The HEALED SBU library of chemically realistic building blocks for construction of hypothetical metal-organic frameworks

Marco Gibaldi, Ohmin Kwon, Andrew White, Jake Burner, Tom K. Woo*

Department of Chemistry and Biomolecular Science, University of Ottawa, 10 Marie Curie Private, Ottawa K1N 6N5, Canada

*email: twoo@uottawa.ca

ABSTRACT

Advancements in hypothetical metal-organic framework (hMOF) databases and construction tools have resulted in a rapidly expanding chemical design space for nanoporous materials. The bulk of these hypothetical structures are constructed using structural building units (SBUs) derived from experimental MOF structures, often collected from the CoRE-MOF database. Recent investigations into the state of these deposited experimental structures’ chemical accuracy identified an array of common structural errors—including omitted protons, missing counterions, and disordered structures. These structural errors propagate into the SBUs mined from experimental MOFs, culminating in inaccurate hMOF structures possessing net charges or missing atoms which were not accounted for previously. This work demonstrates how manual investigation was applied to diagnose structural errors in SBUs obtained from several popular hMOF construction tools and databases. An analysis of the prevailing errors discovered during the examination process is provided along with representative cases to aid with error detection in future studies involving SBU extraction and hMOF construction. A novel repair protocol was established and employed to generate a library of SBUs that are hand-examined and labeled with enhanced detail (HEALED). This repaired library of SBUs contains 952 inorganic SBUs and 568 organic SBUs ideally suited for the generation of hypothetical frameworks that are chemically accurate and properly charge labelled. Additionally, case studies following the effects of SBU errors on electrostatic potential-fitted charges and GCMC-simulated gas adsorption predictions are presented to highlight the significance of using chemically accurate hMOF structures exclusively in all screening efforts going forward.
INTRODUCTION

Metal-organic frameworks (MOFs) are a category of nanoporous materials which have captured extensive attention since their inception.\textsuperscript{1,2} The intrinsic modularity and porosity observed in this family of materials afford flexibility to explore a vast range of chemical and structural features, which has led to their consideration for numerous clean energy and catalytic applications.\textsuperscript{3} The porous MOF networks are formed by periodically-linked chemical building blocks—designated as structural building units (SBUs). A minimum of two multitopic SBUs, consisting of an inorganic metal cluster and an organic linker, assemble to create an individual MOF; however, more complex structures have been prepared by heterogeneous mixing of inorganic and/or organic SBUs.\textsuperscript{4–6} Thus, there exists a nearly inexhaustible set of structural combinations that may be theorized by combining vast libraries of inorganic clusters and organic SBUs available in the literature. The expanding MOF design space simultaneously presents a massive potential boon to strategic materials discovery, and an equally massive challenge for those with ambitions of thoroughly screening the space.

Modern principles of MOF research often revolve around what has been realized in experimental syntheses thus far. Synthetic interest in MOFs has soared in recent years, resulting in an exponential growth\textsuperscript{7,8} in the quantity of experimental MOFs structures deposited into the Cambridge Structural Database\textsuperscript{9} (CSD) each year. Many procedures have been well-established for the preparation and functionalization of MOF materials, with great attention being paid to the areas of modulated syntheses and post-synthetic modifications.\textsuperscript{3,10,11} Access to a wide variety of chemical substructures can be achieved from these experimental MOF structures, thereby leading MOFs being considered for a broad array of catalytic and industrial applications.\textsuperscript{10,12–14} Most notably, MOF materials have been investigated for several gas separation and gas adsorption processes.\textsuperscript{12,15–17} The high degrees of porosity and stability observed by MOFs render them suitable candidates for gas storage and separation materials—processes that are becoming increasingly important with the global need for the reduction of greenhouse gases, primarily CO\textsubscript{2}. Particular interest has focused on screening MOF materials for use in various carbon capture technologies such as pre-combustion, post-combustion, and direct air capture.\textsuperscript{12,17–19} MOFs have demonstrated a remarkable ability to selectively adsorb CO\textsubscript{2} and a number of other gaseous species (\textit{e.g.} SO\textsubscript{x}, NO\textsubscript{x}, H\textsubscript{2}S, CO, etc.),\textsuperscript{14} resulting in the establishment of several promising gas separation systems operating at the bench-scale and pilot-scale.\textsuperscript{20–24} Recently, CALF-20, a MOF which has
been synthesized in 1 tonne batches at ambient conditions, was implemented as the primary sorbent in a pilot scale separation unit that has captured more than 1 tonne of CO₂/day for over a year in a cement manufacturing facility.²⁵

Despite their frequent association with gas separation and capture, MOFs may also be applied to several catalytic processes—including photocatalysis,²⁶–²⁸ electrocatalysis,²⁸–³¹ and biocatalysis³²,³³—and analytical techniques, such as chemical and biological sensing applications.¹³,³⁴–³⁶ MOF materials have occupied diverse roles in many experimental catalytic studies, ranging from catalyst supports, synthetic precursors, substrate storage materials, and active catalysts.²⁶–²⁸,³⁰ Promising photocatalytic conversion of CO₂ to CO has been observed using UiO-66/67 type MOFs functionalized with semiconductors, photosensitizers, and metal nanoparticles (MNPs).²⁷ Evaluation of MOFs and their composites for various H₂-based energy applications, such as H₂ storage,³⁷,³⁸ photocatalytic H₂ production³⁹ and electrochemical H₂ production by water splitting,³¹,⁴⁰ has also garnered significant research attention in recent years. Additionally, enzyme-MOF composites and MOFs constructed from biological molecules—sometimes called “bioMOFs”—have been shown to possess clear potential in biomedical applications (e.g., selective drug delivery³²) and improved catalytic activity in the case of immobilized enzymes, such as catalases, hydrolases, and peroxidases.³³ Thus, while the field continues to advance through its adolescence, MOF research has experienced an explosive uptake by the scientific community and possess a unique capacity for development in an immense collection number of applications.

Prospects for the vast experimental MOF pool comprising tens of thousands of currently known structures are encouraging for the coming decades; however, it would require an impractical number of research-hours to manually explore each promising candidate for all its possible, relevant functions. To combat the combinatorial obstacles that arise from the modularity of MOFs, researchers have adopted high-throughput computational screening techniques to accelerate the materials discovery process.⁴¹,⁴² As the design space of MOF chemistry continues to develop exponentially year over year,⁴¹ it is vital that comprehensive workflows for experimental and computational studies are conceptualized and followed by the community at large. Precedents have been set for the high-throughput syntheses and characterization of porous zeolite materials through implementation of multi-autoclave apparatus and robotic apparatus.⁴²
Microflow reactors coupled to analytical techniques, such as gas chromatography/mass spectroscopy and fluorescence spectroscopy, were also used as a proofs-of-concept for high-throughput catalyst and photoluminescence testing of porous materials. Efficient, parallel characterization protocols are also well-established in the field, with examples including systems employing infrared spectroscopy on Pt-loaded zeolites, and those utilizing X-ray diffraction (XRD) on zeolites manufactured in microreactor systems. High-throughput syntheses and experimentation of porous materials, such as zeolites, has been developed for several decades, and these techniques possess the capacity to analyze hundreds of materials in relatively short timespans, however, this quantity represents a miniscule fraction of the MOF design space that may be imagined. Thus, a great deal of recent work on high-throughput screening of MOFs and other porous materials has been directed towards computational studies to evaluate materials at rates that are orders of magnitude greater than experimentation.

Growth in computational studies surrounding MOFs has mirrored the developments seen in the experimental works. Persistent gains in the power and cost of high-performance computing allow these techniques to properly tackle the combinatorial challenges posed by MOF modularity. Massive efforts towards the formation of open-source materials databases and experimental structure repositories are continually underway. Novel experimental MOF structures are generally deposited into the Cambridge Structural Database (CSD), which serves as an exceptional resource for extracting structures for screening efforts. CSD structures include over a million crystal structures across diverse material subclasses—the quality of which can vary depending on the accuracy of the experimental characterization protocol. As a result, several popular databases have sought to refine CSD entries into subsets consisting only of materials relevant to MOF research. One such subset being the Computation-Ready Experimental (CoRE) database, which sought to prepare experimental MOF structures for use in computational studies by removing uncoordinated solvent molecules, resolving structural disorder and restoring missing atoms in the CSD structures. Further, the CSD curates their own subset of MOF structures that has spawned many subsequent studies. Owing to the innate uncertainty in many characterization techniques employed to study experimental MOF crystal structures, the structures reported to the CSD are often missing protons and/or contain a high degree of disorder resulting in overlapping atoms and hyper-coordinated atoms. Automated schemes for the repair of these problems found in experimental structures have been reported and applied to create sets of
structures that are ideal for gas adsorption and separation studies. Examples include Watanabe et al. who computationally screened over 30 000 experimental MOFs for CO2/N2 separations,52 and Goldsmith et al. screening 20 000 MOF frameworks for hydrogen storage purposes.53 Additionally, databases of experimental MOFs have enabled the study of CH4/N2 separations,54 CO2/N2 separations,55–57 among numerous other applications.

After exhausting exploring the domain of experimentally synthesized MOFs, the next natural advance included the exploration of structures that could be theorized but had not yet been experimentally realized. Hypothetical metal-organic frameworks (hMOFs) represented a massive addition to the chemical and geometric space available for exploitation by high-throughput screenings.41,58 The construction of hMOFs leans on the interchangeability of structural building blocks to generate new MOF structures from libraries of known inorganic and organic SBUs. Various databases of hMOFs have thus far been generated based on representative libraries of hundreds of inorganic and organic SBUs. Wilmer et al.59 prepared a database containing more than 137 000 structures which has become a staple in high-throughput in silico screenings of MOFs for adsorption processes. Boyd et al. generated a vast library of over 300 000 hMOF structures60,61 which was datamined to identify strong CO2 binding sites. Further, Li et al.5 constructed over 10 000 hypothetical multivariate (i.e. containing mixed SBUs) MOFs and presented case studies towards their feasibility in CO2 capture. The synergy between hMOF databases and high-throughput materials screening remains potent, and the scope of these studies has expanded significantly in recent years.

Heightened interest in the theoretical generation of MOF structures led many researchers to prepare new procedures to increase the accessible range of materials. Several approaches have demonstrated the ability to build hMOFs from representative SBUs, including energy-minimization-based techniques such as the automated assembly of secondary building units (AASBU) method proposed by Mellot-Draznieks et al.62, as well as more geometrically-focused topological approaches59,63–68 that have surged into favour as of late. Within these geometric techniques, there exist two primary approaches to hMOF assembly from building blocks—labelled as “bottom-up” and “top-down” approaches. The former involves consecutive additions of building blocks pending formation of a complete periodic structure, while the latter selects a net topology which is used as a blueprint to arrange the structural building blocks into appropriate
Numerous open-source tools for generating periodic frameworks possessing varying net topologies\textsuperscript{72,73} from SBU databases continue to be proposed throughout the literature; examples of which include AuToGraFS,\textsuperscript{63} pormake,\textsuperscript{64,65} ToBaCCo,\textsuperscript{67} and TOBASSCO.\textsuperscript{68} The exact methodology of hMOF construction differs between tools, however, the inputs remain similar and the net benefit to the scientific community are equivalent.

In both approaches the SBU databases are derived from experimental MOF structures, and therefore they are subject to similar advantages and disadvantages as the experimental MOF databases. Building blocks extracted from experiment possess a stronger foundation in “real” chemistry, which indicates that their synthetic prospects may be enhanced. However, these SBUs rely on the chemical accuracy of the deposited MOF structure, which commonly exhibit issues with disorder and missing atoms as previously mentioned.\textsuperscript{74–76} A recent investigation by Chen \textit{et al.}\textsuperscript{74} revealed that thousands of the CoRE-2019\textsuperscript{49} structures possessed some combination of overlapping atoms, isolated atoms, and over-bonded or under-bonded carbon atoms. Comparable structural problems were identified by Keskin \textit{et al.} in their analysis of the simulated adsorption properties of CoRE and CSD MOFs with identical refcodes.\textsuperscript{75} This comparison identified many instances wherein MOFs were reported differently by the two databases with the principal differences involving missing hydrogen atoms, charge balancing ions, solvent molecules and/or other essential MOF components.\textsuperscript{75} The uncertainty caused by these observations warrants an audit of the experimentally derived SBUs used by hypothetical MOF generation tools to identify the presence of similar structural problems and to ensure that future studies explore chemically meaningful regions of the hypothetical design space.

This contribution will introduce the results of a manual investigation of SBU structures intended for use in the construction of hypothetical MOFs. A quick inspection of CoRE structures which were not flagged by Chen \textit{et al.}\textsuperscript{74} as being problematic revealed a considerable number of errors remaining in their chemical accuracy; including, but not limited to, flawed structure cleaning protocols that lead to the removal of critical ligands and counterions, missing protons or other atoms from the framework structure, and incorrect framework charge assignment. These structural errors in the experimental “parent” MOFs propagate into the resultant SBUs that are used to create new hypothetical MOFs by way of the aforementioned structure building algorithms. Even if there are no missing atoms, inaccurate charge assignment on a MOF framework can severely complicate
their use in computational studies and may produce misleading results (e.g. poor agreement between simulated and experimental properties). Therefore, it is vital that charged SBUs be properly labelled, such that they are combined with appropriate SBUs that ensure the overall framework is neutral, as is often assumed, when building and screening MOFs in high-throughput studies.

The following sections will outline the validation and repair processes of SBUs derived from experimental MOF structure databases and hypothetical MOF construction utilities. An updated library of chemically accurate SBUs, we call HEaled (Hand-Examined And Labeled with Enhanced Detail) SBUs, intended for hypothetical MOF construction will be labelled by their respective net charges and made available for public use. Typical errors discovered in experimental MOF structures will be explained, and a straightforward scheme addressing their identification and correction will be outlined fully. This work’s overall objective is to improve the accuracy of the hypothetical MOF construction and screening space by repairing these extensive issues discovered in experimentally derived SBUs.

**METHODS**

**Manual SBU Inspection**

A representative set of 554 inorganic SBUs and 568 organic SBUs were selected from popular topological hMOF construction algorithms and tools, such as pormake\textsuperscript{64,65}, ToBaCCo\textsuperscript{67}, and TOBASSCO\textsuperscript{68}, and several hMOF databases.\textsuperscript{5,69,70,77–81} Further, additional inorganic SBUs were mined from the CSD-MOF\textsuperscript{51} and CoRE-2019\textsuperscript{49,50} experimental databases that were determined to possess diverse metal chemistry based on their calculated revised autocorrelation descriptors.\textsuperscript{58} To account for the widespread issues in the experimental MOFs from which these SBUs are typically mined, manual inspection of each inorganic and organic SBU was required to ensure their chemical accuracy. Visualization of the structures was performed using the tools available within the CCDC’s *Mercury* software.\textsuperscript{82} Determination of an individual building block’s feasibility relied heavily on general chemical knowledge using methods taught in a typical undergraduate chemistry program. For example, oxidation state assignment\textsuperscript{83} and electron counting methods were employed for inorganic SBU species, and bond valence methods\textsuperscript{84} were employed for organic SBUs. Oxidation state assignment and electron counting were executed using ionic approximations\textsuperscript{83} at each distinct metal site within a given inorganic SBU. Structures
that observed conventional oxidation states and bonding interactions were deemed reasonable and added to the SBU library without modification (although even these structures were scrutinized further, as discussed later). Conversely, SBUs were flagged as potentially problematic if they contained any of the following abnormalities: (i) an impossible oxidation state (e.g., Ti(VI), Al(IV), etc.), (ii) a possible, but unlikely oxidation state (e.g., Co(V), Zn(III), etc.), (iii) unsaturated non-metal atoms, and/or (iv) missing or disconnected atom sites. Any of these occurrences suggested the presence of an inaccurate structure or charged SBU, likely due to omission of a key component from the experimental structure data (e.g., missing protons, counterions, ligands, etc.).

All SBUs flagged during the initial, manual inspection were then further examined to classify the exact nature of their structural problem. When available, the CSD$^9$ entry of the corresponding parent MOF was examined to identify a plausible cause for any irregularities in the SBU structure. Furthermore, the primary publication—from which the SBU was derived—was routinely consulted to detect any features which may not appear in the deposited experimental structure. Unsurprisingly, observations made in our experimental MOF investigation and previous works$^{74–76}$ discussing persistent structural issues in experimental MOF databases (e.g., CoRE-2019$^{49}$ and CSD$^9$) held true throughout the manual SBU inspection. As previously seen, the most common issues constituted improper net charge assignment caused by the removal of counterions or long-range charge sharing interactions, improper protonation, and/or missing or mislabelled atoms. These specific structural errors can be largely attributed to difficulties experimentally detecting proton positions and the preprocessing steps used to prepare experimental MOFs for computational studies (e.g., solvent removal$^{49,50}$). Once the explicit cause of the irregularity was determined, a procedure—described later in this work—was implemented to repair each SBU structure prior to its addition to the updated HEALED SBU library. For a minor handful of cases, the exact cause of the erroneous SBU behaviour could not be ascertained, thereby requiring removal of the SBU in question from the library entirely. Furthermore, certain SBUs were determined to possess such a high degree of inaccuracy in their original state that manual re-extraction from the corresponding experimental parent MOF structure was required to yield an agreeable structure.
Computational Details

DFT Calculations: All DFT calculations and geometry optimizations were performed with VASP\textsuperscript{85} version 5.4.4 using the Perdew-Burke-Ernzerhof (PBE)\textsuperscript{86} exchange and correlation functional. Core electron interactions and regions were treated by Blöchl’s projector-augmented wave (PAW)\textsuperscript{87} method with a plane wave cut-off of 400.0 eV. As part of the SBU repair process, some SBUs were found to have missing atoms. Only the “missing” atoms and the capping hydrogen atom positions were optimized while all other atoms remained fixed to their original positions. These positions were relaxed to obtain optimized structures with an electronic and ionic relaxation convergence criterion of $10^{-5}$ eV and 0.02 eV/Å, respectively. Geometry optimizations were performed on isolated SBU molecules contained in cubic simulation cells large enough to fully enclose all atoms with a large vacuum space. When the SBU is removed from the periodic MOF structure, the unfilled valence of the atom(s) at the connection points were capped using hydrogen atoms. Only gamma-point sampling of the Brillouin zone was performed on each SBU structure that underwent geometry optimization.

DFT calculations of periodic MOF structures were performed to determine electrostatic potential fitted point charges using the REPEAT method.\textsuperscript{88} A 3x3x3 Monkhorst-Pack sampling of K-points in the Brillouin zone was used for all MOFs whose smallest cell vector length was less than 14 Å. For MOFs whose cell vectors were all greater than 14 Å only the Γ-point was sampled. A random sampling of 3000 hMOFs from the Boyd-Woo database\textsuperscript{60,61} whose minimum cell-vector length was between 13.5-14.0 Å, showed that the REPEAT charges calculated with a Γ-point sampling had a mean absolute deviation (MAD) from those determined with a 3x3x3 sampling of k-points of only 0.0024 e. The same comparison on a random sampling of 3000 MOFs whose minimum cell-vector length was less than 7 Å was found to have a MAD of 0.0156 e, which is an order of magnitude higher. This shows that the Γ-point only sampling of MOFs with cell vectors greater than 14 Å is justified.

Hypothetical MOF Generation: The open-source MOF constructor \textit{TOBASCCO}\textsuperscript{68} was applied to generate two demonstrative groups of hypothetical MOFs: the first group constructed from the corrected SBUs we present in this publication, while the other group was constructed from the original, unchanged SBUs. A representative set of organic and inorganic SBUs featuring noteworthy edits of the most prevalent errors (\textit{i.e.} missing protons and incorrect charge
assignment) were selected to illustrate the importance of chemical accuracy and charge balance in MOF structure generation and analysis. Within each pair of original and HEALED hMOFs studied, the sole difference was the state of repair for one of the selected SBUs while all other factors in hMOF construction—specifically topology and choice of organic linker—were held constant. These isoreticular, hypothetical MOF pairs were applied to all comparative analyses described hereafter.

**GCMC Simulations**: Grand Canonical Monte Carlo (GCMC) simulations were performed using an internally developed code known as FastMC. Single-component adsorption isotherms were calculated at 298.0 K for CO$_2$, CH$_4$, and N$_2$ guest molecules within a subset of generated hypothetical MOFs. Simulated isotherms were calculated using 30 000 cycles—comprising equally probable guest additions, deletions, translations, and rotations as permissible MC steps—during both the equilibration and production phases. Each calculation was performed using a minimum $3 \times 3 \times 2$ supercell using a Lennard-Jones cut-off radius of 12.5 Å. Fugacity corrections were implemented using the Peng-Robinson equation of state (EOS). Non-bonded interactions between hMOF framework and adsorbate atoms were modelled using the Lennard-Jones (LJ) potential. LJ parameters employed for framework atoms were acquired from the Universal Force Field, and Lorenz-Berthelot mixing rules were used to determine appropriate parameters for interactions between different atom types. Transferable force fields designed for the simulation of MOFs and porous materials were applied to model the adsorbate molecules’ (CO$_2$, CH$_4$-TraPPE, N$_2$-NIMF) non-bonded interactions. Partial atomic charges were implemented to model electrostatic interactions between framework and adsorbate molecules. Electrostatic potential-fitting via the REPEAT method was applied to calculate DFT-derived partial atomic charges for all hMOF framework atoms, while adsorbate molecules derived partial atomic charges directly from their respective force field definitions. The Ewald summation method was used to facilitate convergence during the calculation of long-range electrostatic potentials.

**Diversity Analysis**: To assess the chemical design space represented within the HEALED library, a comparison between the metal chemistry diversity in inorganic HEALED SBUs and experimental MOF structures was performed. Revised autocorrelation (RAC) descriptors—as developed by Kulik et al. and implemented in molSimplify—were calculated to examine the near-metal environments within all experimental MOFs (CSD-MOF and CoRE-2019) and
inorganic HEALED SBUs. All five available metal-centric RACs (e.g. nuclear charge, electronegativity, topology, identity, and atomic radius) were employed with the maximum possible depth (i.e. up to three bonded atoms away from the metal center) during descriptor calculation. The t-stochastic neighbor embedding (t-SNE) method was employed to reduce the multidimensional RAC descriptor space to a two-dimension feature space that allows for more facile visualization. The t-SNE algorithm implemented within the sci-kit learn Python package was applied to complete these diversity analyses. Similar chemical structures become clustered upon analysis of their RACs by t-SNE, which supports the identification of similarities and disparities between the experimental MOF and HEALED SBU datasets.

RESULTS & DISCUSSION

1520 HEALED SBUs—comprised of 952 inorganic SBUs and 568 organic SBUs—were gathered from existing hypothetical MOF and experimental MOF databases, such as CSD-MOF and CoRE-2019. In total, 554 inorganic SBUs and 568 organic SBUs were collected from established hMOF construction tools and databases—many of which were constructed using SBUs extracted from the CoRE MOF database themselves. The remaining 398 inorganic SBUs were extracted directly from a sample of CoRE and CSD MOFs possessing diverse metal chemistry, as determined by their respective RAC descriptors. Each HEALED SBU was manually inspected for structural errors, and the original publication from which they were first characterized was consulted to ensure maximum accuracy. In the sections that follow, we discuss the utilized error identification protocol along with a detailed analysis of the nature and incidence rates of the errors found. Subsequently, we discuss how various categories of structural errors were corrected using a novel repair protocol. We then provide an examination into the effects of these structural errors on results obtained from characteristic molecular simulations. Finally, we analyze the composition and diversity of the HEALED SBU dataset to estimate its effectiveness in representing the experimental MOF chemical design space.

SBU Error Identification: Table 1 presents a summary of common errors identified throughout the manual inspection of organic and inorganic SBUs extracted from experimental MOF structures. Overall, significantly larger rates of error incidence were observed for inorganic building blocks (51.3%) compared to their organic counterparts (13.6%). The bonding motifs within organic linkers and functional groups are generally more well-defined and experimentally
parameterized than their inorganic counterparts. Errors within organic moieties are often more easily identifiable by visual inspection or automated structure validation algorithms, which may provide some justification for the significant discrepancy we observed in the organic and inorganic SBU structural error rates. The dominant observed errors involved improper net charge assignments or missing atoms—commonly protons—compared to the original published structure. Incorrect charge assignment was found to be the most prevalent error, occurring in 36.7% of inorganic SBUs and 10.7% of organic SBUs. Within experimental MOF structures, charged moieties are usually present as non-coordinating counterions inside the pore network (e.g. ammonium, perchlorate, halogens, etc.), and/or as additional charged functional groups or transition metal complexes present within the framework itself. The latter case most frequently involves charge sharing between heterogeneous inorganic sites, but this chemical information may be lost during MOF deconstruction into its constituent SBUs if the interdependent sites become disconnected. Non-coordinating counterions were often correctly found in the original experimental MOF’s characterization; however, these species were regularly mistaken as solvent during preprocessing procedures intended to clean structures for subsequent computations, and hence these counterions were discarded from the “cleaned” experimental MOF framework. While some of the resultant resulting structures are properly labelled as charged, many cleaned frameworks are falsely labelled as neutral. Naturally, we found that these charge assignment errors in the parent MOF structures propagated into its component SBUs if appropriate care was not taken to balance metal and ligand charges during building block extraction.

Table 1. Breakdown of errors identified during the manual inspection of organic and inorganic SBUs derived from experimental MOF structures.

<table>
<thead>
<tr>
<th>SBU Category</th>
<th>Total SBU Count</th>
<th>Errors Identified</th>
<th>Error Type</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Missing protons</td>
<td>Incorrect charge assignment</td>
</tr>
<tr>
<td>Inorganic</td>
<td>952</td>
<td>488 (51.3 %)</td>
<td>201 (21.1 %)</td>
</tr>
<tr>
<td>Organic</td>
<td>568</td>
<td>77 (13.6 %)</td>
<td>10 (1.8 %)</td>
</tr>
</tbody>
</table>
Missing protons were found to be the second-most common source of error in experimentally derived SBUs, with the problem being detected in 21.1% of inorganic and 1.8% of organic SBUs. It is well known that XRD methods typically used to experimentally characterize the structure of MOFs maintain a limited ability to directly elucidate proton positions. Interestingly, this investigation predominantly detected this error type in inorganic building blocks, though the existence of proton omission errors in the organic fragments of experimental MOF frameworks has been conclusively identified by Chen et al. This observation may suggest that experimental corrections to X-ray determined structures which place protons into their expected positions, such as algorithms based on standard bond valence rules and X-H bond lengths, have generally been successful in correcting proton omissions in the organic portions of experimental MOFs. This finding may also indicate that previous efforts in hypothetical MOF construction studies and database creation have carefully selected their organic building blocks in such a way that avoids these proton errors. Omission of protons in the deposited crystallographic information was found to affect metal-bound water, hydroxide, azolate, and/or carboxylate species primarily during our investigation. Diagnosis of absent protons was relatively straightforward as the offending ligands’ contribution to the metal’s oxidation state generally pushed it to an unlikely and/or impossible value.

Illustrative examples of the counterion and proton-induced errors are provided in Figure 1a/b and 1c/d, respectively. The examples provided in Figure 1a/b highlight CoRE-2019 MOFs missing crucial charge-balancing ions—specifically imidazolium cations in the case of DUNXUH04_clean and methylsulfonate anions in WOLMUG_clean. In both instances, the experimental crystal structure contained the necessary counterions to yield a neutral MOF framework which indicates that the ions were removed during the cleaning protocols used to prepare CoRE MOFs for computation. Figure 1c/d feature cases where protons are missing from metal-bound oxygen atoms in CoRE MOFs. Unresolved protons were observed on the bridging water ligands in Fig. 1c (XOTXAG_clean), and the terminal water and hydroxide ligands in Fig. 1d (PAPVAF_freeONLY). Generally, most proton omission errors were also found in the original, deposited crystal structure which supports the argument that these errors are introduced by the experimental crystal structure determination methods’ inability to efficiently resolve all protons positions.
Figure 1. Examples of common errors discovered during the manual validation of experimentally derived SBUs, and their parent MOF structures. The 3D, ball-and-stick models correspond to a representative problematic SBU, below which is the matching parent MOF chemical diagram included with the CSD entry. Frames (a) and (b) demonstrate improper charge assignment errors, while frames (c) and (d) demonstrate missing proton errors—both of which produce SBUs possessing net charges.
Figure 2. Examples of uncommon errors discovered during the manual validation of experimentally derived SBUs, and their parent MOF structures. Frames (a) and (b) present instances where missing atoms were identified in the deposited experimental structures. Frames (c) and (d) provide examples where long-range interactions (i.e. charge balancing or bridging between inorganic SBUs) may be obscured by the reductive nature of SBU extraction methods.

Other structural errors were detected during manual structure investigation, including mislabelled atoms, missing connection points, and missing or incorrect atom identity. These errors were substantially rarer than proton or counterion issues, representing only 1.9% and 1.2% of the total inorganic and organic building block errors, respectively. Figure 2a/b depict examples of deposited MOF structures with missing or incorrectly labelled atoms at one or more positions. Identification and repair of such uncommon errors required precise scrutiny of the original publication to confirm structural details. The error depicted in Fig. 2a relating to terminal oxygen ligands missing from molybdenum atoms—often observed for vanadium and tungsten MOFs as well—can likely be attributed to MOF cleaning protocols. This terminal oxygen ligand could easily be mistaken for deprotonated water molecules during the solvent removal process, leading
to their false elimination. An additional, minor subset of errors was attributed to shortcomings of the “cluster-based” SBU approach in describing particular bridging ligand groups. Highlighted in Figure 2c/d are two such cases where deconstruction of an experimental MOF cannot separate interdependent clusters without introducing a net charge to the resultant SBUs. Figure 2c demonstrates how long-range charge balancing between heterogeneous inorganic sites can be obscured in the SBU extraction process using the CoRE MOF IYICUP_clean as an example. In this example, the Co(III) atom is bound exclusively to neutral ligands, resulting in a localized positive charge of +3 around this inorganic SBU site. Conversely, the La(III) atom is ligated by a surplus of six carboxylate moieties which creates a localized negative charge of −3 at this SBU site. When both inorganic SBUS are present in a MOF framework with the correct proportions, the net SBU charges balance out to yield an overall neutral framework. The charge assignment errors are introduced when these two interdependent charged SBUs are extracted from the experimental structure without consideration for their charge balancing interaction. Many previous efforts in extracting SBUs from experimental MOFs did not carefully examine these charge balancing interactions; thus, they assumed that both the Co(III) and La(III) SBUs derived from IYICUP were neutral despite their net charges which became evident upon manual inspection. Figure 2d shows an analogous error wherein infinite bridging between identical inorganic sites cannot be simply represented by this SBU preparation method. Dicyanamide ligands bridge Nd atoms in this CoRE MOF (BIWYAH10_clean) in all three dimensions infinitely. Any attempt to extract an inorganic SBU from this type of structure will either fail to properly represent the charge on the dicyanamide ligand, or the SBU will continue on forever in one (of more) dimension due to the infinite chain of Nd atoms. Additional examples of these issues representing bridging ligands are exemplified in the CSD entries ADIQEL and QUQPIC (depicted in Figure S1). The vast majority of errors akin to the cases presented in Fig. 2c/d may be corrected by treating the adjacent, charge-bearing SBU cluster(s) as an intraframework counterion and implementing the required charge labelling and/or counterion additions; however, this solution may not accurately reflect the true nature of the original experimental structure due to the lack of these long-range charge-balancing interactions between heterogenous SBU clusters. Unfortunately, current implementations of top-down hMOF construction and SBU extraction algorithms are not designed to conveniently represent these classes of infinitely bridging SBUs. Accordingly, this category of
SBU
SUs were not the focus of this work though they may be included in future HEALED SBU updates if new tools allow for a more accurate representation of infinite chains or “pillared” SBUs.

The initial flagging of problem structures used in the manual inspection protocol assumes that oxidation state and charge counting can pinpoint most structural errors that are commonly found in inorganic and organic building units. All MOFs which were initially flagged in this way possessed structural errors as previously outlined. This suggests that the identification of impossible or unlikely metal oxidation states can be used as an indirect method of screening for problem MOF structures. However, it remains possible that SBUs with reasonable oxidation states and bonding may also have structural problems. An exhaustive examination of all SBUs was undertaken to detect how often the oxidation state-based criterion failed to identify a problem SBU structure. This additional screening found that only ~2% of SBUs with chemically reasonable oxidation states actually contained structural errors. Based on these results, we conclude here that oxidation state assignment is an effective way of identifying structural errors in SBUs and MOF structures.

**SBU Repair Process:** A comprehensive SBU repair protocol was established which concentrates on the predominant errors outlined above. Three main approaches were implemented to repair SBUs flagged with charge balance or atom omission errors, contingent on the origins of the error identified during manual structure inspection:

(i). In cases where the excess charge was balanced by non-coordinating counterions or long-range charge sharing in the experimental structure, SBU structure files were labelled according to their net charge. SBU .xyz files ending in _p(x) or _n(y) tags indicate the presence of a formal positive charge of | x | or formal negative charge of | y |, respectively. These charge labels allow for charge balancing efforts in the subsequent hMOF construction protocol.

(ii). In cases where missing protons were identified, protons were added back to the SBU structure according to the positions existing in the original CSD entry and/or publication. The positions of all new atoms were optimized at the DFT level (with the positions of the original atoms being fixed) as detailed earlier in the computational details section.
In cases where the identified SBU error(s) were more severe (e.g. missing charge-balancing ligands, missing organic linkers, missing non-hydrogen atoms, etc.), the original crystallographic information file of the parent MOF was used to extract a corrected version of the SBU. The new version of the SBU was then manually inspected, and appropriately edited (e.g. addition of charge labels, other missing atoms, etc.) to better reflect the published structure.

These SBU repair protocols were successfully applied to each category of error outlined previously. Figure 3 provides illustrations of the error identification and repair process applied to various distinct SBUs extracted from experimental MOF structures. Figure 3a demonstrates how missing proton errors were corrected for inorganic SBUs, while Figure 3b/c demonstrate how charge assignment errors were handled in both inorganic and organic SBUs, respectively. A comprehensive record of the errors identified, and the revisions made to each SBU is supplied in the Supporting Information. Further information and properties are also included to aid in the exploration of the SBU design space, such as chemical compositions, oxidation states, presence of solvents, metal oxidation state, and so forth. The final library of repaired structures contains 952 inorganic and 568 organic SBUs ready for hypothetical MOF construction applications. Each SBU was given a unique identifier (i.e. m1, o1, etc.) using the following naming conventions: (i) separation of inorganic and organic SBUs using the prefixes “m” or “o” to denote their respective “metal” or “organic” character; and (ii) sequential numbering of structures within each SBU category. The library contains a total of 410 charged SBUs comprised of 349 inorganic and 61 organic SBUs retaining a net charge. All charged species bear charge labels (i.e. m45_n2, o114_p1, etc.) to support further charge accounting and balancing efforts, which may include the addition of counterions (e.g. alkali metal cations, ammonium cations, halogen anions, etc.) and/or selection of a counteracting, charged organic/inorganic SBUs. The remaining SBUs were deemed neutral by the manual survey of each parent MOFs’ structure, and their associated CSD and literature entries. When constructing hMOFs, neutral frameworks are generally desired as this greatly simplifies associated simulations, particularly those involving the prediction of gas adsorption properties. Assuming that charge counting is handled properly using the provided SBU charge labels, any frameworks constructed from HEALED SBUs will possess the expected framework charge with a high degree of confidence.
Figure 3. Demonstration of the SBU repair protocol applied to three SBUs derived from CoRE MOF\textsuperscript{49,50} structures. Example (a) describes the repair of an inorganic SBU with missing proton errors. Examples (b) and (c) demonstrate the repair of charge assignment errors for an inorganic and organic SBU, respectively. The magenta “atoms” constitute the connection points contained in all SBUs to facilitate hMOF construction. Comprehensive details of the SBU repair protocol are provided in the Methodology section, and details concerning the specific repair(s) applied to each SBU are provided in the Supporting Information.

Any SBUs that could not be repaired through one of the above-mentioned methods were not the focus of the present work and discarded from the SBU library. Only a handful of cases were discarded, generally involving experimental structures which possessed multiple possible paths to correction which could not be distinguished through examination of their original publications.
Further investigations may be capable of salvaging the discarded SBUs assuming that a more chemically accurate experimental “parent” MOF structure and/or an analogous, corrected SBU can be harvested from future experimental characterizations.

**Effect of Structural Errors on Computed Properties:** Two representative sets of hMOFs containing 20 MOFs each was created — one constructed from HEALED SBUs and the other from their original, uncorrected SBU counterparts. We will call these sets the **HEALED** hMOF set and the **uncorrected** hMOF set. The HEALED and uncorrected sets were used to evaluate the impact of the repairs presented in this work on the calculated partial atomic charges and gas adsorption properties. Periodic DFT calculations were performed on each structure to extract electrostatic potential (ESP) fitted charges using the REPEAT method. Some of the MOFs in the HEALED set possess a net framework charge, while the corresponding MOFs in the uncorrected set were assigned a neutral charge. A neutral charge was used in order to mimic the result of constructing a MOF with an incorrect charge assignment. For charged frameworks, the periodic DFT calculation was performed on the hMOF with the net framework charge being counterbalanced by a uniform background charge of the same magnitude and the charge fitting procedure was constrained to this net framework charge. The nature of these atomic charges may be used to evaluate changes in the changes to the electrostatic potential brought about by the repairs performed to generate HEALED SBUs. A comparison of REPEAT charges calculated for the HEALED and flawed hMOF sets is provided in Figure 4. Atomic charge distributions for all elements in the two hMOF sets are given in Figure 4a. The uncorrected and HEALED charge distributions exhibit many similarities in their profiles, though shifts in the magnitude and position of several peaks can be easily recognized. These deviations may be partially explained by the addition of new, repaired atoms in the HEALED structures’ distribution; however, it was suspected that charges of existing atoms near the repair sites should also be affected considerably by the repairs.

To evaluate this assertion, the atomic charges calculated prior to, and post-repair were superimposed onto HEALED inorganic SBU structures to examine how the repaired sites are affected. Two such hMOF structures where proton and charge assignment errors in inorganic SBUs were handled are depicted in Figures 4b and 4c, respectively. The example in Figure 4b illustrates a Tb-carboxylate building block that was originally missing protons from its terminal
water ligands. The computed REPEAT charges at these affected terminal oxygen positions varied significantly following HEALED repairs, observing shifts in their atomic charges upwards of 44%. Unsurprisingly, the atomic charges calculated at the Tb atoms were also observed to change alongside this shift in their ligands’ atomic charges. The added atoms’ impact on the calculated ESP, and the ensuing partial charge fitting via the REPEAT method, could be expected by most chemists upon inspection of each hMOF pair’s apparent chemical differences. Conversely, the effect that having the wrong net framework charge on the partial atomic charges may not be as intuitive. The Zn-carboxylate SBU pictured in Figure 4c exemplifies the effect of charge assignment errors on the resulting ESP-fitted atomic charges. If an hMOF containing this SBU was assumed to be neutral (as in the original SBUs’ previous applications), required adjustments would not be employed during quantum-mechanical ESP calculations or adsorption simulations. An excess negative charge is introduced the Zn(COO)₄ cluster due to the surplus of carboxylate ligands, which ultimately amounts to a net charge of –2 per SBU unit. To account for this net charge, the charge labels provided with HEALED SBUs are used to adjust the number of electrons in the quantum-mechanical calculations such that they accurately reflect each frameworks’ overall charge. This key difference between the uncorrected and HEALED SBUs assuredly produces discrepancies in their ensuing hMOFs’ atomic charges and ESP since changes to the number of electrons fundamentally alters the system. The atomic charges shown in Fig 4c observe shifts in excess of 28% for the carboxylate oxygen atoms and 4% for the Zn atoms upon correcting the number of valence electrons in the DFT calculation. These observed discrepancies in atomic charge calculated for uncorrected and HEALED SBUs suggests that the electrostatic potential was significantly impacted by the structural errors described in this work. Moreover, we suggest that these atomic charge differences would be highly correlated with the adsorption simulation results, particularly for more polar adsorbate molecules such as CO₂.
Figure 4. Demonstration of the effect of SBU repairs on DFT-derived charges. Panel (a) shows the charge distributions for all atoms present in the HEALED hMOF set (green line) and the corresponding uncorrected hMOF set (red dashed line) where each set contains 20 MOFs. Panels (b) and (c) show the calculated charges on key atoms on two selected hMOFs, where the bracketed charges belong to the uncorrected hMOF set and those outside the brackets are charges for the corresponding HEALED hMOF. The MOF shown in (b) corresponds to an example where the uncorrected MOF is missing protons, while the MOF shown in (c) is a case where the uncorrected MOF is assigned a neutral framework charge whereas it should have an anionic framework charge.

To test this assertion and further assess the impact of the executed SBU repairs, GCMC simulations were performed on the previously described sets of hMOFs. Again, these simulations focused on qualitatively assessing the influence of the most common HEALED SBU repairs, namely proton omission and incorrect charge assignment errors. In instances where a MOF in the HEALED set possessed a net framework charge, a sufficient number of monovalent counterions (i.e. Na\(^+\) and Cl\(^-\)) were added to neutralize the framework charge. Comparisons of the CO\(_2\), N\(_2\), and CH\(_4\) adsorption behaviour of each pair of hMOFs constructed from the uncorrected and repaired SBU categories are provided in Figure 5. Figure 5a-d demonstrate the adsorption isotherms for hMOFs pairs that are identical in all aspects other than the state of repair of their inorganic SBUs—m36, m450, m45, and m195, respectively. The cases detailed in Figure 5a & 5b feature Mn\(_4\)-tetrazolate and Zn\(_2\)-carboxylate inorganic SBUs that were previously missing protons from terminal water positions. Figure 5c & 5d highlight examples of charged Cd\(_2\)-carboxylate and Zn\(_1\)-carboxylate SBUs that were previously marked as neutral despite net charge accumulation caused by the removal of needed counterions. While the isotherm forms generally remain
unchanged, meaningful discrepancies are observed in the simulated CO\textsubscript{2} uptake in most comparisons between the repaired and original SBUs. The differences in GCMC-simulated uptake appears most pronounced at high adsorbate loadings, typically becoming noticeable at pressures greater than 1.0 bar. A lesser degree of divergence is generally observed in the simulated N\textsubscript{2} and CH\textsubscript{4} uptakes for all pairs of hMOFs presented. This is predictable since electrostatics are expected to play a smaller role in the binding of N\textsubscript{2} and CH\textsubscript{4} as compared to CO\textsubscript{2}. However, in the case of some charged hMOFs, such as the example shown in Figure 5d, greater differences could be observed upon the introduction of counterions to the GCMC simulation. Additionally, similar to the CO\textsubscript{2} adsorption profiles, the differences in N\textsubscript{2} and CH\textsubscript{4} adsorption become increasingly evident as we approach higher adsorbate loadings.

Figure 5. Sample adsorption isotherms demonstrating the effect of HEALED SBU repairs on GCMC simulated uptake. Frames (a)-(d) contain GCMC-simulated adsorption isotherms at 298.0 K for CO\textsubscript{2} (red), N\textsubscript{2} (green), and CH\textsubscript{4} (black) on isoreticular pairs of hMOFs constructed from both the original, uncorrected SBUs (dashed line) and repaired HEALED SBUs (solid line)—constituting m36, m450, m45, and m195, respectively. Frames (a) and (b) compare the simulated uptakes for SBUs which observed proton omission errors prior to repair, while (c) and (d) compare SBUs which observed incorrect charge assignment errors.
Isobaric comparisons of the GCMC-simulated adsorption of CO₂, N₂, and CH₄ on all pairs of hMOFs constructed from the uncorrected and HEALED SBUs are presented in Figure 6a-c—depicting adsorption at 0.1, 1.0, and 5.0 bar, respectively. Adsorption values for N₂ and CH₄ remain close to parity for nearly all pairs, with a slight skew towards higher uptakes calculated for the repaired SBU species at greater adsorbate loadings. The disparities between original and repaired SBUs become much more apparent when examining their simulated CO₂ uptake, which observe substantial deviations from parity across a wider range of adsorbate loadings. The extent of these uptake disparities varied widely between individual MOF pairs as certain materials observed several fold differences in uptake, while others only experienced minor uptake changes between the original and repaired hMOFs. One possible explanation proposes that the extent (i.e. number of added atom positions) and location (i.e. effect of added atoms on local pore environment) of the repair may contribute to how drastically the SBU repair alters the resultant framework structure and its adsorption properties. Other factors in the hMOF construction, such as the choice of organic SBU and topology, will contribute to pore architecture which ultimately affects the adsorption characteristics as well. While this modest case study is not sufficient to quantify the impact of HEALED SBU repairs on the GCMC-simulated adsorption for the entire population of possible hypothetical frameworks; these initial findings validate that the state of SBU repair can have a meaningful effect in specific cases. Thus, employing repaired SBUs will be vital in future high-throughput screenings to prevent results that are not reflective of the true experimental MOF chemistry and to limit drawing false conclusions about the materials’ effectiveness towards gas adsorption applications.

The discrepancies in uncorrected and HEALED hMOFs’ adsorption behaviour appears to agree with observations made in recent study from Keskin et al. that compared the adsorption simulation results for MOFs sourced from two “cleaned” databases—the CoRE-2019 and CSD MOF subsets. Their study looked at entries corresponding to the same published MOF material, ultimately discovering that a significant proportion of these pairs contained differing chemical compositions in their crystallographic information due to their distinct structure processing and solvent removal algorithms. Investigations into their MOF pairs with unmatched formulas identified several classes of structural errors (i.e. missing protons, removal of counterions, etc.) that concurred with the observations made in this work related to the SBUs derived from the two databases at issue. Importantly, they also concluded that the choice of database influenced
conclusions drawn from subsequent GCMC adsorption simulations as the gas uptake was severely impacted by the removal of certain atoms/groups from the framework or certain bulky ion groups from the pore structure. A subsequent study by Keskin and coworkers further corroborated that discrepancies in the presence of coordinated solvent molecules and charge-balancing ions significantly affected the identification of top performing materials in experimental MOF screenings. Though our contribution’s GCMC simulation case study was limited in scope, it suggests agreement with recent literature concerning the effect of structural differences on adsorptive uptake.

![Figure 6](image)

**Figure 6.** Isobaric plots comparing the GCMC simulated uptakes for CO₂ (red), N₂ (green), and CH₄ (black) at 298.0 K for a representative set of isoreticular MOFs constructed from HEALD SBUs and their uncorrected SBU counterparts. Comparison of simulated uptakes are displayed at adsorbate pressures of (a) 0.1 bar, (b) 1.0 bar, and (c) 5.0 bar.

**Composition and Diversity of the HEALED SBUs:** A characterization of the metal chemistry available within the HEALED SBU library is depicted in Figure 7. The CoRE MOF and CSD MOF databases, which we will collectively call the experimental MOF sets, were employed as benchmarks to represent the overall scope and metal diversity of experimental MOFs. Metal chemistry diversity was targeted in this analysis due of the high error rates observed in inorganic SBUs and their limited exploration in previous hMOF databases. The local chemical environment around metal centers in the experimental and HEALED structures were characterized using RAC descriptors, consisting of autocorrelation sums of several key properties (i.e. nuclear charge,
electronegativity, topology, identity, and atomic radius) over atoms neighbouring a given metal center. The following analysis assumes that inorganic SBUs possessing increasingly similar RAC descriptor values correspond to similar structures, and that they will be clustered together during t-SNE dimensionality reduction. Those assertions were largely corroborated by manual comparison of experimental MOFs and HEALED SBUs found to be in close proximity in Figure 7. Upon initial inspection of the relative dispersion of the two datasets depicted in Figure 7, it appears that the experimental MOF set occupies a greater design space than the HEALED library. Experimental MOF structures maintain a much greater area in the resulting t-SNE plot, though this conclusion may be magnified by the fact that the experimental dataset outnumbers HEALED SBUs by over twenty-fold.

**Figure 7.** Overlaid t-SNE plots of the revised autocorrelation (RAC) descriptors to characterize the metal chemistry diversity in HEALED SBU (red) compared to CoRE & CSD MOF (grey) structures. Histograms depicting probability density across the reduced t-SNE space for both structure subsets are also provided along both axes to aid with comparison. Clustering of experimental MOFs in the plot which do not appear to be represented by any HEALED SBU are labelled and subject to additional inspections—the details of which are explained further in the discussion. Regions possessing magenta labels correspond to groups of experimental MOFs containing infinite or “pillared” SBUs, while regions with orange labels correspond to areas where expected structural errors (i.e. missing protons, incorrect charge assignment, etc.) were detected in a considerable proportion of experimental MOF structures.
There are many regions in the t-SNE plot where there is a high density of experimental MOFs (grey points), with no corresponding HEALED SBUs (red points), many of which are circled and labelled A thru K in Figure 7. These regions suggest there may be important or highly represented metal chemistry found in the experimental MOFs that are not available in the HEALED SBUs. However, some of the diversity observed in the experimental MOF sets may be exaggerated by the presence of structural errors such as those that have been established within this work. The absence of essential atoms, such as protons, from the experimental MOF would affect the calculated RAC descriptors as they rely on connectivity within each local environment. Similarly, the presence of disordered atom sites, missing ligands, or other errors within the crystallographic information could produce similar effects in these descriptors’ calculations. Structural errors within members of a given family of SBUs (e.g. Cu paddlewheel, MOF-5, etc.) would falsely indicate the presence of many distinct environments despite them all truly representing identical experimental metal chemistry. Further inspection of clusters in Figure 7 that were represented exclusively in the experimental MOF set confirmed this suspicion. For example, region A contained numerous experimental MOF structures with Zn-paddlewheel inorganic SBUs that were missing protons from terminal water ligands, as exemplified in the CoRE MOFs PURJES_freeONLY and VUJBEI_freeONLY. Similarly, region B of Figure 7 contains a variety of Cu-paddlewheel inorganic SBUs, some of which possessed this same proton error in the terminal water positions—such as the CSD MOF mehmet_p1. In both of these cases, a related HEALED SBU without the erroneous solvent molecule can be identified—specifically m2 and m3 representing the unsolvated Cu-paddlewheel and Zn-paddlewheel SBUs, respectively. In fact, we found that many of these experimental MOF-exclusive regions corresponded to SBUs with structural errors, and these erroneous SBUs typically had a closely related HEALED SBU match. Therefore, the observation of these underrepresented regions of experimental MOFs does not significantly undermine the metal center diversity of HEALED SBUs. In fact, the high rates of error incidence indicated within our manual investigations and previous literature accounts suggest that structural errors may influence a considerable portion of the diversity observed by CoRE and CSD MOFs in Figure 7. The true experimental MOF diversity would likely be considerably reduced if structural errors were completely eliminated from its chemical design space.
Taking these challenges relating to the experimental MOF diversity into account, HEALED SBUs’ metal diversity appears to compare favourably and encompasses a wide range of the accessible experimental chemistry. Histograms of the calculated t-SNE features depicted on the axes of Fig. 7 can be employed to estimate the spreads of the two datasets of interest. Clear disparities can be observed in these distributions which suggests that the experimental MOF and HEALED SBU datasets are not perfect intersections of one another. Notably, the histograms along the y-axis of the t-SNE feature space notice the greatest differences, with approximately uniform and bimodal histogram shapes observed for experimental and HEALED sets, respectively. This observation indicates that the experimental MOFs represent the total available metal design space more evenly, while the majority of HEALED SBUs are more biased towards a narrower portion of the descriptor space. Biasing of the HEALED SBU space relative to experiment was expected owing to the elimination of false diversity caused by the structural errors. Additionally, as previously discussed during error identification, certain classes of SBUs (i.e. infinitely bridging or so-called “pillared” SBUs) are not present in the HEALED library due to limitations in contemporary top-down hMOF construction algorithms which fail to handle SBUs that are infinite in one (or more) dimensions. Several such examples of regions in Figure 7 representing this category of inorganic SBU in experimental MOFs are highlighted in the regions C – K: (i) region C contains Zn_{x}(COO)_{2x} SBU pillars (e.g. CoRE MOFs AVIMIC_clean and HEBTEP_clean); (ii) regions D & I contain various infinite phosphate and hydrogen-phosphate based MOFs (e.g. Be_{x}(HPO_{4})_{x} in VOXVEL_clean and Ga_{x}(PO_{4}) in FALZIC_clean); (iii) region E contains Sc_{x}(COO)_{3x} SBU pillars (e.g. BADHIA_clean and HOJNOL_clean); (iv) region F contains Mg_{x}(COO)_{x} SBU pillars (e.g. RAVWES_clean); (v) region G contains Cu_{x}(COO)_{2x} SBU pillars (e.g. CSD MOF mojpot_p1); and so forth. It is evident that these classes of SBUs can be found in many experimental MOF frameworks, which provides added justification for the observed reduction in chemical space of HEALED relative to the total experimental MOF set. Nonetheless, the comparable spread of these distributions suggest that the HEALED library maintains similar breadth across the design space despite containing fewer structures in certain sections of the range and possessing an order of magnitude fewer structures than its experimental counterparts. While the HEALED SBU library is clearly not an exhaustive collection of all possible experimental SBU structures, it possesses the most diverse collection of SBUs currently available for hMOF
A table summarizing common metal species in HEALED inorganic SBUs is also presented alongside the diversity analysis in Table 2. Trends in HEALED metal identity and oxidation state agree with expectations from experimental MOF structures, which frequently feature inorganic SBUs containing first row transition metals (e.g. Zn, Co, Cd, Cu, etc.). Owing to the undertaken repairs, oxidation states in HEALED SBUs observe greater agreement with the expected values for each metal. Additionally, unlikely (e.g. Co(IV), Cu(III)) and impossible (e.g. Ti(V), Na(II)) oxidation states that were often observed in erroneous experimental MOF structures have been eliminated from HEALED SBUs. Further analysis of the HEALED SBU library, including a full accounting of properties for each individual SBU, are provided in the Supporting Information.

Table 2. Summary of the 10 most common metal-oxidation state species within the HEALED SBU library.

<table>
<thead>
<tr>
<th>Metal (Oxidation State)</th>
<th>HEALED SBU Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zn(II)</td>
<td>159</td>
</tr>
<tr>
<td>Co(II)</td>
<td>125</td>
</tr>
<tr>
<td>Cd(II)</td>
<td>103</td>
</tr>
<tr>
<td>Cu(II)</td>
<td>84</td>
</tr>
<tr>
<td>Ni(II)</td>
<td>61</td>
</tr>
<tr>
<td>Cu(I)</td>
<td>46</td>
</tr>
<tr>
<td>Mn(II)</td>
<td>46</td>
</tr>
<tr>
<td>In(III)</td>
<td>26</td>
</tr>
<tr>
<td>Ag(I)</td>
<td>25</td>
</tr>
<tr>
<td>Fe(III)</td>
<td>17</td>
</tr>
</tbody>
</table>

The HEALED SBU library constitutes the most expansive collection of building blocks for hMOF construction available to date. Users can access well over 1600 distinct 3D topologies using the various SBU geometries found in this library. Given the vast quantity of accessible topologies
and broad SBU pool, we estimate that over 28 million distinct, neutral hMOF frameworks can be realized using the HEALED SBU libraries without functionalization or derivatization of the organic SBUs. This estimate extends even further if multiple inorganic and/or organic SBUs are considered during the hMOF construction. These developments represent a massive expansion to the realistic hMOF chemical space available to the researchers in the screening of MOFs for gas separation and adsorption applications.

CONCLUSIONS

In this work, a detailed manual examination of structural errors present in experimentally derived MOF structural building units was reported. An analysis of typical errors observed in these building blocks was presented, which confirmed that the predominant structural errors involved improper charge assignment and/or proton omission. This investigation uncovered that experimentally derived inorganic and organic SBUs observed substantial error rates of 51.3% and 13.6%, respectively, which doubtlessly culminated in inaccurate hypothetical MOF frameworks being utilized in previous studies. To remedy this obstacle, a protocol for the identification and repair of such errors was illustrated using SBU structures derived from many sources, including the CSD MOF database, CoRE 2019 MOF database, and several prominent hypothetical MOF construction tools/databases (e.g. pormake, TOBASSCO, etc.). Ultimately, the objective of this work pertained to the development of a library of SBUs—dubbed HEALED SBUs—which corrects inaccuracies in SBUs derived from experimental MOF structures and accounts for charge considerations in hypothetical MOF construction. A comparative analysis of the DFT-derived electrostatic potential fitted charges and GCMC-simulated adsorption properties for MOFs constructed from original and the corrected HEALED SBUs was performed to validate the importance of employing accurate building blocks in computational hMOF screenings. Examination of charges calculated for atoms near the structural error sites indicated that atomic charges were significantly impacted by SBU repairs. Adsorption simulations found that many SBUs with proton omission and charge assignment errors observed significant changes to their GCMC-simulated CO$_2$ uptake upon repair. Uptakes of N$_2$ and CH$_4$ were also examined, though fewer instances and lesser extents of deviation were detected when comparing the repaired SBU to its original. Discovery of these meaningful discrepancies between the calculated adsorption properties for many hMOF pairs indicated that failure to account for structural SBU errors may threaten the accuracy of the ensuing gas adsorption screenings. The magnitude of HEALED SBU
repairs’ influence on simulated adsorption behaviour was not comprehensively quantified; however, these initial findings suggest that caution should be exercised when evaluating hypothetical frameworks composed of SBUs in unknown states of repair.

The HEALED SBU library has been made publicly available in the Supporting Information with the complete chemical and connectivity information required by hypothetical MOF construction algorithms. The availability of corrected structures is expected to guide future studies towards the generation of chemically accurate frameworks, and to improve accessibility to the field of hypothetical MOF screening for various potential applications. As an additional consequence, these results may help bridge recorded gaps between experimental and computational structures and results, and further enhance consistency between data produced by each research paradigm. Ideally, these findings will serve as a warning to future studies to exercise caution as the MOF chemical design space continues to rapidly expand. The chemically accuracy and charge of the hypothetical frameworks must be considered to ensure relevance of the massive quantities of data being created in this research area.

Supporting Materials

The Supporting Information is made available free of charge:

Details relating to documentation provided alongside the HEALED SBU Library, and additional analysis of errors introduced during the SBU extraction process. (PDF)

Comprehensive record of HEALED SBUs including their source, composition, repair information, and so on. (XLSX)

Database of HEALED SBUs mol files available for use with TOBASCCO. (ZIP)

Acknowledgements

Financial support from the Natural Sciences and Engineering Research Council of Canada, the University of Ottawa, MITACS, and TotalEnergies is greatly appreciated, as well as the computing resources provided by the Canada Foundation for Innovation, TotalEnergies and Compute Canada.
REFERENCES


17. Stylianou, K. C. & Queen, W. L. Recent advances in carbon capture with Metal-organic


47. Klein, J., Lehmann, C. W., Schmidt, H. W. & Maier, W. F. Combinatorial material libraries on the microgram scale with an example of hydrothermal synthesis. *Angew.*


61. Boyd, P. G. *et al.* Data-driven design of metal–organic frameworks for wet flue gas CO2


95. Provost, B. An Improved N2 Model for Predicting Gas Adsorption in MOFs and Using Molecular Simulation to Aid in the Interpretation of SSNMR Spectra of MOFs. (University of Ottawa, 2014).


