Distinguishing Metal–Organic Frameworks

Senja Barthel,*†‡ Eugeny V. Alexandrov,‡§ Davide M. Proserpio,‡∥ and Berend Smit‡

†Laboratory of Molecular Simulation, Institut des Sciences et Ingénierie Chimiques, Valais, Ecole Polytechnique Fédérale de Lausanne (EPFL), Rue de l’Industrie 17, CH-1951 Sion, Switzerland
‡Samara Center for Theoretical Material Science (SCTMS), Samara University, Moskovskoe shosse 34, 443086 Samara, Russian Federation
§Dipartimento State Technical University, Molodogvardeyskaya street 244, 443100 Samara, Russian Federation
∥Dipartimento di Chimica, Università degli Studi di Milano, Via Golgi 19, 20133 Milano, Italy

Supporting Information

ABSTRACT: We consider two metal–organic frameworks as identical if they share the same bond network respecting the atom types. An algorithm is presented that decides whether two metal–organic frameworks are the same. It is based on distinguishing structures by comparing a set of descriptors that is obtained from the bond network. We demonstrate our algorithm by analyzing the CoRe MOF database of DFT optimized structures with DDEC partial atomic charges using the program package ToposPro.

INTRODUCTION

A primary concern of materials science is the discovery of new materials and the prediction and understanding of their properties. With steadily increasing computer power, computational studies have become an inevitable tool for both analysis and prediction of materials. Large databases contain not only naturally occurring1,2 and synthesized materials but also thousands upon thousands of structures that are generated in silico.3−10 These databases provide the ground for computational studies, in particular screening studies to identify interesting materials for different applications.3,11−15 Less known is that these databases, as we will demonstrate below, can contain many variations of the same structure. Clearly, one would like to avoid spending valuable resources on studying similar structures but, more importantly, having an unspecified number of duplicated structures will make the statistics of any screening study unreliable. Therefore, developing a systematic methodology to identify whether two deposited structures are duplicates not only is an important fundamental question but also is of practical importance. This is in particular the case if the number of structures is so large that manual inspection is out of the question.

To illustrate our approach of comparing structures, we focus on a popular class of materials called metal–organic frameworks (MOFs).16 These are potentially porous 3D, 2D, and 1D crystalline materials, which consist of metal nodes connected by organic ligands.17−19 MOFs have gained much attention during the past decade due to their huge variety. By changing a metal type or substituting the functional group of an organic linker, one can in principle systematically change the properties of a known MOF. This not only makes MOFs and related nanoporous materials such as COFs, ZIFs, PPNs, etc. intriguing material classes for basic research but also suggests them for many potential applications, ranging from gas separation and storage to sensing and catalysis.

For complex compounds such as MOFs, we have to be careful how to define two materials as being equivalent, since similarities exist on different levels. For example, if two crystals do not have the same space groups or similar lattice parameters, they are considered as different materials from a strict crystallographic viewpoint and are listed as two separate entries in most databases. However, from a MOF point of view two structures are considered identical if they share the same bond network, with respect to the atom types and their embedding: i.e., if two structures can in principle be deformed into each other without breaking and forming bonds. We do not consider a particular MOF as a new material after, for example, rotating a ligand. However, such a small change can change the space group and hence can be reported as a new material in these databases. There exist several algorithms to compare crystals, but either they are restricted to structures with the same space groups or similar lattice parameters, or they evaluate the differences between atomic positions29,30 which is useful to detect small differences between crystals due to slightly different experimental conditions. However, while the traditional crystallographic approaches are important for solid-state chemistry, the unit cells of porous MOFs and related materials are much larger and are filled with solvents. This often causes substantial deviations of the crystal parameters for the activated evacuated material.
and the representatives with guest molecules, and a different method is required to compare MOFs.

Most synthesized MOFs are deposited in the Cambridge Structural Database (CSD). These materials often contain remaining solvent molecules, as do their structural files in the CSD. If a material is experimentally obtained under changed conditions, the remaining solvent molecules can differ, ligands can be differently aligned, and the unit cells can be distorted with respect to each other. All these versions of a material are stored independently in the CSD, and different versions of one MOF can have different chemical and physical behavior such as the narrow- and large-pore versions of the highly flexible MIL-53. However, from a fundamental point of view, one is often interested in understanding the properties of the underlying framework: i.e., the material without solvent molecules that are not believed to be part of the true framework. Before computational studies are performed, structures are usually “cleaned”: i.e., solvents are artificially removed and disorders often neglected. This leads to duplications in the resulting databases since many materials, in particular those on which considerable experimental efforts have been spent, are reported in numerous variations: the CSD contains for example more than 50 structures that all describe the famous CuBTC. Clearly, if the number of duplicates is this large, it will bias these databases.

Another postprocess that can cause multiple entries is relaxation: both experimentally known and hypothetical structures are often relaxed to obtain well-defined and energetically most stable representations of the materials before they are studied by simulations. Since it is impossible to ensure that an energetic minimum is global, it is possible that different relaxations find varying local minima that lead to multiple entries in a database.

In this article, we show how to systematically find topological duplicates in these material databases. We demonstrate how to compare frameworks of MOFs, but a small variation of the algorithm can also consider other classes of materials such as molecular crystals by considering the patterns of hydrogen bonds and van der Waals interactions. Similarly, it is possible to distinguish different versions of flexible MOFs by including van der Waals interactions in the bond network. In a representative study, we analyze a subset of the so-called “computationally ready” MOFs of experimentally known structures (CoRe MOF database): namely, the database of 502 frameworks (502 CoRe MOF database) that contains the structures of that are relaxed using density functional theory (DFT) and to which density derived electrostatic and chemical (DDEC) partial atomic charges are assigned. The files stored in the CoRe MOF database are mainly derived from the CSD by removing solvents and sometimes adding missing hydrogens. The results are of interest in their own right, since this database is frequently used for screening studies. Alternative databases of cleaned MOFs can be obtained by applying the MOF detection and the user-adopted solvent-removal algorithms that have been made available by the CSD. The prospective generation of databases of existing and new MOFs made the development of a tool for removing duplicates relevant and urgent.

The issue of duplication in databases is well-known. For example, the authors of the CoRe MOF databases already eliminated some duplicates: two cleaned CSD structures were considered equivalent if they share the numbers and type of atoms and if the root-mean-square deviation of the atomic positions of their Niggli cells is smaller than 0.1 Å. While this approach is intuitive, it is neither necessary nor sufficient to determine duplicates. Clearly, all duplicates have the same number of atoms and atom types. However, the atomic positions can vary largely between different representations of the same material. Indeed, we still find many duplicates in the CoRe MOF databases. The fundamental problem is, that allowing larger root-mean-square deviation does not address the problem of detecting duplicates correctly. Increasing the limit allows us to find more duplicates but also falsely identifies more nonidentical structures as duplicates.

We present a systematic and rigorous way of distinguishing structures that describe different materials, by introducing a set of descriptors that each give the same value for identical structures (invariants). Therefore, two structures with different descriptors are necessarily different. According to our notion of equivalence, atom types and atom numbers as well as all properties that are derived from the graph describing the bond network are invariants. We consider the following invariants: atom types, ligand graph, ligand coordination mode, and properties derived from the bond network and from several of its simplified versions, such as the dimensionality of the net, its topological indices, and possible interpenetration. In contrast to atomic positions, symmetries, cell parameters, volumes, or surface areas, our methodology is independent from distortion, which makes it very robust and reliable. Our set of invariants does not provide a complete invariant, meaning that there might exist different structures that cannot be distinguished by the set of descriptors. Such an example would be a pair of structures whose bond networks were practically indistinguishable by their topological indices (e.g., net topology, vertex symbol, point symbol), which is the case for stereoisomers. Excluding one couple of enantiomers, we have not come across an example in the 502 CoRe MOF database where our invariants wrongly identify two structures as identical.

All analyses have been performed using the software package ToposPro. We found that the 502 CoRe MOF database of 502 relaxed structures with DDEC partial atomic charges contains 48 structures with duplicates, some of them being reported several times, leading to 78 redundant entries. MOF-5 is the most often listed structure with 17 entries.

**SIMILARITIES OF REPORTED MATERIALS**

Given the large number of deposited structures, it is inevitable to use an algorithm to automatically detect similar structures and duplicates. The results of our representative study of the 502 CoRe MOF database, the CSD refcodes of each structure (with all bibliographic references), all chemical data, the analyses of the nets, and a list of all duplicates is given in the Supporting Information.

At present, we often rely on visual inspection to determine whether a newly reported crystal structure is similar to one of the existing materials, which is, given the ever-increasing number of reported MOF structures, close to impossible. Interestingly, even for a single pair of frameworks visual inspection might not be sufficient to determine with confidence whether they are identical or not. To illustrate this point, we consider three structures with the simple composition [Li(isonicotinate)], XUNGOD, XUNHAQ, and XUNGJ, which contain only C, H, Li, N, and O atomic species and which all form three-dimensional porous networks. The experimental structures contain different solvent molecules (morpholine, N-methylpyrrolidinone, and dimethylformamide, respectively) and have different space groups (P1, P21, and P21/n). They

**DOI:** 10.1021/acs.cgd.7b01663

**Cryst. Growth Des.** 2018, 18, 1738−1747
are reported in the same publication\textsuperscript{38} and are correctly stored as different structures in the CSD. Figure 1 shows a striking similarity, and one could easily conclude that the frameworks have the same topology.

Figure 1. XUNGOD (left) and XUNGUJ (right) in [100] projection (top) and [010] projection (bottom). The two frameworks have different topologies, as can be seen by simplifying the adjacency matrix.

Indeed, the authors assigned to all frameworks the same topological type \textit{sra} (with Li$_2$O$_2$ dimers as 4-c node), without naming it. (We use the RCSR three letter names\textsuperscript{29} for net topologies, when available, or else ToposPro TTD names.\textsuperscript{40}) However, the ligands of XUNGOD have a connectivity to the metals different from that of the ligands of XUNHAQ and XUNGUJ. All three frameworks contain infinite rod-shaped structural units aligned parallel to each other, which is clearly seen after removing dangling atoms (1-c vertices) and suppressing 2-coordinated atoms (2-c vertices). (Figure 2).

The analysis of the coordination sequence (CS) of atoms in the simplified net shows that XUNHAQ and XUNGUJ have the same CS for all atoms and share the net topology, while the CS of XUNGOD is different. For example, the CS of the O atom differs for the fifth coordination sphere (Figure 2). The frameworks of XUNHAQ and XUNGUJ are duplicates but are consequently different from XUNGOD. These subtleties cannot be found by visual inspection but are only detectable by a more sensitive graph analysis using the simplified adjacency matrix and topological indexes such as the coordination sequences (CS).

An example of two structures that have identical frameworks is the pair AMILUE and AMIMEP,\textsuperscript{41} two versions of [Zn$_4$(urotropin)$_3$(2,6-naphtalenedicarboxylato)$_4$]. They arise from a study of different framework–host interactions: AMIMEP contains guest ferrocene molecules that are not present in AMILUE. However, the frameworks (Figure 3) are too complicated to be reliably identified as identical by visual inspection, which is additionally hindered by the difference in the cell parameters and a shift of the unit cells.

Two identical frameworks of [Zn$_3$(bpdc)$_3$bpyp] (bpdc$^{2-}$ = biphenyldicarboxylate dianion, bpyp = 4,4′-bipyridine), which were originally reported as two different structures, are HEGJUZ\textsuperscript{42} and XUVHEB.\textsuperscript{43} The two publications do not refer to each other. This is not surprising, since HEGJUZ has space group $P2_1/n$ and some disorder on the solvated dimethylformamide, while XUVHEB has space group $Pbcn$ and no disorder on the solvate molecules but instead contains two additional uncoordinated waters (Figure 4).

Finally, we illustrate that an analysis of the net topology alone is also not sufficient in general to distinguish frameworks, since frameworks with different composition can share their net topologies. Clearly, substituting one atom type with another will change the structure but not the net. An example is IBICED\textsuperscript{44} (or its analogue IBIDAA\textsuperscript{44}), which differs from IBICAZ\textsuperscript{44} only by the type of halogen atom in the [Zn(Hal)(mpmab)] framework (Figure 5). A more complex reason for two different structures to share the same net can be that they are formed from enantiomeric ligands. An example is...
IBICON, which is the enantiomeric isomer of IBICED and IBIDAA. While IBICED and IBIDAA are constructed with the chiral L ligand and belong to the chiral space group $P6_1$, IBICON has space group $P6_5$ using the D ligand. Comparing the space groups of chiral structures (e.g., $P6_1$ and $P6_5$) will tell enantiomeric pairs apart, but this is a difficult task for frameworks taken from the CoRe MOF databases, since all relaxed structures are stored in the space group $P1$ and the original information on the space group is lost.

## METHODS

To automatically search for duplicates, we first compare the atom types of networks and the composition and the graph of linkers and subsequently analyze topological properties of the bond network and its simplifications as described below. This analysis is very robust in distinguishing networks of different topologies as well as in detecting skeleton isomers. In principle, it is also possible to find stereoisomers (enantiomers, cis/trans isomers, conformers) using information about crystal symmetry and geometrical fingerprints.45,46

The bond network of a structure is the graph whose vertices correspond to the atoms and whose edges correspond to interatomic bonds. A network, net, or graph is a particular combinatorial structure that consists of vertices and edges attached to the vertices. The degree of a vertex is the number of end points of edges connected to it. The degree of a vertex corresponds to the coordination of an atom. The bond network is equivalent to the adjacency matrix of a structure: i.e., the matrix that lists all atoms and the bonds between them. An underlying net of a structure is a simplified version of the bond network. It is constructed by adding a vertex for each structural group and connecting a pair of vertices with an edge if the corresponding structural groups have a bond between them.40,47 We perform three different simplifications on the bond network, which we further analyze. They are illustrated for MOF-5 in Figure 6a.

1. The simplified adjacency matrix is obtained by deleting isolated and dangling atoms and suppressing atoms that have only two
bonds. Every vertex of the underlying net with degree 1 is
removed together with its adjacent edge, and edges with an end
point of degree 2 are contracted iteratively until the minimal
degree of the graph is 3 (the resulting graph is independent of
the order in which the deletions and edge contractions are
performed) (Figure 6b).

(2) The standard simplification considers metal atoms and organic
ligands of a MOF as its structural units and substitutes the
atoms of each ligand by one dummy atom, usually placed at the
center of mass. In more general terms, anything that is not a
metal is contracted to its center of mass. That applies not only
to organic ligands but also to single nonmetal atoms, such as
oxygen, halogen, or multatomic noncoordinated species
(anion, cation, solvent) (Figure 6c).

(3) The motivation of the cluster simplification is to recognize
clusters of atoms by decomposing the structure into pieces with
high connectivity. For each bond, the smallest ring of bonds is
found that contains the bond. The ring sizes are sorted by
increasing values into the sequence \(a_1 \leq a_2 \leq \ldots \leq a_N\) where \(N\)
is the number of bonds in the structure. If the sequence contains a pair \(a_i, a_{i+1}\) such that \(a_i < a_{i+1} > 2\), bonds whose
smallest rings are formed with fewer than \(i + 1\) bonds are
considered to belong to a cluster while the other bonds connect
two clusters (Figure 6d). The cluster simplification for \(i\) is
obtained by substituting each cluster with a dummy atom and
keeping the bonds between clusters (Figure 6e). If there exist
several gaps in the sequence \(a_i\), the structure permits several
different cluster simplifications and one cluster simplification is
obtained for each index. Note that identical structures have
the same sets of cluster simplifications.

An unlimited number of simplifications can be performed on top of
each other, and it clearly matters in which order the simplifications are
performed. However, only finitely many nonidentical simplified nets of
a given structure can be obtained, as at some point it is impossible to
further simplify a net. To facilitate the analysis of the network topology,
we perform an adjacency matrix simplification on top of
both the standard simplification and the cluster simplification.
While the topology of the net obtained from simplifying the adjacency matrix
is often too specific to match one of the common three-letter
topologies, the net obtained by the cluster simplification is the most
simplified one and usually carries the topology that is commonly
assigned to a structure. For example, the topology of the net obtained
by simplifying the adjacency matrix of MOF-5 got its own name "MOF-5"
only because it is such a famous structure. However, one would usually
consider MOF-5 to be of primitive cubic topology "pcu", which indeed
is the topology of the net obtained by performing a cluster
simplification and subsequently simplifying the adjacency matrix.
Simplifying the adjacency matrix of the standard simplified MOF-5
yields a net with topological type "ff".

The standard and cluster descriptions coincide in many cases (239
from 488 structures in the 502 CoRe MOF database: 49%): namely, if
the structure building unit is a single metal atom and the ligand is not
branched, which prevents the underlying net from splitting into several
vertices with degree greater than 2. For example, both simplified
versions of \([\text{Cd}(-\text{isonicotinate})_2]\) AVAQIX50 have dia (diamond)
topology.

The topological type of a net is an invariant, as are (extended) point
and vertex symbols. These are weaker notions than the net topology,50
but the combination of the extended point symbol and the vertex symbol is in praxis able to distinguish different topologies. If a
net is not identified because it is not contained in the ToposPro
database of topologies, the point symbol and vertex symbol can still be
used to compare two structures. However, two nets with the same net
topology might have different structural building units. For example,
KAYBIX and KAYBU11 have the same composition C-CaH4NO4 and
their standard simplified nets both have 5,5T7 topology. However,
they are not duplicates since their ligands are isomers: pyridine-2,5-
dicarboxylate anion and pyridine-2,4-dicarboxylate anion, respectively.
Such a difference can be detected by comparing the graphs of ligands,
which were analyzed by computing the coordination modes of ligands
and metals following the approach of Serezhkin et al.32 A difference in
one of the obtained graph descriptors, namely the coordination mode
of the ligand (in brackets) and an identifier for their composition (in
braces), is sufficient to conclude that two structures are chemically
different. Examples are given in Table 2. Coordination isomers and
illustrated in Figures S11 and S12.

The topological type of a framework contains no information on
interpenetration, but ToposPro is able to determine the degree of
interpenetration. We add this check to our analysis and distinguish
differently interpenetrated versions of a structure. For example,
HIFTOG49 is a 2-fold interpenetrated version of MOF-5 (Figure 6f).
It is also possible to detect rare cases of entanglement isomers by
using the extended ring net.35

To analyze the 502 CoRe MOF database, we performed the steps
given below. They turned out to give a test, which is not only sufficient
but also necessary to distinguish MOFs up to enantiomers.

We did not compare the exact number of atoms, since the CoRe
MOF database contains structures given in multiples of the unit cell
(e.g., Figure 7a,b), but the ratios between elements and between
central atoms and ligands were determined. At each step, uniquely
determined structures were filtered out and sets of indistinguishable
structures compared during the following steps.

(1) composition (atom types and stoichiometry), i.e. empirical
formula
(2) central atom type: ligand graph, composition, and coordination
(3) topological type of the net obtained by standard simplification
(4) topological type of the net obtained by simplifying the
adjacency matrix
(5) topological type of the net obtained by cluster simplification
(6) degree of interpenetration

Clearly, the order of the steps can be interchanged. In particular, the
cost of computing the net type of a more complicated net competes
with the cost of highly simplifying a net. Therefore, interchanging
We investigated the 502 CoRe MOF database with 502 DFT relaxed structures with assigned DDEC partial atomic charges as an example. Of these, 488 were considered to be reliable for comparative analysis. While searching for duplicates, we performed some simple tests on the integrity of the 502 CoRe MOF database, such as searching for too short interatomic contacts and wrongly coordinated atoms. That flagged 66 entries with potential problems. Before we performed our analysis, we replaced in 46 structures erroneous atom coordinates by their positions before relaxation to maintain the net. We furthermore detected errors in 14 structures that were mainly caused by the removal of solvents that are structural building blocks or attached to the structure and chemically important or by the removal of charged anions without balancing the charges. In these cases, it is not surprising that the DFT optimization dramatically changes the network by breaking and rejoining valence bonds. We excluded 14 structures, for which hydrogenic (CISMAT01, CUNXIS, CUNX10, GHIHII, XUWVE), anions (AVEMOE, BICDAU, SENWAL, SENWIT, SENWOZ), or cations (VAHSHI, MODNIC) were missed or excess atoms were present (IJIROY, YIWMA). Among the removed structures is AVEMOE, from which a bridging coordinate sulfate anion was removed together with a terminal water ligand. As a result, the removed charge is not balanced and the DFT relaxed structure has not only a very different cell but even different multiples of the original one. The atomic charges are also incorrect for BICDAU, where terminal acetate ligands were excluded from the structure and thus could not be taken into account in the DFT calculations. Details and the list of problematic structures are given in the Supporting Information.

**RESULTS AND DISCUSSION**

We investigated the 502 CoRe MOF database with 502 DFT relaxed structures with assigned DDEC partial atomic charges as an example. Of these, 488 were considered to be reliable for comparative analysis. While searching for duplicates, we performed some simple tests on the integrity of the 502 CoRe MOF database, such as searching for too short interatomic contacts and wrongly coordinated atoms. That flagged 66 entries with potential problems. Before we performed our analysis, we replaced in 46 structures erroneous atom coordinates by their positions before relaxation to maintain the net. We furthermore detected errors in 14 structures that were mainly caused by the removal of solvents that are structural building blocks or attached to the structure and chemically important or by the removal of charged anions without balancing the charges. In these cases, it is not surprising that the DFT optimization dramatically changes the network by breaking and rejoining valence bonds. We excluded 14 structures, for which hydrogenic (CISMAT01, CUNXIS, CUNX10, GHIHII, XUWVE), anions (AVEMOE, BICDAU, SENWAL, SENWIT, SENWOZ), or cations (VAHSHI, MODNIC) were missed or excess atoms were present (IJIROY, YIWMA). Among the removed structures is AVEMOE, from which a bridging coordinate sulfate anion was removed together with a terminal water ligand. As a result, the removed charge is not balanced and the DFT relaxed structure has not only a very different cell but even different multiples of the original one. The atomic charges are also incorrect for BICDAU, where terminal acetate ligands were excluded from the structure and thus could not be taken into account in the DFT calculations. Details and the list of problematic structures are given in the Supporting Information.

**Duplicates.** As can be expected from the generation of the CoRe MOF database, most of its duplicates originate from structures in the CSD that differ only by their clathrate solvents. The CSD refcode of each structure, all chemical data, the results from the analyses of the nets, and a list of all duplicates are contained in the Supporting Information. In addition, a list of structures that should be removed to obtain a duplicate-free version of the database is given in the Supporting Information. Here, whenever one representative was correct and the other erroneous, the correct one is kept, and if all representatives were correct but were reported with different multiples of the unit cell, the representative with larger cell volume is removed. Examples are discussed below.

We followed the procedure outlined in Methods. In the first step, we found 325 materials uniquely determined by their composition and detected a further 163 structures distributed among 39 unique empirical formulas. We then examined the structures with the same empirical formula separately by comparing them in the next step. The second step found a further 28 uniquely determined structures from the 163, and the resulting 135 structures with duplicate ligand sets were distributed among 47 representatives.

### Table 1. Compounds with Isomeric Ligands

| refcode   | compound                              | Table 1. Compounds with Isomeric Ligands
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>MIMVEJ</td>
<td>[Zn(nicotinato)$_2$]</td>
<td></td>
</tr>
<tr>
<td>WIDZOA</td>
<td>[Cd(cucicato)$_3$] ($^3$-hydroxazyl-1,3,2-diylenedithyro-1-y1-yl)</td>
<td></td>
</tr>
<tr>
<td>UBACOR</td>
<td>$[Zn_2(1,1-‘biphenyl-2,2‘,6,6‘-tetrahydroxyanil)/(4,4‘-bipyridyl)]$</td>
<td></td>
</tr>
<tr>
<td>BERGAI</td>
<td>$[Zn_2(3-amino-1,2,4-triazolato)_(terephthalato)]$</td>
<td></td>
</tr>
<tr>
<td>KAYBIX</td>
<td>$[Ca(pyridine-2,5-dicarboxylato)]$</td>
<td></td>
</tr>
<tr>
<td>ESEVIH</td>
<td>$<a href="benzene-1,3,5-tricarboxylato">Zn_2(OH)</a>]$</td>
<td></td>
</tr>
</tbody>
</table>

**The differences are highlighted in **boldface.**

**Table 2. Compounds with the Same Stoichiometric Compositions and Ligands but Different Modes of Ligand Coordination**

<table>
<thead>
<tr>
<th>compound</th>
<th>refcode</th>
<th>ligand</th>
</tr>
</thead>
<tbody>
<tr>
<td>$[\text{Cd}_3(\mu_3-biphenyl-3,4,4‘-tricarboxylato)]_2$</td>
<td>HEKTUO</td>
<td>$C_{15}H_{10}O_{13}[G42]{196}$</td>
</tr>
<tr>
<td>$[\text{Y}(\text{benzene-1,3,5-tricarboxylato})_2]$</td>
<td>QEKILD, QEKILD01</td>
<td>$C_{19}H_{14}O_{14}[G51]{196}$</td>
</tr>
<tr>
<td>$[\text{Y}(\text{terephthalato})_2]$</td>
<td>SEHTEF</td>
<td>$C_{19}H_{12}O_{12}[G22]{158}$</td>
</tr>
<tr>
<td>$[\text{Y}(\text{pyridine-3,5-dicarboxylato})_2]$</td>
<td>LAGNOY</td>
<td>$C_{20}H_{12}O_{12}[K22]{78}$</td>
</tr>
<tr>
<td>$[\text{Y}(\text{pyridine-3,5-dicarboxylato})_2]$</td>
<td>LAGNEU</td>
<td>$C_{20}H_{12}O_{12}[K4]{78}$</td>
</tr>
<tr>
<td>$[\text{Y}(\text{pyridine-3,5-dicarboxylato})_2]$</td>
<td>SERJUV</td>
<td>$C_{18}H_{20}O_{12}[K22]{230}$</td>
</tr>
<tr>
<td>$[\text{Y}(\text{pyridine-3,5-dicarboxylato})_2]$</td>
<td>SERKAR</td>
<td>$C_{18}H_{20}O_{12}[K22]{230}$</td>
</tr>
</tbody>
</table>

For example, there are two types of $[\text{Cd}_3(\mu_6-biphenyl-3,4,4‘,5-tricarboxylato)]_2$ complexes, in which the hexadentate ligand is either coordinated in $G^{22}$ mode (HEKTU057) or in $G^{22}$ mode (QEKILD, QEKILD01) see Figure SII1). The difference in the coordination mode also leads to different underlying topologies of the standard simplified nets, 4,4,6T38 and 4,4,6T24, respectively. The original structures (in the CSD) differ in addition by terminal ligands, namely dimethylacetae- (HEKTU0) and dimethylformamide (QEKILD, QEKILD01), and water solvates contained in QEKILD and QEKILD01 but not in HEKTU0. Other examples are the three different clathrate structures of $[\text{Y}(\text{benzene-1,3,5-tricarboxylato})_2]$ that are distinguished by the coordination of the tricarboxylate: $G^{22}$ (SEHTEF, $C_{19}H_{12}O_{12}[G22]{158}$ dimethylformamide and dimethyl sulfoxide), $G^{22}$ (AVESLAY, $C_{18}H_{20}O_{12}[K22]{230}$ dimethylformamide), and $G^9$ (NADZEB, $C_{18}H_{20}O_{12}[K22]{230}$ dimethylformamide and water).
Table 3. Skeleton Isomers Revealed at the Third Step of the Analysis (Comparing the Topologies of the Nets Obtained from Standard Simplification)\textsuperscript{42}

<table>
<thead>
<tr>
<th>compound</th>
<th>refcode</th>
<th>net</th>
<th>isomeric net</th>
</tr>
</thead>
<tbody>
<tr>
<td>([\text{AlPO}_4])</td>
<td>LOFZUB</td>
<td>SAV</td>
<td>GOMRAC, GOMREG</td>
</tr>
<tr>
<td>([\text{Zn}(\text{imidazol-1,3-diyd})_2])</td>
<td>HIPVOI</td>
<td>df</td>
<td>GIJOP, VEJYUF01, VEJYUF02</td>
</tr>
<tr>
<td>([\text{Zn}(	ext{HCO}_3)_2])</td>
<td>KAVROQ</td>
<td>3,3,6,6,6T1S</td>
<td>RATDAS02, TESGOO, TESGUU</td>
</tr>
<tr>
<td>([\text{Cu}(3,4'-\text{biphenyl dicarboxylato})])</td>
<td>MOYYEF</td>
<td>4,4T69</td>
<td>MOYUI</td>
</tr>
<tr>
<td>([\text{Zn}(4'-(tetrazol-5-ylicarboxylato})])</td>
<td>WENDIE</td>
<td>4,4,4,4T59</td>
<td>FECWOB01</td>
</tr>
</tbody>
</table>

The seven unique structures are underlined.

Chart 1. Detecting Duplicates in the 502 CoRe MOF Database

(see Figure S12). The topologies of the underlying nets obtained by standard simplification are also different: namely 4-
}

c_sra, 6,6T2, and 6-c_htp, respectively. One more striking example is the pair SERJUV\textsuperscript{62} and SERKEG,\textsuperscript{62} which can be distinguished by the coordination modes of their ligands as well as by the topologies of their simplified adjacency matrices, while the topological type of their nets obtained from standard simplification is step for both.

Examination of the remaining 135 structures in step 3, i.e. comparison of their nets obtained from standard simplification, identifies an additional 7 structures as unique (see Table 3). The 128 structures so obtained with potential duplicates occur in 45 unique combinations of composition, ligand symbol, and topology of the net obtained from standard simplification.

Comparing the topological types of the nets obtained from the matrix simplification in step 4 detects two more unique structures (NUTQEZ and XUNGOD), and one quartet of \([\text{Ca}(4,4'-\text{biphenyl dicarboxylato})]\) structures (ZERQOE,\textsuperscript{63} KAXQOR,\textsuperscript{64} KAXQIL,\textsuperscript{64} KAXQOR01),\textsuperscript{65} originally containing different clathrates, is split into two pairs of isomers (ZERQOE-KAXQOR, KAXQIL-KAXQOR01). KAXQOR and KAXQOR01 are the only examples of real polymorphs in the 502 CoRe MOF database. Therefore, the number of possible duplicated structures reduces to 126 and the number of unique representatives increases to 46.

Comparing the topological types of the nets obtained from cluster simplification in step 5 does not distinguish any additional structures, which can be explained by this commonly used representation being the simplest notion of underlying nets.\textsuperscript{47} Even on analyzing the large CoRE MOF database with more than 4700 structures, we did not find any structures that step 5 distinguishes but which were not already differentiated by the previous steps. However, we include the cluster representation in our analysis for it captures the net topology that is usually used to classify and describe the topology of a MOF. Furthermore, the order in which the steps of the algorithm can be performed is a matter of choice as described in Methods.

Counting the degree of interpenetration in step 6 allows differentiation of four isomers. The \([\text{Zn}_2(2,2'-\text{biphenylene-5,5'-dicarboxylato})(4,4'-\text{bipyridyl})]\) framework is twice listed as 2-fold interpenetrated (GUYLOC, GUYMAP) and the 3-fold interpenetrated analogue is given two times as well (GUYLUI, GUYLUI01). The well-known MOF-5 framework of the composition \([\text{Zn}_4\text{O}(\text{benzene-1,4-dicarboxylato})_3]\) is found twice in its 2-fold interpenetrated version (HIFTOG, HIFTOG02), and is listed 17 times as single framework. Consequently, the number of possibly distinct structures in the previous list of 126 structures is now 46 + 2 = 48. The remaining 126 structures cannot be uniquely described by applying our set of invariants. Indeed, all of the indistinguishable structures have multiple entries: we find 39 pairs, 5 triples, 2 quadruples, one structure that is deposited 8 times, and MOF-5 with 17 entries. Most duplicates are caused by the removal of different clathrates/solvent molecules from the original structure. For example, KAXQOR\textsuperscript{64} and ZERQOE\textsuperscript{63} (Figure 7b) only differ in the CSD by the CO\textsubscript{2} adsorbed in ZERQOE. Similarly, WOWGET\textsuperscript{66} and GUUXLI\textsuperscript{67} are independently listed in the CSD only because they contain different numbers of clathrate water molecules in the pores of the framework \([\text{Al}_2\text{F}_2(\text{ethylene dichloridato})]\). Two structures, JAVNIE\textsuperscript{59} and FUSWIA,\textsuperscript{69} differ by water coordinated to the copper atoms of the framework \([\text{Cu}_2\text{Cl}_2(5-(4-pyridyl)-tetrazolato)]\), which is present in JAVNIE but absent in FUSWIA, as well as by the clathrate molecules dimethylformamide and methanol in FUSWIA and dimethylformamide and water in JAVNIE. More examples of duplicates caused by different solvent molecules are the pairs AMIMEP\textsuperscript{41} and AMILUE\textsuperscript{41} and HEGIJUZ\textsuperscript{32} and XUVHEB,\textsuperscript{33} which are discussed in Methods (Figures 3 and 4), SAKRED\textsuperscript{70} and SEFBOV\textsuperscript{71} and KAXQOR\textsuperscript{64} and ZERQOE\textsuperscript{65} (Figure 7a,b). However, the pair GOMRAC\textsuperscript{72} and GOMREG\textsuperscript{72} of AlPO\textsubscript{4} is a duplicate due to neglect of the metal disorder (Figure 7c). In both materials, a third of the aluminum sites are substituted, but while GOMRAC contains zinc, GOMREG contains manganese. Although the two materials are different, they are both stored with full occupation of aluminum in the 502 CoRe MOF database and must therefore be counted as duplicates.

In Chart 1 we summarize the structures that are distinguished during the six steps of our algorithm applied to...
the 502 CoRe MOF database, leaving 126 duplicates in 48 groups (UNIQUE structures: 325 + 28 + 7 + 2 + 48 = 410): 16% of 488 structures are redundant.

**Statistical Errors Caused by Multiple Entries in a Database.** We close with an example of the significance of cleaning databases from duplicates before drawing statistical conclusions. The following examination of interpenetration gives a simple example for a misleading statistical analysis caused by multiple entries: if we consider all 502 structures of the 502 CoRe MOF database, we find 58 2-fold, 16 3-fold, 9 4-fold, 3 5-fold, 1 6-fold, 3 7-fold, and 2 8-fold interpenetrated structures. However, if duplicates are removed, we find that there is only one 7-fold interpenetrated structure of [Zn(4-(2-(pyridin-4-yl) vinyl)benzoato)]$_2^-$; namely, UVARIT = UVAROZ there is only one 7-fold interpenetrated structure of [Zn(4-(2-

As an example, the 502 CoRe MOF database of 502 DFT relaxed MOFs with assigned DDEC partial atomic charges was investigated, showing that 15.5% (78) of the structures are redundant duplicates. A total of 9.2% (46) structural files contain incorrect atomic coordinates that affect the network topology and were replaced before the study, and 2.8% (14) structures have wrong framework compositions. In all, 502 − 78 − 14 = 364 structures are reliable, which is 72.5% of the database. The analysis was performed using ToposPro.

**ASSOCIATED CONTENT**

Supporting Information
The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.cgd.7b01663.

Results of our study of 488 structures from the 502 CoRe MOF database of DFT optimized structures with DDEC partial atomic charges, the CSD refcodes of each structure (with all bibliographic references), all chemical data, analyses of the nets (XLSX)

List of structures that need to be removed from the 502 CoRe MOF database to obtain a duplicate-free database (XLS)

List of the structures with problematic structural files, detailed information, and the replaced coordinates that were used for the analysis (XLSX)

Coordination spheres (PDF)

**AUTHOR INFORMATION**

Corresponding Author
*E-mail for S.B.: senja.barthel@epfl.ch.

ORCID
Senja Barthel: 0000-0002-9175-5067
Eugeny V. Alexandrov: 0000-0001-9229-8892
Davide M. Proserpio: 0000-0001-6597-9406
Berend Smit: 0000-0003-4653-8562

Notes
The authors declare no competing financial interest.

**ACKNOWLEDGMENTS**

S.B. thanks the National Center of Competence in Research (NCCR) "Materials Revolution: Computational Design and Discovery of Novel Materials (MARVEL)" of the Swiss National Science Foundation (SNSF). D.M.P. thanks the Russian Government (grant 14.B25.31.0005). E.V.A. is grateful to the Russian Science Foundation for financial support (Grant No. 16-13-10158). B.S. was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No 666983, MaGic).

**ABBREVIATIONS**

MOF, metal−organic framework; COF, covalent organic framework; ZIF, zeolitic imidazolate framework; PPN, porous polymer network; CSD, Cambridge Structural Database; CoRe MOF, computationally ready MOFs of experimentally known structures that are DFT relaxed with assigned DDEC partial atomic charges; CoRe MOF, computationally ready MOFs of experimentally known structures; DFT, density functional theory; DDEC, density derived electrostatic and chemical; CS, coordination sequence.
REFERENCES


