scientific data

Check for updates

OPEN

DATA DESCRIPTOR

MOFSimplify, machine learning models with extracted stability data of three thousand metal–organic frameworks

Aditya Nandy[1,2,4], Gianmarco Terrones[1,4], Naveen Arunachalam[1], Chenru Duan[1,2], David W. Kastner[1,3] & Heather J. Kulik[1] ✉

We report a workflow and the output of a natural language processing (NLP)-based procedure to mine the extant metal–organic framework (MOF) literature describing structurally characterized MOFs and their solvent removal and thermal stabilities. We obtain over 2,000 solvent removal stability measures from text mining and 3,000 thermal decomposition temperatures from thermogravimetric analysis data. We assess the validity of our NLP methods and the accuracy of our extracted data by comparing to a hand-labeled subset. Machine learning (ML, i.e. artificial neural network) models trained on this data using graph- and pore-geometry-based representations enable prediction of stability on new MOFs with quantified uncertainty. Our web interface, MOFSimplify, provides users access to our curated data and enables them to harness that data for predictions on new MOFs. MOFSimplify also encourages community feedback on existing data and on ML model predictions for community-based active learning for improved MOF stability models.

Background & Summary

Metal–organic frameworks (MOFs) have reticular chemistry and well-defined, isolated metal sites[1,2] and are comprised of molecular secondary building units that can then impose directionality. This potential for exquisite control makes them promising for applications in gas adsorption[3,4], sensing[5,6], separations[7,8], and catalysis[9–14]. The modular nature of MOFs enables the design of hypothetical materials libraries amenable to virtual high throughput screening (VHTS) by combining distinct organic linkers, inorganic building blocks, and topologies to make a MOF[2,7,15–19]. The number of experimentally realizable MOFs with varying metals, linkers, and pore size has grown rapidly[20], despite challenges in synthesis[21,22] and post-synthetic modification[23]. After synthesis, MOFs must also undergo activation (i.e., solvent removal from pores) to enable their practical use. Despite advances in experimental methods[24,25] for MOF activation, many MOFs are unstable upon activation[22,26,27] and thus unusable[28]. For practical use as catalysts or functional materials, MOFs must also sustain their porosity and structural integrity at elevated temperatures[29–34].

VHTS efforts for screening hypothetical MOFs typically rely heavily on expert intuition for identifying candidate materials that are then synthesized[2,7,35]. Although heuristics such as pore size[36] or hard-soft acid base theory[37] for predicting metal–linker bond strength are frequently invoked to predict MOF stability, numerous exceptions exist, limiting the broad applicability of heuristics for stability prediction[36,38–41]. Rules for thermal stability derived from subsets of MOFs do not extrapolate well to new MOFs outside of those subsets[42]. Molecular mechanics models that are useful for VHTS with MOFs also cannot predict activation stability[19,43–45].

Limitations in using computation[19,43,44] or heuristics[36,37] to predict stability motivates data driven machine learning (ML) models trained on large experimental data sets. Gaining experimental solvent removal and thermal stability data in sufficient quantities to train ML models, however, remains a formidable challenge. Although a few studies have gathered experimental data from a single source[42,46] to reveal stability trends, the unified

[1]Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA, 02139, USA. [2]Department of Chemistry, Massachusetts Institute of Technology, Cambridge, MA, 02139, USA. [3]Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA, 02139, USA. [4]These authors contributed equally: Aditya Nandy, Gianmarco Terrones. ✉e-mail: hjkulik@mit.edu

efforts of thousands of researchers over multiple decades represents an untapped source of knowledge[47] for the factors that govern MOF stability. Natural language processing[48] (NLP) is a promising approach to leverage this data from the literature. Many studies have combined NLP of the extant literature with ML to identify synthesis conditions for inorganic materials[49–51]. NLP has been used to quantify the role of organic structure directing agents in governing zeolite topology[52]. However, a lack of systematic naming[53,54] in MOF chemistry (e.g., HKUST-1 and Cu-BTC are the same MOF) has limited the use of NLP-based named entity recognition for the design of new MOFs. While NLP has worked well for identifying MOF properties such as surface area through their unique units[55,56], human interpretation of the structure name is required to relate extracted properties back to the original structure[55,56]. Due to challenges in mapping MOF names to structures[53,54], coupled with the lack of unique units or measurements for stability assessments, no efforts have collated data on MOF stability.

We recently leveraged[57] the extant literature to identify how MOF linker and inorganic secondary building unit (SBU) composition as well as MOF connectivity govern MOF stability. We utilized NLP to curate stability-related experimental properties for structurally characterized MOFs. From the curated data, we trained artificial neural network (ANN) models that achieve high prediction accuracies for solvent removal stability (accuracy: 0.76, area under the receiver operating curve: 0.79) and thermal stability (mean absolute error: 47 °C). These models use revised autocorrelations[18,58] (RACs) as fingerprints for each MOF that are derived from the MOF's clean (e.g., without solvent or disorder) crystallographic information file (CIF). Models trained on this data revealed the importance of both linker and metal features, demonstrating why solely metal-based or linker-based heuristics fail to predict MOF stability. The power of NLP and automated extraction has also been recently demonstrated for predicting MOF synthesis recipes[59].

Here, we tabulate data on solvent removal stability for 2,179 structures and thermal stability for 3,132 structures of MOFs reported in the experimental literature. Our data set is the first source to map MOF experimental stabilities to well-defined experimental structures. We also provide representative linkers and SBUs from each structure, and the fingerprints we used to construct our data-driven models. We demonstrate how users can utilize our data sets, make predictions on new materials, or improve the quality of labels for our experimental stability data set. Our dataset and methodology will enable the curation of more reports of MOF stability, paving the way for the design of stable MOFs.

Methods

Data mining. The starting point for data curation was the all solvent removed portion of the 2019 Computation-ready, experimental (CoRE) MOF database v1.1.2, which contains 10,143 non-disordered structures[60]. Of this set, the 9,597 MOFs that were compatible with the generation of graph-based revised autocorrelation[18,58] (RAC) and geometric[61,62] features were retained for further filtering steps (see Data Records). A subset of 9,202 MOFs were sanitized[60] structures from the Cambridge Structural Database[63,64] (CSD) that could be associated with a unique CSD refcode. We used these refcodes to obtain the digital object identifier (DOI) of the manuscript associated with each structure in the CSD[63] v5.41, released in November 2019. In total, 8,809 refcodes had associated DOI entries, which corresponded to 5,152 unique manuscripts (Fig. 1).

We used the ArticleDownloader[65] package to automatically obtain manuscripts from the Royal Society of Chemistry (RSC), Wiley-VCH, the American Association for the Advancement of Science (AAAS), Springer, and Nature. Articles from the American Chemical Society (ACS) were obtained via a direct download agreement between ACS and the Massachusetts Institute of Technology. Through this procedure, 3,809 manuscripts were downloaded, which corresponded to 7,004 structures, in HTML or XML format for subsequent text extraction and parsing (Fig. 1). For the 1,343 manuscripts associated with 1,805 structures that could not be automatically downloaded in HTML or XML format, roughly two-thirds were only available as PDFs (938 manuscripts, 1,307 structures) and one-third could not be automatically downloaded (405 manuscripts, 498 structures).

Next, we used text parsing on our corpus to determine labels (i.e. unstable or stable) for the solvent removal stability of MOFs and to identify manuscripts that contain thermogravimetric analysis (TGA) data (Table 1). We tokenized all manuscripts into sentences using the ChemDataExtractor[66] package. Because information about MOF solvent removal and thermal stability does not appear in a specific experimental methods section (e.g., as is the case for synthesis), text search of the entire manuscript is necessary. We avoided false positives (i.e., from introductory text) by excluding sections labeled as introductions, and only analyzed the last 60% of the manuscript text for letters or communications that lacked section headers (Fig. 1).

Complex sentence structure limited the utility of sentiment-based models (e.g., VADER[67] sentiment) in identifying stability. We employed syntactic dependency parsing to extract labels for MOF solvent removal stability. First, we pattern matched (i.e., used regular expressions) for keywords pertaining to common MOF solvents, MOF structural integrity, and the process of MOF activation, identifying a set of sentences relating to activation stability. We used additional regular expressions to eliminate sentences relating to air or water stability, or activation processes that are not MOF activation (e.g., catalytic C–H activation). Next, we performed dependency parsing using the Stanza[68] NLP toolkit. Through dependency parsing, we analyzed pairwise mappings of words and disambiguated negations that are challenging to distinguish with regular expressions (e.g., "no crystallinity" vs. "no loss of crystallinity") for the manuscripts containing relevant sentences (2,649 out of 3,809). We then assigned each sentence a label of unstable (0) or stable (1). Because most manuscripts report on more than one MOF, we only assigned labels to manuscripts where all sentences had the same label (1,209 out of 2,649 manuscripts, Fig. 1). We then assigned all MOFs from a given labeled manuscript the text-mined manuscript label. Finally, we eliminated 111 MOFs that had identical connectivity (e.g., same RACs), but conflicting text-mined labels from different manuscripts. Some compounds can be identical in RAC representation but have distinct connectivity. We also calculate the Weisfeiller-Lehman graph hash for each of these 111 MOFs and determine that the majority (i.e., 66) have the same Weisfeiller-Lehman graph hash for the atomic-number attributed graph. Both RAC features and the Weisfeller-Lehman graph hash are provided for these cases in the

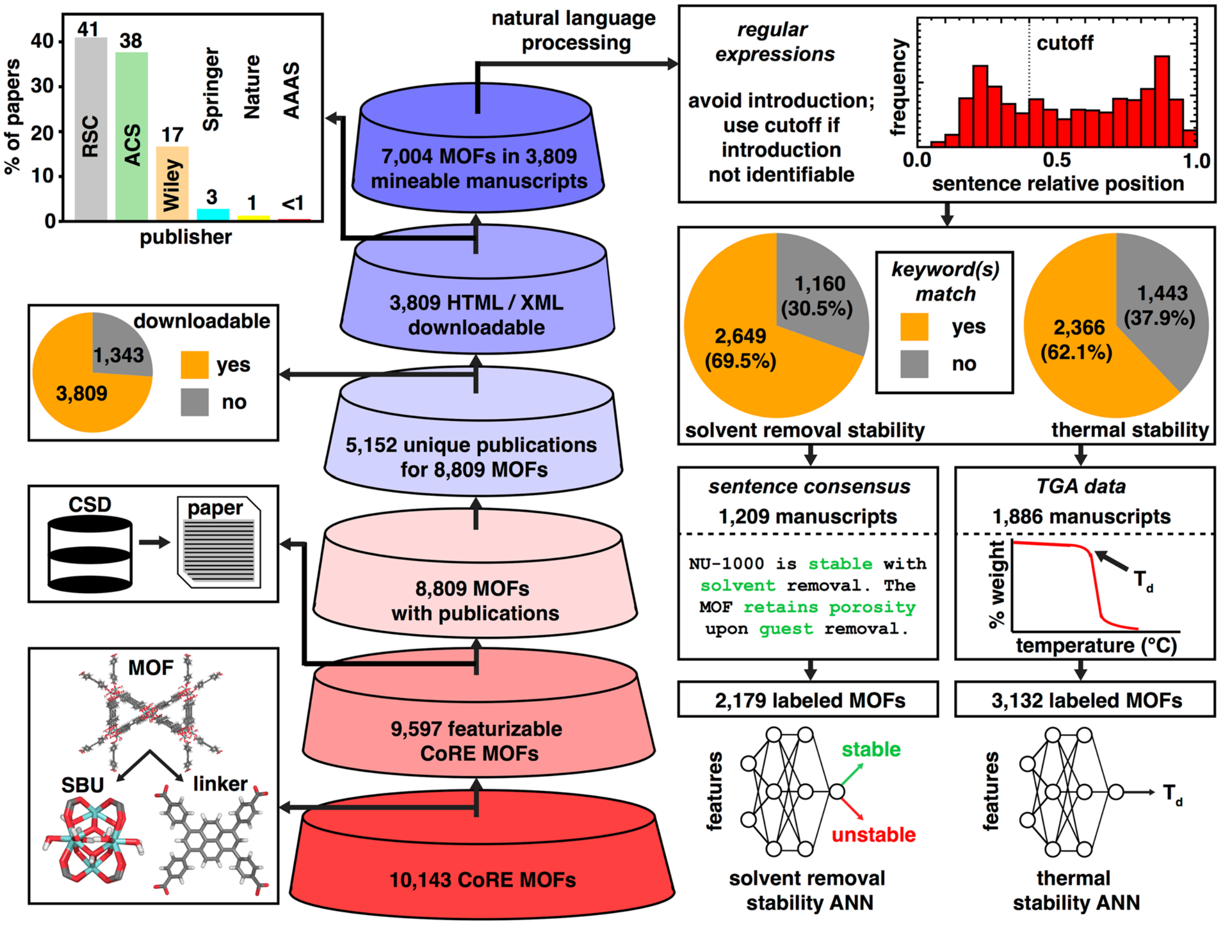

% of papers
publisher
41
RSC
38
ACS
17
Wiley
3
Springer
1
Nature
<1
AAAS
natural language processing
7,004 MOFs in 3,809 mineable manuscripts
3,809 HTML / XML downloadable
5,152 unique publications for 8,809 MOFs
8,809 MOFs with publications
9,597 featurizable CoRE MOFs
10,143 CoRE MOFs
downloadable
1,343
3,809
yes
no
CSD
paper
MOF
SBU
linker
regular expressions
avoid introduction; use cutoff if introduction not identifiable
frequency
cutoff
sentence relative position
keyword(s) match
1,160 (30.5%)
2,649 (69.5%)
1,443 (37.9%)
2,366 (62.1%)
solvent removal stability
thermal stability
sentence consensus
1,209 manuscripts
NU-1000 is stable with solvent removal. The MOF retains porosity upon guest removal.
TGA data
1,886 manuscripts
% weight
temperature (°C)
T_d
2,179 labeled MOFs
3,132 labeled MOFs
features
stable
unstable
solvent removal stability ANN
thermal stability ANN

Fig. 1 Workflows for curating datasets for solvent removal and thermal stability. First, we use sanitized MOFs from published works, filter by structures that can be featurized, obtain manuscripts corresponding to structures, download these manuscripts to prepare them for natural language processing, and finally text mine the manuscripts to identify mentions of solvent removal stability or thermogravimetric analysis data. We identify thermogravimetric analysis traces from manuscripts with thermogravimetric analysis keywords. The two sets of data gathered during this workflow are then used to train machine learning models.

online repository[69]. In total, we identified 2,179 labels for solvent removal stability corresponding to structures from the CoRE MOF 2019 dataset (Fig. 1).

For thermal stability analysis, we performed regular expression searches to identify a subset of 2,366 manuscripts (out of the 3,809 downloadable manuscripts) that could be expected to contain a TGA trace (Table 1). Of this set, 1,886 contain one or more TGA traces corresponding to CoRE MOF 2019 structures. The remainder either lack a TGA trace or only contain TGA traces for structures not deposited in the CoRE MOF 2019 database. Because TGA decomposition temperatures (T_d) are reported in a number of ways that could refer to the onset temperature or temperature of complete collapse, we extracted all critical TGA trace temperatures following a consistent protocol using WebPlotDigitizer[70]. We obtained two lines representing the TGA data before and after the decomposition step reported by the authors (i.e., in cases with more than one step) from four points on the TGA trace and calculated the intersection point of the two lines to obtain T_d (Fig. 1). An example of this approach on a representative MOF is shown in Fig. 2. For papers where more than one TGA was reported, unit cell parameters were used to identify the relevant MOF CSD entry and map the TGA trace to the corresponding CSD structure. Our extracted T_ds always correspond to the intersection of these two lines and can thus differ from T_d values determined in a different manner (e.g., from a value extracted only after complete decomposition). Overall, we identified 3,132 thermal decomposition temperatures that corresponded to featurized CoRE MOF 2019 structures.

Building blocks and descriptors. First, we obtained the primitive unit cell for each MOF in the CoRE MOF 2019 database using pymatgen[71]. We divided each MOF into its constituent inorganic SBUs and organic linkers during the generation of RACs[58,72,73] using molSimplify[18] (Table 2). For the special case of metal-coordinating linkers (e.g., porphyrinic) linkers, the metal serves as a starting point for SBU RACs but the nitrogen in the porphyrin is also the metal-coordinating atom in a linker-centered RAC (Table 2). To identify unique SBUs and linkers in each MOF, we computed the atom-weighted molecular graph deter minants[74] and obtained the relevant subgraphs in the MOF components with unique determinants (see Data

stemmed keyword	keyword category	words identified	stability type
collaps	collapse	collaps(e/ed/es/ing)	solvent removal
deform	collapse	deform(ed/s/ing/ation)	solvent removal
amorph	collapse	amorph(ous/ize)	solvent removal
blockage	collapse	blockage	solvent removal
degrad	collapse	degrad(e/ed/es/ing/ation)	solvent removal
unstable	collapse	unstable	solvent removal
instability	collapse	instability	solvent removal
destroy	collapse	destroy(ed/s/ing)	solvent removal
one step weight	collapse	one(−)step weight loss	solvent removal
single step weight	collapse	single(−)step weight loss	solvent removal
stable	stable	stable	solvent removal
stability	stable	stability	solvent removal
preserv	stable	preserv(e/ed/es/ing)	solvent removal
crystallinity	stable	crystallinity	solvent removal
coordinatively unsaturat	stable	coordinatively unsaturat(ed/ing)	solvent removal
porosity	stable	(micro)porosity	solvent removal
retain	stable	retain(ed/s/ing)	solvent removal
maintain	stable	maintain(ed/s/ing)	solvent removal
two step weight	stable	two(-)step weight loss	solvent removal
solvent	solvent	solvent(s)	solvent removal
guest	solvent	guest(s)	solvent removal
desolvat	solvent	desolvat(e/ed/es/ing)	solvent removal
remov	solvent	remov(e/ed/es/ing)	solvent removal
activat	solvent	activat(e/ed/es/ing)	solvent removal
evacuat	solvent	evacuat(e/ed/es/ing)	solvent removal
dehydrat	solvent	dehydrat(e/ed/es/ing)	solvent removal
eliminat	solvent	eliminat(e/ed/es/ing)	solvent removal
water, H2O	solvent	water, H2O	solvent removal
DMF, formamide	solvent	DMF, formamide	solvent removal
DMA, methylamine, diamine	solvent	DMA, methylamine, diamine	solvent removal
EtOH, MeOH, ethanol, methanol	solvent	EtOH, MeOH, ethanol, methanol	solvent removal
pyrrolidone	solvent	pyrrolidone	solvent removal
TG	thermal	TG(A)	thermal
thermogravimetric	thermal	thermogravimetric analysis	thermal
thermal gravimetric	thermal	thermal(−)gravimetric analysis	thermal

Table 1. Keywords used for regular expression searches for solvent removal and thermal stabilities. Stemmed forms of each word were used to identify keywords that have different tenses or forms. We label each word with a category and the type of stability that it identified.

Records). In addition, we computed geometric properties (e.g. maximum included sphere) with Zeo++ using a nitrogen probe molecule with a radius of 1.86 Å (Table 3)[61,62]. We used 10,000 Monte Carlo samples per unit cell to obtain the geometric quantities in conjunction with the -sa command to compute surface areas and -volpo for probe-occupiable volumes. We also used the command -ha -res to obtain pore diameters. All ANN models use RACs and geometric features as inputs to make predictions and were trained using keras[75] with a tensorflow[76] backend (see Data Records). Additional criteria could be used to reduce either the solvent removal stability or thermal decomposition datasets. For example, multiple refinements of a MOF could have been carried out by multiple groups with distinct geometric properties or labels leading to either distinct refcodes or the same base refcode with different numbering. All data is provided with these refcodes, calculated RACs, and geometric properties[69]. Users may down select to eliminate similar structures starting from the larger data set.

Data Records

We provide two JSON files, one for MOFs with solvent removal stability labels (solvent_removal_stability.json) and the other for MOFs with thermal stability labels (thermal_stability.json). The solvent removal stability JSON file contains 2,179 entries, and the thermal stability JSON file contains 3,132 entries.

Within the JSON files, each MOF structure is tabulated as a separate entry. In the solvent removal stability JSON file each entry contains the refcode of a MOF (i.e., as in the CoRE MOF 2019 database[60] and the CSD[63,64]), the DOI of the associated manuscript, sentences identified during regular expression matching and their corresponding locations in the manuscript, and the data partition for ANN usage[57] (i.e., train, validation, or test). Additionally,

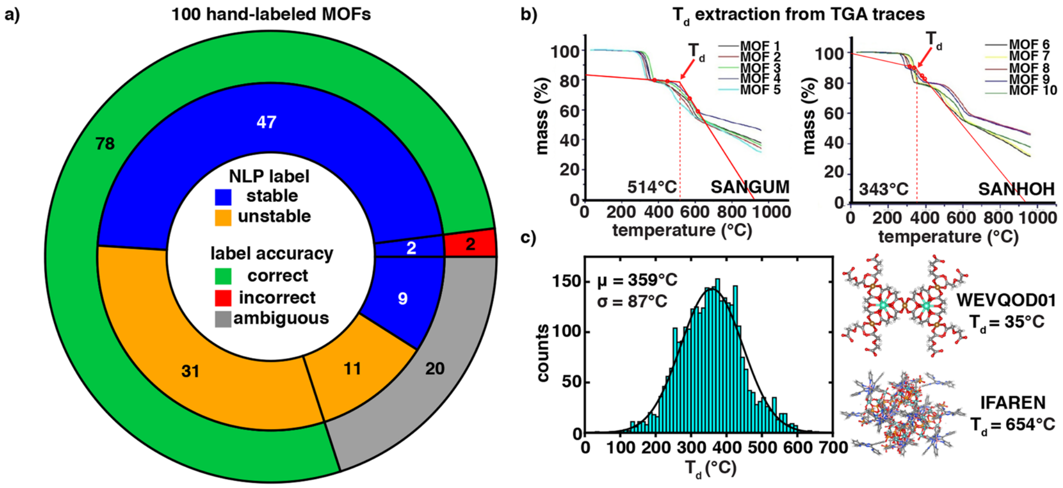


Fig. 2 Validation of the solvent removal and thermal stability data sets. (**a**) Comparison of NLP-assigned labels to hand-assigned labels over a 100 MOF subset, with stable NLP-assigned stable labels in blue and NLP-assigned unstable labels in orange. Cases that were correctly assigned are shown with a green outer ring, those that were incorrect are shown with a red outer ring, and ambiguous cases are shown with a gray outer ring. (**b**) Assignment of T_d from TGA traces (top right, TGA traces adapted from ref. [80]) shown for two MOFs (SANGUM and SANHOH), with T_d values inset. (**c**) The distribution of T_d over the full thermal stability dataset is shown, with the MOF containing the lowest (WEVQOD01) and highest (IFAREN) thermal decomposition temperatures shown inset.

start	scope	operation	features removed	feature count
mc	all	product	1 (mc-I-0-all)	19
mc	all	difference	8 (D_{mc}-I-0-all, D_{mc}-I-1-all, D_{mc}-I-2-all, D_{mc}-I-3-all, D_{mc}-S-0-all, D_{mc}-T-0-all, D_{mc}-Z-0-all, D_{mc}-χ-0-all)	12
lc	linker	product	1 (lc-I-0-linker)	19
lc	linker	difference	8 (D_{lc}-I-0-linker, D_{lc}-I-1-linker, D_{lc}-I-2-linker, D_{lc}-I-3-linker, D_{lc}-S-0-linker, D_{lc}-T-0-linker, D_{lc}-Z-0-linker, D_{lc}-χ-0-linker)	12
func	linker	product	0	20
func	linker	difference	8 (D_{func}-I-0-linker, D_{func}-I-1-linker, D_{func}-I-2-linker, D_{func}-I-3-linker, D_{func}-S-0-linker, D_{func}-T-0-linker, D_{func}-Z-0-linker, D_{func}-χ-0-linker)	12
full	all	product	0	20
full	linker	product	0	20
			26	134

Table 2. Description of revised autocorrelation (RAC) features with start/scope, operation performed, count of features removed, and total feature count. Five heuristic atom-wise quantities are used to perform all product and difference operations: nuclear charge (Z), electronegativity (χ), topology (T), identity (I), and covalent radius (S). MOF RACs contain four possible starts and two possible scopes: metal-centered (mc) start, linker coordinating atom centered (lc) start, functional group centered (func) start, every atom (full) start, all atom in primitive cell (all) scope, or all atom in linker (linker) scope. All starts, scopes and operations use bond depths of 0, 1, 2, and 3 to generate autocorrelations (for a total of 20 possible features for each scope). Cases that are invariant across all MOFs are listed in the "features removed" column. RAC features are given using the notation: <operation/start>-<atomic property>-<depth>-<scope>. RAC features using products are indicated by their start (e.g. mc), and those using differences contain a "D" prefix with a subscripted start (e.g. D_{mc}).

in each entry we report RAC[18,58] and geometric features[62]; ANN prediction probabilities, which are float values between 0 and 1, with values <0.5 (≥0.5) corresponding to instability (stability) upon solvent removal, respectively; and ANN latent space entropy[77] measurements (which have a maximum value of 0.693 for binary classification) from training data. We also provide blocks for each unique inorganic SBU and organic linker in TRIPOS mol2 format, which can be automatically loaded into molSimplify[78] for structure manipulation. We determine whether or not each linker or SBU is unique by computing atom-weighted molecular graph determinants[74], and we keep only one representative example of each linker and SBU with a unique molecular graph determinant.

In addition to the entry information provided in the solvent removal stability JSON file, the thermal stability JSON file contains the four extracted points from the TGA trace of each MOF with a thermal stability label. We provide ANN predictions (T_d*) in units of degrees Celsius, and we provide latent space distance (i.e., both scaled and unscaled) measurements that can be used for uncertainty quantification[79] in regression models. The scaled

variable name	explanation	units
D_f	maximum free sphere	Å
D_i	maximum included sphere	Å
D_{if}	maximum included sphere in the free sphere path	Å
GPOAV	gravimetric pore accessible volume	cm^3/g
GPONAV	gravimetric pore non-accessible volume	cm^3/g
GPOV	gravimetric pore volume	cm^3/g
GSA	gravimetric surface area	m^2/g
POAV	pore accessible volume	$Å^3$
PONAV	pore non-accessible volume	$Å^3$
POAVF	pore accessible volume fraction	unitless
PONAVF	pore non-accessible volume fraction	unitless
VPOV	volumetric pore volume	cm^3/cm^3
VSA	volumetric surface area	m^2/cm^3
ρ	crystal density	g/cm^3

Table 3. Description of geometric features generated by Zeo++ with definitions and units.

latent space distances have a maximum distance of 1 with respect to the training data, in accordance with prior work[57].

As an alternative to the JSON files, we provide CSV files for the solvent removal stability and thermal stability data sets. These CSV files contain 2,179 and 3,132 entries respectively, and they contain the same information as the JSON files. We also provide TRIPOS mol2 files for the representative extracted inorganic SBUs and organic linkers separately.

We provide the refcodes, DOIs, and extracted sentences as a CSV file for the structures for which we could identify keywords but could not assign a unique label. For solvent removal stability, multiple sentences may have different labels, preventing the assignment of an unambiguous final label (e.g., both positive and negative stability identified or challenging disambiguation of MOF structures). For thermal stability, TGA may be mentioned within the manuscript, but a TGA trace corresponding to the MOF in the CoRE MOF 2019 database may not be identifiable (e.g., when there are multiple structures corresponding to a manuscript).

Lastly, we provide our two ANN models (solvent removal stability classification and thermal stability regression) from prior work[57] as .h5 files that can be used with our open-source Python scripts, found on our GitHub repository (see Code Availability). We provide all JSON files, Excel sheets, SBU and linker structures, and models at our Zenodo repository[69].

On the MOFSimplify website (see Usage Notes), the user can download information on latent space nearest neighbor (LSNN) MOFs to a MOF input by the user. These LSNN MOFs are drawn from model training data, and the user can download information on them in the form of TXT, CIF, and CSV files. The TXT files each describe one MOF and include the experimentally observed stability for the MOF, the associated DOI, and its latent space distance to the MOF input by the user. In addition, LSNN CoRE MOF 2019 structures can be downloaded as CIF files. For thermal stability LSNN MOFs, the user can download their simplified TGA data as CSV files.

Technical Validation

We obtained a random sample of 100 MOFs from our solvent removal stability dataset to assess the quality of our NLP-assigned solvent removal stability labels in comparison to manual interpretation by a scientist. Over this set, there are only two cases of MOFs that are incorrectly labeled as unstable upon solvent removal but are stable upon solvent removal. The majority (i.e., 78 MOFs) are correctly labeled, 47 of which are stable and 31 unstable upon solvent removal (Fig. 2). For the remaining 20 MOFs, the extracted sentences do not make a definitive statement about solvent removal stability, with 9 cases labeled as stable and 11 unstable (Fig. 2). Analyzing the cases where the NLP workflow definitively assigns stability but the sentences are more ambiguous, these cases either mention another aspect of stability (e.g. stable coordination environment) while mentioning solvents or mention that solvent removal stability was evaluated without stating the outcome.

To extract T_d for thermal stability labels, we used NLP only to identify the presence of the TGA trace, which we then systematically digitized. Because thermal stability is not reported consistently across manuscripts (e.g., some manuscripts report decomposition onset temperatures, and others decomposition completion temperatures), we extracted T_d from TGA traces consistently, using the start and the end of the decomposition step (see Methods). This process makes the thermal stability quantitative data less sensitive than solvent removal stability to either the NLP protocol or the method of reporting by the researcher.

As an example of the benefit of systematic extraction of T_d, we select a representative manuscript[80] (DOI: 10.1002/slct.201600844) that contains ten MOFs. Only six of these MOFs (SANGEW, SANGUM, SANHIB, SANHOH, SANHUN, and SANJAV) are present in the CoRE MOF 2019 dataset, whereas the remaining four MOFs (SANGIA, SANGOG, SANHAT, and SANHEX) are not. As a result, the latter four MOFs are excluded from our dataset. The manuscript reports all ten MOFs "remain thermally stable until 553 K" and states that the first step of the TGA trace corresponds to the loss of a guest molecule, while the second step corresponds to decomposition. Our procedure uses the unit cell parameters provided in the supporting information to identify the CSD refcodes corresponding to the MOF labels in the manuscript (SANGEW: MOF1, SANGUM: MOF4, SANHIB: MOF7, SANHOH: MOF8, SANHUN: MOF9, SANJAV: MOF10) and then uses these name–structure mappings to associate a TGA trace with each MOF. The digitization procedure uses the beginning and end of the second step of each TGA trace to quantify decomposition temperatures. For SANHOH and SANHUN, manual inspection of the TGA trace reveals that decomposition starts near 300 °C and ends near 400 °C. In contrast, for SANGEW, SANGUM, SANHIB, and SANJAV, decomposition also starts near 300 °C but does not conclude until 600 °C (Fig. 2). Although these MOFs begin decomposing at similar temperatures, all are below the value reported in the text by the authors, and some MOFs decompose more slowly than others. This case study demonstrates the differences in how TGA trace results are reported and quantified, motivating a systematic analysis and labeling. From the systematically labeled data, our final distribution of extracted T_d values over the thermal stability dataset is a normal distribution centered around 359 °C with an 87 °C standard deviation (Fig. 2).

As an additional blinded test, we hand-labeled the solvent removal stability and thermal stability of 40 MOFs from manuscripts that could not be automatically downloaded from the publisher (i.e., from Elsevier in this case). These data points are not present within the entire (i.e., train or test) solvent removal stability or thermal stability data sets. From these 40 MOFs, 20 were assigned stable and 20 unstable with respect to solvent removal by our solvent removal stability ANN. Over this hand-labeled set, we find that the ANN correctly predicted the stability of the majority (i.e., 31 out of 40) of this set of MOFs. For the remaining 9 MOFs, 7 stable MOFs were predicted to be unstable, while 2 unstable MOFs were predicted to be stable. This 78% accuracy is comparable to the ANN test set performance. We find that the mean absolute error (MAE) of the T_d predictions generated by the thermal stability ANN on the hand-labeled MOFs is 55 °C, which is comparable to the test set performance (MAE: 47 °C) of the thermal stability ANN. The comparable performances on unseen data demonstrate that we can use our models to screen unseen MOFs to quantitatively predict their activation and thermal stabilities.

Usage Notes

We introduce the MOFSimplify website mofsimplify.mit.edu, a tool for analyzing and comparing the data provided in our data set as well as making property predictions on MOFs (Fig. 3). As an alternative to using the MOFSimplify web interface, users may download compiled data for solvent removal stabilities and thermal stabilities in JSON or CSV formats. All MOF extracted labeled properties are reported with both the Zeo++−computed and RAC features (see Data Records). MOFSimplify includes a means of visualizing the curated data set and separating these MOFs into constituent inorganic SBUs and organic linkers in the Component Analysis tab. MOFSimplify uses both 3Dmol.js[81] and code from the MOFid[53] website for MOF unit cell visualization along with molSimplify to separate MOFs into their constituent parts[18]. MOFSimplify also lets the user provide feedback to the curated experimental data, providing an assessment of data fidelity.

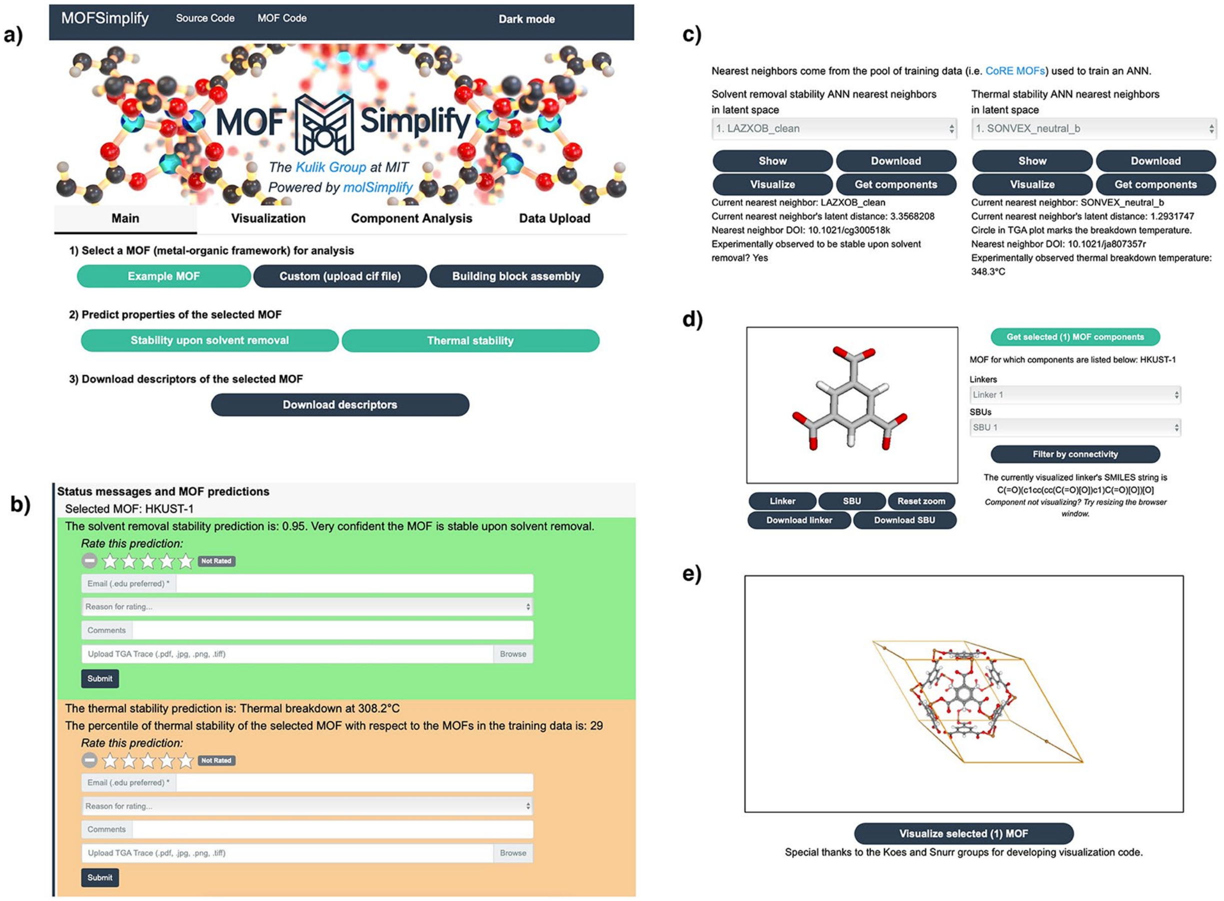
To use the MOFSimplify web interface, the user selects a MOF for analysis in CIF file format in the Main tab. This can be done either by uploading a solvent-free CIF file of a MOF without partial occupancies or by constructing a CIF file for a hypothetical MOF from linkers and SBU building blocks selected by the user. For the latter option, MOFSimplify uses the Topologically Based Crystal Constructor (ToBaCCo) 3.0 code[15,82]. Prior to assembly, the user must select a compatible linker, SBU, and MOF net combination from dropdown menus. Incompatible combinations are rejected by MOFSimplify.

Once the user selects a MOF for analysis, MOFSimplify generates RAC features and geometric descriptors of the selected MOF (Fig. 3). If the selected MOF is present in the relevant solvent removal stability or thermal stability training data for which a prediction is requested, MOFSimplify returns the data set value for the selected MOF. Otherwise, it provides an ML model prediction with quantified uncertainty[79]. The web server determines the presence or absence of the selected MOF in the dataset by comparing RAC and geometric descriptors generated for the selected MOF to the descriptors previously generated for the training data. MOFSimplify reports the latent space nearest neighbors (LSNNs), which are the MOFs in training data that appear most proximal in the ANN latent space to the loaded MOF for either thermal stability or solvent removal stability. The user can display and download information about the identified LSNN MOFs, which are also provided in the online repository (see Data records)[69]. This includes structures for LSNNs, which can be downloaded as CIF files, along with LSNN metadata such as the latent space distance to the selected MOF, DOI of the associated manuscript, and experimentally determined stability, which can be downloaded in TXT file format. Once a prediction is requested and either the ground truth or an ML model prediction is returned, the user can also download the RAC and Zeo++ descriptors generated for the MOF.

For TGA data, MOFSimplify displays a simplified experimental TGA plot for thermal stability ANN LSNNs generated from four TGA trace points and allows the user to to download the same data in CSV format (see Methods). MOFSimplify reports the prediction/ground truth temperature for the selected MOF (T_d*) and the percentile rank of T_d* relative to the training data T_d (Fig. 3). If a solvent removal stability prediction is requested, MOFSimplify will either report a ground truth (i.e., stable or unstable) for the selected MOF, or it will display a prediction between 0 (confidently unstable) and 1 (confidently stable) and a sentence reflecting model confidence.

MOFSimplify lets the user identify MOF components and allows the user to filter these MOF components by their atom-weighted molecular graph determinants to isolate unique components as determined by graph connectivity. By default, MOFSimplify does not apply the filter and instead displays all copies identified in the CIF unit cell. MOFSimplify can visualize component structures using 3Dmol.js[81] and display their SMILES codes that are generated with OpenBabel v2.4.1[83,84]. The user can download these MOF components as XYZ files (Fig. 3).

Additionally, the MOFSimplify interface encourages community engagement by enabling the user to add new MOF data to our database by uploading MOF CIF files and TGA traces in the Data Upload tab. MOFSimplify also lets the user indicate whether they agree with an ANN prediction or curated experimental data and support



1) Select a MOF (metal-organic framework) for analysis

2) Predict properties of the selected MOF

3) Download descriptors of the selected MOF

Status messages and MOF predictions

Selected MOF: HKUST-1

The solvent removal stability prediction is: 0.95. Very confident the MOF is stable upon solvent removal.

Rate this prediction:

The thermal stability prediction is: Thermal breakdown at 308.2°C

The percentile of thermal stability of the selected MOF with respect to the MOFs in the training data is: 29

Rate this prediction:

Nearest neighbors come from the pool of training data (i.e. CoRE MOFs) used to train an ANN.

Solvent removal stability ANN nearest neighbors in latent space

Current nearest neighbor: LAZXOB_clean
Current nearest neighbor's latent distance: 3.3568208
Nearest neighbor DOI: 10.1021/cg300518k
Experimentally observed to be stable upon solvent removal? Yes

Thermal stability ANN nearest neighbors in latent space

Current nearest neighbor: SONVEX_neutral_b
Current nearest neighbor's latent distance: 1.2931747
Circle in TGA plot marks the breakdown temperature.
Nearest neighbor DOI: 10.1021/ja807357r
Experimentally observed thermal breakdown temperature: 348.3°C

MOF for which components are listed below: HKUST-1

Linkers

SBUs

The currently visualized linker's SMILES string is C(=O)(c1cc(cc(C(=O)[O])c1)C(=O)[O])[O]
Component not visualizing? Try resizing the browser window.

Special thanks to the Koes and Snurr groups for developing visualization code.

Fig. 3 Sections of the MOFSimplify web interface. (**a**) Interface for selecting a MOF for analysis and predicting properties of the selected MOF using ANNs trained on experimental data mined from the literature. The default MOF loaded upon selecting "Example MOF" is HKUST-1, a well-studied MOF[85]. (**b**) The feedback interface for evaluating model predictions. (**c**) The interface listing similar (i.e., LSNN) MOFs to the selected MOF as determined by the ANNs. (**d**) Visualization of the selected MOF's components. (**e**) Visualization of the selected MOF's unit cell.

their position by uploading a TGA trace (Fig. 3). These TGA traces will be digitized by us to extract T_d data in a manner consistent with our previous thermal stability data. User input will be used to improve our ANN models through community-based active learning. Users can opt out of uploading data or providing feedback. If users wish to remove data after the fact, an email form is provided for removal requests.

Code availability

All scripts used to mine the extant literature corresponding to the CoRE MOF 2019 database are commented and are available on a public GitHub repository at https://github.com/hjkgrp/text_mining_tools. Manuscript copyrights are retained by the publishers, preventing the complete dissemination of full-length articles, but the mined data is provided with an open source CC-BY license and is available on Zenodo[69] (see also Data Records).

The MOFSimplify website is located at https://mofsimplify.mit.edu. The code backend for the MOFSimplify website is available in a public GitHub repository at https://github.com/hjkgrp/MOFSimplify. The repository contains a user manual for the website.

Received: 30 September 2021; Accepted: 17 January 2022;
Published online: 11 March 2022

References

1. Furukawa, H., Cordova, K. E., O'Keeffe, M. & Yaghi, O. M. The chemistry and applications of metal-organic frameworks. *Science* **341**, 1230444 (2013).
2. Wilmer, C. E. *et al.* Large-scale screening of hypothetical metal–organic frameworks. *Nat. Chem.* **4**, 83–89 (2011).
3. Simon, C. M. *et al.* The materials genome in action: Identifying the performance limits for methane storage. *Energy Environ. Sci.* **8**, 1190–1199 (2015).
4. Sumida, K. *et al.* Carbon dioxide capture in metal–organic frameworks. *Chem. Rev.* **112**, 724–781 (2011).
5. Kreno, L. E. *et al.* Metal–organic framework materials as chemical sensors. *Chem. Rev.* **112**, 1105–1125 (2011).
6. Campbell, M. G., Sheberla, D., Liu, S. F., Swager, T. M. & Dincă, M. Cu_3(hexaiminotriphenylene)$_2$: An electrically conductive 2d metal-organic framework for chemiresistive sensing. *Angew. Chem. Int. Ed.* **54**, 4349–4352 (2015).

7. Boyd, P. G. *et al*. Data-driven design of metal–organic frameworks for wet flue gas co$_2$ capture. *Nature* **576**, 253–256 (2019).
8. Gonzalez, M. I. *et al*. Separation of xylene isomers through multiple metal site interactions in metal–organic frameworks. *J. Am. Chem. Soc.* **140**, 3412–3422 (2018).
9. Yang, D. & Gates, B. C. Catalysis by metal organic frameworks: Perspective and suggestions for future research. *ACS Catal.* **9**, 1779–1798 (2019).
10. Lee, J. *et al*. Metal–organic framework materials as catalysts. *Chem. Soc. Rev.* **38**, 1450 (2009).
11. Wang, Z., Bilegsaikhan, A., Jerozal, R. T., Pitt, T. A. & Milner, P. J. Evaluating the robustness of metal–organic frameworks for synthetic chemistry. *ACS Appl. Mater. Interfaces* **13**, 17517–17531 (2021).
12. Barona, M. *et al*. Computational predictions and experimental validation of alkane oxidative dehydrogenation by fe$_2$m mof nodes. *ACS Catal.* **10**, 1460–1469 (2019).
13. Simons, M. C. *et al*. Structure, dynamics, and reactivity for light alkane oxidation of fe(ii) sites situated in the nodes of a metal–organic framework. *J. Am. Chem. Soc.* **141**, 18142–18151 (2019).
14. Xiao, D. J. *et al*. Oxidation of ethane to ethanol by n$_2$o in a metal–organic framework with coordinatively unsaturated iron(ii) sites. *Nat. Chem.* **6**, 590–595 (2014).
15. 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.* **17**, 5801–5810 (2017).
16. Gómez-Gualdrón, D. A. *et al*. Evaluating topologically diverse metal–organic frameworks for cryo-adsorbed hydrogen storage. *Energy Environ. Sci.* **9**, 3279–3289 (2016).
17. Rosen, A. S., Notestein, J. M. & Snurr, R. Q. Structure–activity relationships that identify metal–organic framework catalysts for methane activation. *ACS Catal.* **9**, 3576–3587 (2019).
18. Moosavi, S. M. *et al*. Understanding the diversity of the metal-organic framework ecosystem. *Nat. Commun.* **11** (2020).
19. Moghadam, P. Z. *et al*. Structure-mechanical stability relations of metal-organic frameworks via machine learning. *Matter* **1**, 219–234 (2019).
20. Long, J. R. & Yaghi, O. M. The pervasive chemistry of metal–organic frameworks. *Chem. Soc. Rev.* **38**, 1213 (2009).
21. Stock, N. & Biswas, S. Synthesis of metal-organic frameworks (mofs): Routes to various mof topologies, morphologies, and composites. *Chem. Rev.* **112**, 933–969 (2011).
22. Farha, O. K. & Hupp, J. T. Rational design, synthesis, purification, and activation of metal—organic framework materials. *Acc. Chem. Res.* **43**, 1166–1175 (2010).
23. Wang, Z. & Cohen, S. M. Postsynthetic modification of metal–organic frameworks. *Chem. Soc. Rev.* **38**, 1315–1329 (2009).
24. Ma, J., Kalenak, A. P., Wong-Foy, A. G. & Matzger, A. J. Rapid guest exchange and ultra-low surface tension solvents optimize metal–organic framework activation. *Angew. Chem. Int. Ed.* **56**, 14618–14621 (2017).
25. Mondloch, J. E., Karagiaridi, O., Farha, O. K. & Hupp, J. T. Activation of metal–organic framework materials. *CrystEngComm* **15**, 9258 (2013).
26. Dodson, R. A., Wong-Foy, A. G. & Matzger, A. J. The metal–organic framework collapse continuum: Insights from two-dimensional powder x-ray diffraction. *Chem. Mater.* **30**, 6559–6565 (2018).
27. Zhang, X. *et al*. A historical overview of the activation and porosity of metal-organic frameworks. *Chem. Soc. Rev.* **49**, 7406–7427 (2020).
28. Sun, L., Campbell, M. G. & Dincă, M. Electrically conductive porous metal-organic frameworks. *Angew. Chem. Int. Ed.* **55**, 3566–3579 (2016).
29. Yuan, S. *et al*. Stable metal-organic frameworks: Design, synthesis, and applications. *Adv. Mater.* **30**, 1704303 (2018).
30. Hendon, C. H., Rieth, A. J., Korzyński, M. D. & Dincă, M. Grand challenges and future opportunities for metal–organic frameworks. *ACS Cent. Sci.* **3**, 554–563 (2017).
31. Osadchii, D. Y. *et al*. Isolated fe sites in metal organic frameworks catalyze the direct conversion of methane to methanol. *ACS Catal.* **8**, 5542–5548 (2018).
32. Li, H., Eddaoudi, M., O'Keeffe, M. & Yaghi, O. M. Design and synthesis of an exceptionally stable and highly porous metal-organic framework. *Nature* **402**, 276–279 (1999).
33. Eddaoudi, M., Li, H. & Yaghi, O. M. Highly porous and stable metal—organic frameworks: Structure design and sorption properties. *J. Am. Chem. Soc.* **122**, 1391–1397 (2000).
34. Howarth, A. J. *et al*. Chemical, thermal and mechanical stabilities of metal–organic frameworks. *Nat. Rev. Mater.* **1**, 15018 (2016).
35. Gómez-Gualdrón, D. A., Wilmer, C. E., Farha, O. K., Hupp, J. T. & Snurr, R. Q. Exploring the limits of methane storage and delivery in nanoporous materials. *J. Phys. Chem. C* **118**, 6941–6951 (2014).
36. Ayoub, G., Islamoglu, T., Goswami, S., Friščić, T. & Farha, O. K. Torsion angle effect on the activation of uio metal-organic frameworks. *ACS Appl. Mater. Interfaces* **11**, 15788–15794 (2019).
37. Lv, X.-L. *et al*. Ligand rigidification for enhancing the stability of metal–organic frameworks. *J. Am. Chem. Soc.* **141**, 10283–10293 (2019).
38. Healy, C. *et al*. The thermal stability of metal-organic frameworks. *Coord. Chem. Rev.* **419**, 213388 (2020).
39. Wei, Z., Lu, W., Jiang, H.-L. & Zhou, H.-C. A route to metal–organic frameworks through framework templating. *Inorg. Chem.* **52**, 1164–1166 (2013).
40. Feng, L., Wang, K.-Y., Day, G. S., Ryder, M. R. & Zhou, H.-C. Destruction of metal–organic frameworks: Positive and negative aspects of stability and lability. *Chem. Rev.* **120**, 13087–13133 (2020).
41. Dincǎ, M., Dailly, A. & Long, J. R. Structure and charge control in metal-organic frameworks based on the tetrahedral ligand tetrakis(4-tetrazolylphenyl)methane. *Chem. Eur. J.* **14**, 10280–10285 (2008).
42. Mu, B. & Walton, K. S. Thermal analysis and heat capacity study of metal–organic frameworks. *J. Phys. Chem. C* **115**, 22748–22754 (2011).
43. Coudert, F.-X. & Fuchs, A. H. Computational characterization and prediction of metal–organic framework properties. *Coord. Chem. Rev.* **307**, 211–236 (2016).
44. Bouëssel du Bourg, L., Ortiz, A. U., Boutin, A. & Coudert, F.-X. Thermal and mechanical stability of zeolitic imidazolate frameworks polymorphs. *APL Mater.* **2**, 124110 (2014).
45. Moosavi, S. M., Boyd, P. G., Sarkisov, L. & Smit, B. Improving the mechanical stability of metal–organic frameworks using chemical caryatids. *ACS Cent. Sci.* **4**, 832–839 (2018).
46. Batra, R., Chen, C., Evans, T. G., Walton, K. S. & Ramprasad, R. Prediction of water stability of metal–organic frameworks using machine learning. *Nat. Mach. Intell.* **2**, 704–710 (2020).
47. Tshitoyan, V. *et al*. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature* **571**, 95–98 (2019).
48. Krallinger, M., Rabal, O., Lourenço, A., Oyarzabal, J. & Valencia, A. Information retrieval and text mining technologies for chemistry. *Chem. Rev.* **117**, 7673–7761 (2017).
49. Kim, E. *et al*. Inorganic materials synthesis planning with literature-trained neural networks. *J. Chem. Inf. Model.* **60**, 1194–1201 (2020).
50. Kim, E. *et al*. Materials synthesis insights from scientific literature via text extraction and machine learning. *Chem. Mater.* **29**, 9436–9444 (2017).

51. Jensen, Z. *et al*. A machine learning approach to zeolite synthesis enabled by automatic literature data extraction. *ACS Cent. Sci.* **5**, 892–899 (2019).
52. Jensen, Z. *et al*. Discovering relationships between osdas and zeolites through data mining and generative neural networks. *ACS Cent. Sci.* **7**, 858–867 (2021).
53. Bucior, B. J. *et al*. Identification schemes for metal–organic frameworks to enable rapid search and cheminformatics analysis. *Cryst. Growth Des.* **19**, 6682–6697 (2019).
54. Weston, L. *et al*. Named entity recognition and normalization applied to large-scale information extraction from the materials science literature. *J. Chem. Inf. Model.* **59**, 3692–3702 (2019).
55. Park, S. *et al*. Text mining metal–organic framework papers. *J. Chem. Inf. Model.* **58**, 244–251 (2018).
56. Datar, A., Chung, Y. G. & Lin, L.-C. Beyond the bet analysis: The surface area prediction of nanoporous materials using a machine learning method. *J. Phys. Chem. Lett.* **11**, 5412–5417 (2020).
57. Nandy, A., Duan, C. & Kulik, H. J. Using machine learning and data mining to leverage community knowledge for the engineering of stable metal-organic frameworks. (2021).
58. Janet, J. P. & Kulik, H. J. Resolving transition metal chemical space: Feature selection for machine learning and structure-property relationships. *J. Phys. Chem. A* **121**, 8939–8954 (2017).
59. Luo, Y. *et al*. Mof synthesis prediction enabled by automatic data mining and machine learning. (2021).
60. Chung, Y. G. *et al*. Advances, updates, and analytics for the computation-ready, experimental metal–organic framework database: Core mof 2019. *J. Chem. Eng. Data* **64**, 5985–5998 (2019).
61. Martin, R. L., Smit, B. & Haranczyk, M. Addressing challenges of identifying geometrically diverse sets of crystalline porous materials. *J. Chem. Inf. Model.* **52**, 308–318 (2011).
62. Willems, T. F., Rycroft, C. H., Kazi, M., Meza, J. C. & Haranczyk, M. Algorithms and tools for high-throughput geometry-based analysis of crystalline porous materials. *Micropor. Mesopor. Mat.* **149**, 134–141 (2012).
63. Groom, C. R., Bruno, I. J., Lightfoot, M. P. & Ward, S. C. The cambridge structural database. *Acta Crystallogr., Sect. B: Struct. Sci.* **72**, 171–179 (2016).
64. Allen, F. H. The cambridge structural database: A quarter of a million crystal structures and rising. *Acta Crystallogr., Sect. B: Struct. Sci.* **58**, 380–388 (2002).
65. Kim, E., Huang, K., Jegelka, S. & Olivetti, E. Virtual screening of inorganic materials synthesis parameters with deep learning. *npj Comput. Mater*. **3** (2017).
66. Swain, M. C. & Cole, J. M. Chemdataextractor: A toolkit for automated extraction of chemical information from the scientific literature. *J. Chem. Inf. Model.* **56**, 1894–1904 (2016).
67. Hutto, C. J. & Gilbert, E. in *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*.
68. Qi, P., Zhang, Y., Zhang, Y., Bolton, J. & Manning, C. D. in *In Association for Computational Linguistics (ACL) System Demonstrations* (2020).
69. Nandy, A. *et al*. Mofsimplify: Machine learning models with extracted stability data of three thousand metal- organic frameworks. *zenodo* https://doi.org/10.5281/zenodo.5737968 (2021).
70. Rohatgi, A. *Webplotdigitizer: Version 4.4*, https://automeris.io/WebPlotDigitizer (2020).
71. Ong, S. P. *et al*. Python materials genomics (pymatgen): A robust, open-source python library for materials analysis. *Comput. Mater. Sci.* **68**, 314–319 (2013).
72. Moreau, G. & Broto, P. The autocorrelation of a topological structure: A new molecular descriptor. *Nouv. J. Chim.* **4**, 359–360 (1980).
73. Broto, P., Moreau, G. & Vandycke, C. Molecular structures: Perception, autocorrelation descriptor and sar studies: System of atomic contributions for the calculation of the n- octanol/water partition coefficients. *Eur. J. Med. Chem.* **19**, 71–78 (1984).
74. Taylor, M. G. *et al*. Seeing is believing: Experimental spin states from machine learning model structure predictions. *J. Phys. Chem. A* **124**, 3286–3299 (2020).
75. Keras (2015).
76. Tensorflow: Large-scale machine learning on heterogeneous systems (2015).
77. Duan, C., Janet, J. P., Liu, F., Nandy, A. & Kulik, H. J. Learning from failure: Predicting electronic structure calculation outcomes with machine learning models. *J. Chem. Theory Comput.* **15**, 2331–2345 (2019).
78. Ioannidis, E. I., Gani, T. Z. H. & Kulik, H. J. Molsimplify: A toolkit for automating discovery in inorganic chemistry. *J. Comput. Chem.* **37**, 2106–2117 (2016).
79. Janet, J. P., Duan, C., Yang, T., Nandy, A. & Kulik, H. J. A quantitative uncertainty metric controls error in neural network-driven chemical discovery. *Chem. Sci.* **10**, 7913–7922 (2019).
80. Kariem, M., Yawer, M., Sharma, S. & Sheikh, H. N. Syntheses, crystal structure, luminescence, porosity and magnetic properties of three-dimensional lanthanide coordination polymers with 2-aminoterepthalic acid. *ChemistrySelect* **1**, 4489–4501 (2016).
81. Rego, N. & Koes, D. 3dmol. Js: Molecular visualization with webgl. *Bioinformatics* **31**, 1322–1324 (2015).
82. Anderson, R. & Gómez-Gualdrón, D. A. Increasing topological diversity during computational "synthesis" of porous crystals: How and why. *CrystEngComm* **21**, 1653–1665 (2019).
83. O'Boyle, N. M. *et al*. Open babel: An open chemical toolbox. *Journal of cheminformatics* **3**, 1–14 (2011).
84. O'Boyle, N. M., Morley, C. & Hutchison, G. R. Pybel: A python wrapper for the openbabel cheminformatics toolkit. *Chemistry Central Journal* **2**, 1–7 (2008).
85. Agrawal, M., Han, R., Herath, D. & Sholl, D. S. Does repeat synthesis in materials chemistry obey a power law? *Proc. Natl. Acad. Sci. USA* **117**, 877–882 (2020).

Acknowledgements

The authors acknowledge support by DARPA (grant number D18AP00039) for the text extraction, machine learning, and website development efforts (to A.N., G.T., C.D. and N.A.), and some of the database curation efforts were supported by the Office of Naval Research under grant number N00014-20-1-2150 (to C.D. and H.J.K.). This work is also partially supported as part of the Inorganometallic Catalysis Design Center, an Energy Frontier Research Center funded by the U.S. Department of Energy, Office of Science, Basic Energy Sciences under Award DE-SC0012702 (to A.N. and G.T.). This work was also partially supported by a National Science Foundation Graduate Research Fellowship under Grant #1122374 (to A.N., N.A. and D.W.K.). H.J.K. holds a Career Award at the Scientific Interface from the Burroughs Wellcome Fund and an AAAS Marion Milligan Mason Award, which supported this work. The authors thank Adam H. Steeves and Akash Bajaj for providing a critical reading of the manuscript. The authors also acknowledge helpful conversations with members of the Elsa Olivetti lab.

Author contributions

A.N. curated the data and performed analyses on data validity. A.N. and C.D. worked on the ML model training. G.T. constructed the MOFSimplify website, with contributions from N.A., A.N., C.D., and D.W.K.. A.N., G. T., and H.J.K. wrote the manuscript. All authors contributed to revising the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to H.J.K.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.