CRYSTAL GROWTH & DESIGN

Article

Cite This: Cryst. Growth Des. 2019, 19, 6682−6697

pubs.acs.org/crystal

Identification Schemes for Metal−Organic Frameworks To Enable Rapid Search and Cheminformatics Analysis

Benjamin J. Bucior,[†] Andrew S. Rosen,[†] Maciej Haranczyk,[‡] Zhenpeng Yao,[□] Michael E. Ziebel,[§] Omar K. Farha,[∥,†] Joseph T. Hupp,[∥] J. Ilja Siepmann,[⊥,#] Alán Aspuru-Guzik,[□,◆,△,▽] and Randall Q. Snurr*[,†]

[†]Department of Chemical and Biological Engineering and [∥]Department of Chemistry, Northwestern University, Evanston, Illinois 60208, United States

[‡]IMDEA Materials Institute, C/Eric Kandel 2, 2890 6 Getafe, Madrid, Spain

[□]Department of Chemistry and Chemical Biology, Harvard University, Cambridge, Massachusetts 02138, United States

[§]Department of Chemistry, University of California, Berkeley, California 94720, United States

[⊥]Department of Chemistry and Chemical Theory Center, University of Minnesota, 207 Pleasant Street SE, Minneapolis, Minnesota 55455, United States

[#]Department of Chemical Engineering and Materials Science, University of Minnesota, 421 Washington Avenue SE, Minneapolis, Minnesota 55455, United States

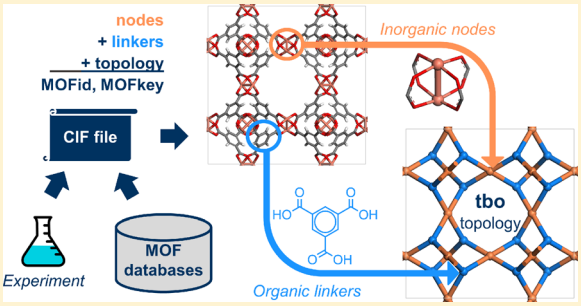
[◆]Department of Chemistry and Department of Computer Science, University of Toronto, Toronto, Ontario M5S 3H6, Canada

[△]Vector Institute for Artificial Intelligence, Toronto, Ontario M5S 1M1, Canada

[▽]Canadian Institute for Advanced Research (CIFAR) Senior Fellow, Toronto, Ontario M5S 1M1, Canada

Supporting Information

ABSTRACT: The modular nature of metal−organic frameworks (MOFs) leads to a very large number of possible structures. High-throughput computational screening has led to a rapid increase in property data that has enabled several potential applications for MOFs, including gas storage, separations, catalysis, and other fields. Despite their rich chemistry, MOFs are typically named using an ad hoc approach, which can impede their searchability and the discovery of broad insights. In this article, we develop two systematic MOF identifiers, coined MOFid and MOFkey, and algorithms for deconstructing MOFs into their building blocks and underlying topological network. We review existing cheminformatics formats for small molecules and address the challenges of adapting them to periodic crystal structures. Our algorithms are distributed as open-source software, and we apply them here to extract insights from several MOF databases. Through the process of designing MOFid and MOFkey, we provide a perspective on opportunities for the community to facilitate data reuse, improve searchability, and rapidly apply cheminformatics analyses.



INTRODUCTION

Metal−organic frameworks (MOFs) are a class of nanoporous materials with well-defined pore shape and chemistry. Their modular construction via the self-assembly of inorganic "nodes" (metal ions or small metal oxide clusters) and organic "linkers" in different framework topologies leads to a combinatorial design space. In principle, by judicious selection of nodes and linkers (including defects, such as missing nodes and/or linkers), one can design a MOF that is well-suited for an application of interest (e.g., catalysis, separation).[1−3] The challenge is identifying the ideal and viable combination of MOF building blocks and their configuration from this near-unlimited design space. Some of us have previously been involved in the development of databases for the MOF community to manage and utilize this complexity, such as the Computation-Ready, Experimental (CoRE) MOF databases.[4−7] Other efforts, such as the NIST/ARPA-E Database of Novel and Emerging Adsorbent Materials,[8] have also demonstrated the tremendous potential of reusing data to unlock new insights, such as running meta-analyses to examine the reproducibility of adsorption isotherms.[9] More generally, best practices in data stewardship can be characterized by four key attributes: findability, accessibility, interoperability, and reusability, which have collectively been coined the FAIR Guiding Principles.[10] In this work, we will highlight current

Received: August 6, 2019
Published: September 13, 2019

DOI: 10.1021/acs.cgd.9b01050
Cryst. Growth Des. 2019, 19, 6682−6697

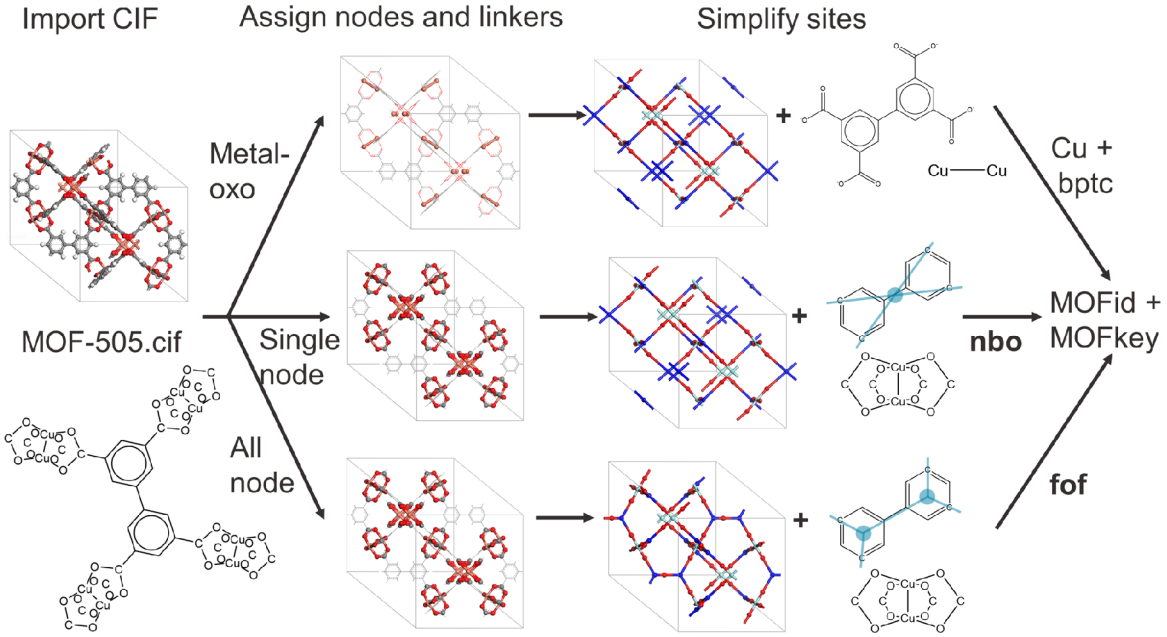


Figure 1. A scheme sketching the algorithms for deconstructing MOF-505[52]/NOTT-100[53] into its basic building blocks, which provide information about the MOF chemistry and underlying topological net. The "metal-oxo" algorithm describes the MOF chemistry, "single node" provides the underlying topology, and "all node" provides an alternative topological representation, if available. The light blue circles indicate the vertices in the simplified net: there is a single four-coordinated vertex in the "single node" topology and two three-coordinated branch points in the "all node" topology. In this example, the MOF contains a copper paddlewheel node, 3,3′,5,5′-biphenyltetracarboxylate (bptc) linker, and topological nets represented by the RCSR symbols **nbo** and **fof**.

limitations in MOF metadata and present a new scheme to facilitate value-added analysis using existing data.

Some of the largest barriers for materials informatics currently include the "diversity of research areas within materials, lack of data standards, and missing incentives for sharing."[11] Unfortunately, the MOF literature also faces similar challenges. MOFs have an ad hoc system of naming, typically numbered (e.g., MOF-5) and/or derived from the university of origin (e.g., NU-1000). As such, multiple names can refer to the same MOF. For example, the names Cu-BTC (BTC = benzene-1,3,5-tricarboxylic acid), $Cu_3(BTC)_2$, $Cu_2(BTC)_3$, HKUST-1 (HKUST = Hong Kong University of Science and Technology), and MOF-199 all refer to the same material. M-MOF-74 is another common MOF family, which for a given metal (M) has notations including M-MOF-74, M-CPO-27 (CPO = Coordination Polymer of Oslo), M_2(DOBDC) (DOBDC = 2,5-dioxido-1,4-benzenedicarboxylate), M_2(DHBDC) (DHBDC = 2,5-dihydroxybenzenedicarboxylate), M_2(DHTP) (DHTP = 2,5-dihydroxyterephthalate), M_2(DOT) (DOT = 2,5-dioxidoterephthalate), and variants on these notations with different punctuation. The current difficulty in MOF discoverability also raises the possibility of naming two distinct MOFs using the same identifier, such as two unrelated structures[12,13] named MOF-48 a decade apart.

Besides the benefits in nomenclature, a systematic MOF identification scheme would also facilitate data mining efforts. In a recent paper, text mining algorithms were used to extract data about MOF pore volumes and surface areas from the research literature.[14] In this work, the authors employed a six-step process to classify whether a given string of text was likely to represent the name of a MOF. By contrast, if a MOF identification standard were adopted by the research community, users could search for a specific keyword to rapidly identify publications or other data sources containing information for a specific MOF. MOFs with the same molecular formula but different connectivity, such as families of zeolitic imidazolate frameworks (ZIFs), could also be differentiated by their topologies using the identifier.

Ideally, standards for systematically naming MOFs could be established like those for small organic molecules. We draw inspiration from the field of cheminformatics, which has had many promising advances for small molecules.[15−20] Common cheminformatics representations are transferable across different platforms for chemical search and analysis, including Reaxys,[21] ChemDraw,[22] PubChem,[23] and chemical supplier Web sites. Historically, cheminformatics developments have been primarily focused on small organic molecules instead of polymers, metal-containing compounds, or framework topologies. One challenge is that MOFs combine all three of these domains. More broadly, the MOF and coordination polymer community is also still navigating ambiguous definitions for the field as a whole, let alone how to describe individual crystal structures.[24] The IUPAC task group on "Coordination Polymers and Metal−Organic Frameworks: Terminology and Nomenclature Guidelines" released a report[25] encouraging researchers to continue describing framework topology using symbols from the Reticular Chemistry Structure Resource (RCSR) database.[26] This report led to a follow-up IUPAC task group on topological representations,[27] which will be discussed more thoroughly in the Methods. There has been some related progress developing IUPAC conventions and computer formats for polymers, which represent their chemistry by the monomer(s) or constitutional repeat unit(s).[28−32] These polymer methods have not been formally extended to MOFs.

Ultimately, we would like more powerful methods to integrate multiple sources of data together, to easily search for MOFs, and to better manage the vast amounts of data being generated by the MOF community. The Cambridge

DOI: 10.1021/acs.cgd.9b01050
Cryst. Growth Des. 2019, 19, 6682−6697

Structural Database (CSD) is a great success story of data sharing. At the time of writing, this database includes over one million crystal structures[33,34] and has built a culture where authors in the MOF community publicly deposit crystal structure data prior to the publication of a new material (increasingly with a journal mandate). Similarly, the RCSR[26] is a database for framework topologies that has been broadly adopted by the community to compactly communicate the connectivity and overall framework structure for MOFs and related crystalline, periodic materials. Researchers can currently search the CSD for a specific crystal substructure[35,36] or the RCSR for a given framework topology[26,37−39] as independent queries, but there is an opportunity to consider these two properties in a single resource.

The primary objective of this work is to expand the capabilities for MOF data reuse. Much like the successes of the RCSR for topologies and the CSD for crystal structures, new cheminformatics conventions for MOFs could similarly transform the research capabilities for finding and labeling MOFs in the literature. The proposed standards could supplement the current system of human-readable colloquial names (e.g., Cu-BTC) with machine-readable identification schemes (e.g., `Cu.QMKYBPDZANOJGF.MOFkey-v1.tbo`) to improve MOF search, data management, and discovery of structure−property relationships. In this work, we present the development of two complementary MOF representation schemes, their validation, and their utility for analysis and insights through a few examples. One is a compact representation of a MOF's composition and topology built upon InChIKeys;[40] the other is a more verbose format providing additional information on bond connectivity that builds upon SMILES strings.[41,42] In both cases, we deconstruct a MOF into its individual building blocks and represent them using these modified cheminformatics formats. After discussing the details of the proposed format and our underlying software implementation, we review some challenges and limitations in our canonical MOF identifiers. We test the code using "known" MOF structures and demonstrate applications in database statistics, duplicate detection, and identification of polymorphic families. Finally, we comment on the progress we have made to facilitate analysis between MOF databases and future opportunities in the field.

METHODS

Topological Deconstruction Algorithm. The overall premise of the proposed MOF identification scheme is to combine information about the chemistry of MOF building blocks and information on how they assemble topologically. Given the complexity of deconstructing MOFs into their building blocks, we have designed a workflow that combines the results from three different algorithms to assign our proposed MOF identifiers. The general steps in our approach are depicted in Figure 1. We have developed a new "metal-oxo" algorithm focused on MOF chemistry, which keeps the organic linkers intact as discrete building blocks (including any carboxylate groups). From a topological perspective, the published "single node" and "all node" algorithms[27,43,44] provide a better representation of the MOF connectivity. Each algorithm provides different information about the MOF, thus leading to our hybrid approach. In this section, we describe the differences between these methods and highlight related work in the literature.

We start with a brief review of MOF deconstruction concepts from the literature and then discuss the approach we used in this work. Multiple methods for decomposing MOFs into their building blocks have been developed, though they have not yet been applied for the purpose of a MOF identification system. Coupry et al. assigned MOF building blocks for force field assessment by classifying the atoms into three main classes: metal atoms, adjacent oxygen atoms bonded to metals within the SBUs, and all other atoms, which were assumed to represent the linkers.[45] Another approach is to enumerate possible MOF templates, such as single metal ions bonded to linkers or carboxylate linkers binding to metal nodes.[46] Perhaps the simplest algorithm is disconnecting any bonds to metal atoms and leaving the rest of the molecular graph intact. This algorithm is called the "standard simplification" algorithm[47] and has the same approach as IUPAC's International Chemical Identifier (InChI) standard.[48] A more complex "cluster simplification" algorithm in ToposPro breaks apart crystal structures by examining the minimal ring sizes for each bond to differentiate between *intra-* and *inter*cluster bonds.[37,38] Other algorithms, which take into consideration the linker shape, will be described in the "single node" and "all node" methods below.

In this work, we developed a new "metal-oxo" simplification algorithm to describe MOF chemistry by dividing MOF structures into distinct inorganic and organic building blocks. First, we assign the bond adjacency matrix from the crystal structure using a simple distance cutoff method implemented in Open Babel,[49] which uses the CSD covalent radii to determine expected bond distances.[50] We adopt the convention from InChI for classifying elements as metals versus nonmetals.[40] For the "metal-oxo" algorithm, we generally define the inorganic building blocks as metal-oxo clusters, including oxides and bound hydroxide, peroxide, and water species (e.g., the $Zr_6(\mu_3\text{-}O)_4(\mu_3\text{-}OH)_4(OH)_4(OH_2)_4$ node of NU-1000[51] or the Zn_4O cluster of MOF-5). Remaining fragments are classified as organic building blocks, approximately described as any larger nonmetal cluster. Thus, a hydroxyl group bound to a metal atom would be considered as part of the inorganic cluster, but a methoxy group in the same position would instead be treated as an organic building block and may be classified as a bound solvent molecule with a crystallographically invisible hydrogen atom for charge neutrality. Carboxylate functional groups, including the oxygen atoms, are considered as part of the organic building blocks, because they are covalently bonded to the rest of the linker molecule.

Once we have determined the MOF chemistry, we also need to assign a topological net, which describes the underlying connectivity of the MOF building blocks. Describing the MOF topology is often more apparent using a shape- or connectivity-based building block convention as opposed to the chemistry-based "metal-oxo" algorithm. We represent the building blocks as secondary building units (SBUs),[54,55] which are characterized by their points of extension[56] connecting to other building blocks in the topological net. By using SBUs in the "single node" and "all node" algorithms, we can consider the inorganic "nodes" and organic "linkers" as abstract shapes (polygons and polyhedra) linked together in the simplified net.[43,54] Unlike the "metal-oxo" algorithm, the other two approaches generally consider coordinated carboxylates (and certain heteroaromatic rings) as part of the "node" (see Figure 1 and Section S1.2). This convention can be important in cases such as MFU-4*l*, which is represented as a Kuratowski-type pentametallic SBU instead of five discrete metal atoms (see Figure S1).[57]

After the MOF atoms have been assigned as "nodes" and "linkers," we simplify each cluster by replacing it with a single pseudoatom at its centroid (geometric center). Additional simplification steps, such as solvent detection, checking for interpenetrated nets, and handling infinite rod-like SBUs, are detailed in Section S1.2.

We adopt two standard conventions for reporting the topology, in accordance with prior literature and recommendations from an IUPAC task group on the subject.[27,43,44] Most importantly, we determine the basic net from the "single node" simplification algorithm, which considers each SBU and polytopic linker as a single vertex in the topological net. We also run a topological simplification using the "all node" approach, which explicitly identifies branch points within the linker. As shown by the MOF-505 example in Figure 1, this algorithm can provide additional information about the underlying shape of the MOF building blocks. However, there can be ambiguity in assigning the locations of branch points, so we recommend reporting the parent "single node" topology as well.

DOI: 10.1021/acs.cgd.9b01050
Cryst. Growth Des. 2019, 19, 6682−6697

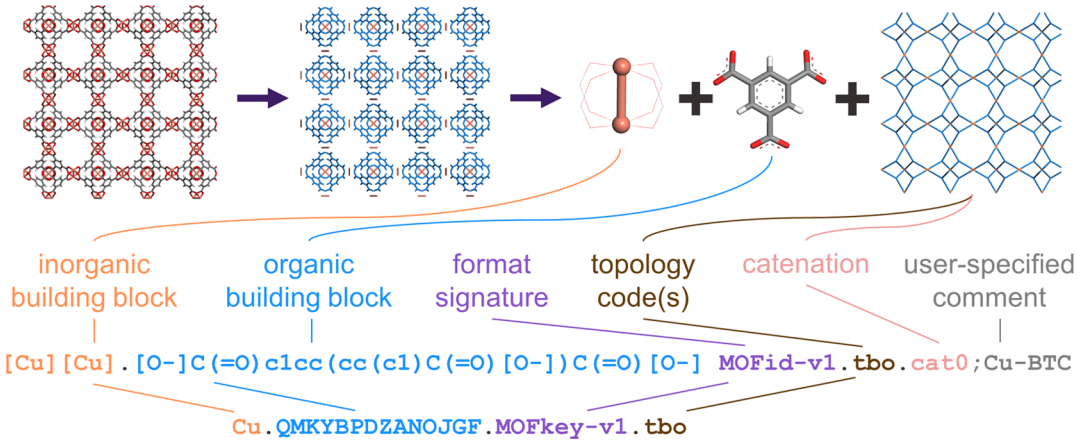


Figure 2. Example of identifying the Cu-BTC MOF using the SMILES-derived MOFid format and InChIKey-based MOFkey format.

We describe the simplified topological net by using the abbreviations tabulated in the online RCSR database.[26] Our analysis code exports a crystal geometry data (.cgd) file containing the vertices in the simplified net and their periodic connectivity. Then, the open-source program Systre[58] is invoked as a command-line Java program to assign the RCSR topology. Briefly, Systre detects the underlying topological net by calculating a fingerprint called the Systre Key, which describes how vertices connect in the labeled quotient graph.[59] We exclusively use topology codes from the RCSR database in this work, using the archive published on June 1, 2019. The use of RCSR codes is the only part of our identifier schemes that relies on a central registry. In a future iteration of the schemes, it may be possible to decouple the topological description from the RCSR (i.e., to describe MOFs having non-RCSR or new topologies) by hashing the Systre Key into a somewhat compact fingerprint.[59]

Adapting Cheminformatics Formats for MOFs. In this section, we motivate the selection and modification of two cheminformatics formats for our MOF identifier schemes. Once a MOF is deconstructed into its representative building blocks and underlying topological net, the chemical structure needs to be represented in a standardized format. One of the objectives of cheminformatics is compactly representing the chemical structures of molecules.[60] A popular format is the Simplified Molecular Input Line Entry System (SMILES), which compactly encodes a molecular graph as a string of bonded elements and how they branch.[41,42] Some work has extended SMILES for supramolecular structures and nanodevices, although not specifically applied to MOFs.[61] A major disadvantage of SMILES is that multiple SMILES strings can refer to the same compound: the assigned SMILES can vary by implementation and sometimes even the order of atoms in the molecular graph. Multiple canonicalization algorithms have been proposed, though different cheminformatics toolkits (and sometimes versions) will not produce the same canonical SMILES.[30,49,62] In the past decade, the InChI format[40,48] has also been widely adopted due to its standardized, open-source implementation.[48,63] This format is more similar to a barcode of the molecular graph and has a compact hashed version called the InChIKey.[64] Even though InChI is a standardized format, metals are represented as nonbonded atoms, so additional flags or nonstandard implementations[65] would be required to describe the molecular graph of MOF nodes. A format called Universal SMILES combines the readability of SMILES with the canonical ordering from InChI.[66] Universal SMILES may be a good approach for future descriptions of MOFs, but we did not use it in this work due to the complicated, unusual bond valence within MOF nodes. In addition, one of us recently introduced SELFIES[67] as a novel semantic graph representation tool that has been shown to be more robust than SMILES, especially for purposes of machine learning. A possible extension of MOFid would be creating a SELFIES variant.

Even though cheminformatics methods are well-established for small organic molecules, they cannot be used without significant modifications to adequately describe MOFs. As shown in Section S3, the periodic, crystalline structure of MOFs has undesirable effects on traditional cheminformatics formats, such as long, uninterpretable strings and a dependence on the number of unit cells. However, the SMILES and InChIKey formats work well once the MOF is split up into its building blocks. Decomposing MOF structures into their nodes and linkers enables compatibility with existing cheminformatics software (e.g., ChemDraw[22]) and enables rapid analysis based on the individual building blocks. This approach also parallels current MOF names like M_2(DOT), which describe MOFs based on their building blocks (though these names contain nonstandard abbreviations, special characters, and generally no information on the overall framework topology).

The SMILES format has only seen limited use in the literature for MOFs and other metal−organic compounds. SMILES strings have primarily been used in the MOF literature to enumerate linker structures, such as automating the design of flexible linkers.[68] Searches through chemical databases, such as PubChem,[23] can identify compatible linker molecules for constructing new MOFs in silico, which has enabled the systematic analysis of MOFs for methane storage[69,70] and high surface areas,[71] and the design of MOF-74 analogues.[72] Some reports on organometallics and inorganic complexes have analyzed the metal core and organic ligands separately, similar to our MOF decomposition strategy. Using this "divide-and-conquer" approach, they can more successfully generate 3D geometries and machine learning descriptors for these structures.[73−76] Parallel efforts using SMILES to describe chemical connectivity are being undertaken by the Crystallography Open Database to improve searchability and facilitate the identification of structure−property relationships.[77] In their work, the database authors specifically focused on smaller species like metallocenes and explicitly excluded large polymeric structures like MOFs. There is also increasing interest and a request for proposal from the InChI trust for developing new capabilities for organometallics,[78] so this field remains an active area of research. We have constructed MOF identifier schemes derived from the SMILES and InChIKey formats; their specifications and anticipated use cases are detailed in the next section.

MOF Identifier Formats. We propose two related identifiers to represent MOFs, adapting existing infrastructure for cheminformatics and topological representations. The MOFid format is derived from SMILES and provides detailed chemical information about the MOF building blocks as well as metadata about the overall simplified topology. The MOFkey format is derived from the InChIKey and serves as a canonical identifier for a MOF's linkers, identity of the metal(s), and overall framework topology. Generally, the MOFkey would be useful as a compact identifier and barcode for MOFs (e.g., in a research article and/or abstract), whereas the MOFid trades brevity and searchability for reversibility between the identifier and

DOI: 10.1021/acs.cgd.9b01050
Cryst. Growth Des. 2019, 19, 6682−6697

Table 1. MOFid and MOFkey Identifiers for Common MOFs in the Literature[a]

common name(s)	MOFid	MOFkey	duplicates
Cu-BTC, HKUST-1, MOF-199	[Cu][Cu].[O-]C(=O)c1cc(cc(c1)C(=O)[O-])C(=O)[O-] MOFid-v1.tbo.cat0;	Cu.QMKYBPDZANOJGF.MOFkey-v1.tbo	84
IRMOF-1, MOF-5	[O-]C(=O)c1ccc(cc1)C(=O)[O-].[Zn][O]([Zn])([Zn])[Zn] MOFid-v1.pcu.cat0;	Zn.KKEYFWRCBNTPAC.MOFkey-v1.pcu	59
MIL-47	[O-]C(=O)c1ccc(cc1)C(=O)[O-].[O].[V] MOFid-v1.rna.cat0;	V.KKEYFWRCBNTPAC.MOFkey-v1.rna	26
ZIF-8	CC1=NC=C[N]1.[Zn] MOFid-v1.sod.cat0;	Zn.YFFQUDCLMWOYCW.MOFkey-v1.sod	26
UiO-67	[O-]C(=O)c1ccc(cc1)c1ccc(cc1)C(=O)[O-].[O]12[Zr]34[OH]5[Zr]62[OH]2[Zr]71[OH]4[Zr]14[O]3[Zr]35[O]6[Zr]2([O]71)[OH]43 MOFid-v1.fcu.cat0;	Zr.NEQFBGHQPUXOFH.MOFkey-v1.fcu	11
MIL-53	[Cr].[O-]C(=O)c1ccc(cc1)C(=O)[O-].[OH] MOFid-v1.rna.cat0;	Cr.KKEYFWRCBNTPAC.MOFkey-v1.rna	8
UiO-66	[O-]C(=O)c1ccc(cc1)C(=O)[O-].[O]12[Zr]34[OH]5[Zr]62[OH]2[Zr]71[OH]4[Zr]14[O]3[Zr]35[O]6[Zr]2([O]71)[OH]43 MOFid-v1.fcu.cat0;	Zr.KKEYFWRCBNTPAC.MOFkey-v1.fcu	8
IRMOF-10	[O-]C(=O)c1ccc(cc1)c1ccc(cc1)C(=O)[O-].[Zn][O]([Zn])([Zn])[Zn] MOFid-v1.pcu.cat0;	Zn.NEQFBGHQPUXOFH.MOFkey-v1.pcu	6
NU-1000	O[Zr]123([OH2])[OH]4[Zr]56[O]3[Zr]37([OH]2[Zr]28[O]1[Zr]14([O]6[Zr]([OH]53)([OH]21)([O]78)([OH2])O)([OH2])O)([OH2])O.[O-]C(=O)c1ccc(cc1)c1cc(c2ccc(cc2)C(=O)[O-])c2c3c1ccc1c3c(cc2)c(cc1c1ccc(cc1)C(=O)[O-])c1ccc(cc1)C(=O)[O-] MOFid-v1.csq.cat0;	Zr.HVCDAMXLLUJLQZ.MOFkey-v1.csq	3
MIL-101	F[Cr][O]([Cr])[Cr].[O-]C(=O)c1ccc(cc1)C(=O)[O-] MOFid-v1.mtn-e.cat0;	Cr.KKEYFWRCBNTPAC.MOFkey-v1.mtn-e	1
MOF-177	[O-]C(=O)c1ccc(cc1)c1cc(cc(c1)c1ccc(cc1)C(=O)[O-])c1ccc(cc1)C(=O)[O-].[Zn][O]([Zn])([Zn])[Zn] MOFid-v1.qom.cat0;	Zn.SATWKVZGMWCXOJ.MOFkey-v1.qom	1
MOF-525	[O-]C(=O)c1ccc(cc1)C1=C2C=CC3=[N]2[Fe]24(n5c1ccc5C(=C1[N]2=C(C=C1)C(=c1n4c(=C3c2ccc(cc2)C(=O)[O-])cc1)c1ccc(cc1)C(=O)[O-])c1ccc(cc1)C(=O)[O-])(Cl)Cl.[OH2][Zr]123([OH2])[O]4[Zr]56[O]3[Zr]37([O]2[Zr]28[O]1[Zr]14([O]6[Zr]([O]53)([O]21)([O]78)([OH2])[OH2])([OH2])[OH2])([OH2])[OH2] MOFid-v1.csq.cat0;	Zr.ZBSKGGJJCPDFRA.MOFkey-v1.csq	1
ZIF-67	CC1=NC=C[N]1.CC1=N[CH]C=N1.[Co] MOFid-v1.sod.cat0;	Co.YFFQUDCLMWOYCW.MOFkey-v1.sod	1
Mg-MOF-74	[Mg].[O-]C(=O)c1cc([O])c(cc1[O])C(=O)[O-] MOFid-v1.UNKNOWN.cat0;	Mg.YXUXCIBWQAOXRL.MOFkey-v1.UNKNOWN	N/A
MIL-100	F[Cr][O]([Cr])[Cr].F[Cr][O]([Cr]F)[Cr].[Cr][O]([Cr])[Cr].[O-]C(=O)c1cc(cc(c1)C(=O)[O-])C(=O)[O-] MOFid-v1.moo.cat0;	Cr.QMKYBPDZANOJGF.MOFkey-v1.moo	
UiO-66-NH_2	[O-]C(=O)c1ccc(c(c1)N)C(=O)[O-].[O]12[Zr]34[OH]5[Zr]62[OH]2[Zr]71[OH]4[Zr]14[O]3[Zr]35[O]6[Zr]2([O]71)[OH]43 MOFid-v1.fcu.cat0;	Zr.GPNNOCMCNFXRAO.MOFkey-v1.fcu	

[a]The "duplicates" column was calculated by counting the number of structures in the CoRE MOF 2019-ASR database[7,80] with a matching MOFkey. See section on database statistics.

DOI: 10.1021/acs.cgd.9b01050
Cryst. Growth Des. 2019, 19, 6682−6697

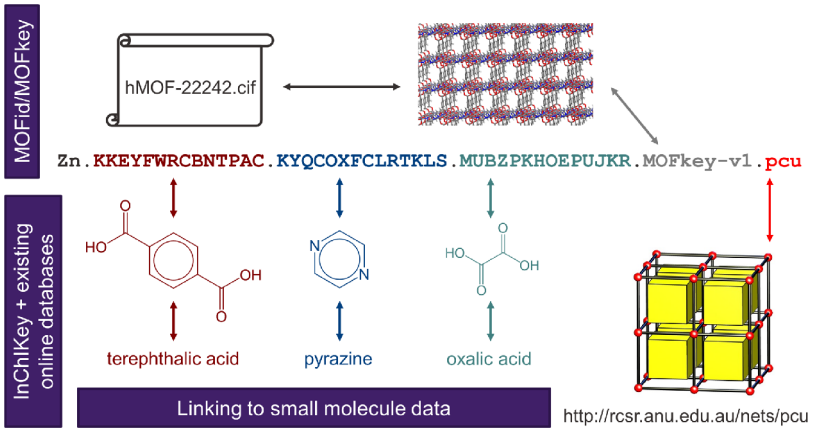


Figure 3. Example of linking the hypothetical structure GA-hMOF-22242 to publicly available chemical data. All three distinct linkers are searchable on Google via the InChIKey portion of the MOFkey. For more information about the composition of the GA hMOFs, see Section S2.3 and refs 87 and 88.

structure (e.g., useful for Supporting Information or a detailed database).

We first present examples of the MOFid and MOFkey schemes and then discuss these formats in more detail. Figure 2 shows how the formats are derived from the composition of the MOF Cu-BTC. Table 1 provides additional examples of identifiers for common MOFs identified by Anderson and Gómez-Gualdrón.[79] We note that the current version of the MOFid code cannot successfully determine the topology of certain MOFs with rod-like metal nodes, such as Mg-MOF-74, although it may be possible in future versions using the algorithms outlined in Section S1.4.

MOFid uses a specialized form of the SMILES format. Both formats represent the chemical structure(s) using the elemental symbols and a series of special characters to represent bonds, rings, branches, and other features in the molecular graph. Distinct chemical components are denoted by a standard dot-separated notation (a "dot-bond").[81] We use the canonical SMILES output format implemented in Open Babel to export the nodes and linkers. The current implementation sorts all of the building blocks together in alphabetical ASCII order, though other schemes such as nodes then linkers may be possible as part of a future specification. We disabled notation of stereochemistry to avoid complicated SMILES strings, which denote unnecessary details such as octahedral stereochemistry within metal SBUs. Like the Crystallography Open Database,[77] we denote connectivity within inorganic clusters using single bonds (i.e., a metal−oxygen connection is always represented by a bond order of one). We avoid assigning formal charges to metal atoms, because we do not want to imply any assumptions or guesses about the metal oxidation state, especially given the difficulty of assigning bonds to metal atoms. (See also the section on Limitations and Challenges.) The system does not currently handle charge-balancing cations or anions as special cases. Before exporting a linker substructure to SMILES, InChIKey, or other formats, we adjust the charges on linker molecules by pattern-matching against commonly charged sub-structures (e.g., assigning a formal charge of −1 to anionic carboxylate groups). Many cheminformatics representations of small organic molecules impose bond valence requirements, so explicitly assigning charges avoids an unexpected number of implicitly represented hydrogen atoms.

After defining the building blocks, we use the molecule name or comment field from the SMILES format to pack metadata about the parent MOF structure, such as its topology and interpenetration, separated by periods. In some cheminformatics toolkits (e.g., Open Babel), the comment field is defined as anything following the chemical structure and a tab or other whitespace. First, there is a `MOFid-v1` string that flags the SMILES as a valid MOFid (version 1 of the scheme). Next, the MOFid contains the three-letter RCSR code for the base "single-node" topology. If the "all node" representation is different, or the user wishes to specify other topological representations, they can be appended using commas. If a topology cannot be assigned, the code reports a placeholder value of `ERROR` or `UNKNOWN`. The catenation, or degree of interpenetration, of the MOF is specified with the keyword `cat` and number of additional disconnected nets (e.g., IRMOF-1 is represented as `cat0`, but interpenetrated IRMOF-1 and IRMOF-9 are both represented by `cat1`). Finally, the MOFid ends in a semicolon followed by a user-specified comment field, which authors could use to provide a common name, structure number, or other identifying information in a paper. Since the MOFid metadata repurposes a comment field, there is the possibility to extend the MOFid specification later to include more metadata, such as the ratio of linkers in multivariate MOFs or the source of the MOF crystal structure (e.g., hypothetical, SHELXL, CCDC deposit).

MOFkey compactly contains similar information but is based on the InChIKey format. Nodes are identified by a list of unique metals, specified by their elemental symbol and sorted by atomic number. For example, the Zn_4O node of MOF-5 would be identified by `Zn`. Nonmetal elements, such as oxygen and hydrogen, are excluded from MOFkey's representation of a node to avoid inconsistencies in how crystal structures are reported. For example, a copper paddlewheel could be represented using two coordinatively unsaturated metal atoms alone, with a bound water, or with an oxygen atom (without hydrogen atoms explicitly included). Thus, the MOFid could contain any of these three representations, whereas MOFkey avoids this ambiguity by only reporting the metal(s) and MOF topology. Then, each unique linker molecule is specified by a truncated 14-character InChIKey, which describes its "molecular skeleton" or "connectivity layer."[40,82] A standard InChIKey contains 27 characters in three layers; however, in the databases we tested, the second two layers contain the identical 13 characters `-UHFFFAOYSA-N` *per linker*. We safely removed these layers for brevity: these layers would only be different if they contained information like linker stereochemistry or rare cases like nonstandard isotopes, but these cases are not considered by our code. Like MOFid, dots are used to separate the MOF metadata and linker InChIKeys. Next, the MOFkey contains an easily searchable format identifier, in this case `MOFkey-v1`. Finally, the base "single node" RCSR topology is specified, if available. We intend for MOFkey to be searchable verbatim, so we do not include alternate or user-specified topologies. Similarly, we do not include the MOF catenation for brevity and simplicity. In MOFs with a single type of metal and linker, the full MOFkey will be 31 characters in length, considerably shorter than the MOFid in most instances. We note that 10 characters of the MOFkey format are devoted to the `MOFkey-v1.` label. We kept a verbose version flag (`-v1`) as part of our MOFkey prototype, but a future, formal specification could omit this information or represent it using a single character (similar to InChIKey).

DOI: 10.1021/acs.cgd.9b01050
Cryst. Growth Des. 2019, 19, 6682−6697

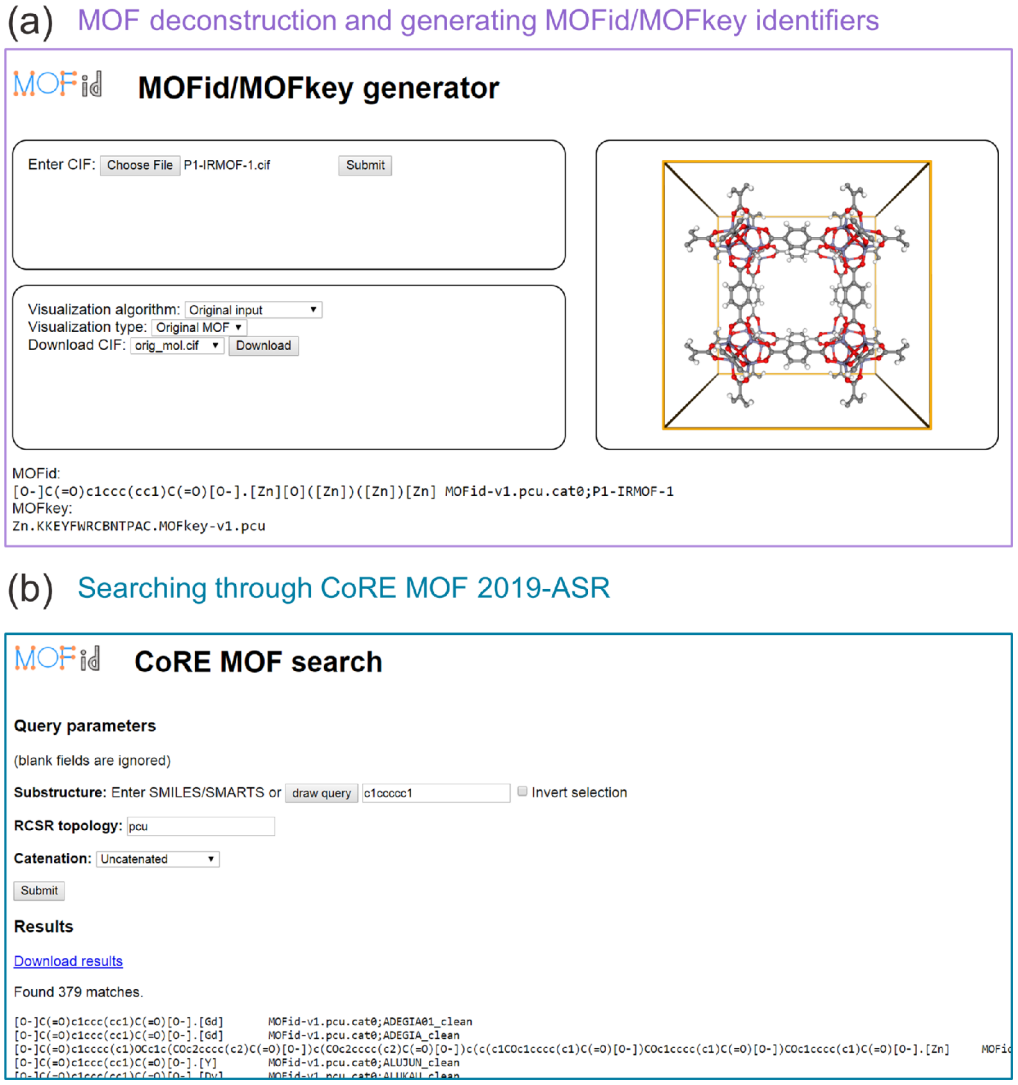


Figure 4. Overview of the MOFid platform's online user interface, enabling (a) web-based generation of MOFid/MOFkey identifiers and (b) searches through CoRE MOF 2019-ASR by topology and/or chemical substructure.

We note that only identifying the inorganic nodes by the metal elements in the MOFkey offers some advantages and disadvantages compared with the graph-based approach in the MOFid. It is difficult to assign the precise proton topology on the metal nodes for some MOFs, requiring detailed study to elucidate the precise isomer.[51] Differences in assigning the locations of the protons within the Zr_6 node, as seen in MOF-545 and PCN-222,[51,83,84] lead to differences in the SMILES and ambiguity in the identifier, whereas MOFkey classifies these nodes as the same Zr metal and **csq** topology. Given that MOFkey is intended to be a compact, canonical identification scheme for a parent MOF family, we see the potential loss of information as an acceptable trade-off to gain a stable identification scheme. Along the same lines, we note that MOFkey is more robust to the underlying cheminformatics toolkit than MOFid. As noted earlier, several canonicalization algorithms and implementations exist for SMILES (recently leading to an IUPAC SMILES+ project[85]). SMILES is a useful format for reversibly representing molecules within a single project (especially after recanonicalizing the identifiers for self-consistency), but an InChI-based solution is more suitable for establishing database keys and search terms that remain constant long-term.

The MOFid and MOFkey schemes could facilitate the communication and automated linking of data in the literature and within databases. MOFid would likely be most useful as Supporting Information to provide detailed structural data about the node and linker building blocks. Unlike the InChIKey, which uses a one-way hash function, canonical SMILES is generally a reversible one-to-one notation (ignoring special cases). Thus, if a paper has multiple structures, the MOFid's could be tabulated and labeled to unambiguously tag individual structures of interest. The MOFkey, on the other hand, would be most useful as an abbreviated barcode to place in the abstract, keywords, and/or main text of a paper to enhance MOF searchability and linking data together. Its base InChIKey format is compatible with common web search engines[64] like Google by minimizing the use of special characters or lengthy queries, which can be difficult for search engines to parse.[86] For example, a hypothetical MOF structure[87,88] containing three linker molecules can be linked to online data for all three building blocks via the MOFkey, as shown in Figure 3. While the InChIKey hash is not inherently reversible, indexed databases and dedicated InChIKey resolvers (e.g., CACTUS[89]) can resolve some of these hashes against a list of known chemicals. It is also important to note that some aspects of the MOFid/MOFkey are not directly compatible with every cheminformatics software package, although we attempted to maximize compatibility when possible. For example, some software (e.g., ChemDraw) does not permit the SMILES comment field, but this incompatibility can be readily resolved by removing nonstructural information (i.e., the metadata and comment fields) from a given MOFid. Some InChIKey resolvers will not operate unless the full InChIKey is specified, so the user would need to adapt the linker data

DOI: 10.1021/acs.cgd.9b01050
Cryst. Growth Des. 2019, 19, 6682−6697

from the MOFkey before searching these databases. For example, to search for the linker in Cu-BTC (`Cu.QMKYBPDZANOJGF.MOFkey-v1.tbo`), the InChIKey would be `QMKYBPDZANOJGF-UHFFFAOYSA-N`.

If MOF identification format(s) are adopted in the literature, we anticipate that the MOFkey would be more generally useful for quickly finding MOF structures, whereas the MOFid would provide detailed structural information and metadata about the MOF. Said another way, these schemes have similar objectives as the underlying SMILES and InChI formats: "InChI is not a replacement for any existing internal structure representations [e.g., SMILES]. InChI is in *addition* to what one uses internally. Its value to chemists is *finding* and *linking* information."[90] Likewise, the MOFid and MOFkey are not intended to replace the existing MOF nomenclature (e.g., Cu-BTC) but rather to provide another method for finding and processing MOF data. If MOFid and/or MOFkey are adopted by the community, we suggest that a standards authority formally assigns a specification. Left unchecked, standards have a tendency to proliferate.[91] The primary goal of MOFkey is to provide a consistent identifier for linking together databases, and multiple implementations or adaptations could lead to incompatibilities with one another and existing software toolkits. However, if managed properly, MOFkey is a powerful tool for analyzing MOF data and linking it to the broader chemical literature.

Software Environment. As part of this project, we are releasing an open-source code on GitHub (https://github.com/snurr-group/mofid), which includes an implementation of the MOFid and MOFkey schemes as well as tools for analysis. The code is comprised of three overall parts: a main C++ code for deconstructing MOF structures into their building blocks, Python code to assemble the MOFid/MOFkey identifiers, and various analysis utilities. The C++ code is responsible for reading a MOF structure in CIF format, deconstructing the MOF into its building blocks, simplifying the underlying topology, and exporting the simplified components. The code extensively uses the Open Babel library, a popular open-source framework for analyzing chemical structures as molecular graphs and handling chemical file formats.[49,92] We have modified the library[93] to include new features such as periodic boundary conditions, which are necessary to properly calculate bond lengths, angles, and torsion in repeating crystal structures. The final version of these modifications includes contributions from Giovanni Garberoglio, who previously implemented a similar feature independently as part of the OBGMX project.[94] The Python part of our code assembles the MOFid/MOFkey identifiers by wrapping other utilities: it parses the SMILES/InChI output from the C++ code and calls Systre to perform topology assignment. We also provide a Python API to allow users to quickly integrate MOFid/MOFkey analysis into their workflows and Bash scripts for common analyses on supercomputing clusters.

Our code generates the MOF identifiers and produces several outputs for analysis. For each of the topology simplification algorithms, we write the atomic coordinates for the individual MOF building blocks, the overall simplified net, solvent molecules (if detected), and related structures. Once the MOF has been deconstructed into its basic building blocks, it is considerably easier to perform complex analyses and searches based on the MOF chemistry. For example, composition data for a collection of MOF crystal structures can be organized into a spreadsheet or relational database, allowing for the use of complex queries.[95] Substructure searching methods from the cheminformatics literature, such as the SMARTS language, can also rapidly filter through databases to find chemical moieties of interest. These querying techniques are powerful, although they require awareness of certain subtleties, such as implicitly handling hydrogen atoms and syntax that differentiates between any carbon atom (`[#6]`), aromatic (`c`), or aliphatic (`C`).[96]

We also developed a client-side Web site available at https://snurr-group.github.io/web-mofid/. The web interface allows users to interactively run the MOFid/MOFkey code while avoiding the compilation and installation requirements. Users can import their own CIFs to the tool, which runs locally in the user's web browser without uploading files to a web server. The screenshots in Figure 4 highlight the Web site's capabilities for assigning MOF identifiers and running database searches. In the literature, web apps have been shown to increase molecular understanding and allow users to quickly explore content and calculations interactively.[97] By reducing the barrier to setup and use of our code, we hope to facilitate user adoption of the MOFid/MOFkey analysis and identifier tools.

The use of open-source software for the MOFid and MOFkey schemes is advantageous. Users can inspect the source code of the underlying algorithms, and repositories such as GitHub facilitate distribution of the code to the research community. Researchers can also revisit old data and assign MOFid/MOFkey identifiers to previously reported structures. A nice feature of the MOFid/MOFkey schemes is that they do not inherently require a crystal structure. If clean crystallographic results are infeasible, or if a user wishes to propose a new hypothetical structure, it is possible to manually assemble the MOFid or MOFkey formats using the simplified topological net and corresponding SMILES or InChIKeys.

Limitations and Challenges. One challenge with applying cheminformatics methods to metal-containing compounds, such as MOFs, is that bond assignment is still an active area of research and debate. There are several ways to assign bonds to construct a molecular graph, but all have limitations.[98] Geometric methods are common, where two atoms are considered bonded if their distance is less than the sum of their covalent radii,[50] plus an extra skin distance (0.3−0.45 Å is common[4,99]). Some bonding algorithms, such as those in Open Babel and Materials Studio, additionally include a minimum distance cutoff by default.[49,100] Bond orders are particularly ambiguous.[101] Some algorithms assign a formal valence of 1/2 to metal−oxygen bonds,[45,102] while others treat everything as single bonds.[101] A zero-order bond has also been proposed for denoting coordination bonds in cheminformatics software.[103] The CSD performs bond perception using Bayesian methods that combine information about the geometry and known structures in the database.[104] Many subtleties in MOF structures could cause issues for bond perception routines, in particular, bonding within metal nodes, atoms near the cutoff distances, and disordered structures. Regardless, even approximate representations would be generally useful to narrow down the number of possible matches in a search.

The MOFid and MOFkey represent MOFs as their idealized structures, because the formats must strike a balance between a simple versus comprehensive identifier. The complexity of the MOF field quickly leads to many corner cases that the MOFid and MOFkey schemes do not consider. Some examples of these special cases include hetero-interpenetrated topological nets,[105] framework defects,[106,107] and composites, including metal nanoparticles encapsulated in MOFs[108,109] and hybrid MOF−polymer materials.[110] The MOFid code will retrieve all unique linkers in multivariate (MTV) MOFs, though it does not (currently) list a ratio of compositions. Likewise, postsynthetic modification[111] and/or building block replacement[112] techniques could lead to ambiguity if the eligible sites are partially substituted. Covalent organic frameworks (COFs)[113,114] are currently out of the project scope: the code requires a boundary between inorganic and organic building blocks to deconstruct framework structures. Neglecting these special cases, there is generally a one-to-one correspondence between a MOF identifier and a MOF's composition, provided that an RCSR framework topology can be assigned. By representing MOFs as their idealized structures, the proposed MOF identifiers substantially narrow down the search space and improve data organization.

Framework topologies are another challenge. Some MOFs have multiple possible topological representations and ways to assign the "topologically significant" connection points.[115] In fact, even abstract topology definitions can themselves sometimes be reduced to other underlying nets by clustering vertices together in the network.[37] The merged nets approach can expand the topological space by combining two compatible edge-transitive nets.[116] To date, newly discovered merged nets have been reported to the RCSR, and thus these topologies are likely compatible with the MOFid/MOFkey scheme automatically. Another approach for introducing topological complexity is by building metal−organic polyhedra into a hierarchy of

DOI: 10.1021/acs.cgd.9b01050
Cryst. Growth Des. 2019, 19, 6682−6697

secondary or tertiary building units.[117] Since the MOFid/MOFkey schemes represent MOFs as their simplest isolated inorganic and organic clusters, they cannot explicitly capture topological hierarchies. Overall, the diversity of MOF structures is excellent for providing a large design space, although challenging for rigorous nomenclature protocols. However, after accepting some level of arbitrariness, analyzing MOF structures as molecular graphs can be rather effective as long as the criteria are well-documented and applied systematically.

Even given the same chemical building blocks, many classes of "framework isomers" have been defined for MOFs.[118−120] Topological isomers arise when the same MOF SBUs are connected into different topologies. Catenation isomers result when multiple copies of the topology are interpenetrated, typically resulting in a low void fraction within the material. These interpenetrated nets can take many forms,[115] including partially interpenetrated nets,[121] distinct symmetry classes,[122] and woven periodic knots.[123] Isoreticular isomers can be formed when SBUs have lower symmetry than vertices in the underlying net, typically showing up as different relative orientations of nodes[119,124] or rotations of pseudosymmetric tetratopic linkers[125,126] (e.g., rotating the rectangular linkers in NU-1100).[127] Finally, conformational isomers have the same molecular graph but different conformations within the SBUs due to bond bending or related transformations. Examples include "breathing MOFs" such as MIL-53,[128] pressure-induced phase transitions such as PCN-250 and PCN-250′,[129] and stimuli-responsive materials.[130,131] One approach for distinguishing these isomers could be including an additional field into the MOFid about the MOF volume per metal atom. Overall, the proposed naming scheme can be viewed as a family of answers, depending on the level of detail required for comparison. Of these isomer classes, the MOFkey only considers the overall framework topology, and MOFid considers topology and number of interpenetrated nets.

APPLICATIONS

Validation against MOF Databases. Given the complexity of deconstructing MOFs, we developed a set of test cases to validate the MOFid code. We started with databases of hypothetical materials constructed in silico from presumably well-defined SBUs and topologies, which would be a cleaner source of labeled data than experimentally reported crystal structures. The goal was to benchmark the calculated MOFid against a ground truth, which we defined from the recipe used to construct the crystal structure. We extracted a test set from two databases of hypothetical MOFs, which were constructed by crystal generation algorithms. First, we considered a database of hypothetical MOFs studied using a genetic algorithm,[87] which we denote the "GA hMOFs" in this work. This database was adapted from the "bottom-up" hypothetical MOF database of Wilmer et al.[88] to have well-defined, unique chromosomes describing the composition. We also considered the topologically diverse ToBaCCo database constructed by Colón, Gómez-Gualdrón, and Snurr using a "top-down" assembly method.[132,133] Together, these databases contain more than 10 000 MOFs for testing.

Figure 5a shows an example validation by comparing the input recipe of MOF-5 fed into a "top-down" crystal generator against the MOFid identifier calculated from the output crystal structure. In theory, any mismatches would be attributable to errors in the MOFid code, thus directly indicating the code's robustness. However, in practice, some of the building blocks and topologies were ill-suited for this purpose, leading to an inexact mapping between the crystal generator input and expected output. For example, Figure 5b highlights a case where we had expected an **rna** topology based on the V_3O_3 node. Because of a geometry misalignment, instead of completing the 4,4′-(ethyne-1,2-diyl)dibenzoic acid linker

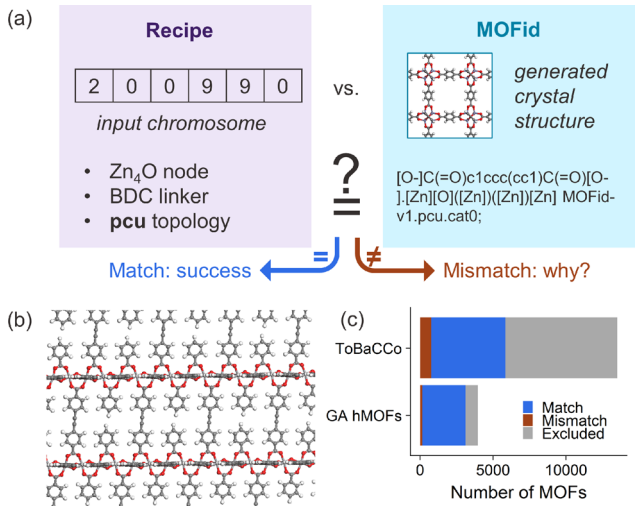


Figure 5. Validating the MOFid code against MOF structures generated in silico. (a) Overview of the validation procedure using the example of MOF-5, (b) example mismatch between an expected crystal structure and the calculated MOFid, (c) validation results, detailed in Section S2.

between two nodes, the crystal generator capped the sites with benzoic acid groups, which prevented our MOFid code from finding the expected topology. Additionally, some bond geometries in the ToBaCCo and GA hMOF databases are incompatible with a simple distance-based bond perception algorithm; some of the atom positions, while suitable for high-throughput screening of the design space, are technically unphysical due to limitations in the settings for the construction algorithms and/or the force field optimization. In both of these examples, the validator should not flag an error with the MOFid code: although there is a mismatch between the MOFid and our expectations, the underlying cause was misunderstanding the crystal structures, not necessarily an issue with the underlying simplification algorithms. To account for these discrepancies, we selected a subset of the full GA hMOF and ToBaCCo MOFs, excluding cases known to be challenging for the validator. See Section S2.4 for details.

In order to estimate the success rate for the MOFid code, we excluded mismatches where we could identify an underlying systematic disagreement between an expected and actual crystal structure, labeled as "Excluded" in Figure 5. We excluded 839 of the 3952 unfunctionalized GA hMOFs (and 47 211 GA hMOFs containing functional groups) and 7679 of the 13 511 ToBaCCo MOFs. Out of the MOF subset remaining, we estimate a 95.4% success rate for the GA hMOFs and 86.9% success rate for the ToBaCCo MOFs, shown in Figure 5c. Given the challenges in interpreting reported MOF crystal structures, we surmise that the code was sufficiently robust for analysis, subject to a few known issues documented in Section S2.4.5.

Analysis of MOF Structure Databases. In this section, we demonstrate a few examples of using the MOFid/MOFkey code for analysis of large MOF databases. Beyond hypothetical structures or a small set of common MOFs, we extend the analysis to thousands of experimentally reported crystal structures that have been collected in the CoRE MOF 2019-ASR (all solvent removed) database.[7,80] In this work, we only

DOI: 10.1021/acs.cgd.9b01050
Cryst. Growth Des. 2019, 19, 6682−6697

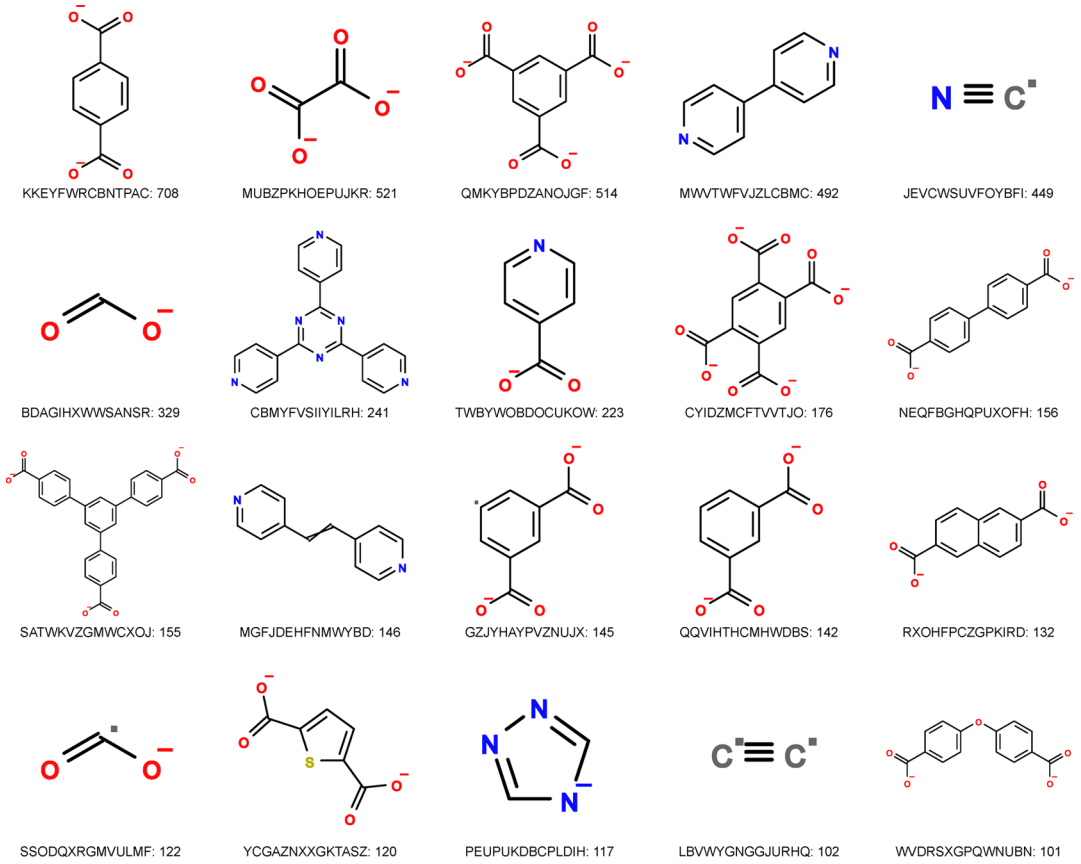


Figure 6. Most common linkers in CoRE MOF 2019-ASR according to the assigned MOFkeys. Each structure is labeled with the truncated InChIKey and number of MOFs identified.

include crystal structures that contain carbon, yielding 14 026 MOFids/MOFkeys under consideration.

MOF identification algorithms enable exploration of many MOF properties based on their SBUs. Certain questions, such as "what linkers are the most common?" or "how many unique MOFs exist?" can be answered directly using MOFid. The MOFid/MOFkey scheme reports MOFs deconstructed into their building blocks, which enables rapid aggregation or comparison between structures. Looking at unique nodes within the MOFids (Table S1), we find that the most common inorganic structures identified by the "metal-oxo" algorithm are isolated metal atoms. The most common node is a zinc cation, characteristic of many ZIFs. Some other common nodes contain pairs of metal atoms (characteristic of metal paddlewheels), the Zn_4O cluster of MOF-5, and the Zr cluster of UiO-66.

Using the InChIKey information embedded in the MOFkeys, we extracted the most common linkers in the CoRE MOF database, shown in Figure 6. We identify many of the same linkers, albeit with some differences, compared with the CoRE MOF 2019 analysis,[7] which was run by searching a subset of the database for common linkers using the Conquest[35] program. Repeating our analysis using the SMILES structures from MOFid instead of the InChIKey from MOFkey yields different numbers. As shown in Figure S10, there are sometimes multiple SMILES structures corresponding to the same InChIKey. Deriving the chemical graph of bonds from a crystal structure is a challenging problem, especially in cases of disorder or missing hydrogen atoms, which leads to (incorrectly) assigning carbon atoms in the linker as radicals with missing valence. The InChI normalization procedures remove some but not all of this sensitivity to the structure definition. Within Figure 6, there are two pairs of linkers with multiple InChIKeys: formate (BDAGIHXWWSANSR and SSODQXRGMVULMF) and isophthalic acid (GZJYHAYPVZNUJX and QQVIHTHCMHWDBS). These cases reinforce that high-throughput analysis can be useful for identifying promising targets or overall trends, but conclusions about individual MOFs merit another detailed look at the crystal structure.

Duplicate MOF structures. MOFid also enables the rapid identification of multiple copies of the same MOF in a database, provided that a MOF is sufficiently defined by its building blocks and topology. Identifying duplicate MOF structures has been used in the literature to avoid redundant density functional theory (DFT) calculations,[5,134] recalculate database statistics,[47] and compare MOF databases,[7] among other applications. More generally, deduplicating crystal structures or identifiers presents a data curation challenge for many groups, such as the Crystallography Open Database[135] and NIST/ARPA-E Adsorption Database.[8] Duplicates analysis can be carried out by several methods, including a direct comparison of the atomic positions,[5] information about the molecular formula and unit cell,[135] textural property fingerprints,[7] and bond network descriptors.[47] As hinted at in the earlier Challenges section, there is no single definition for what constitutes a unique structure because it can be context dependent. For example, when eliminating redundant DFT calculations, atomic coordinates are well-suited for reporting geometry-dependent results, but for the application of

DOI: 10.1021/acs.cgd.9b01050
Cryst. Growth Des. 2019, 19, 6682−6697

aggregating database statistics, the parent bond structure is likely more relevant and excludes extraneous complications such as conformational isomers.

In this section, we consider a similar definition for duplicates as Barthel et al.:[47] "two structures [that] can in principle be deformed into each other without breaking and forming bonds" (in our case, with minor caveats excluding stereochemistry and the exact node composition). We implement this approach by comparing MOFs by their MOFkeys, which encode information about the metal, linker bond network, and how the building blocks are assembled topologically. We assume that the MOFkey is a one-to-one representation of a MOF: each unique MOF chemistry/structure corresponds with exactly one MOFkey, and vice versa. Our workflow is unable to assign topologies to all of the MOFs in the databases for various reasons (importing the structure fails, non-RCSR topology, etc.), so we exclude any MOFkeys containing an ill-defined topology (see Section S2.2). After filtering the CoRE MOF 2019-ASR database, we are left with 6259 out of the 14 026 structures for analysis. Within this subset, we identify 4104 unique MOFkeys. Twelve of these MOFkeys are aggregated earlier in Table 1 as common MOFs in the literature, and we report the number of duplicate structures in the last column of the table. On the basis of the MOFkey, we also find 42 MOFs that have been reported at least 10 times in CoRE MOF 2019-ASR, which are collected in Table S2.

In the literature, bond network descriptors have been applied[47] to deduplicate a slightly modified version of 502 DFT-optimized CoRE MOF structures with DDEC partial atomic charges.[6] We have benchmarked this method against our MOFkey deduplication approach on this same set of structures. As before, we only excluded MOFkeys without a well-defined RCSR topology, leaving 242 MOF structures under consideration. On the basis of this subset, 209 of the structures were classified into 186 groups (167 singletons) that identically matched the network bond classifications. The remainder of the MOFs came from five families with distinct structures, but the same MOFkey and the network bond method correctly distinguished these families. Section S4.2 details the underlying causes, which include catenation and detailed node information found in the Barthel/MOFid methods but not the MOFkey. The overall agreement between the bond network descriptor and MOFkey approaches provide confidence in the method for duplicate detection while also showing its limitations, such as the inability to report non-RCSR topologies.

Overlap between MOF databases. Running duplicate searches across multiple databases can indicate structures that are common between them: in other words, the overlap between the databases. Using this approach, we calculated the overlap among the CoRE MOF 2019-ASR,[7,80] GA hMOF,[87] and ToBaCCo databases,[132,133] and identified dozens of MOFs in common between the databases, as shown in Figure 7. Using the MOFkeys, we identify 10 MOFs in common among all three databases: Cu-BTC, IRMOF-1, IRMOF-8, IRMOF-9, IRMOF-14, IRMOF-61, DUT-34, TCM-8, UiO-66, and refcode LIHFAK.

We benchmarked our MOFkey-based deduplication approach against the textural property fingerprinting method from the CoRE MOF 2019 analysis.[7] In searching for structures in common between the unfunctionalized GA hMOFs and CoRE MOF 2019-ASR, we identify the same list of MOFs with a few exceptions. The MOFkey-based

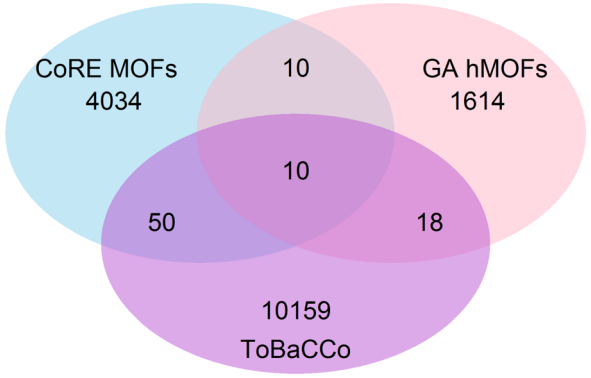


Figure 7. Overlap among three databases: CoRE MOF 2019-ASR, unfunctionalized GA hMOFs, and ToBaCCo MOFs. Databases were deduplicated before calculation and compared using the MOFkeys. These numbers include one false positive in the GA-CoRE overlap (Zn(bpe)(muco)) and four in the CoRE-ToBaCCo overlap.

overlap approach does not identify IRMOF-3 (equivalent to amine-functionalized IRMOF-1), because we had only searched through unfunctionalized hMOF structures. The two IRMOF-3 structures identified in the CoRE MOF 2019 paper, VURMOL and hMOF-69, generate identical MOFkeys (`Zn.GPNNOCMCNFXRAO.MOFkey-v1.pcu`). Therefore, the discrepancy between the fingerprint and MOFkey methods was caused by the list of MOFs under consideration, not something inherent to the MOFkey approach. There may be other functionalized GA hMOFs that we have missed, but in general these structures tend to be less compatible with the MOFid/MOFkey workflow (Section S2.4). We also note that the MOFkey does not account for catenation, only considering the building block chemistry and topological connectivity. There are also some overlapping structures that the MOFkey-based approach flagged as overlapping MOFs, but the fingerprinting method did not: UiO-66, TCM-8, DUT-34, COMOC-3, Zn(bpe)(muco), and Cu_2(dicarboxylate)$_2$(amine). UiO-66 was not identified by the fingerprint algorithm due to differences in the hydrogen topology of the Zr_6 node: the GA hMOF structure has a carbon-to-hydrogen element ratio of 2:1, whereas the CoRE MOF structures have ratios of 3:2 or 12:7. The other MOFs are not identified by the fingerprinting method due to differences in the structure densities (Table S4). The reported crystal structure of COMOC-3 has narrower pores[136] than the square-like open pores of the analogous GA hMOF structure 1002260. Zn(bpe)(muco) is a false positive flagged by the MOFkey method: the GA hMOF structure 5049620 is built with a Zn paddlewheel, whereas the CoRE MOF structure SUJQOE has a different inorganic node built from Zn$_2$(muco).[137] DUT-34, TCM-8, and Cu_2(dicarboxylate)$_2$(amine) are found as their catenated versions in the CoRE MOF database but not the GA hMOFs, causing their densities to differ by a factor of 2. The fingerprint algorithm considers catenated structures as distinct MOFs, whereas the MOFkey does not, leading to the difference in reporting. The MOFkey approach is a more direct method of comparing structures than fingerprint-based deduplication, resulting in far fewer false positives to parse.

Polymorphism. Storing MOF composition information in an easily queried format (such as a relational database) enables rapid search and exploration of new questions, such as the identification of MOF polymorphs. Figure 8a shows an

DOI: 10.1021/acs.cgd.9b01050
Cryst. Growth Des. 2019, 19, 6682−6697

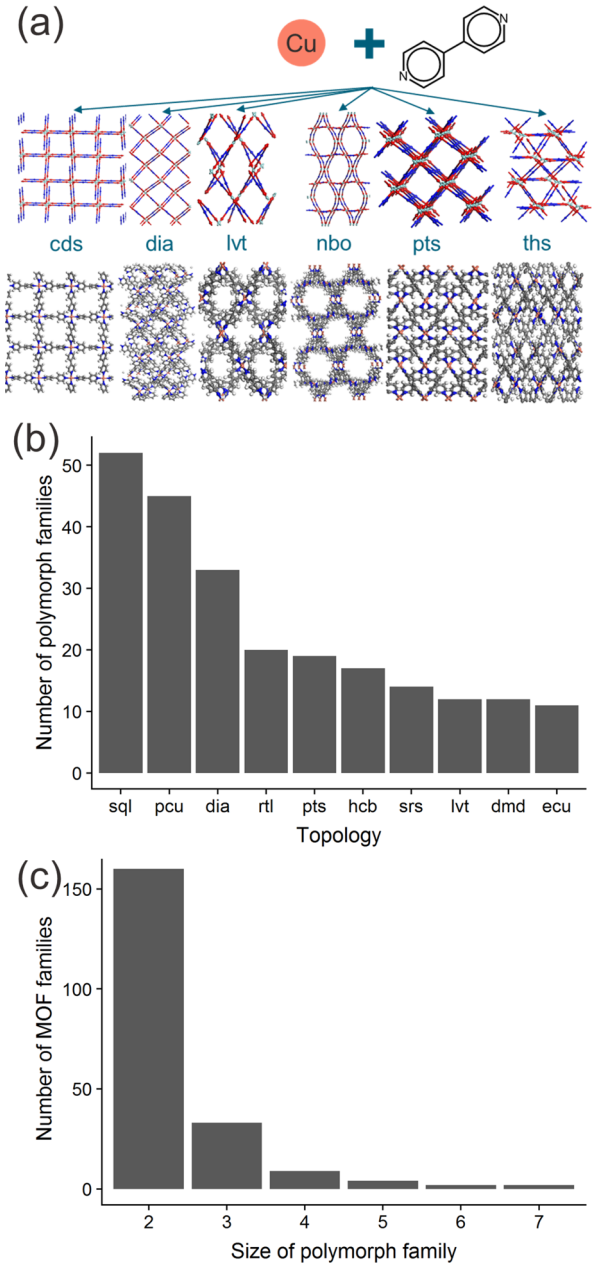


Figure 8. Properties of polymorph families in the CoRE MOF 2019-ASR database. (a) Example visualizing a polymorph family with a copper metal, 4-4′-bipyridine linker, and six MOF topologies. (b) Polymorph count by topology (frequency of topologies among polymorphic MOFs), for the top 10 most frequent polymorph topologies. (c) Size of polymorph families (number of topologies per family).

example polymorphic MOF family, defined as MOFs with the same building blocks but different topologies. There has been a growing literature on MOF polymorphs, which explores the effect of topology on gas adsorption,[70,71,133,138] directly compares the energetics of MOF polymorphs,[139] and achieves topologically controlled syntheses.[140−142] In this section, we use information from MOFid to rapidly extract and analyze families of MOF polymorphs from the CoRE MOF 2019-ASR database.

Querying the database for polymorphic MOF families has a similar process as the duplicates analysis. The main difference is that we only aggregate by the chemical composition instead of both the chemistry and topology (and also require at least two distinct topologies per polymorph family). Such an analysis would be considerably more difficult to complete at scale without automated methods that tabulate both the building blocks and topology of the MOF structures. MOFid and MOFkey are natural choices for this application.

Looking at Figure 8b, we see that **sql**, **pcu**, and **dia** are the most frequent topologies found among polymorph families. From Figure 8c, the vast majority of polymorph families consist of two topologies represented. There are 17 polymorph families containing at least 4 distinct topologies (Figure S12). One of the largest families is a set of ZIFs with seven topologies for the same building blocks. The actual number of polymorphic ZIFs in the CoRE MOF database is even larger, because this analysis does not account for aromaticity errors in the linker (Section S2.4.5). Using MOFid and cheminformatics tools like SMARTS could enable additional insight into the role of functional groups in MOF polymorphs in future work.

CONCLUSIONS

In this work, we have proposed and taken first steps toward assigning MOF identifiers using automated cheminformatics algorithms. The MOFkey scheme provides a compact description of a MOF's composition and topology, building on the open InChIKey format. The MOFid scheme provides more detailed information, building on a SMILES description of the MOF building blocks. We have provided an open-source code and web interface that analyze MOFs and generate these identifiers given a MOF crystal structure. We emphasize that short, colloquial names are a convenient approach for describing MOFs, but their usage in the literature should be augmented with a systematic identifier, such as MOFkey, to avoid miscommunication and facilitate easier linking between data sources.

The proposed MOFid/MOFkey formats are not converged specifications: the goal of this paper is to initiate the development of a tool that addresses a current gap in MOF data management. The proposed formats provide a foundation for the community to work toward an open standard describing MOFs, which could be potentially extended with more complexity later. Using an open-source implementation for these identifiers, we can efficiently assign a barcode (systematic nomenclature) for MOFs, rapidly search through MOF structures, and analyze broad categories of the MOF space as a whole. Looking to the future, we hope that IUPAC and/or CCDC takes a leadership role in formalizing these standards to facilitate their mainstream adoption in the MOF community.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.cgd.9b01050.

Additional algorithm details, validation results, and database statistics (PDF)

Data files with MOFid and MOFkey identifiers for CoRE MOF 2019-ASR, duplicate information, and polymorph families (ZIP)

AUTHOR INFORMATION

Corresponding Author

*E-mail: snurr@northwestern.edu.

DOI: 10.1021/acs.cgd.9b01050
Cryst. Growth Des. 2019, 19, 6682−6697

ORCID

Benjamin J. Bucior: 0000-0002-8545-3898
Andrew S. Rosen: 0000-0002-0141-7006
Maciej Haranczyk: 0000-0001-7146-9568
Zhenpeng Yao: 0000-0001-8286-8257
Omar K. Farha: 0000-0002-9904-9845
Joseph T. Hupp: 0000-0003-3982-9812
J. Ilja Siepmann: 0000-0003-2534-4507
Alán Aspuru-Guzik: 0000-0002-8277-4434
Randall Q. Snurr: 0000-0003-2925-9246

Notes

The authors declare the following competing financial interest(s): O.K.F., J.T.H., and R.Q.S. have a financial interest in NuMat Technologies, a startup company that is seeking to commercialize MOFs.

■ ACKNOWLEDGMENTS

This work was primarily supported by the U.S. Department of Energy, Office of Basic Energy Sciences, Division of Chemical Sciences, Geosciences and Biosciences through the Nanoporous Materials Genome Center under Award Numbers DE-FG02-12ER16362 and DE-FG02-17ER16362. B.J.B. also acknowledges support from a National Science Foundation Graduate Research Fellowship under Grant No. DGE-1324585. A.S.R. acknowledges government support under Contract FA9550-11-C-0028 and awarded by the Department of Defense (DOD), Air Force Office of Scientific Research, National Defense Science and Engineering Graduate (NDSEG) Fellowship, 32 CFR 168a. B.J.B. and A.S.R. acknowledge support from Ryan Fellowships via the International Institute for Nanotechnology at Northwestern University. This research was supported in part through the computational resources and staff contributions provided for the Quest high-performance computing facility at Northwestern University, which is jointly supported by the Office of the Provost, the Office for Research, and Northwestern University Information Technology. The authors thank Jeffrey R. Long for helpful conversations about the identifiers.

■ REFERENCES

(1) Furukawa, H.; Cordova, K. E.; O'Keeffe, M.; Yaghi, O. M. The Chemistry and Applications of Metal-Organic Frameworks. *Science* **2013**, *341*, 1230444−1230444.

(2) Kaskel, S., Ed. *The Chemistry of Metal-Organic Frameworks: Synthesis, Characterization, and Applications*; Wiley-VCH Verlag GmbH & Co. KGaA: Weinheim, 2016.

(3) Sturluson, A.; Huynh, M. T.; Kaija, A.; Laird, C.; Yoon, S.; Hou, F.; Feng, Z.; Wilmer, C. E.; Colon, Y. J.; Chung, Y. G.; Siderius, D.; Simon, C. *Role of Molecular Modeling & Simulation in the Discovery and Deployment of Metal-Organic Frameworks for Gas Storage and Separation*. 2019, DOI: 10.26434/chemrxiv.7854980.v1.

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

(5) Nazarian, D.; Camp, J. S.; Sholl, D. S. A Comprehensive Set of High-Quality Point Charges for Simulations of Metal−Organic Frameworks. *Chem. Mater.* **2016**, *28*, 785−793.

(6) Nazarian, D.; Camp, J. S.; Chung, Y. G.; Snurr, R. Q.; Sholl, D. S. Large-Scale Refinement of Metal-Organic Framework Structures Using Density Functional Theory. *Chem. Mater.* **2017**, *29*, 2521−2528.

(7) Chung, Y. G.; Haldoupis, E.; Bucior, B. J.; Haranczyk, M.; Lee, S.; Zhang, H.; Vogiatzis, K. D.; Milisavljevic, M.; Ling, S.; Camp, J. S.; Slater, B.; Siepmann, J. I.; Sholl, D.; Snurr, R. Q. Advances, Updates, and Analytics for the Computation-Ready, Experimental Metal-Organic Framework Database: CoRE MOF 2019. Submitted.

(8) Siderius, D., Shen, V., Johnson, R., van Zee, R., Eds. *NIST/ARPA-E Database of Novel and Emerging Adsorbent Materials*; National Institute of Standards and Technology: Gaithersburg MD, DOI: 10.18434/T43882.

(9) Park, J.; Howe, J. D.; Sholl, D. S. How Reproducible Are Isotherm Measurements in Metal−Organic Frameworks? *Chem. Mater.* **2017**, *29*, 10487−10495.

(10) Wilkinson, M. D.; et al. The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Sci. Data* **2016**, *3*, sdata201618.

(11) Hill, J.; Mulholland, G.; Persson, K.; Seshadri, R.; Wolverton, C.; Meredig, B. Materials Science with Large-Scale Data and Informatics: Unlocking New Opportunities. *MRS Bull.* **2016**, *41*, 399−409.

(12) Vodak, D. T.; Braun, M. E.; Kim, J.; Eddaoudi, M.; Yaghi, O. M. Metal−Organic Frameworks Constructed from Pentagonal Antiprismatic and Cuboctahedral Secondary Building Units. *Chem. Commun.* **2001**, 2534−2535.

(13) Phan, A.; Czaja, A. U.; Gándara, F.; Knobler, C. B.; Yaghi, O. M. Metal−Organic Frameworks of Vanadium as Catalysts for Conversion of Methane to Acetic Acid. *Inorg. Chem.* **2011**, *50*, 7388−7390.

(14) Park, S.; Kim, B.; Choi, S.; Boyd, P. G.; Smit, B.; Kim, J. Text Mining Metal−Organic Framework Papers. *J. Chem. Inf. Model.* **2018**, *58*, 244−251.

(15) Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent. Sci.* **2018**, *4*, 268−276.

(16) Segler, M. H. S.; Kogej, T.; Tyrchan, C.; Waller, M. P. Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks. *ACS Cent. Sci.* **2018**, *4*, 120−131.

(17) Kadurin, A.; Aliper, A.; Kazennov, A.; Mamoshina, P.; Vanhaelen, Q.; Khrabrov, K.; Zhavoronkov, A. The Cornucopia of Meaningful Leads: Applying Deep Adversarial Autoencoders for New Molecule Development in Oncology. *Oncotarget* **2017**, *8*, 10883−10890.

(18) Jin, W.; Barzilay, R.; Jaakkola, T. *Junction Tree Variational Autoencoder for Molecular Graph Generation. ArXiv180204364 Cs Stat* 2018, http://arxiv.org/abs/1802.04364.

(19) Elton, D. C.; Boukouvalas, Z.; Fuge, M. D.; Chung, P. W. Deep Learning for Molecular Design−a Review of the State of the Art. *Mol. Syst. Des. Eng.* **2019**, *4*, 828.

(20) Dimitrov, T.; Kreisbeck, C.; Becker, J. S.; Aspuru-Guzik, A.; Saikin, S. K. Autonomous Molecular Design: Then and Now. *ACS Appl. Mater. Interfaces* **2019**, *11*, 24825.

(21) Reaxys, Elsevier Life Sciences, 2019; https://www.reaxys.com/#/search/quick.

(22) *ChemDraw Prime 15*; PerkinElmer Informatics, Inc.: Waltham, MA, 1998.

(23) Kim, S.; Thiessen, P. A.; Bolton, E. E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B. A.; Wang, J.; Yu, B.; Zhang, J.; Bryant, S. H. PubChem Substance and Compound Databases. *Nucleic Acids Res.* **2016**, *44*, D1202−D1213.

(24) Batten, S. R.; Champness, N. R.; Chen, X.-M.; Garcia-Martinez, J.; Kitagawa, S.; Öhrström, L.; O'Keeffe, M.; Suh, M. P.; Reedijk, J. Coordination Polymers, Metal−Organic Frameworks and the Need for Terminology Guidelines. *CrystEngComm* **2012**, *14*, 3001−3004.

(25) Batten, S. R.; Champness, N. R.; Chen, X.-M.; Garcia-Martinez, J.; Kitagawa, S.; Öhrström, L.; O'Keeffe, M.; Paik Suh, M.; Reedijk, J. Terminology of Metal−Organic Frameworks and Coordination Polymers (IUPAC Recommendations 2013). *Pure Appl. Chem.* **2013**, *85*, 1715−1724.

DOI: 10.1021/acs.cgd.9b01050
Cryst. Growth Des. 2019, 19, 6682−6697

(26) O'Keeffe, M.; Peskov, M. A.; Ramsden, S. J.; Yaghi, O. M. The Reticular Chemistry Structure Resource (RCSR) Database of, and Symbols for, Crystal Nets. *Acc. Chem. Res.* **2008**, *41*, 1782−1789.
(27) Bonneau, C.; O'Keeffe, M.; Proserpio, D. M.; Blatov, V. A.; Batten, S. R.; Bourne, S. A.; Lah, M. S.; Eon, J.-G.; Hyde, S. T.; Wiggin, S. B.; Öhrström, L. Deconstruction of Crystalline Networks into Underlying Nets: Relevance for Terminology Guidelines and Crystallographic Databases. *Cryst. Growth Des.* **2018**, *18*, 3411−3418.
(28) Donaruma, L. G.; Block, B. P.; Loening, K. L.; Plate, N.; Tsuruta, T.; Buschbeck, K. C.; Powell, W. H.; Reedijk, J. Nomenclature for Regular Single-Strand and Quasi Single-Strand Inorganic and Coordination Polymers (Recommendations 1984): International Union of Pure and Applied Chemistry (IUPAC) Macromolecular Division, Commission on Macromolecular Nomenclature Inorganic Chemistry Division, Commission on Nomenclature of Inorganic Chemistry. *Polym. Sci. U.S.S.R.* **1986**, *28*, 1240−1260.
(29) Adams, N.; Murray-Rust, P. Engineering Polymer Informatics: Towards the Computer-Aided Design of Polymers. *Macromol. Rapid Commun.* **2008**, *29*, 615−632.
(30) Warr, W. A. Representation of Chemical Structures. *WIREs Comput. Mol. Sci.* **2011**, *1*, 557−579.
(31) Zhang, T.; Li, H.; Xi, H.; Stanton, R. V.; Rotstein, S. H. HELM: A Hierarchical Notation Language for Complex Biomolecule Structure Representation. *J. Chem. Inf. Model.* **2012**, *52*, 2796−2806.
(32) Kim, C.; Chandrasekaran, A.; Huan, T. D.; Das, D.; Ramprasad, R. Polymer Genome: A Data-Powered Polymer Informatics Platform for Property Predictions. *J. Phys. Chem. C* **2018**, *122*, 17575−17585.
(33) CSD One Million - The Cambridge Crystallographic Data Centre (CCDC), https://www.ccdc.cam.ac.uk/csd-1-million/.
(34) Taylor, R.; Wood, P. A. A Million Crystal Structures: The Whole Is Greater than the Sum of Its Parts. *Chem. Rev.* **2019**, *119*, 9427.
(35) Bruno, I. J.; Cole, J. C.; Edgington, P. R.; Kessler, M.; Macrae, C. F.; McCabe, P.; Pearson, J.; Taylor, R. New Software for Searching the Cambridge Structural Database and Visualizing Crystal Structures. *Acta Crystallogr., Sect. B: Struct. Sci.* **2002**, *58*, 389−397.
(36) Moghadam, P. Z.; Li, A.; Wiggin, S. B.; Tao, A.; Maloney, A. G. P.; Wood, P. A.; Ward, S. C.; Fairen-Jimenez, D. Development of a Cambridge Structural Database Subset: A Collection of Metal−Organic Frameworks for Past, Present, and Future. *Chem. Mater.* **2017**, *29*, 2618−2625.
(37) Alexandrov, E. V.; Blatov, V. A.; Kochetkov, A. V.; Proserpio, D. M. Underlying Nets in Three-Periodic Coordination Polymers: Topology, Taxonomy and Prediction from a Computer-Aided Analysis of the Cambridge Structural Database. *CrystEngComm* **2011**, *13*, 3947.
(38) Blatov, V. A.; Shevchenko, A. P.; Proserpio, D. M. Applied Topological Analysis of Crystal Structures with the Program Package ToposPro. *Cryst. Growth Des.* **2014**, *14*, 3576−3586.
(39) TopCryst, https://topcryst.com/.
(40) Heller, S. R.; McNaught, A.; Pletnev, I.; Stein, S.; Tchekhovskoi, D. InChI, the IUPAC International Chemical Identifier. *J. Cheminf.* **2015**, *7*. DOI: 10.1186/s13321-015-0068-4
(41) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Model.* **1988**, *28*, 31−36.
(42) Weininger, D.; Weininger, A.; Weininger, J. L. SMILES. 2. Algorithm for Generation of Unique SMILES Notation. *J. Chem. Inf. Model.* **1989**, *29*, 97−101.
(43) O'Keeffe, M.; Yaghi, O. M. Deconstructing the Crystal Structures of Metal−Organic Frameworks and Related Materials into Their Underlying Nets. *Chem. Rev.* **2012**, *112*, 675−702.
(44) Li, M.; Li, D.; O'Keeffe, M.; Yaghi, O. M. Topological Analysis of Metal−Organic Frameworks with Polytopic Linkers and/or Multiple Building Units and the Minimal Transitivity Principle. *Chem. Rev.* **2014**, *114*, 1343−1370.
(45) Coupry, D. E.; Addicoat, M. A.; Heine, T. Extension of the Universal Force Field for Metal−Organic Frameworks. *J. Chem. Theory Comput.* **2016**, *12*, 5215−5225.
(46) Ockwig, N. W.; Delgado-Friedrichs, O.; O'Keeffe, M.; Yaghi, O. M. Reticular Chemistry: Occurrence and Taxonomy of Nets and Grammar for the Design of Frameworks. *Acc. Chem. Res.* **2005**, *38*, 176−182.
(47) Barthel, S.; Alexandrov, E. V.; Proserpio, D. M.; Smit, B. Distinguishing Metal−Organic Frameworks. *Cryst. Growth Des.* **2018**, *18*, 1738−1747.
(48) Warr, W. A. Many InChIs and Quite Some Feat. *J. Comput.-Aided Mol. Des.* **2015**, *29*, 681−694.
(49) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An Open Chemical Toolbox. *J. Cheminf.* **2011**, *3*, 33.
(50) Cordero, B.; Gómez, V.; Platero-Prats, A. E.; Revés, M.; Echeverría, J.; Cremades, E.; Barragán, F.; Alvarez, S. Covalent Radii Revisited. *Dalton Trans.* **2008**, 2832−2838.
(51) Planas, N.; Mondloch, J. E.; Tussupbayev, S.; Borycz, J.; Gagliardi, L.; Hupp, J. T.; Farha, O. K.; Cramer, C. J. Defining the Proton Topology of the Zr6-Based Metal−Organic Framework NU-1000. *J. Phys. Chem. Lett.* **2014**, *5*, 3716−3723.
(52) Chen, B.; Ockwig, N. W.; Millward, A. R.; Contreras, D. S.; Yaghi, O. M. High H2 Adsorption in a Microporous Metal−Organic Framework with Open Metal Sites. *Angew. Chem., Int. Ed.* **2005**, *44*, 4745−4749.
(53) Lin, X.; Jia, J.; Zhao, X.; Thomas, K. M.; Blake, A. J.; Walker, G. S.; Champness, N. R.; Hubberstey, P.; Schröder, M. High H2 Adsorption by Coordination-Framework Materials. *Angew. Chem., Int. Ed.* **2006**, *45*, 7358−7364.
(54) Yaghi, O. M.; O'Keeffe, M.; Ockwig, N. W.; Chae, H. K.; Eddaoudi, M.; Kim, J. Reticular Synthesis and the Design of New Materials. *Nature* **2003**, *423*, 705−714.
(55) Kalmutzki, M. J.; Hanikel, N.; Yaghi, O. M. Secondary Building Units as the Turning Point in the Development of the Reticular Chemistry of MOFs. *Sci. Adv.* **2018**, *4*, eaat9180.
(56) Kim, J.; Chen, B.; Reineke, T. M.; Li, H.; Eddaoudi, M.; Moler, D. B.; O'Keeffe, M.; Yaghi, O. M. Assembly of Metal-Organic Frameworks from Large Organic and Inorganic Secondary Building Units: New Examples and Simplifying Principles for Complex Structures. *J. Am. Chem. Soc.* **2001**, *123*, 8239−8247.
(57) Denysenko, D.; Grzywa, M.; Tonigold, M.; Streppel, B.; Krkljus, I.; Hirscher, M.; Mugnaioli, E.; Kolb, U.; Hanss, J.; Volkmer, D. Elucidating Gating Effects for Hydrogen Sorption in MFU-4-Type Triazolate-Based Metal−Organic Frameworks Featuring Different Pore Sizes. *Chem. - Eur. J.* **2011**, *17*, 1837−1848.
(58) Delgado-Friedrichs, O.; O'Keeffe, M. Identification of and Symmetry Computation for Crystal Nets. *Acta Crystallogr., Sect. A: Found. Crystallogr.* **2003**, *59*, 351−360.
(59) Delgado-Friedrichs, O.; Hyde, S. T.; O'Keeffe, M.; Yaghi, O. M. Crystal Structures as Periodic Graphs: The Topological Genome and Graph Databases. *Struct. Chem.* **2017**, *28*, 39−44.
(60) Leach, A. R.; Gillet, V. J. *An Introduction to Chemoinformatics*; Springer: Dordrecht, 2007.
(61) Drefahl, A. CurlySMILES: A Chemical Language to Customize and Annotate Encodings of Molecular and Nanodevice Structures. *J. Cheminf.* **2011**, *3*, 1.
(62) Schneider, N.; Sayle, R. A.; Landrum, G. A. Get Your Atoms in Order−An Open-Source Implementation of a Novel and Robust Molecular Canonicalization Algorithm. *J. Chem. Inf. Model.* **2015**, *55*, 2111−2120.
(63) Tropsha, A.; Williams, A. How Many Miles Have We Gone, InChI by InChI? *Chemistry International − Newsmagazine for IUPAC* **2012**, *34*.
(64) Southan, C. InChI in the Wild: An Assessment of InChIKey Searching in Google. *J. Cheminf.* **2013**, *5*, 10.
(65) Mbue, S. P.; Cho, K.-H. Identification of Isomers of Organometallic Compounds. *Bull. Korean Chem. Soc.* **2015**, *36*, 1569−1574.
(66) O'Boyle, N. M. Towards a Universal SMILES Representation - A Standard Method to Generate Canonical SMILES Based on the InChI. *J. Cheminf.* **2012**, *4*, 22.

DOI: 10.1021/acs.cgd.9b01050
Cryst. Growth Des. 2019, 19, 6682−6697

(67) Krenn, M.; Häse, F.; Nigam, A.; Friederich, P.; Aspuru-Guzik, A. *SELFIES: ARobust Representation of Semantically Constrained Graphs with an Example Application in Chemistry*. *arXiv* 2019, arXiv:1905.13741.
(68) Manion, C.; Arlitt, R.; Campbell, M. I.; Tumer, I.; Stone, R.; Greaney, P. A. Automated Design of Flexible Linkers. *Dalton Trans* **2016**, *45*, 4338−4345.
(69) Martin, R. L.; Lin, L.-C.; Jariwala, K.; Smit, B.; Haranczyk, M. Mail-Order Metal−Organic Frameworks (MOFs): Designing Isoreticular MOF-5 Analogues Comprising Commercially Available Organic Molecules. *J. Phys. Chem. C* **2013**, *117*, 12159−12167.
(70) Bao, Y.; Martin, R. L.; Simon, C. M.; Haranczyk, M.; Smit, B.; Deem, M. W. In Silico Discovery of High Deliverable Capacity Metal−Organic Frameworks. *J. Phys. Chem. C* **2015**, *119*, 186−195.
(71) Bao, Y.; Martin, R. L.; Haranczyk, M.; Deem, M. W. In Silico Prediction of MOFs with High Deliverable Capacity or Internal Surface Area. *Phys. Chem. Chem. Phys.* **2015**, *17*, 11962−11973.
(72) Witman, M.; Ling, S.; Anderson, S.; Tong, L.; Stylianou, K. C.; Slater, B.; Smit, B.; Haranczyk, M. In Silico Design and Screening of Hypothetical MOF-74 Analogs and Their Experimental Synthesis. *Chem. Sci.* **2016**, *7*, 6263−6272.
(73) Foscato, M.; Venkatraman, V.; Occhipinti, G.; Alsberg, B. K.; Jensen, V. R. Automated Building of Organometallic Complexes from 3D Fragments. *J. Chem. Inf. Model.* **2014**, *54*, 1919−1931.
(74) Ioannidis, E. I.; Gani, T. Z. H.; Kulik, H. J. molSimplify: A Toolkit for Automating Discovery in Inorganic Chemistry. *J. Comput. Chem.* **2016**, *37*, 2106−2117.
(75) Janet, J. P.; Gani, T. Z. H.; Steeves, A. H.; Ioannidis, E. I.; Kulik, H. J. Leveraging Cheminformatics Strategies for Inorganic Discovery: Application to Redox Potential Design. *Ind. Eng. Chem. Res.* **2017**, *56*, 4898−4910.
(76) Grajciar, L.; Heard, C. J.; Bondarenko, A. A.; Polynski, M. V.; Meeprasert, J.; Pidko, E. A.; Nachtigall, P. Towards Operando Computational Modeling in Heterogeneous Catalysis. *Chem. Soc. Rev.* **2018**, *47*, 8307−8348.
(77) Quirós, M.; Gražulis, S.; Girdzijauskaitė, S.; Merkys, A.; Vaitkus, A. Using SMILES Strings for the Description of Chemical Connectivity in the Crystallography Open Database. *J. Cheminf.* **2018**, *10*, 23.
(78) Heller, S. [InChI-Discuss] InChI Organometallics RFP, 2019; https://sourceforge.net/p/inchi/mailman/message/36602177/.
(79) Anderson, R.; Gómez-Gualdrón, D. A. Increasing Topological Diversity during Computational "Synthesis" of Porous Crystals: How and Why. *CrystEngComm* **2019**, *21*, 1653−1665.
(80) Chung, Y. G. Public Release of the Computation-Ready, Experimental (CoRE) Metal-Organic Frameworks Database. http://gregchung.github.io/CoRE-MOFs/, 2019.
(81) James, C. A. OpenSMILES Specification. http://opensmiles.org/opensmiles.html, 2016.
(82) Technical FAQ — InChI Trust. https://www.inchi-trust.org/technical-faq-2/.
(83) Morris, W.; Volosskiy, B.; Demir, S.; Gándara, F.; McGrier, P. L.; Furukawa, H.; Cascio, D.; Stoddart, J. F.; Yaghi, O. M. Synthesis, Structure, and Metalation of Two New Highly Porous Zirconium Metal−Organic Frameworks. *Inorg. Chem.* **2012**, *51*, 6443−6445.
(84) Feng, D.; Gu, Z.-Y.; Li, J.-R.; Jiang, H.-L.; Wei, Z.; Zhou, H.-C. Zirconium-Metalloporphyrin PCN-222: Mesoporous Metal−Organic Frameworks with Ultrahigh Stability as Biomimetic Catalysts. *Angew. Chem., Int. Ed.* **2012**, *51*, 10307−10310.
(85) Scalfani, V. F. IUPAC SMILES+ Specification (Project No. 2019-002-2-024). https://iupac.org/projects/project-details/?project_nr=2019-002-2-024, 2019.
(86) Day, N. E. *Automated Analysis and Validation of Open Chemical Data*. Thesis, University of Cambridge, 2009.
(87) Chung, Y. G.; Gómez-Gualdrón, D. A.; Li, P.; Leperi, K. T.; Deria, P.; Zhang, H.; Vermeulen, N. A.; Stoddart, J. F.; You, F.; Hupp, J. T.; Farha, O. K.; Snurr, R. Q. In Silico Discovery of Metal-Organic Frameworks for Precombustion CO_2 Capture Using a Genetic Algorithm. *Sci. Adv.* **2016**, *2*, e1600909.
(88) Wilmer, C. E.; Leaf, M.; Lee, C. Y.; Farha, O. K.; Hauser, B. G.; Hupp, J. T.; Snurr, R. Q. Large-Scale Screening of Hypothetical Metal−Organic Frameworks. *Nat. Chem.* **2012**, *4*, 83−89.
(89) NCI/CADD Chemical Identifier Resolver, https://cactus.nci.nih.gov/chemical/structure.
(90) Heller, S. *A Short History of the IUPAC InChI Algorithm*, 2019; http://www.hellers.com/steve/pub-talks/orlando-4-19.pdf.
(91) Munroe, R. Xkcd: Standards. https://xkcd.com/927/.
(92) O'Boyle, N. M.; Morley, C.; Hutchison, G. R. Pybel: A Python Wrapper for the OpenBabel Cheminformatics Toolkit. *Chem. Cent. J.* **2008**, *2*, 5.
(93) *Open Babel - Pull Request #1853 Implement Periodic Boundary Conditions*. *Open Babel*, 2019; https://github.com/openbabel/openbabel/pull/1853.
(94) Garberoglio, G. OBGMX: AWeb-based Generator of GROMACS Topologies for Molecular and Periodic Systems Using the Universal Force Field. *J. Comput. Chem.* **2012**, *33*, 2204−2208.
(95) Chamberlin, D. D.; Boyce, R. F. *SEQUEL: A Structured English Query Language*. Proceedings of the 1974 ACM SIGFIDET (Now SIGMOD) Workshop on Data Description, Access and Control. New York, NY, USA, 1974; pp 249−264.
(96) *Daylight Theory Manual*, 2011; http://www.daylight.com/dayhtml/doc/theory/.
(97) Abriata, L. A. Web Apps Come of Age for Molecular Sciences. *Informatics* **2017**, *4*, 28.
(98) Zimmermann, N. E. R.; Horton, M. K.; Jain, A.; Haranczyk, M. Assessing Local Structure Motifs Using Order Parameters for Motif Recognition, Interstitial Identification, and Diffusion Path Characterization. *Front. Mater.* **2017**, *4*, 34.
(99) Artemova, S.; Jaillet, L.; Redon, S. Automatic Molecular Structure Perception for the Universal Force Field. *J. Comput. Chem.* **2016**, *37*, 1191−1205.
(100) *Materials Studio*; Accelrys Software Inc.: San Diego, CA, 2001.
(101) Isayev, O.; Oses, C.; Curtarolo, S.; Tropsha, A. *Universal Fragment Descriptors for Predicting Electronic Properties of Inorganic Crystals*. ArXiv160804782 Cond-Mat, 2016, https://arxiv.org/abs/1608.04782.
(102) Tranchemontagne, D. J.; Mendoza-Cortés, J. L.; O'Keeffe, M.; Yaghi, O. M. Secondary Building Units, Nets and Bonding in the Chemistry of Metal−Organic Frameworks. *Chem. Soc. Rev.* **2009**, *38*, 1257.
(103) Clark, A. M. Accurate Specification of Molecular Structures: The Case for Zero-Order Bonds and Explicit Hydrogen Counting. *J. Chem. Inf. Model.* **2011**, *51*, 3149−3157.
(104) Bruno, I. J.; Shields, G. P.; Taylor, R. Deducing Chemical Structure from Crystallographically Determined Atomic Coordinates. *Acta Crystallogr., Sect. B: Struct. Sci.* **2011**, *67*, 333−349.
(105) Sezginel, K. B.; Feng, T.; Wilmer, C. E. Discovery of Hypothetical Hetero-Interpenetrated MOFs with Arbitrarily Dissimilar Topologies and Unit Cell Shapes. *CrystEngComm* **2017**, *19*, 4497−4504.
(106) Sholl, D. S.; Lively, R. P. Defects in Metal−Organic Frameworks: Challenge or Opportunity? *J. Phys. Chem. Lett.* **2015**, *6*, 3437−3444.
(107) DeStefano, M. R.; Islamoglu, T.; Garibay, S. J.; Hupp, J. T.; Farha, O. K. Room-Temperature Synthesis of UiO-66 and Thermal Modulation of Densities of Defect Sites. *Chem. Mater.* **2017**, *29*, 1357−1361.
(108) Rösler, C.; Fischer, R. A. Metal−Organic Frameworks as Hosts for Nanoparticles. *CrystEngComm* **2015**, *17*, 199−217.
(109) Whitford, C. L.; Stephenson, C. J.; Gómez-Gualdrón, D. A.; Hupp, J. T.; Farha, O. K.; Snurr, R. Q.; Stair, P. C. Elucidating the Nanoparticle−Metal Organic Framework Interface of Pt@ZIF-8 Catalysts. *J. Phys. Chem. C* **2017**, *121*, 25079−25091.
(110) Kitao, T.; Zhang, Y.; Kitagawa, S.; Wang, B.; Uemura, T. Hybridization of MOFs and Polymers. *Chem. Soc. Rev.* **2017**, *46*, 3108−3133.

DOI: 10.1021/acs.cgd.9b01050
Cryst. Growth Des. 2019, 19, 6682−6697

(111) Lollar, C. T.; Qin, J.-S.; Pang, J.; Yuan, S.; Becker, B.; Zhou, H.-C. Interior Decoration of Stable Metal−Organic Frameworks. *Langmuir* **2018**, *34*, 13795−13807.
(112) Deria, P.; Mondloch, J. E.; Karagiaridi, O.; Bury, W.; Hupp, J. T.; Farha, O. K. Beyond Post-Synthesis Modification: Evolution of Metal−Organic Frameworks via Building Block Replacement. *Chem. Soc. Rev.* **2014**, *43*, 5896−5912.
(113) Côté, A. P.; Benin, A. I.; Ockwig, N. W.; O'Keeffe, M.; Matzger, A. J.; Yaghi, O. M. Porous, Crystalline, Covalent Organic Frameworks. *Science* **2005**, *310*, 1166−1170.
(114) Diercks, C. S.; Yaghi, O. M. The Atom, the Molecule, and the Covalent Organic Framework. *Science* **2017**, *355*, eaal1585.
(115) Carlucci, L.; Ciani, G.; Proserpio, D. M. Polycatenation, Polythreading and Polyknotting in Coordination Network Chemistry. *Coord. Chem. Rev.* **2003**, *246*, 247−289.
(116) Jiang, H.; Jia, J.; Shkurenko, A.; Chen, Z.; Adil, K.; Belmabkhout, Y.; Weselinski, L. J.; Assen, A. H.; Xue, D.-X.; O'Keeffe, M.; Eddaoudi, M. Enriching the Reticular Chemistry Repertoire: Merged Nets Approach for the Rational Design of Intricate Mixed-Linker Metal−Organic Framework Platforms. *J. Am. Chem. Soc.* **2018**, *140*, 8858−8867.
(117) Kim, D.; Liu, X.; Lah, M. S. Topology Analysis of Metal−Organic Frameworks Based on Metal−Organic Polyhedra as Secondary or Tertiary Building Units. *Inorg. Chem. Front.* **2015**, *2*, 336−360.
(118) Makal, T. A.; Yakovenko, A. A.; Zhou, H.-C. Isomerism in Metal−Organic Frameworks: "Framework Isomers. *J. Phys. Chem. Lett.* **2011**, *2*, 1682−1689.
(119) Bureekaew, S.; Schmid, R. Hypothetical 3D-Periodic Covalent Organic Frameworks: Exploring the Possibilities by a First Principles Derived Force Field. *CrystEngComm* **2013**, *15*, 1551.
(120) Karmakar, A.; Paul, A.; Pombeiro, A. J. L. Recent Advances on Supramolecular Isomerism in Metal Organic Frameworks. *CrystEngComm* **2017**, *19*, 4666−4695.
(121) Ferguson, A.; Liu, L.; Tapperwijn, S. J.; Perl, D.; Coudert, F.-X.; Van Cleuvenbergen, S.; Verbiest, T.; van der Veen, M. A.; Telfer, S. G. Controlled Partial Interpenetration in Metal−Organic Frameworks. *Nat. Chem.* **2016**, *8*, 250−257.
(122) Blatov, V. A.; Carlucci, L.; Ciani, G.; Proserpio, D. M. Interpenetrating Metal−Organic and Inorganic 3D Networks: A Computer-Aided Systematic Investigation. Part I. Analysis of the Cambridge Structural Database. *CrystEngComm* **2004**, *6*, 377−395.
(123) Liu, Y.; O'Keeffe, M.; Treacy, M. M. J.; Yaghi, O. M. The Geometry of Periodic Knots, Polycatenanes and Weaving from a Chemical Perspective: A Library for Reticular Chemistry. *Chem. Soc. Rev.* **2018**, *47*, 4642−4664.
(124) Amirjalayer, S.; Schmid, R. Conformational Isomerism in the Isoreticular Metal Organic Framework Family: A Force Field Investigation. *J. Phys. Chem. C* **2008**, *112*, 14980−14987.
(125) Gomez-Gualdron, D. A.; Gutov, O. V.; Krungleviciute, V.; Borah, B.; Mondloch, J. E.; Hupp, J. T.; Yildirim, T.; Farha, O. K.; Snurr, R. Q. Computational Design of Metal−Organic Frameworks Based on Stable Zirconium Building Units for Storage and Delivery of Methane. *Chem. Mater.* **2014**, *26*, 5632−5639.
(126) Øien-Ødegaard, S.; Shearer, G. C.; Wragg, D. S.; Lillerud, K. P. Pitfalls in Metal−Organic Framework Crystallography: Towards More Accurate Crystal Structures. *Chem. Soc. Rev.* **2017**, *46*, 4867−4876.
(127) Gutov, O. V.; Bury, W.; Gomez-Gualdron, D. A.; Krungleviciute, V.; Fairen-Jimenez, D.; Mondloch, J. E.; Sarjeant, A. A.; Al-Juaid, S. S.; Snurr, R. Q.; Hupp, J. T.; Yildirim, T.; Farha, O. K. Water-Stable Zirconium-Based Metal−Organic Framework Material with High-Surface Area and Gas-Storage Capacities. *Chem. - Eur. J.* **2014**, *20*, 12389−12393.
(128) Serre, C.; Millange, F.; Thouvenot, C.; Noguès, M.; Marsolier, G.; Louër, D.; Férey, G. Very Large Breathing Effect in the First Nanoporous Chromium(III)-Based Solids: MIL-53 or CrIII(OH)·[O2C-C6H4-CO2]·[HO2C-C6H4-CO2H]X·H2Oy. *J. Am. Chem. Soc.* **2002**, *124*, 13519−13526.
(129) Yuan, S.; Sun, X.; Pang, J.; Lollar, C.; Qin, J.-S.; Perry, Z.; Joseph, E.; Wang, X.; Fang, Y.; Bosch, M.; Sun, D.; Liu, D.; Zhou, H.-C. PCN-250 under Pressure: Sequential Phase Transformation and the Implications for MOF Densification. *Joule* **2017**, *1*, 806−815.
(130) Shivanna, M.; Yang, Q.-Y.; Bajpai, A.; Patyk-Kazmierczak, E.; Zaworotko, M. J. A Dynamic and Multi-Responsive Porous Flexible Metal−Organic Material. *Nat. Commun.* **2018**, *9*, 3080.
(131) Katsoulidis, A. P.; Antypov, D.; Whitehead, G. F. S.; Carrington, E. J.; Adams, D. J.; Berry, N. G.; Darling, G. R.; Dyer, M. S.; Rosseinsky, M. J. Chemical Control of Structure and Guest Uptake by a Conformationally Mobile Porous Material. *Nature* **2019**, *565*, 213.
(132) Gómez-Gualdrón, D. A.; Colón, Y. J.; Zhang, X.; Wang, T. C.; Chen, Y.-S.; Hupp, J. T.; Yildirim, T.; Farha, O. K.; Zhang, J.; Snurr, R. Q. Evaluating Topologically Diverse Metal−Organic Frameworks for Cryo-Adsorbed Hydrogen Storage. *Energy Environ. Sci.* **2016**, *9*, 3279−3289.
(133) Colón, Y. J.; Gómez-Gualdrón, D. A.; Snurr, R. Q. Topologically Guided, Automated Construction of Metal−Organic Frameworks and Their Evaluation for Energy-Related Applications. *Cryst. Growth Des.* **2017**, *17*, 5801−5810.
(134) Rosen, A. S.; Notestein, J. M.; Snurr, R. Q. Identifying Promising Metal−Organic Frameworks for Heterogeneous Catalysis via High-Throughput Periodic Density Functional Theory. *J. Comput. Chem.* **2019**, *40*, 1305−1318.
(135) Gražulis, S.; Daškevič, A.; Merkys, A.; Chateigner, D.; Lutterotti, L.; Quirós, M.; Serebryanaya, N. R.; Moeck, P.; Downs, R. T.; Le Bail, A. Crystallography Open Database (COD): An Open-Access Collection of Crystal Structures and Platform for World-Wide Collaboration. *Nucleic Acids Res.* **2012**, *40*, D420−427.
(136) Liu, Y.-Y.; Leus, K.; Grzywa, M.; Weinberger, D.; Strubbe, K.; Vrielinck, H.; Van Deun, R.; Volkmer, D.; Van Speybroeck, V.; Van Der Voort, P. Synthesis, Structural Characterization, and Catalytic Performance of a Vanadium-Based Metal−Organic Framework (COMOC-3). *Eur. J. Inorg. Chem.* **2012**, *2012*, 2819−2827.
(137) Mir, M. H.; Koh, L. L.; Tan, G. K.; Vittal, J. J. Single-Crystal to Single-Crystal Photochemical Structural Transformations of Interpenetrated 3 D Coordination Polymers by [2 + 2] Cycloaddition Reactions. *Angew. Chem., Int. Ed.* **2010**, *49*, 390−393.
(138) Zhu, N.; Lennox, M. J.; Düren, T.; Schmitt, W. Polymorphism of Metal−Organic Frameworks: Direct Comparison of Structures and Theoretical N2-Uptake of Topological Pto- and Tbo-Isomers. *Chem. Commun.* **2014**, *50*, 4207−4210.
(139) Keupp, J.; Schmid, R. TopoFF: MOF Structure Prediction Using Specifically Optimized Blueprints. *Faraday Discuss.* **2018**, *211*, 79−101.
(140) Müller, P.; Grünker, R.; Bon, V.; Pfeffermann, M.; Senkovska, I.; Weiss, M. S.; Feng, X.; Kaskel, S. Topological Control of 3,4-Connected Frameworks Based on the Cu2-Paddle-Wheel Node: Tbo or Pto, and Why? *CrystEngComm* **2016**, *18*, 8164−8171.
(141) Frahm, D.; Hoffmann, F.; Fröba, M. Two Metal−Organic Frameworks with a Tetratopic Linker: Solvent-Dependent Polymorphism and Postsynthetic Bromination. *Cryst. Growth Des.* **2014**, *14*, 1719−1725.
(142) Dau, P. V.; Tanabe, K. K.; Cohen, S. M. Functional Group Effects on Metal−Organic Framework Topology. *Chem. Commun.* **2012**, *48*, 9370−9372.

DOI: 10.1021/acs.cgd.9b01050
Cryst. Growth Des. 2019, 19, 6682−6697