* top-down approach, in which the nodes of the topology are decorated one-by-one. In each step, the optimal configuration of the SBU of the selected node, *i.e.*, the configuration that minimizes both the RMSD and the deformation energy of the linkages with the already inserted SBUs, is identified. By following a breadth-first iteration through the topological graph, all nodes at a certain distance from the starting vertex are decorated before continuing with the next layer. As such, the number of linkages with neighboring building blocks is increased for each SBU as compared to a random iteration or a depth-first iteration. As indicated in Fig. 2, step 3 can be omitted for the starting node, as no neighboring SBUs are present yet and the deformation energy E_{def} is the same for all configurations. The final SBU configuration for vertex V_{11} is thus chosen from the set of configurations that minimizes the RMSD in step 2. In the next step of the breadth-first iteration, a neighboring node to the ones already occupied is decorated with an SBU. In the example of Fig. 2, the vertex V_{21} is selected once a configuration for the SBU that is put on vertex V_{11} is determined. From the SBU configurations that minimize the RMSD in step 2, the deformation energy E_{def} of the linkage from this SBU with the already inserted SBU on vertex V_{11} is calculated. Finally, the SBU configuration that minimizes the deformation energy E_{def} is selected and inserted on the vertex V_{21}. As such, the deformation energy E_{def} has to be calculated only once for every SBU configuration, instead of for every material configuration, and the computational complexity of the structure assembly process is reduced from exponential to linear in terms of the number of nodes in the topology. For linear linkers, which only have two points of extension lying on the same axis as the center of the SBU, there is still a degree of freedom that step 2 can not describe: the rotation around their axis. By using the deformation energy E_{def} in step 3, this degeneracy is lifted to the point that only the internal symmetries of the linker remain.

Step 4. After step 3 is completed for each node in the topology and all SBUs are inserted and connected, an initial periodic structure is obtained. However, up to now, the SBUs remained rigid. As the threshold for the maximally allowed RMSD in step 2 is quite high, this means that some mismatch between SBUs is allowed, which is motivated by the fact that COF building blocks can have an appreciable degree of flexibility that can accommodate for the introduced mismatch. Therefore, after the initial material is assembled, it is relaxed in step 4 using its system-specific force field that is generated according to the procedure described in Section 2.4 to find an optimal structure. The mismatch between the SBUs, which was localized around the linkages, will now be released after a full geometry optimization. However, as some SBUs are more flexible than others, the final optimized structure can have a low synthetic likelihood. Therefore, the last filter uses the deformation energy E_{def} again to check whether the strain that was introduced between the rigid SBUs is sufficiently released. The threshold is here defined as 14 kJ mol^{-1} (see Section S1.5 of the ESI†). When the deformation energy of the optimized structure does not exceed this threshold, it is finally included in the ReDD-COFFEE database, which contains 268 687 structures, distributed over 1114 topologies. In total, there are 5856 2D structures and 262 831 3D structures. A detailed overview of the distribution of the linkages and dimensionality found in our database is given in Table S1 of the ESI.†

2.3 Diversity metrics

It is important to compare the diversity of our database with already existing COF databases to verify that a large part of material space is covered and that these regions are sampled equally to avoid focus on certain subclasses.[82,83,86,89] Inspired by a previous MOF study by Moosavi *et al.*, three metrics are defined below to assess the diversity of a subset of the material space: the variety V, the balance B, and the disparity D.[37] We will calculate these diversity metrics on four domains: (i) the pore geometry, and the chemical environment of the (ii) linkages, (iii) linker cores, and (iv) functional groups. Whereas the pore geometry can be described by eight structural parameters, the chemical environments of the COF are assessed with revised autocorrelation functions (RACs),[109] which were proven to be useful in multiple applications.[37,109–112] Further details about these diversity metrics are described in Section S3 of the ESI.†

The RACs are calculated between two sets of atoms, which are defined for each chemical environment (ii–iv). Initially, the linkages (ii) that hold together the COF are identified by scanning the material graph for linkage patterns. By removing these linkages from the material graph, the linker cores (iii) that constitute the COF are retrieved. The functional groups (iv) are detected as parts of the linker cores that are attached to its skeleton with exactly one bond and do not exist of a single hydrogen atom. Once the start and scope atom lists are determined for each chemical environment (see Section S3.2 of the ESI†), the difference and product RACs are defined in eqn (1) and (2), respectively.

$$ {}^{\text{start}}_{\text{scope}}P_d^{\text{diff}} = \sum_i^{\text{start}} \sum_j^{\text{scope}} (P_i - P_j)\delta(d_{ij}, d) \quad (1) $$

$$ {}^{\text{start}}_{\text{scope}}P_d^{\text{prod}} = \sum_i^{\text{start}} \sum_j^{\text{scope}} P_i P_j \delta(d_{ij}, d) \quad (2) $$

Each RAC investigates one out of six properties P: the atom identity (I), connectivity (T), Pauling electronegativity (χ), covalent radius (S), nuclear charge (Z), or polarizability (α). The depth d indicates the number of bonds that must be present between the considered atoms in the start and the scope list. Together with the definition of the start and scope atom lists, which depend on the chemical environment, this depth specifies the atom pairs over which to iterate. With two types of RACs, six properties, and a maximum depth of three bonds, a total of 48 descriptors are obtained for each environment (linkages, linker cores, and functional groups).

Once all materials are featurized, either with structural parameters or with RACs, the diversity metrics that determine how well a material set covers the material space can be defined. The material space is in this context described by the union of all available databases: the four existing databases and our database presented here. Before calculating the diversity metrics, the material space is subdivided into a specific number of bins, here chosen to be 1 000, defined through k-means clustering. These bins partition the material space into subclasses. The variety V of the considered database checks if each of the subclasses is sampled by measuring the number of bins that are examined. To make sure that each subclass is sampled equally, the balance B indicates the evenness of the distribution of materials among the sampled bins. In the ideal case, all covered subclasses are represented with the same number of structures. Lastly, the disparity D quantifies again the fraction of material space that is covered by the considered database, using a distance-based approach instead of the clustering into bins. Therefore, it is also a measure of the spread of the bins. The mathematical definition of these diversity metrics is given in Section S3.3 of the ESI.† Together, the variety V, the balance B, and the disparity D summarize the diversity of a subset of the material space for each investigated domain.

2.4 Force field generation

The ReDD-COFFEE database aims to report on both material geometries and system-specific force fields. When compared to *ab initio* methods, these force fields can be directly adopted to perform high-throughput simulations on a larger number of structures and for longer time scales. Moreover, it can do so with a higher accuracy than generic force fields. Most system-specific force fields are derived from periodic *ab initio* data. However, for a database containing more than 100 000 structures, such procedure becomes unfeasible. To reduce the amount of necessary *ab initio* data, we exploit the building block nature of reticular materials to partition each COF in SBUs and associated clusters. For each of these clusters, a system-specific cluster force field is derived. Following our QuickFF procedure,[38,39] a system-specific force field for the periodic structure is obtained by combining the cluster force fields to account for all SBUs in a specific material. As many materials can be generated with a limited set of SBUs, this proven approach circumvents the need for expensive periodic *ab initio* reference data and requires only a modest set of cluster calculations.[38–50] Section S2 of the ESI† contains supplementary details about this approach.

The accuracy of these cluster force fields is to a large extent defined by the choice of termination. A larger termination mimics the environment of the SBU in the material in more detail but is only applicable for a smaller set of materials as the termination has to represent the material correctly. Herein, we have chosen to define a single termination for each observed linkage section (blue in Fig. 1). The cluster termination is therefore independent of the SBU environment and can be adopted in each material containing that SBU. As such, there are as many clusters as there are combinations of linker cores and linkage sections. These terminations are depicted in Fig. S3 of the ESI.†

Once a cluster force field is obtained for each SBU in the material, the parameters of the periodic force field are derived. As the cluster approximation has the smallest impact on covalent terms that are fully embedded in the SBU, these terms are directly adopted from the cluster force field. In contrast, the overlap terms that span multiple SBUs are most accurately described in the cluster force field of the SBU with the majority of the atoms. Hence, these parameters are obtained by taking a weighted average over the respective terms in both constituent SBUs. For each term, the weights associated with each cluster involved in the term are proportional to the number of atoms embedded in the SBU core of the cluster. Also bond charge increments of bonds spanning two SBUs are obtained following this procedure. Charge neutrality is satisfied by using bond charge increments to derive the partial charges in the periodic material.[113] The averaging is only allowed when all covalent bonds involved in the term have the same bond order in both clusters involved, which is mostly the case in this paper. When this is not the case, the parameters are directly obtained from the building block that mimics the environment correctly.

Inserting the SBUs in a periodic material introduces topological constraints to its environment, which can impose strain between the SBUs that is not present when considering an isolated cluster. The equilibrium geometry of the SBU in the material is therefore not necessarily the same as in the cluster model. This is for example illustrated in COF-108,[2] where the tetrahedral TBPM building block, with point group T_d, is inserted in the **bor** topology and placed on a vertex with point group D_{2d}. Due to the geometric mismatch, which results in an RMSD of 0.09 Å before optimization, the SBUs are slightly reshaped with respect to their equilibrium structure.

To quantify the energy penalty for inserting the SBUs in a suboptimal environment, the deformation energy E_{def} is introduced as the energy difference between the periodic material and the sum of the energies of the isolated clusters it is composed of. Only interactions that are present both in the periodic material and the isolated clusters are included.

Although they differ only slightly from the cluster force field parameters, the force field parameters of the periodic structure are also used for the calculation of the energies of the different clusters to ensure a consistent comparison. More specifically, the covalent and non-covalent interactions in the periodic material are divided into interactions between atoms within the same SBU ($E_{\text{cov}}^{\text{per,intra}}$ and $E_{\text{non-cov}}^{\text{per,intra}}$) and in different SBUs ($E_{\text{cov}}^{\text{per,inter}}$ and $E_{\text{non-cov}}^{\text{per,inter}}$). In the periodic material, the interactions within the same SBU, *i.e.*, $E_{\text{cov}}^{\text{per,intra}}$ and $E_{\text{non-cov}}^{\text{per,intra}}$, are also described in the cluster model by $E_{\text{cov}}^{\text{clust,intra}}$ and $E_{\text{non-cov}}^{\text{clust,intra}}$, respectively. The covalent interactions between different SBUs in the periodic material, *i.e.*, $E_{\text{cov}}^{\text{per,inter}}$, exactly correspond to the overlap terms between the core and the terminations of the appropriate clusters $E_{\text{cov}}^{\text{clust,inter}}$. To avoid that these overlap terms would be counted in both corresponding clusters, they are weighted with the same rescaling factors as in the generation of the periodic force field. However, no cluster counterpart for the long-range non-covalent interactions between different SBUs $E_{\text{non-cov}}^{\text{per,inter}}$ can be defined, as these interactions are not integrated into the isolated cluster model. They are therefore neglected in the calculation of the deformation energy. As the topological constraints scale with the number of linkages between SBUs, N, the deformation energy is normalized with this number, resulting in the definition of E_{def} in eqn (3).

$$E_{\text{def}} = \frac{E_{\text{cov}}^{\text{per,intra}} + E_{\text{cov}}^{\text{per,inter}} + E_{\text{non-cov}}^{\text{per,intra}}}{N} - \frac{E_{\text{cov}}^{\text{clust,intra}} + E_{\text{cov}}^{\text{clust,inter}} + E_{\text{non-cov}}^{\text{clust,intra}}}{N} \quad (3)$$

Since the deformation energy indicates the energy change upon inserting an SBU in a topology, it describes how easily the SBUs can accommodate the mismatches introduced due to topological constraints, *i.e.*, if they are sufficiently flexible. High values of the deformation energy indicate that the SBUs are too rigid to accommodate the introduced mismatches and result in highly contorted structures. As described in Section 2.2, COFs with a high deformation energy have a low synthetic likelihood and are therefore discarded from the database *via* the third filter.

3 Computational details

For each SBU, a cluster force field is derived that is fitted to the *ab initio* Hessian and equilibrium structure. An initial cluster geometry is assembled with Avogadro (v1.2.0)[114] and afterwards optimized with Gaussian 16 (Revision C.01)[115] using the B3LYP functional[116–118] with Grimme D3 dispersion correction.[119] The Hessian of the relaxed structure is calculated using the same level of theory. When imaginary frequencies are observed, indicating that the cluster is optimized to a saddle point on the potential energy surface, a small perturbation is added to the geometry and the procedure is repeated until all frequencies are positive. All calculations adopt the 6-311++G(d,p) Pople basis set[120] and the NoSymm flag is used. Once the *ab initio* reference data is obtained, a cluster force field is derived comprising three main parts. The van der Waals interactions are described with a Buckingham potential using the MM3 parameters derived by Allinger *et al.*[121] The partial charges for the electrostatic interactions are obtained with the MBIS procedure[122] as implemented in HORTON (v2.0.0).[123] Bond charge increments[113] are used to divide the charges over the different bonds and Gaussian smearing is applied to obtain a more accurate description of the charge distribution. Finally, the covalent part of each cluster force field is fitted to the *ab initio* Hessian and equilibrium structure using QuickFF (v2.2.4),[38,39] where both the electrostatic and van der Waals interactions are provided as reference force fields. Each set of equivalent atoms is given a unique force field atom type. Specific details about the triazine dihedral angles and the out-of-plane terms are provided in Section S2.1 of the ESI.† Besides the QuickFF derived force field, also a UFF force field[124] is generated for each SBU to compare their accuracy with system-specific force fields. The covalent and van der Waals interactions are defined according to the rules mentioned in the original UFF paper,[124] which are implemented in our in-house force field generator software. As no partial charge fitting scheme is imposed by the UFF method, and as our focus is to describe the covalent part of the force field, the electrostatic interactions are defined identically to the QuickFF derived force field. For the validation of the cluster force fields, each cluster is optimized with both force fields, using our in-house force field engine Yaff (v1.6.0),[125] and the vibrational frequencies are calculated with a normal mode analysis (NMA) using TAMkin (v1.2.6).[126]

The initial periodic structure is generated by placing the SBUs on the nodes of the topology and its periodic force field is derived from the cluster force fields of the underlying building blocks, as described in Sections 2.2 and 2.4, respectively. For 2D COFs, a $1 \times 1 \times 2$ supercell is used to include two layers in the simulation cell that can better describe the specific type of stacking. The initial structure is a perfect eclipsed structure with an interlayer distance of 10 Å. This interlayer distance is chosen sufficiently large to avoid overlap between the layers and decreases during the optimization. The structures are relaxed with Yaff,[125] following a three-step procedure. During the structure assembly process, nearby SBUs may contain atoms that almost overlap. The potential well of the Buckingham potential that describes the van der Waals interactions of the QuickFF derived force field would diverge to minus infinity when two such atoms approach each other. To push these atoms apart and generate a more physical geometry, the first step of the optimization procedure is to optimize the atomic positions with the UFF force field for 50 steps. The van der Waals interaction of the UFF force field is described by a Lennard-Jones potential, which is repulsive at short distances.[124] Subsequently, the system-specific QuickFF force field is applied to fully optimize the atomic positions. Finally, the unit cell parameters are added to the degrees of freedom that are relaxed. For each optimization, the conjugate gradient optimizer as implemented in Yaff is adopted. A real-space cutoff r_{cut} of 11 Å is used for the nonbonded interactions. Tail corrections are used to estimate the van der Waals interactions beyond this cutoff distance. Furthermore, the electrostatic interactions are

calculated with an Ewald summation, with scaling factor $\alpha = 0.26$ Å^{-1} and reciprocal space cutoff $k_{max} = 0.26$ Å^{-1}. A truncation model is used to smooth these interactions. The textural properties of the optimized structures are derived with Zeo++ (v0.3).[127] To calculate the accessible surface area and volume, a probe radius of 1.84 Å, coinciding with the kinetic radius of nitrogen,[128] and 3000 Monte Carlo samples are adopted. In Section S5 of the ESI,† benchmark studies for all simulation parameters are provided.

Molecular dynamics (MD) simulations are performed to compute the powder X-ray diffraction (PXRD) patterns and single crystal structures at operating conditions[50] using the Yaff software package[125] with the same force field parameters as determined for the optimizations. The MD simulations sample the (N, P, $\boldsymbol{\sigma}_a = \mathbf{0}$, T) ensemble at operating conditions, *i.e.*, a pressure of 1 atm and a temperature of 100 K for COF-300 and LZU-111, a pressure of 1 atm and a temperature of 89 K and 298 K for the distinct phases of COF-320, respectively, and a pressure of 1 bar and a temperature of 300 K for the calculation of the PXRD patterns. They are controlled by a Nosé–Hoover chain thermostat[129–131] with three beads and a relaxation time of 100 fs, and a Martyna–Tuckerman–Tobias–Klein barostat[132,133] with a relaxation time of 1000 fs, respectively. The velocity Verlet integration scheme is used, with a timestep of 0.5 fs. For the computation of the crystal structures, 9900 snapshots are collected from an MD trajectory of 500 ps, where the first 5 ps are considered as equilibration run. An MD trajectory of 200 ps is created to compute the PXRD patterns and 50 snapshots are extracted from the last 100 ps. 3D COFs in the test set are simulated using a $2 \times 2 \times 2$ supercell, whereas a $1 \times 1 \times 5$ supercell is adopted for the 2D COFs, resulting in simulation cells of ten layers to account for the inherent freedom in layer movement. All PXRD patterns are computed using the *pyobjcryst* python package, based on the ObjCryst++ Object-Oriented Crystallographic Library.[134] A Cu K_α wavelength of 1.54056 Å is adopted. The peak shape is computed with a pseudo-Voigt shape function, where the mixing parameters are set to $\eta_0 = 0.5$, $\eta_1 = \eta_2 = 0$, and a fixed width W of 0.02° in Caglioti's formula together with the U and V parameters equal to 0°.

In Section 4.4, we perform grand-canonical Monte Carlo (GCMC) simulations[135] using RASPA[136] to calculate methane storage capacities at pressures of 5.8 bar and 65 bar. A cutoff radius of 14 Å is used for all interactions and tail corrections are applied, as required by the TraPPE model.[137] The simulation cell contains as many unit cells as needed to ensure a distance of twice the cutoff radius in every direction. The simulations are run for 10 000 cycles, from which the first 5000 are discarded for equilibration. As outlined in Section S5.3 of the ESI,† this is sufficient to let the system equilibrate. For the subset of 10 000 COFs that are contained in our test set, 1600 structures are discarded from the GCMC analysis as their pore sizes are prohibitively large, preventing the calculation of the interactions between the framework and the methane molecules at 65 bar within a reasonable timeframe. As explained in Section 4.4, it is expected that these structures have a low volumetric deliverable capacity and are therefore unviable candidates for vehicular methane storage. Six additional COFs are discarded as they require too much memory.

4 Results

4.1 Accuracy of the system-specific force fields

As explained in Section 2.4, we have derived a system-specific QuickFF force field for each material in the ReDD-COFFEE database from the cluster force fields of its constituent SBUs. In this section, we will verify that these QuickFF force fields indeed achieve a higher accuracy than universal force fields typically used in high-throughput screenings. Here, we will compare the system-specific force fields with the widely adopted UFF force field.[124] To this end, we will assess the ability of the cluster force fields in reproducing the vibrational frequencies and internal coordinates of the *ab initio* optimized clusters, and we will quantify to what extent the periodic force fields can reproduce experimental powder X-ray diffraction (PXRD) patterns and single crystal structures.

Starting from the *ab initio* optimized geometry, both the QuickFF and UFF force fields are used to relax the SBU clusters. Subsequently, the adjustments of the internal coordinates, *i.e.*, the bonds, bends, dihedral angles, and out-of-plane distances, are measured. The QuickFF force field successfully reproduces the *ab initio* optimized geometry, with RMSD errors on the bonds, bends, and out-of-plane distances as small as 4.73×10^{-3} Å, 7.18×10^{-1}°, and 4.12×10^{-2} Å, respectively (see Fig. S26 in Section S2.3 of the ESI†). These are substantially lower than the RMSD errors obtained for the UFF optimized structures, which amount to 3.56×10^{-2} Å, 2.87°, and 4.50×10^{-2} Å, respectively. The most challenging internal coordinates to describe with force fields are the dihedral angles of non-planar COF building blocks, as they are primarily dictated by long-range electrostatic and van der Waals interactions. The RMSD error of these internal coordinates in the QuickFF relaxed structure is relatively large, namely 9.40°, which is still improved substantially compared to the UFF relaxed structure, for which the RMSD error is 22.27°.

For these optimized clusters, the force field vibrational frequencies are derived and compared with the ones obtained from the *ab initio* Hessian. Similar to the internal coordinates, also the vibrational frequencies are described more accurately by the QuickFF force field (see Fig. S27 in Section S2.3 of the ESI†). The RMSD error on the frequencies is lower than the errors obtained for the UFF force field in all frequency regions. While they are still comparable in the low-frequency regime (<500 cm^{-1}), the QuickFF force fields outperform UFF by an order of magnitude for the higher frequencies (>500 cm^{-1}), as the QuickFF parameters are fitted to reproduce the *ab initio* Hessian.

This quantitative comparison between the force field and *ab initio* derived values of both the optimized cluster geometry and the vibrational frequencies provides an indication of the accuracy our QuickFF force fields can reach. However, we are more interested in the capacity of our periodic force fields to predict experimentally measured, structural properties. For COFs, a key macroscopic descriptor is the PXRD pattern, capturing the

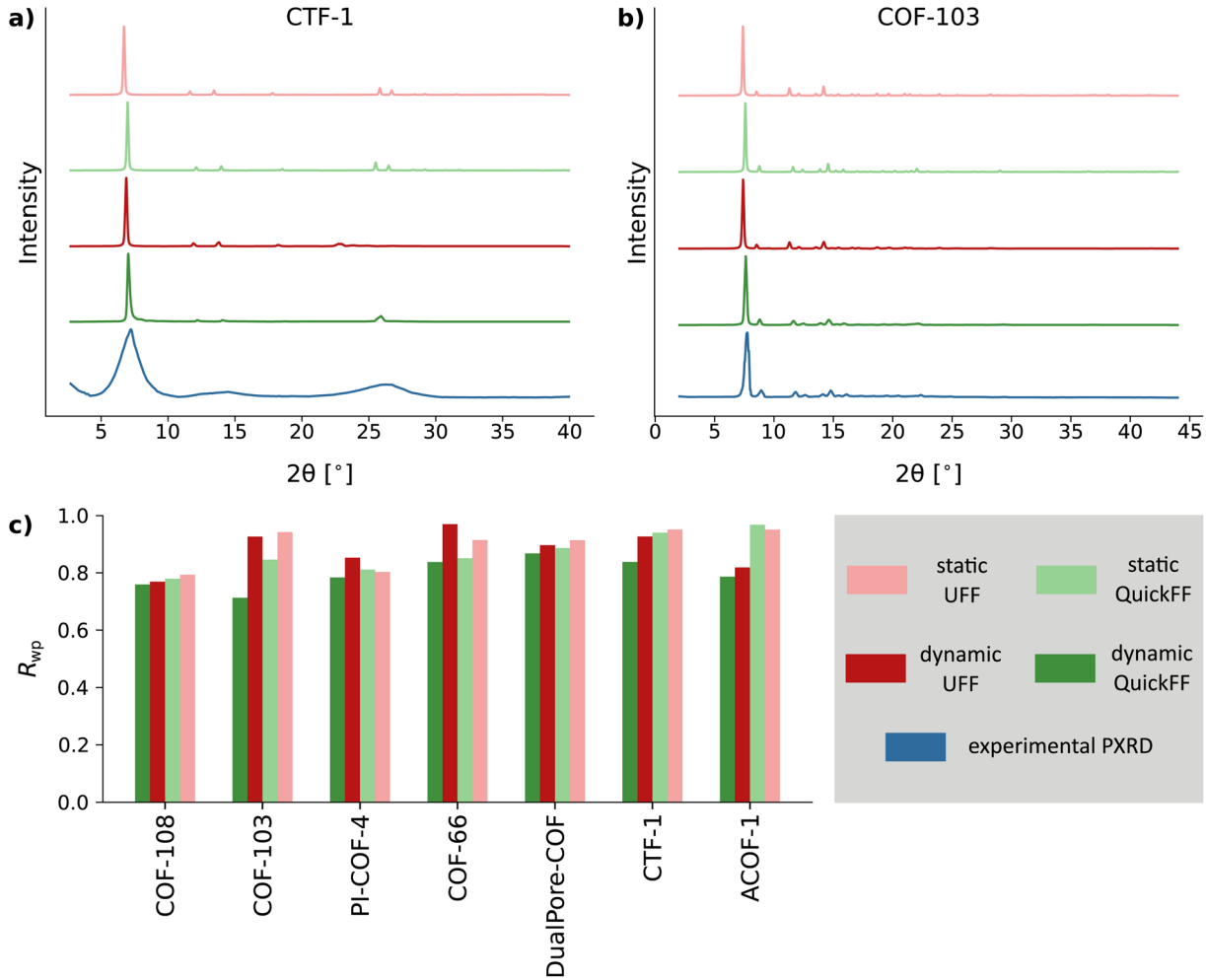

Fig. 3 Comparison of the experimental PXRD patterns of seven COFs with four computational patterns for each material. Static and dynamically averaged PXRD patterns are derived for both the system-specific QuickFF force fields derived in this work and the transferable UFF force field. Diffractograms are offset vertically for clarity. The PXRD patterns of (a) CTF-1 and (b) COF-103 are visualized. (c) Weighted profile residual R_{wp} quantifying the comparison of each of the calculated patterns with the experimental pattern for all selected COFs. The lower the value of R_{wp}, the better the agreement with experiment.

atomic structure of the material, typically used due to the challenging synthesis of single crystal COFs.[5] In Fig. 3c, the weighted profile residual R_{wp} is used as a heuristic metric to compare the calculated PXRD patterns of a diverse set of seven COFs with their experimentally measured pattern. For each material, a static and dynamically averaged PXRD pattern is derived both with the QuickFF and UFF force fields, following our procedure outlined in ref. 50. Dynamical averaging at *operando* conditions is necessary for COFs to account for the inherent temporal character of experimental measurements.[50] In contrast to the static approach, during which the PXRD pattern is calculated for the optimized structure, the dynamic approach starts from an MD trajectory performed at operating conditions. The resulting PXRD pattern is an ensemble average of the pattern calculated for different snapshots from this trajectory.

For the large majority of the examined materials, the R_{wp} metric is lower for the QuickFF than for the UFF force field for both the static and dynamically averaged patterns, indicating a better agreement of the former with the experimental pattern. Of the materials shown in Fig. 3, the static PXRD is predicted better by UFF than QuickFF only for PI-COF-4 and ACOF-1.[12,26] However, once experimental *operando* conditions are taken into account to obtain higher accuracy, our system-specific force fields again outperform the generic ones. As an example, the PXRD patterns for CTF-1 and COF-103 are provided in Fig. 3a and b. It can be observed that the peak that is detected at 27° and 23° in the experimental patterns of CTF-1 and COF-103,[2,138] respectively, is correctly reproduced in the dynamically averaged QuickFF PXRD pattern. However, the poor description of the dihedral angles between aromatic rings in the UFF force field enlarges the unit cell parameters and shifts these peaks to lower angles. PXRD patterns for the additional COFs can be found in Fig. S28 of the ESI.†

When single crystals are available, the structure of a COF can be determined with an even higher resolution compared to

analyzing the PXRD pattern. While the synthesis of single crystal COFs is challenging, some studies have succeeded in experimentally determining the framework's atomic structure with single crystal X-ray diffraction (SCXRD) or 3D rotation electron diffraction (RED).[139,140] To verify that our system-specific force fields also predict the COF structure more accurately than generic ones at these higher resolutions, additional MD calculations at *operando* conditions are performed to reproduce the crystal structures of COF-300 and LZU-111. Furthermore, simulations are executed to distinguish between two distinct phases of COF-320, which are observed at different temperatures and both have a specific pore structure and unit cell volume.

As can be observed in Tables S6–S9 of the ESI,† our system-specific QuickFF force fields indeed achieve an overall better agreement with experiment than the generic UFF force field in describing the crystal structures of COF-300 and LZU-111 and the distinct phases of COF-320. Especially for the bonds, a consistently better description is achieved by our force fields, with relative differences being half of the ones obtained by UFF. Whereas some bends are more accurately described by the generic force field, most of them achieve a significantly higher precision in the simulations performed with our system-specific force fields. Also the dihedral angles, which are most difficult to reproduce due to the importance of the nonbonded interactions, are better described by our *ab initio* force fields. Finally, the unit cell volume and the unit cell lengths are reproduced more accurately by the system-specific force fields, again with the relative difference only being half the one obtained with UFF.

4.2 Diversity of the database

To faithfully describe the material space a database represents, all regions in this space must be sampled adequately and equally. Whereas experimental databases have several regions that are overrepresented, hypothetical databases often lack structures in specific areas of the material space. Such databases are characterized by a low variety V or a low balance B, respectively. Furthermore, it is better to include regions in material space that are well-separated, to obtain a higher disparity D and maximize the information contained in the database. Due to our specific database generation approach, many materials in the ReDD-COFFEE database strongly resemble each other. Therefore, it would also be interesting to identify a subset of the database that has a similar diversity and covers a comparable region of material space as the whole database but with much fewer structures.

The diversity of experimental and hypothetical databases is illustrated in Fig. 4, where the distributions of selected frequently observed linkages of four COF databases are compared with the one presented in this work. Typically, the COF material class is divided into several subclasses according to the formed linkage. Each of these subclasses has its unique characteristics and advantages or disadvantages for several applications.[6,141] While the majority of materials in the CoRE and CURATED databases contain imine and boronate ester

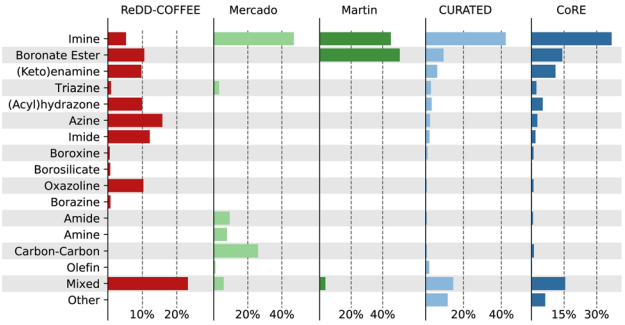

Fig. 4 Distribution of frequently occurring linkage types in five COF databases. The CoRE[89] and CURATED[86] databases are experimental ones, while the databases of Martin[82] and Mercado[83] contain hypothetical structures, similar to our ReDD-COFFEE database.

linkages, other linkage types are less frequently observed in these experimental databases, indicating that they are more difficult to synthesize. To broaden the scope of investigated COF structures, Martin *et al.*[82] and Mercado *et al.*[83] built hypothetical databases using a geometric top-down approach. They chose to focus on a diverse set of linker cores but limited the versatility of linkage types. Both hypothetical databases contain the experimentally abundant imine linkages. In addition, the database of Martin *et al.* also focuses on boronate ester COFs, the second most observed linkage type, and to a lesser extent on borosilicate COFs, while other linkages are not described. Mercado *et al.* chose not to focus on boronate ester COFs, but to describe some less frequently observed linkage types, *i.e.*, the amide, amine, and carbon–carbon linkages. The amide and carbon–carbon linked COFs represent only 1.79% and 1.46% of the experimental CoRE and CURATED databases, respectively, while the amine linked COFs are not present in either of the two databases at all. We chose to continue on this path and include a large variety of linkage types in the ReDD-COFFEE database. We also include linkages that are less frequently observed than the abundant imine and boronate ester COFs but occupy a larger fraction of the experimental databases than amide and carbon–carbon linked COFs. These linkages cover regions in material space that are experimentally more relevant but remain largely uncharted. Furthermore, the structures are well distributed over most linkage types as the only linkage dependent criterium in the (topology, SBUs) combination is that the linkage sections of neighboring SBUs have to combine to form their specified linkage. The structures that are assembled using a boroxine, triazine, borazine, or borosilicate linkage are present less frequently than the other COF subclasses, as the natural constraint that each SBU has to be connected with a three-connected vertex for their linkage largely limits the possible topologies.

Moosavi *et al.* developed a systematic approach to assess the diversity of a MOF database,[37] which we extended to COF databases in Section 2.3. The resulting diversity metrics, *i.e.*, variety V, balance B, and disparity D, for each domain and each of the five COF databases are plotted in Fig. 5. The geometric properties of the structures in hypothetical databases are more

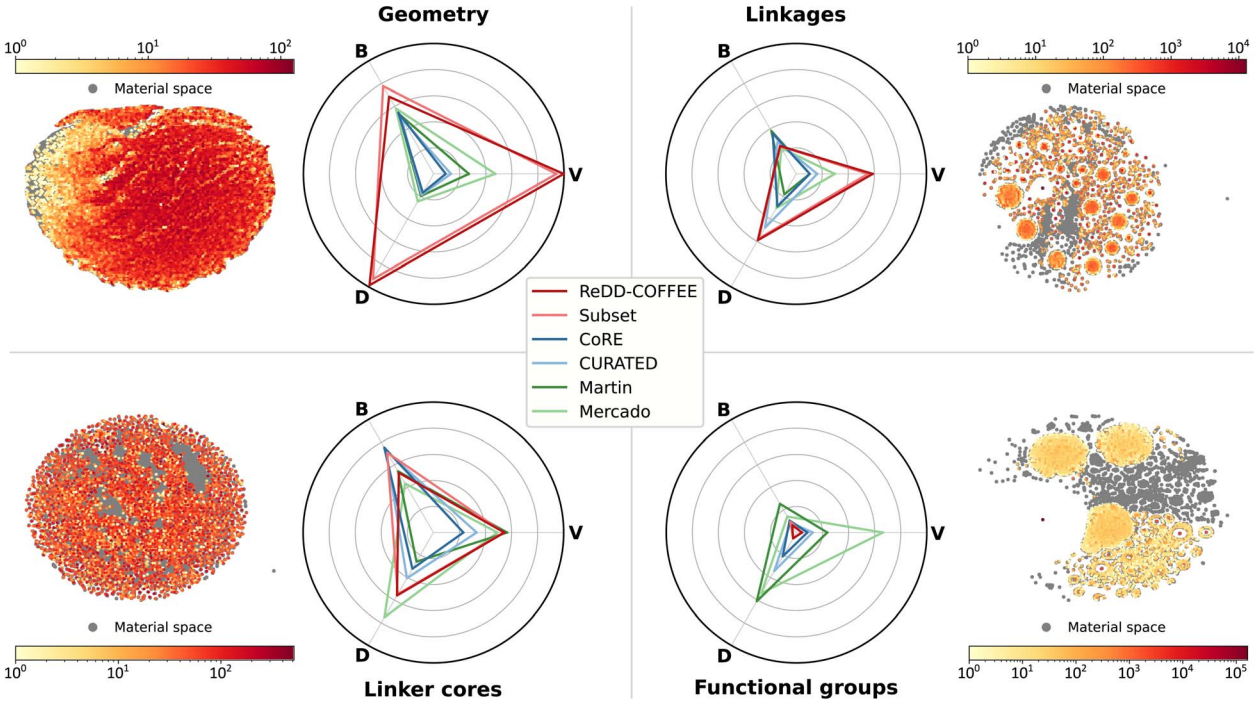

Fig. 5 Diversity metrics (variety *V*, balance *B*, and disparity *D*) for the five domains in COF chemistry. All five databases discussed in the text are considered, together with the diverse subset of 10 000 materials. A *t*-SNE plot visualizes the material space in gray. Overlaid on this plot are the density of the structures from our ReDD-COFFEE database.

diverse than the ones in experimental databases. This comes as no surprise, as *in silico* materials assembly algorithms provide more structural freedom than experimental synthesis procedures. Compared to the other hypothetical databases, the structures generated in this work can deviate more strongly from the perfect topological graph as provided in the RCSR, because the deformation filter only removes those structures for which the deformation energy is prohibitively large. Therefore, the structures in the ReDD-COFFEE database are assembled in a larger number of topologies than the ones in the databases of Martin *et al.* and Mercado *et al.*[82,83] Furthermore, while the distribution of the linker cores in experimental databases is more balanced, their variety is higher in hypothetical databases due to the combinatorial freedom in choosing the linker cores that constitute the COF. Moreover, while the experimental materials and the structures in the databases of Martin *et al.* and Mercado *et al.* are mostly restricted to containing two different SBUs at maximum, the structures generated in this work can combine more than two building blocks in one structure. The diversity of the linkages present in the ReDD-COFFEE database is better in terms of variety and disparity, but the linkages are more balanced in the CoRE database and the database of Martin *et al.* There is no clear preference between the hypothetical databases on the one hand and the experimental databases on the other hand. Lastly, the diversity of the functional groups is the largest in the hypothetical databases of Martin *et al.* and Mercado *et al.* They exploited a large database of linker cores with various functional groups, which are even more diverse than the ones used in experimental structures. In contrast, since we chose to focus on describing different linkage types, we started from a limited number of linker cores, to which a limited set of functional groups were attached (see Fig. S1 of the ESI†). However, the ReDD-COFFEE database can be used to extend the scope towards more functional groups. To this end, a two-step procedure could be used. Our nonfunctionalized database could first be reduced to a set of frameworks with a high potential for the targeted application, after which they can be easily functionalized *a posteriori*.

While our ReDD-COFFEE database combines a high diversity in terms of geometric properties and linker core and linkage chemistry with accurate system-specific force fields, the large number of structures can also form a barrier to its adoption in high-throughput screenings. Therefore, we created a smaller set of 10 000 structures, using the same descriptors that were used to derive the diversity metrics, *i.e.*, geometric properties and RACs to characterize the linker core, linkage, and functional group chemical environments. Starting from a random initial structure, an iterative procedure is followed that selects the structure that has the largest minimal distance to the set of already selected structures in each step. As materials can only be added to the selected set, the variety *V* and disparity *D* of each domain will continue to grow as the selected set becomes larger. However, the balance *B* will initially show a sharp peak when the first chosen structures are added to empty bins. Upon adding further structures, the balance will start to gradually drop. The selected subset, therefore, occupies a slightly smaller region in the material space than the full database, but the structures are better balanced, as can be observed in Fig. 5. Within the full

database, many structures sample similar regions in material space and one structure hardly provides additional information that cannot be learned from the other structures. Computational high-throughput screenings that require expensive calculations can, therefore, use this subset to reduce the computational cost, while still achieving a comparable accuracy with respect to screening the whole database.

4.3 Textural features and property–property relationships

Textural properties are relatively easy to calculate and provide a first insight into the characteristics of the COF material class and the performance of individual materials. Furthermore, they can serve as criteria to filter out structures that do not meet the target design criteria. Therefore, we calculate the textural properties of all COFs in the ReDD-COFFEE database. These properties consist of the mass density and the diameters of the largest included sphere, free sphere, and included sphere along the free path. Also the gravimetric and volumetric accessible surface areas are calculated, as well as the gravimetric accessible volume and the pore fraction. In Fig. 6, property–property relationships between them are established. Additional relationships are visualized in Section S4 of the ESI.† Due to the porous nature of COFs and the fact that they are built up from organic, lightweight atoms, they possess a very low mass density, which allows for exceptionally high gravimetric properties. 2D and 3D COFs show a distinct behavior. Fig. 6c shows that the majority of the 3D COFs in our database have mass densities lower than 200 kg m^{-3} and gravimetric accessible surface areas in the range of 6000 to 10 000 m^2 g^{-1}. In contrast, the mass densities of the more densely packed 2D COFs are mostly within 200 to 600 kg m^{-3} and reach gravimetric accessible surface areas varying from 1750 to 3000 m^2 g^{-1}. On the one hand, COFs with the highest mass density are observed for structures with small pore diameters which do not accept guest molecules entering the material and therefore have no accessible surface area or volume, as is evidenced by Fig. 6a and b. The COFs with the smallest mass density, on the other hand, are obtained for the most porous materials with the largest pores. For these structures, the unit cell grows rapidly, while the accessible surface area increases to a more modest extent. Therefore, the volumetric accessible surface area of Fig. 6a and b drops to zero also for these structures. In between these two regimes of very light and very heavy materials, COFs with an ideal balance between pore volume and accessible surface area are found, which combine high volumetric and gravimetric accessible surface areas with large pore volumes.

The pore diameter, and in turn also the mass density, is not only influenced by the topology and the building blocks of the material, as one could expect, but also by the linkage that connects these building blocks (see Fig. 6d). A linkage type that has a larger spatial extent, such as the azine and (acyl)hydrazone linkages, increases the space between the linker cores and,

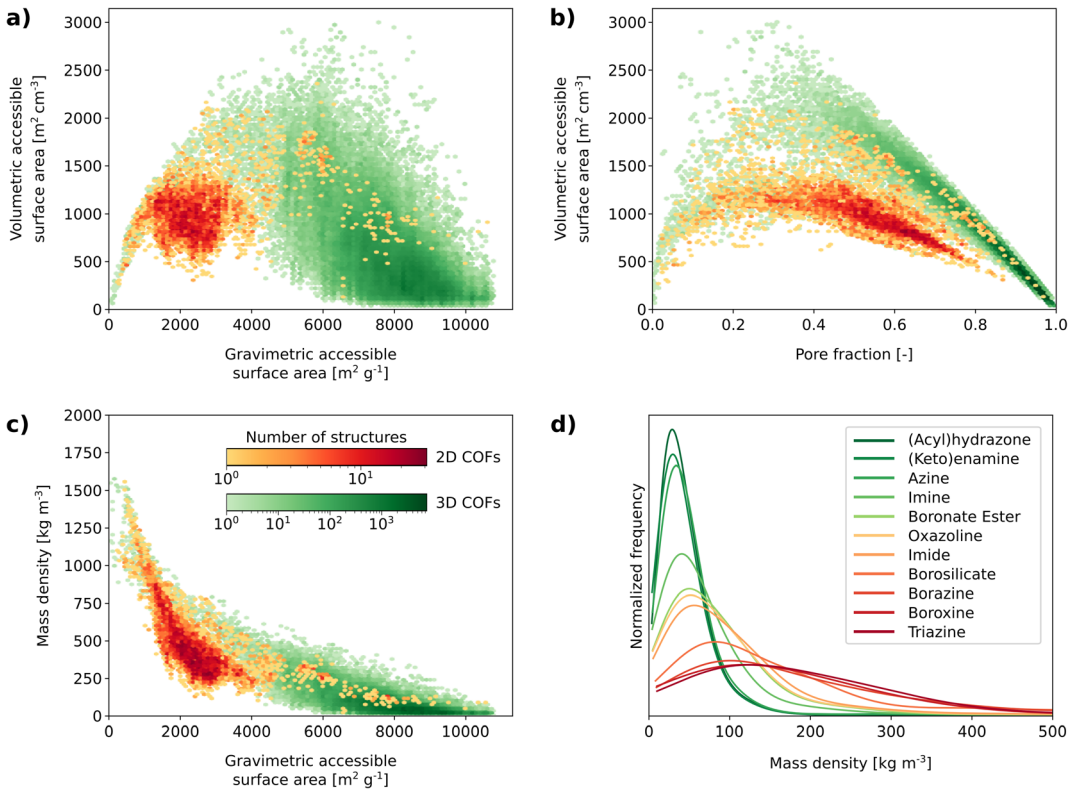

Fig. 6 Textural property–property relations of COFs in the ReDD-COFFEE database. Volumetric accessible surface area in function of (a) the gravimetric accessible surface area and (b) the pore fraction. (c) Relation between the gravimetric accessible surface area and the mass density. (d) Histograms of the mass density for each class of linkage types in the database.

therefore, also the pore diameter. This results in materials with a lower mass density as opposed to boronate ester or imide COFs, which have linkages with a smaller spatial extent. Also the linkage types that emerge during synthesis, such as the triazine and borosilicate linkages, result in higher mass densities as the linker cores of the SBUs are only separated by a single bond and are thus placed closely together. This reasoning does not take into account that interpenetrated nets can form during experimental synthesis upon increasing the pore volume, as these are described to a lesser extent in our database.[32]

These considerations are not unique to COFs. Similar trends in textural properties can be observed for other classes of nanoporous materials, such as MOFs and zeolites. In Fig. 7, a density map of the textural property–property relationships is visualized for our ReDD-COFFEE database, together with three MOF databases, *i.e.*, the experimental QMOF database[68] and the hypothetical hMOF[53] and ToBaCCo databases,[55] as well as for the database of zeolite structures of the International Zeolite Association (IZA).[97] Each relationship shows the same qualitative behavior, independent of the material class. As observed in Fig. 6a and b, the highest volumetric accessible surface areas of almost 2500 m^2 cm^{-3} are found for materials with pore fractions of 0.4 and gravimetric accessible surface areas around 4000 m^2 g^{-1}. Fig. 6c shows that these have a mass density of around 500 kg m^{-3}. As is illustrated in Fig. 6d, a higher pore fraction results in an increased gravimetric accessible surface area. COFs that approach a pore fraction of 1.0 can reach gravimetric accessible surface areas up to 9000 m^2 g^{-1}. While Fig. 6 illustrates that the COFs in our database cover the whole range of textural properties, Fig. 7 emphasizes that most of them are present in the region with the lowest mass density among the three material classes. More precisely, as is illustrated in Fig. S45 of the ESI,† 71.11% of the structures in the ReDD-COFFEE database have a pore fraction above 0.85 and a gravimetric accessible surface area larger than 7000 m^2 g^{-1}. This results in a unique combination of exceptionally high gravimetric accessible surface areas and pore volumes, together with a very large porosity as compared to MOFs and zeolites. This endows them with a large potential in adsorption applications. However, while there are COFs that have a comparable volumetric accessible surface area as the two aforementioned materials classes, most of them cover the smaller accessible surface region. Our property–property relationships show that the choice of topology, building blocks, and, importantly, linkage type, allows for a substantial freedom in the porosity reached in the synthesized COF.

4.4 High-throughput COF screening for vehicular methane storage

One of the applications for which COFs are highly promising is vehicular storage of natural gas,[142,143] due to their low mass

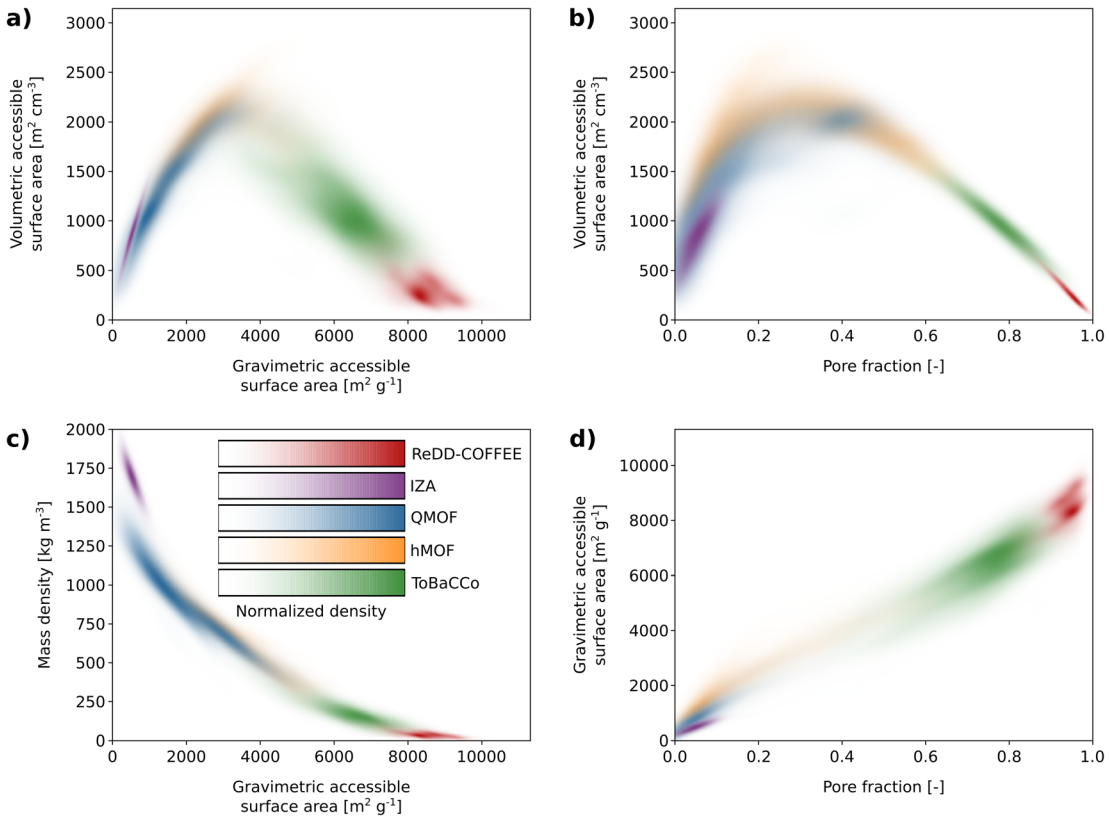

Fig. 7 Textural property–property relations for databases of different material classes. (a) Volumetric accessible surface area as a function of (a) the gravimetric accessible surface area and (b) the pore fraction. (c) Relation between the gravimetric accessible surface area and the mass density. (d) Gravimetric accessible surface area as a function of the pore fraction.

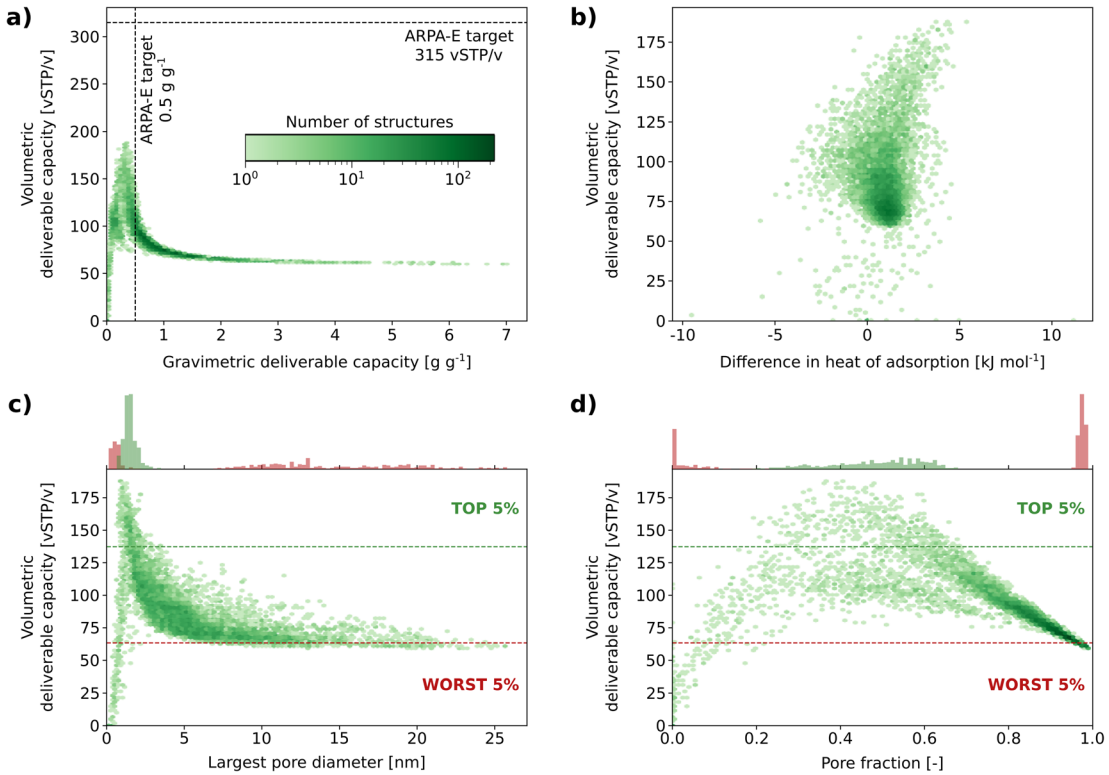

Fig. 8 Relation between vehicular methane uptake descriptors for 8394 COFs in the subset of our database. (a) Gravimetric and volumetric deliverable capacities, with the dotted lines indicating the ARPA-E targets. (b) The volumetric deliverable capacity as a function of the difference between the isosteric heat of adsorption at 65 bar and 5.8 bar. (c) and (d) The volumetric deliverable capacity in function of the largest pore diameter and pore fraction. Histograms on top of the plots indicate the regions where the top 5% (green) and worst 5% (red) performing structures in terms of the volumetric deliverable capacity are observed.

density and high internal surface (see also Fig. 7).[144,145] Natural gas, mainly consisting of methane, is an environmentally more friendly alternative than the traditional petroleum and gasoline based fuels.[142,143] However, the main challenge is that the energy density of natural gas is too low for practical applications. Several approaches to densify natural gas are proposed, such as compressed natural gas (CNG),[146] liquid natural gas (LNG),[147] and adsorbed natural gas (ANG).[148,149] In the latter approach, the gas is stored in the pores of a nanoporous material at a higher storage pressure, after which it is gradually released until the tank reaches the lower depletion pressure, when a refill is needed. The amount of gas that can be released every cycle is defined as the deliverable capacity and is a characteristic of the adopted nanoporous material. To compete with a 250 bar CNG tank, the MOVE program of the Advanced Research Projects Agency-Energy (ARPA-E) of the US Department of Energy has set targets on this deliverable capacity for nanoporous materials.[150] Candidate materials should have a volumetric deliverable capacity of at least 315 vSTP/v, whereas the gravimetric deliverable capacity should exceed 0.5 g g^{-1}.[150]

To check the performance limits of COFs for vehicular methane storage, we perform GCMC simulations on the 10 000 structures in the diverse subset of the ReDD-COFFEE database defined in Section 2.3. A storage pressure of 65 bar and a depletion pressure of 5.8 bar are applied, as imposed by the MOVE program.[51,150] For the electrostatic contribution, MBIS charges[122] are obtained from the periodic force fields and an *ab initio* calculation on a single methane molecule. The choice of van der Waals interactions is thoroughly benchmarked by reproducing the experimental adsorption isotherms of COF-1, COF-5, COF-102, and COF-103 (see Section S5.3 of the ESI†).[8] This benchmark study showed that UFF overestimates the methane uptake, while it is underestimated when the MM3 van der Waals host–guest interactions are adopted. Combining the DREIDING force field[151] to describe the host–guest interactions and the TraPPE united atom model[137] for the guest–guest interactions gives the best accuracy while maintaining a reasonable computational efficiency. Since these GCMC simulations adopt a rigid framework structure, no host–host interaction model is required.

Fig. 8a shows that the majority of the COFs in the subset (70.6%) meet the ARPA-E target of 0.5 g g^{-1} for gravimetric deliverable capacity due to their extremely low mass density. However, the lightest materials also have the largest pores and encompass the largest volumes, resulting in low volumetric deliverable capacities of about 60 vSTP/v. As discussed in Section 4.3, the heaviest materials have no accessible pores and both the volumetric and gravimetric deliverable capacities approach zero. This can also be observed in Fig. 8c and d, where the dependency of the volumetric deliverable capacity on the

largest pore diameter and the pore fraction is plotted, respectively. Histograms of the structures with the 5% highest and lowest volumetric deliverable capacity support the previous claims and identify that the COFs with the highest volumetric deliverable capacity have a pore diameter between 0.7 nm and 3.4 nm and exhibit quite a broad pore fraction within the range of 0.2 to 0.7. While these materials do not have the highest gravimetric deliverable capacity, there are still structures with high volumetric deliverable capacity that meet the ARPA-E target for gravimetric deliverable capacity. The highest volumetric deliverable capacity, 187.4 vSTP/v, is observed for the boronate ester based COF ths-c3_11-01-01_06-08-01_06-08-01, which has a gravimetric deliverable capacity of 0.37 g g^{-1}. Among the structures that meet the gravimetric deliverable capacity ARPA-E target, the COF with the highest volumetric deliverable capacity is the imine COF ths-c3_11-02-04_04-03-04_04-03-04, which has a deliverable capacity of 141.1 vSTP/v and 0.50 g g^{-1}.

Whereas the methane molecules at low pressures are mainly located at the adsorption sites of the framework, they are more distributed over the pore volume at higher pressures. Therefore, the adsorption behavior is dominated by methane–framework interactions at low pressures and methane–methane interactions at high pressures, respectively. Thus, as evidenced in Fig. 8b, the deliverable capacity, *i.e.*, the difference between the methane uptake at high and low pressure, is an interplay between the two interaction types. When methane–framework interactions are less important than methane–methane interactions, the uptake at low pressures will be small and therefore, a higher deliverable capacity will be obtained. Contrarily, when the methane–framework interactions are more dominant, a lot of guests will remain bound to the framework at low pressures, decreasing the deliverable capacity. In Fig. S49 of the ESI,† the dependency of the methane uptake on the dimensionality of the framework is visualized. Despite that no COF in the database meets the ARPA-E target for the volumetric deliverable capacity, many of the here identified structures have volumetric deliverable capacities comparable with record-holding materials, such as MOF-5 (182 vSTP/v),[152] HKUST-1 (183 vSTP/v),[153] Co(bdp) (197 vSTP/v),[154] or NJU-Bai 43 (198 vSTP/v).[155] However, the hypothetical COF proposed by Mercado *et al.*[83] that consists of a carbon–carbon linked triazine framework and achieves a volumetric deliverable capacity of 216 vSTP/v, is not matched in this work, as we did not consider carbon–carbon linkages in our database. Several studies have proven that the targets imposed by the ARPA-E are too ambitious and that physical limits are preventing these objectives from being achieved by the current state-of-the-art materials.[51,156] These observations are valid also for the COFs in our ReDD-COFFEE database. However, upon functionalization, as discussed in Section 2.3, or by anchoring alkali metals to the framework,[157] an enhanced volumetric deliverable capacity could be obtained.

5 Conclusions

In this paper, the ReDD-COFFEE database of 268 687 COF structures and system-specific force fields is presented. ReDD-COFFEE is a Ready-to-use and Diverse Database of Covalent Organic Frameworks with system-specific Force field based Energy Evaluation. This database has been generated using an additive top-down approach, taking both geometric and energetic criteria into account. The structures with the lowest synthetic likelihood were filtered out using a deformation energy criterium. The combination of the structure and a system-specific force field for each COF makes the database ready-to-use in molecular simulations—for instance, for high-throughput screenings. We demonstrated the improved accuracy of the system-specific force fields over fully transferable force fields in predicting the equilibrium and single crystal geometries and experimental PXRD patterns, which provides confidence in the predictions made by our database. Furthermore, as our database is diverse in terms of geometric properties, linker cores, and linkages, it gives a representative picture of the COF material space as a whole—although we explicitly did not take a large variety of functional groups into account. Additional functionalization can be introduced to the frameworks *a posteriori*. Next to the large database, we also derived a diverse subset of 10 000 COFs with diversity metrics that are comparable or even higher than those of the full database, which allows for an efficient initial screening of the database.

We screened the textural properties of the COFs in the ReDD-COFFEE database and highlighted interesting property–property relations. Furthermore, our database is compared to databases of other material classes, *i.e.*, MOFs and zeolites, which showed that COFs possess the lowest mass densities among the studied materials. They combine high gravimetric accessible surface areas and pore fractions with low volumetric accessible surface areas. Finally, we screened the diverse subset for attractive COF storage materials for ANG in vehicular transport. Property–property relations were determined and the structural characteristics of the top candidates were selected. The highest observed volumetric deliverable capacity is comparable with the current record-holding materials. This study forms the basis for future, more elaborate screening studies focussing on gas storage and separation processes, such as carbon capture from flue gasses, as well as the mechanical stability of COFs, which will maximally benefit from the increased accuracy of the derived force fields. We hope our ReDD-COFFEE database—which is freely available on the Materials Cloud (**https://doi.org/10.24435/materialscloud:nw-3j**)—may also encourage other researchers to perform high-throughput simulations on these materials and further tap into the potential of functional COF materials.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

This work is supported by the Research Board of Ghent University (BOF) through a Concerted Research Action (GOA010-17). J. S. D. V. and S. M. J. R. acknowledge the Fund for Scientific Research-Flanders (FWO) for a strategic basic (SB) research fellowship

(grant no. 1S94521N) and a postdoctoral fellowship (grant no. 12T3522N), respectively. V. V. S. acknowledge the Research Board of Ghent University (BOF). The computational resources (Stevin Supercomputer Infrastructure) and services used in this work were provided by VSC (Flemish Supercomputer Center), funded by Ghent University, FWO, and the Flemish Government—department EWI. The authors wish to thank Massimo Bocus for the design of the ReDD-COFFEE logo.

Notes and references

1 A. P. Côté, A. I. Benin, N. W. Ockwig, M. O'Keeffe, A. J. Matzger and O. M. Yaghi, *Science*, 2005, **310**, 1166–1170.
2 H. M. El-Kaderi, J. R. Hunt, J. L. Mendoza-Cortés, A. P. Côté, R. E. Taylor, M. O'Keeffe and O. M. Yaghi, *Science*, 2007, **316**, 268–272.
3 X. Feng, X. Ding and D. Jiang, *Chem. Soc. Rev.*, 2012, **41**, 6010–6022.
4 S.-Y. Ding and W. Wang, *Chem. Soc. Rev.*, 2013, **42**, 548–568.
5 C. S. Diercks and O. M. Yaghi, *Science*, 2017, **355**, eaal1585.
6 M. S. Lohse and T. Bein, *Adv. Funct. Mater.*, 2018, **28**, 1705553.
7 Y. Zeng, R. Zou and Y. Zhao, *Adv. Mater.*, 2016, **28**, 2855–2873.
8 H. Furukawa and O. M. Yaghi, *J. Am. Chem. Soc.*, 2009, **131**, 8875–8883.
9 H. Wei, S. Chai, N. Hu, Z. Yang, L. Wei and L. Wang, *Chem. Commun.*, 2015, **51**, 12178–12181.
10 N. Huang, X. Chen, R. Krishna and D. Jiang, *Angew. Chem., Int. Ed.*, 2015, **54**, 2986–2990.
11 H. Ma, H. Ren, S. Meng, Z. Yan, H. Zhao, F. Sun and G. Zhu, *Chem. Commun.*, 2013, **49**, 9773–9775.
12 Z. Li, X. Feng, Y. Zou, Y. Zhang, H. Xia, X. Liu and Y. Mu, *Chem. Commun.*, 2014, **50**, 13825–13828.
13 Q. Sun, B. Aguila, J. Perman, L. D. Earl, C. W. Abney, Y. Cheng, H. Wei, N. Nguyen, L. Wojtas and S. Ma, *J. Am. Chem. Soc.*, 2017, **139**, 2786–2793.
14 N. Huang, L. Zhai, H. Xu and D. Jiang, *J. Am. Chem. Soc.*, 2017, **139**, 2428–2434.
15 S. He, T. Zeng, S. Wang, H. Niu and Y. Cai, *ACS Appl. Mater. Interfaces*, 2017, **9**, 2959–2965.
16 S.-Y. Ding, J. Gao, Q. Wang, Y. Zhang, W.-G. Song, C.-Y. Su and W. Wang, *J. Am. Chem. Soc.*, 2011, **133**, 19816–19822.
17 Q. Fang, S. Gu, J. Zheng, Z. Zhuang, S. Qiu and Y. Yan, *Angew. Chem., Int. Ed.*, 2014, **53**, 2878–2882.
18 H.-S. Xu, S.-Y. Ding, W.-K. An, H. Wu and W. Wang, *J. Am. Chem. Soc.*, 2016, **138**, 11489–11492.
19 V. S. Vyas, F. Haase, L. Stegbauer, G. Savasci, F. Podjaski, C. Ochsenfeld and B. V. Lotsch, *Nat. Commun.*, 2015, **6**, 8508.
20 S. Chandra, T. Kundu, S. Kandambeth, R. Babarao, Y. Marathe, S. M. Kunjir and R. Banerjee, *J. Am. Chem. Soc.*, 2014, **136**, 6570–6573.
21 C. R. DeBlase, K. E. Silberstein, T.-T. Truong, H. D. Abruña and W. R. Dichtel, *J. Am. Chem. Soc.*, 2013, **135**, 16821–16824.
22 M. Dogru and T. Bein, *Chem. Commun.*, 2014, **50**, 5531–5546.
23 S. Wan, J. Guo, J. Kim, H. Ihee and D. Jiang, *Angew. Chem., Int. Ed.*, 2008, **47**, 8826–8830.
24 X. Ding, J. Guo, X. Feng, Y. Honsho, J. Guo, S. Seki, P. Maitarad, A. Saeki, S. Nagase and D. Jiang, *Angew. Chem., Int. Ed.*, 2011, **50**, 1289–1293.
25 C. Zhang, S. Zhang, Y. Yan, F. Xia, A. Huang and Y. Xian, *ACS Appl. Mater. Interfaces*, 2017, **9**, 13415–13421.
26 Q. Fang, J. Wang, S. Gu, R. B. Kaspar, Z. Zhuang, J. Zheng, H. Guo, S. Qiu and Y. Yan, *J. Am. Chem. Soc.*, 2015, **137**, 8352–8355.
27 H.-C. Zhou and S. Kitagawa, *Chem. Soc. Rev.*, 2014, **43**, 5415–5418.
28 H. Furukawa, K. E. Cordova, M. O'Keeffe and O. M. Yaghi, *Science*, 2013, **341**, 1230444.
29 R. W. Tilford, S. J. Mugavero III, P. J. Pellechia and J. J. Lavigne, *Adv. Mater.*, 2008, **20**, 2741–2746.
30 R.-R. Liang, F.-Z. Cui, R.-H. A, Q.-Y. Qi and X. Zhao, *CCS Chem.*, 2020, **2**, 139–145.
31 X. Ding, X. Feng, A. Saeki, S. Seki, A. Nagai and D. Jiang, *Chem. Commun.*, 2012, **48**, 8952–8954.
32 S. Borgmans, S. M. J. Rogge, J. S. De Vos, P. Van Der Voort and V. Van Speybroeck, *Commun. Chem.*, 2023, **6**, 5.
33 O. M. Yaghi, M. O'Keeffe, N. W. Ockwig, H. K. Chae, M. Eddaoudi and J. Kim, *Nature*, 2003, **423**, 705–714.
34 M. Eddaoudi, D. B. Moler, H. Li, B. Chen, T. M. Reineke, M. O'Keeffe and O. M. Yaghi, *Acc. Chem. Res.*, 2001, **34**, 319–330.
35 A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder and K. A. Persson, *APL Mater.*, 2013, **1**, 011002.
36 Y. J. Colón and R. Q. Snurr, *Chem. Soc. Rev.*, 2014, **43**, 5735–5749.
37 S. M. Moosavi, A. Nandy, K. M. Jablonka, D. Ongari, J. P. Janet, P. G. Boyd, Y. Lee, B. Smit and H. J. Kulik, *Nat. Commun.*, 2020, **11**, 4068.
38 L. Vanduyfhuys, S. Vandenbrande, T. Verstraelen, R. Schmid, M. Waroquier and V. Van Speybroeck, *J. Comput. Chem.*, 2015, **36**, 1015–1027.
39 L. Vanduyfhuys, S. Vandenbrande, J. Wieme, M. Waroquier, T. Verstraelen and V. Van Speybroeck, *J. Comput. Chem.*, 2018, **39**, 999–1011.
40 P. Z. Moghadam, S. M. J. Rogge, A. Li, C.-M. Chow, J. Wieme, N. Moharrami, M. Aragones-Anglada, G. Conduit, D. A. Gomez-Gualdron, V. Van Speybroeck and D. Fairen-Jimenez, *Matter*, 2019, **1**, 219–234.
41 L. Vanduyfhuys, T. Verstraelen, M. Vandichel, M. Waroquier and V. Van Speybroeck, *J. Chem. Theory Comput.*, 2012, **8**, 3217–3231.
42 S. Bureekaew, S. Amirjalayer, M. Tafipolsky, C. Spickermann, T. K. Roy and R. Schmid, *Phys. Status Solidi B*, 2013, **250**, 1128–1141.
43 M. Tafipolsky, S. Amirjalayer and R. Schmid, *J. Comput. Chem.*, 2007, **28**, 1169–1176.
44 M. Tafipolsky and R. Schmid, *J. Phys. Chem. B*, 2009, **113**, 1341–1352.

45 M. Tafipolsky, S. Amirjalayer and R. Schmid, *J. Phys. Chem. C*, 2010, **114**, 14402–14409.
46 J. Wieme, L. Vanduyfhuys, S. M. J. Rogge, M. Waroquier and V. Van Speybroeck, *J. Phys. Chem. C*, 2016, **120**, 14934–14947.
47 S. M. J. Rogge, A. Bavykina, J. Hajek, H. Garcia, A. I. Olivos-Suarez, A. Sepúlveda-Escribano, A. Vimont, G. Clet, P. Bazin, F. Kapteijn, M. Daturi, E. V. Ramos-Fernandez, F. X. Llabrés i Xamena, V. Van Speybroeck and J. Gascon, *Chem. Soc. Rev.*, 2017, **46**, 3134–3184.
48 R. Schmid and M. Tafipolsky, *J. Am. Chem. Soc.*, 2008, **130**, 12600–12601.
49 S. Amirjalayer, R. Q. Snurr and R. Schmid, *J. Phys. Chem. C*, 2012, **116**, 4921–4929.
50 S. Borgmans, S. M. J. Rogge, J. S. De Vos, C. V. Stevens, P. Van Der Voort and V. Van Speybroeck, *Angew. Chem., Int. Ed.*, 2021, **60**, 8913–8922.
51 C. M. Simon, J. Kim, D. A. Gomez-Gualdron, J. S. Camp, Y. G. Chung, R. L. Martin, R. Mercado, M. W. Deem, D. Gunter, M. Haranczyk, D. S. Sholl, R. Q. Snurr and B. Smit, *Energy Environ. Sci.*, 2015, **8**, 1190–1199.
52 P. G. Boyd, A. Chidambaram, E. García-Díez, C. P. Ireland, T. D. Daff, R. Bounds, A. Gładysiak, P. Schouwink, S. M. Moosavi, M. M. Maroto-Valer, J. A. Reimer, J. A. R. Navarro, T. K. Woo, S. Garcia, K. C. Stylianou and B. Smit, *Nature*, 2019, **576**, 253–256.
53 C. E. Wilmer, M. Leaf, C. Y. Lee, O. K. Farha, B. G. Hauser, J. T. Hupp and R. Q. Snurr, *Nat. Chem.*, 2012, **4**, 83–89.
54 Y. G. Chung, J. Camp, M. Haranczyk, B. J. Sikora, W. Bury, V. Krungleviciute, T. Yildirim, O. K. Farha, D. S. Sholl and R. Q. Snurr, *Chem. Mater.*, 2014, **26**, 6185–6192.
55 Y. J. Colón, D. A. Gómez-Gualdrón and R. Q. Snurr, *Cryst. Growth Des.*, 2017, **17**, 5801–5810.
56 R. Anderson and D. A. Gómez-Gualdrón, *CrystEngComm*, 2019, **21**, 1653–1665.
57 J. Goldsmith, A. G. Wong-Foy, M. J. Cafarella and D. J. Siegel, *Chem. Mater.*, 2013, **25**, 3373–3382.
58 Y. J. Colón, D. Fairen-Jimenez, C. E. Wilmer and R. Q. Snurr, *J. Phys. Chem. C*, 2014, **118**, 5383–5389.
59 N. S. Bobbitt, J. Chen and R. Q. Snurr, *J. Phys. Chem. C*, 2016, **120**, 27328–27341.
60 D. A. Gómez-Gualdrón, Y. J. Colón, X. Zhang, T. C. Wang, Y.-S. Chen, J. T. Hupp, T. Yildirim, O. K. Farha, J. Zhang and R. Q. Snurr, *Energy Environ. Sci.*, 2016, **9**, 3279–3289.
61 S. Li, Y. G. Chung and R. Q. Snurr, *Langmuir*, 2016, **32**, 10368–10376.
62 G. Avci, S. Velioglu and S. Keskin, *ACS Appl. Mater. Interfaces*, 2018, **10**, 33693–33706.
63 Z. Qiao, C. Peng, J. Zhou and J. Jiang, *J. Mater. Chem. A*, 2016, **4**, 15904–15912.
64 C. Altintas, G. Avci, H. Daglar, A. N. V. Azar, I. Erucar, S. Velioglu and S. Keskin, *J. Mater. Chem. A*, 2019, **7**, 9593–9608.
65 C. Gu, J. Liu and D. S. Sholl, *J. Phys. Chem. C*, 2021, **125**, 20076–20086.
66 H. Daglar and S. Keskin, *Coord. Chem. Rev.*, 2020, **422**, 213470.
67 Y. He, E. D. Cubuk, M. D. Allendorf and E. J. Reed, *J. Phys. Chem. Lett.*, 2018, **9**, 4562–4569.
68 A. S. Rosen, S. M. Iyer, D. Ray, Z. Yao, A. Aspuru-Guzik, L. Gagliardi, J. M. Notestein and R. Q. Snurr, *Matter*, 2021, **4**, 1578–1597.
69 A. S. Rosen, V. Fung, P. Huck, C. T. O'Donnell, M. K. Horton, D. G. Truhlar, K. A. Persson, J. M. Notestein and R. Q. Snurr, *npj Comput. Mater.*, 2022, **8**, 112.
70 A. S. Rosen, J. M. Notestein and R. Q. Snurr, *J. Comput. Chem.*, 2019, **40**, 1305–1318.
71 F. H. Allen, *Acta Crystallogr., Sect. B: Struct. Sci.*, 2002, **58**, 380–388.
72 Y. G. Chung, E. Haldoupis, B. J. Bucior, M. Haranczyk, S. Lee, H. Zhang, K. D. Vogiatzis, M. Milisavljevic, S. Ling, J. S. Camp, B. Slater, J. I. Siepmann, D. S. Sholl and R. Q. Snurr, *J. Chem. Eng. Data*, 2019, **64**, 5985–5998.
73 P. Z. Moghadam, A. Li, S. B. Wiggin, A. Tao, A. G. P. Maloney, P. A. Wood, S. C. Ward and D. Fairen-Jimenez, *Chem. Mater.*, 2017, **29**, 2618–2625.
74 C. Mellot Draznieks, J. M. Newsam, A. M. Gorman, C. M. Freeman and G. Férey, *Angew. Chem., Int. Ed.*, 2000, **39**, 2270–2275.
75 M. Fernandez, P. G. Boyd, T. D. Daff, M. Z. Aghaji and T. K. Woo, *J. Phys. Chem. Lett.*, 2014, **5**, 3056–3060.
76 B. J. Sikora, R. Winnegar, D. M. Proserpio and R. Q. Snurr, *Microporous Mesoporous Mater.*, 2014, **186**, 207–213.
77 R. L. Martin and M. Haranczyk, *Cryst. Growth Des.*, 2014, **14**, 2431–2440.
78 M. A. Addicoat, D. E. Coupry and T. Heine, *J. Phys. Chem. A*, 2014, **118**, 9607–9614.
79 S. Bureekaew, V. Balwani, S. Amirjalayer and R. Schmid, *CrystEngComm*, 2015, **17**, 344–352.
80 P. G. Boyd and T. K. Woo, *CrystEngComm*, 2016, **18**, 3777–3792.
81 M. Tong, Y. Lan, Z. Qin and C. Zhong, *J. Phys. Chem. C*, 2018, **122**, 13009–13016.
82 R. L. Martin, C. M. Simon, B. Medasani, D. K. Britt, B. Smit and M. Haranczyk, *J. Phys. Chem. C*, 2014, **118**, 23790–23802.
83 R. Mercado, R.-S. Fu, A. V. Yakutovich, L. Talirz, M. Haranczyk and B. Smit, *Chem. Mater.*, 2018, **30**, 5069–5086.
84 M. Tong, W. Zhu, J. Li, Z. Long, S. Zhao, G. Chen and Y. Lan, *Chem. Commun.*, 2020, **56**, 6376–6379.
85 E. Gülçay and İ. F. Eruçar, *J. Turk. Chem. Soc., Sect. A*, 2020, **7**, 65–76.
86 D. Ongari, A. V. Yakutovich, L. Talirz and B. Smit, *ACS Cent. Sci.*, 2019, **5**, 1663–1675.
87 K. S. Deeg, D. Damasceno Borges, D. Ongari, N. Rampal, L. Talirz, A. V. Yakutovich, J. M. Huck and B. Smit, *ACS Appl. Mater. Interfaces*, 2020, **12**, 21559–21568.
88 O. F. Altundal, C. Altintas and S. Keskin, *J. Mater. Chem. A*, 2020, **8**, 14609–14623.
89 M. Tong, Y. Lan, Q. Yang and C. Zhong, *Chem. Eng. Sci.*, 2017, **168**, 456–464.
90 Y. Lan, M. Tong, Q. Yang and C. Zhong, *CrystEngComm*, 2017, **19**, 4920–4926.

91 W. Li, X. Xia and S. Li, *ACS Appl. Mater. Interfaces*, 2019, **12**, 3265–3273.
92 S. Kandambeth, A. Mallick, B. Lukose, M. V. Mane, T. Heine and R. Banerjee, *J. Am. Chem. Soc.*, 2012, **134**, 19524–19527.
93 P. Puthiaraj, Y.-R. Lee, S. Zhang and W.-S. Ahn, *J. Mater. Chem. A*, 2016, **4**, 16288–16311.
94 S.-Y. Ding, M. Dong, Y.-W. Wang, Y.-T. Chen, H.-Z. Wang, C.-Y. Su and W. Wang, *J. Am. Chem. Soc.*, 2016, **138**, 3031–3037.
95 R. Anderson and D. A. Gómez-Gualdrón, *Chem. Mater.*, 2020, **32**, 8106–8119.
96 M. W. Deem, R. Pophale, P. A. Cheeseman and D. J. Earl, *J. Phys. Chem. C*, 2009, **113**, 21353–21360.
97 *Database of zeolite structures*, **http://www.iza-structure.org/databases/**, Accessed: 2022-08-01.
98 M. O'Keeffe, M. A. Peskov, S. J. Ramsden and O. M. Yaghi, *Acc. Chem. Res.*, 2008, **41**, 1782–1789.
99 M. Li, D. Li, M. O'Keeffe and O. M. Yaghi, *Chem. Rev.*, 2014, **114**, 1343–1370.
100 F. J. Uribe-Romo, J. R. Hunt, H. Furukawa, C. Klöck, M. O'Keeffe and O. M. Yaghi, *J. Am. Chem. Soc.*, 2009, **131**, 4570–4571.
101 P. Kuhn, M. Antonietti and A. Thomas, *Angew. Chem., Int. Ed.*, 2008, **47**, 3450–3453.
102 F. J. Uribe-Romo, C. J. Doonan, H. Furukawa, K. Oisaki and O. M. Yaghi, *J. Am. Chem. Soc.*, 2011, **133**, 11478–11481.
103 S. Dalapati, S. Jin, J. Gao, Y. Xu, A. Nagai and D. Jiang, *J. Am. Chem. Soc.*, 2013, **135**, 17310–17313.
104 Q. Fang, Z. Zhuang, S. Gu, R. B. Kaspar, J. Zheng, J. Wang, S. Qiu and Y. Yan, *Nat. Commun.*, 2014, **5**, 4503.
105 J. R. Hunt, C. J. Doonan, J. D. LeVangie, A. P. Côté and O. M. Yaghi, *J. Am. Chem. Soc.*, 2008, **130**, 11872–11873.
106 D. A. Pyles, J. W. Crowe, L. A. Baldwin and P. L. McGrier, *ACS Macro Lett.*, 2016, **5**, 1055–1058.
107 K. T. Jackson, T. E. Reich and H. M. El-Kaderi, *Chem. Commun.*, 2012, **48**, 8823–8825.
108 W. Kabsch, *Acta Crystallogr., Sect. A: Cryst. Phys., Diffr., Theor. Gen. Crystallogr.*, 1976, **32**, 922–923.
109 J. P. Janet and H. J. Kulik, *J. Phys. Chem. A*, 2017, **121**, 8939–8954.
110 A. Nandy, C. Duan, J. P. Janet, S. Gugler and H. J. Kulik, *Ind. Eng. Chem. Res.*, 2018, **57**, 13973–13986.
111 J. P. Janet, F. Liu, A. Nandy, C. Duan, T. Yang, S. Lin and H. J. Kulik, *Inorg. Chem.*, 2019, **58**, 10592–10606.
112 A. Nandy, J. Zhu, J. P. Janet, C. Duan, R. B. Getman and H. J. Kulik, *ACS Catal.*, 2019, **9**, 8243–8255.
113 B. L. Bush, C. I. Bayly and T. A. Halgren, *J. Comput. Chem.*, 1999, **20**, 1495–1516.
114 M. D. Hanwell, D. E. Curtis, D. C. Lonie, T. Vandermeersch, E. Zurek and G. R. Hutchison, *J. Cheminf.*, 2012, **4**, 17.
115 M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery Jr, J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman and D. J. Fox, *Gaussian 16 Revision C.01*, Gaussian Inc. Wallingford CT, 2016.
116 A. D. Becke, *J. Chem. Phys.*, 1993, **98**, 5648–5652.
117 C. Lee, W. Yang and R. G. Parr, *Phys. Rev. B*, 1988, **37**, 785–789.
118 P. J. Stephens, F. J. Devlin, C. F. Chabalowski and M. J. Frisch, *J. Phys. Chem.*, 1994, **98**, 11623–11627.
119 S. Grimme, J. Antony, S. Ehrlich and H. Krieg, *J. Chem. Phys.*, 2010, **132**, 154104.
120 M. J. Frisch, J. A. Pople and J. S. Binkley, *J. Chem. Phys.*, 1984, **80**, 3265–3269.
121 N. L. Allinger, Y. H. Yuh and J. H. Lii, *J. Am. Chem. Soc.*, 1989, **111**, 8551–8566.
122 T. Verstraelen, S. Vandenbrande, F. Heidar-Zadeh, L. Vanduyfhuys, V. Van Speybroeck, M. Waroquier and P. W. Ayers, *J. Chem. Theory Comput.*, 2016, **12**, 3894–3912.
123 T. Verstraelen, P. Tecmer, F. Heidar-Zadeh, C. E. González-Espinoza, M. Chan, T. D. Kim, K. Boguslawski, S. Fias, S. Vandenbrande, D. Berrocal and P. W. Ayers, *HORTON*, 2017, **2**(1), **https://github.com/theochem/horton**.
124 A. K. Rappe, C. J. Casewit, K. S. Colwell, W. A. Goddard and W. M. Skiff, *J. Am. Chem. Soc.*, 1992, **114**, 10024–10035.
125 T. Verstraelen, L. Vanduyfhuys, S. Vandenbrande and S. M. J. Rogge, *Yaff, yet another force field (v1.6.0)*, **https://molmod.ugent.be/software/**.
126 A. Ghysels, T. Verstraelen, K. Hemelsoet, M. Waroquier and V. Van Speybroeck, *J. Chem. Inf. Model.*, 2010, **50**, 1736–1750.
127 T. F. Willems, C. H. Rycroft, M. Kazi, J. C. Meza and M. Haranczyk, *Microporous Mesoporous Mater.*, 2012, **149**, 134–141.
128 A. F. Ismail, K. C. Khulbe and T. Matsuura, in *Fundamentals of gas permeation through membranes*, Springer International Publishing, Cham, 2015, pp. 11–35.
129 S. Nosé, *Mol. Phys.*, 1984, **52**, 255–268.
130 W. G. Hoover, *Phys. Rev. A*, 1985, **31**, 1695–1697.
131 G. J. Martyna, M. L. Klein and M. Tuckerman, *J. Chem. Phys.*, 1992, **97**, 2635–2643.
132 G. J. Martyna, D. J. Tobias and M. L. Klein, *J. Chem. Phys.*, 1994, **101**, 4177–4189.
133 G. J. Martyna, M. E. Tuckerman, D. J. Tobias and M. L. Klein, *Mol. Phys.*, 1996, **87**, 1117–1157.
134 V. Favre-Nicolin and R. Černý, *J. Appl. Crystallogr.*, 2002, **35**, 734–743.
135 D. Frenkel and B. Smit, *Understanding molecular simulation: from algorithms to applications*, Elsevier, 2001, vol. 1.

136 D. Dubbeldam, S. Calero, D. E. Ellis and R. Q. Snurr, *Mol. Simul.*, 2016, **42**, 81–101.
137 M. G. Martin and J. I. Siepmann, *J. Phys. Chem. B*, 1998, **102**, 2569–2577.
138 Z. Yang, H. Chen, S. Wang, W. Guo, T. Wang, X. Suo, D.-e. Jiang, X. Zhu, I. Popovs and S. Dai, *J. Am. Chem. Soc.*, 2020, **142**, 6856–6860.
139 T. Ma, E. A. Kapustin, S. X. Yin, L. Liang, Z. Zhou, J. Niu, L.-H. Li, Y. Wang, J. Su, J. Li, X. Wang, W. D. Wang, W. Wang, J. Sun and O. M. Yaghi, *Science*, 2018, **361**, 48–52.
140 Y.-B. Zhang, J. Su, H. Furukawa, Y. Yun, F. Gándara, A. Duong, X. Zou and O. M. Yaghi, *J. Am. Chem. Soc.*, 2013, **135**, 16336–16339.
141 K. Geng, T. He, R. Liu, S. Dalapati, K. T. Tan, Z. Li, S. Tao, Y. Gong, Q. Jiang and D. Jiang, *Chem. Rev.*, 2020, **120**, 8814–8933.
142 S. Yeh, *Energy Policy*, 2007, **35**, 5865–5875.
143 M. Q. Wang and H. S. Huang, *A full fuel-cycle analysis of energy and emissions impacts of transportation fuels produced from natural gas*, 2000, **https://www.osti.gov/biblio/750803**.
144 J. L. Mendoza-Cortés, S. S. Han, H. Furukawa, O. M. Yaghi and W. A. Goddard, *J. Phys. Chem. A*, 2010, **114**, 10824–10833.
145 J. L. Mendoza-Cortes, T. A. Pascal and W. A. Goddard, *J. Phys. Chem. A*, 2011, **115**, 13852–13857.
146 M. I. Khan, T. Yasmin and A. Shakoor, *Renewable Sustainable Energy Rev.*, 2015, **51**, 785–797.
147 W. Lim, K. Choi and I. Moon, *Ind. Eng. Chem. Res.*, 2013, **52**, 3065–3088.
148 V. Menon and S. Komarneni, *J. Porous Mater.*, 1998, **5**, 43–58.
149 J. Wegrzyn and M. Gurevich, *Appl. Energy*, 1996, **55**, 71–83.
150 DE-FOA-0000672, *Methane Opportunities For Vehicular Energy (Move)*, **https://arpa-e-foa.energy.gov/Default.aspx?Search=DE-FOA-0000672**, Accessed: 2022-09-21.
151 S. L. Mayo, B. D. Olafson and W. A. Goddard, *J. Phys. Chem.*, 1990, **94**, 8897–8909.
152 J. A. Mason, M. Veenstra and J. R. Long, *Chem. Sci.*, 2014, **5**, 32–51.
153 J. Möllmer, A. Möller, F. Dreisbach, R. Gläser and R. Staudt, *Microporous Mesoporous Mater.*, 2011, **138**, 140–148.
154 J. A. Mason, J. Oktawiec, M. K. Taylor, M. R. Hudson, J. Rodriguez, J. E. Bachman, M. I. Gonzalez, A. Cervellino, A. Guagliardi, C. M. Brown, P. L. Llewellyn, N. Masciocchi and J. R. Long, *Nature*, 2015, **527**, 357–361.
155 M. Zhang, W. Zhou, T. Pham, K. A. Forrest, W. Liu, Y. He, H. Wu, T. Yildirim, B. Chen, B. Space, Y. Pan, M. J. Zaworotko and J. Bai, *Angew. Chem., Int. Ed.*, 2017, **56**, 11426–11430.
156 Y. Peng, V. Krungleviciute, I. Eryazici, J. T. Hupp, O. K. Farha and T. Yildirim, *J. Am. Chem. Soc.*, 2013, **135**, 11887–11894.
157 J. Lan, D. Cao and W. Wang, *Langmuir*, 2010, **26**, 220–226.