Journal of Materials Chemistry A

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: J. Mater. Chem. A, 2025, 13, 32255

Generalizable classification of crystal structure error types using graph attention networks

Marco Gibaldi, [a] Jun Luo,[a] Andrew J. White,[a] R. Alex Mayo,[a] Cécile Pereira[b] and Tom K. Woo *[a]

Modern chemical applications of machine learning rely on massive training datasets collected through computational simulations or data mining. The quality of such datasets is increasingly challenged due to the discovery of errors in the most popular crystal structure databases. While methods exist to determine error presence, determining an error's cause is not straightforward. We propose a graph neural network-based approach to classify the presence of crystal structure errors, including proton omissions, charge balancing errors, and crystallographic disorder. A training dataset comprising >11k metal–organic frameworks (MOFs) labelled by error type was generated through domain expert inspection. Chemically intuitive features, such as atomic number and oxidation state, were found to achieve high classification accuracies ranging from 85 to 95%. Despite only training on MOFs, classification was generalizable towards unseen databases of molecules and metal complexes, observing accuracies eclipsing 96% in proton and disorder error classification in random samples of drug molecules and metal complexes. Further, graph explainability analysis indicated that these models frequently identify chemically-problematic subgraph structures—analogous to those a chemist would flag—as important towards the error label prediction.

Received 5th July 2025
Accepted 26th August 2025

DOI: 10.1039/d5ta05426e

rsc.li/materials-a

Introduction

As improvements in the domains of machine learning (ML) and artificial intelligence (AI) drive innovation, a growing number of researchers from diverse disciplines aspire to engage these powerful tools to their specific application. Recently, drastic advancements have been spurred on by the proliferation of models trained on enormous quantities of data—such as the billions and trillions of tokens utilized in training of state-of-the-art large language models[1,2] or the millions of molecules and crystal structures applied in small molecule and materials design[3–6]—which achieve incredible predictive and generative accuracy. The exact specifications of these datasets are application-dependent; for instance in the context of science, collections of molecular structures and their experimental bioactivity may be desired by medical scientists to identify effective drug molecules while materials chemists may require simulated gas adsorption or experimental stability data to select appropriate sorbent materials. Sourcing a sufficient quantity of high-quality scientific data, which must be either extracted from experimental studies or calculated at great additional computational cost, remains a significant hurdle in developing ML- and AI-guided materials evaluation and discovery models. Contemporary approaches pairing the vast repositories of experimental and/or hypothetical chemical structures with high-throughput simulation or experimental data-mining techniques yielded large materials databases that shrink these gaps in the required data. One example of many can be seen in the ChEMBL[7] database which assembles millions of chemical compounds alongside experimental and simulated bioactivity data from various sources thereby facilitating influential molecular ML studies.[4,8–10] Comparable endeavours led to the construction of numerous small molecule (*e.g.*, ZINC,[11] QM9,[12] PubChem,[13] *etc.*) and crystal structure (*e.g.* Materials Project,[14] ICSD,[15] CSD,[16] *etc.*) databases serving vital functions in the ML and AI-driven innovations of modern computational chemistry research.

While the establishment of these materials databases has allowed researchers to apply deep learning techniques to scientifically relevant tasks with remarkable accuracy, the validity of the underlying data is increasingly called into question. Beyond the discrepancies between simulated and real properties that may be expected, the constituent chemical structures may possess inadequacies affecting performance in many cases. Solely within the field of metal–organic frameworks (MOFs), our recent investigations into the major materials databases applied to ML tasks found that upwards of 40% of the MOFs composing these databases are chemically invalid due to errors in their crystal structure.[17] Furthermore, the structural building units (SBUs) that may be applied to construct novel hypothetical

[a]*Department of Chemistry and Biomolecular Sciences, University of Ottawa, 10 Marie Curie Private, Ottawa K1N 6N5, Canada. E-mail: twoo@uottawa.ca*
[b]*TotalEnergies OneTech SE, Palaiseau, France*

structures in generative materials discovery schemes—such as those employing diffusion models,[18] genetic algorithms,[19] variational autoencoders,[20] reinforcement learning,[21] and so on—were also found to contain approximately 13% and 51% chemically invalid organic and inorganic structures, respectively, upon manual inspection of the matching experimental publications.[22] This chemical invalidity manifests as a consequence of various structural errors that emerge during the experimental crystallographic analysis or due to decisions made during the dataset construction process. Generally speaking, these errors result in incorrect atomic position data within the crystallographic information (*e.g.*, omission of key atoms or molecules, duplication of atom sites, *etc.*) which is relied upon for dataset creation and consecutive ML model training. For example, unseen structure errors have been shown to produce deviations in the simulated adsorption properties,[22] and they would be expected to alter nearly all simulated properties relying on accurate atomic positions such as textural properties and electronic structure properties. Additionally, application of graph-based techniques—commonly used in chemistry tasks due to the natural description of molecules and periodic crystal structures as graphs—is expected to be severally hindered by these erroneous structures as the resulting graph representation will not be chemically reasonable owing to the inappropriate insertion or deletion of essential graph nodes and edges.

As the research community has become increasingly aware of these problems in chemical datasets, efforts to discover and potentially remedy these issues continue to advance. A handful of specialized datasets—for instance, the QMOF[23,24] and ARC-MOF[25] databases—were constructed with strict structure processing and validation guidelines to improve data reliability and facilitate immediate use in machine learning and simulation. Moreover, the availability of automated error detection algorithms such as MOSAEC[17] enables renewal of datasets through elimination of any data related to erroneous crystal structures. Though these developments boost user confidence by narrowing existing datasets to only chemically valid entries, they do not attack the dataset quality and structure error crisis at its core. A restorative approach that deals with structural errors directly instead would be preferred to retain the maximal amount of data possible. Envisioning such a framework remains difficult as current approaches are generally only capable of providing binary evaluations regarding the presence of any errors but are not capable of clarifying the errors' origin. While this distinction may appear minor, the diverse categories of structural errors necessitate unique repair solutions and as such exact knowledge regarding error types present in a given structure is necessary to prevent further introduction of errors—as observed in various efforts to repair crystal structures in the CoRE[26,27] and CSD[28] computational MOF databases.

In this work, we address the rampant structural errors in computational materials databases through the development of SETC (Structure Error Type Classification), a graph neutral network (GNN) model which accurately classifies the types of error present within MOF crystal structures. To that end, a manual investigation into over 17k MOF crystal structures and their associated publications was conducted to generate a dataset containing mappings between chemical structures and their relevant structural error type labels. This one-of-a-kind dataset was subsequently employed in a supervised learning approach to train binary-relevance and multi-label classification models which diagnosed the presence of errors in crystal structures' charge, disorder, and protonation. This technique proves to be a powerful tool in the assessment of crystal structure validity, achieving excellent classification accuracies for all studied crystal structure error categories. These findings represent a significant feat in the automation of database refinement for high-throughput computational screenings and ML/AI-guided studies as it achieves comparable performance to state-of-the-art error detection methods[17] which themselves only determine when an error is present and cannot classify error type or source. We further explore how various model architectures and graph featurization approaches affected the performance of SETC, including the introduction of a new approach to implement formal charges and metal oxidation states as an advantageous atomic node feature. Finally, an analysis of the models' explainability and generalizability is undertaken to postulate how their underlying processes may be analogous to the protocols utilized by an expert chemist when manually evaluating chemical structure validity and disentangling the possibility of various coincident errors.

Methodology

Error type categories

Numerous errors may arise in the experimental characterization and computational preprocessing of a crystal structure resulting in failure to match its true chemical composition and connectivity. These issues manifest in the crystallographic information in many ways, including but not limited to atomic overlap, overbonded atoms, atom site omission, and charge imbalances. Tools already exist to capture the simplest cases such as simple atomic distance calculations to detect atomic overlap or bond connectivity criterion to detect overbonded or underbonded components, thus this work aims to identify the more complex categories of structural errors. Three major categories of crystal structure errors were defined during manual inspection which currently observe no consistent remedy, namely hydrogen atom omission, improper charge balancing, and crystallographic disorder. Fig. 1 demonstrates each category of crystal structure error considered during the dataset labelling procedure using specific examples identified by their structure filename (*e.g.*, "WAHMEY_clean", "BAXFUD_clean", *etc.*) in the source database. Fig. 1a demonstrates how hydrogen atom omission—herein simply denoted as proton errors—manifest in a MOF crystal structure. This class of error is expected given the difficulty of detecting light (*i.e.* low electron density) atoms using X-ray crystallography,[29] and the relative lack of access to neutron diffraction instruments. While algorithms exist to place hydrogen atoms in expected positions when they cannot be pinpointed from the crystallographic data, our previous investigations in MOF and SBU structures determined that these correction procedures frequently produced

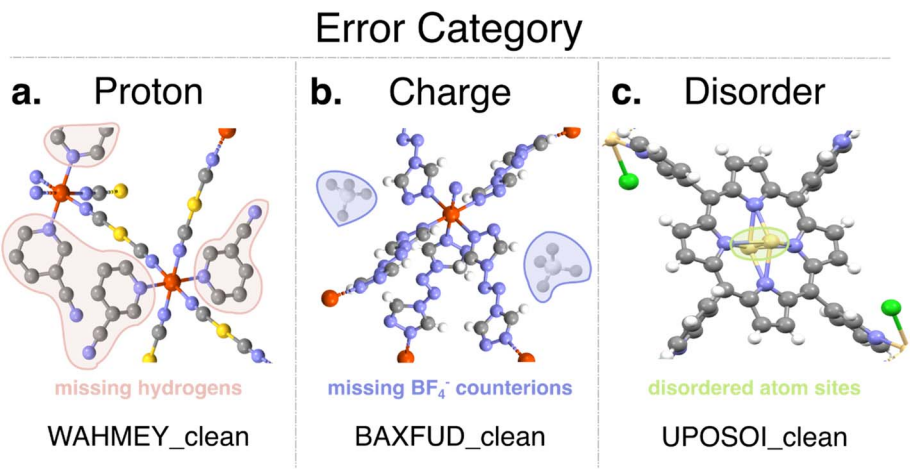

Fig. 1 Illustrative examples of the three categories of structural errors observed in MOF crystal structures present in the CoRE-2019 database. Instances possessing a single error relating to their (a) protons, (b) charge, or (c) disorder are depicted for clarity, though assorted combinations are possible. Grayscale models of the counterion missing from case (b) are included for visual clarity.

erroneous protonation and/or failed to protonate atoms near metal atom sites. Next, the impact that charge imbalance errors—herein simply denoted as charge errors—presents in a crystal structure is highlighted in Fig. 1b. Issues with the charge balance in a given structure most often arise due to the improper modelling of non-coordinating counterions and/or the omission of necessary charged ligands. These charge-balancing issues may originate from difficulty in the experimental determination of charged counterions present in porous materials; however, it was observed that these errors were regularly introduced by shortcomings in the preprocessing algorithms employed to generate 'computation-ready' structures from raw, experimental crystal structures. Failures to identify charged counterions or ligands during solvent removal can lead to unreliable outcomes with numerous crystal structures possessing undisclosed framework charges. Lastly, Fig. 1c presents a case where crystallographic disorder errors—herein simply denoted as disorder errors—produce a structure which does not accurately reflect the material's true chemistry. Crystallographic disorder is a physical reality of the imperfect processes in experimental crystal growth and crystallography which presents as uncertainty in atomic positions and/or composition of the solved crystal structure. These experimental artefacts stem from various sources, such as the dynamic nature of the crystal (*i.e.* presence of molecular vibrations and rotations), symmetry, imperfections in a real crystal (*i.e.* mosaicity, impurities, *etc.*), and so on. Certain instances of crystallographic disorder are simple to detect due to their symptomatic overlapping atoms; however, this is not always the case as disordered structures may pass even the most rigorous structural validity examinations. These three classifications of crystal structure error were determined to encompass the vast majority of errors found within the selected class of materials, and intuitively these groupings conformed with chemical intuition regarding both the sources and repercussions of such errors. Furthermore, these labels were selected as we envision that each distinct error category would be handled quite differently in any future efforts to restore these erroneous crystal structures to a chemically sensible state.

Crystal structure error dataset

Contemporary understanding of crystal structure errors and methods of their determination are thus far underdeveloped in computational screening and ML communities. Recently, we outlined a method of automatically identifying the presence of such errors through consideration of various metal oxidation state-based criteria;[17] however, this method may only identify the occurrence of an error within crystal structures and not the exact type or origin of the error(s) at issue. Consequently, due to the dearth of validated and accurate tools available for determining structural error type, considerable manual intervention was necessary during data curation to generate the requisite crystal structure to error type mappings. This innovative dataset was constructed through manual inspection of over 17k total MOF crystal structures—in conjunction with the content of their associated publications—to yield precise information on the cause of any constituent errors. It is worth noting that the ground-truth, error-free crystal structure was assumed to be the pristine crystal structure reported in the published crystallographic information, excluding any structural defects that may exist such as missing inorganic node or organic linkers in the case of MOFs. The distinction between pristine and defective crystal structures is fundamentally different from the concepts of erroneous *versus* error-free crystal structures. To obtain reasonable results in atomistic simulations or machine learning applications, crystal structures representing either pristine or defective materials should be free from structural errors otherwise the physical model will contain chemical motifs which are not stable. In the event that the exact nature of the synthesized material was unclear from the information contained in the experimental text, the crystal structure was omitted from the dataset entirely to prevent uncertainty in the error labelling. Examples of such structures are presented in the

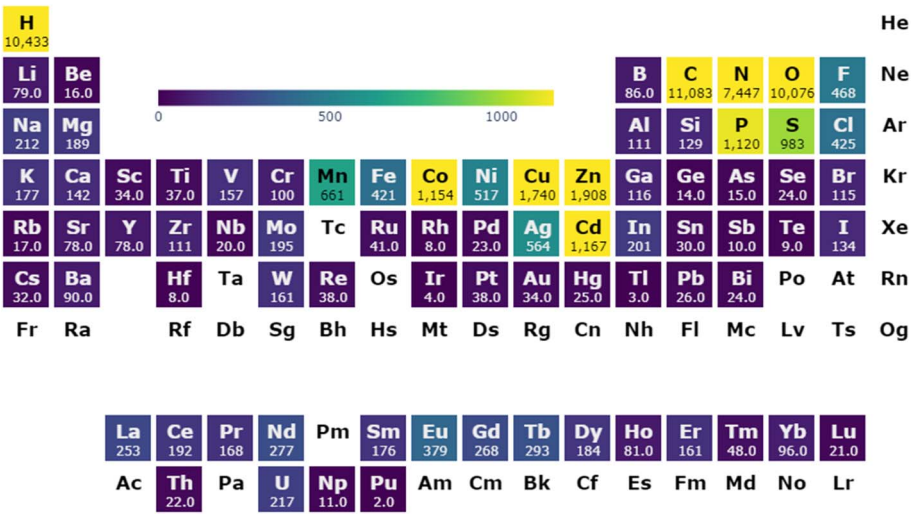

Fig. 2 Periodic table visualization of the crystal structure counts by element in the error-labelled dataset. The maximum of the colour gradient is set equal to 10% of the total dataset size (generated *via* pymatviz[34]).

SI (Fig. S1), but ultimately only approximately 99 crystal structures were eliminated from the dataset for this ambiguity. Though an error could be theoretically assigned based on structural error trends seen in similar crystal structures and analysis of their metal oxidation states, the complicated interplay between co-occurring error types would render any such error label estimates questionable at best. In addition to uncertainties stemming from the reported crystallographic information, systematic errors in the manual labelling protocol relating to the experts' own biases when extracting published details remain possible despite our best efforts. To combat this, the labels of more than 1000 crystal structures were independently revisited during SETC model development and performance validation to certify the veracity of the previous error type designations. This analysis generally found low absolute incidence of labels requiring correction (<2% of those sampled), however, we also provide the full error label data alongside this work for further consideration.

The investigated MOF crystal structures were collected primarily from the CoRE-2019 (ref. 26 and 27) MOF database (version 1.1.3) which has been previously shown to contain extensive structural errors.[17] The dataset was augmented using a second collection of structures to counteract the relative underrepresentation of the disorder error type discovered within original dataset sampled solely from CoRE MOFs. This observation is thought to stem from the CoRE-2019 MOF databases' decision to includes a text-based disordered atom removal step and to separate disordered and non-disordered samples during crystallographic information preprocessing.[26] These structures in the "augmented" set were directly retrieved from the CSD MOF Subset[30] with disordered sites intentionally retained, and a recently reported solvent removal scheme[31] (SAMOSA) was applied to produce "computation-ready" crystal structures akin to those found in CoRE-2019. Additionally, a high degree of structure duplication was observed upon inspection of the CoRE-2019 MOF database, thereby requiring implementation of a duplicate identification and removal protocol to prevent data leakage during model training and testing. This duplicate evaluation protocol is developed on the basis of differences between simulated powder diffractograms to produce crystal similarity scores.[32,33] Full details regarding the process employed to identify and eliminate duplicated structure examples are presented in the SI (Fig. S4). Overall, this dataset curation step yielded a final dataset containing *ca.* 11.4k non-equivalent crystal structures composed of 9.9k CoRE-2019 MOFs and 1.5k CSD MOFs manually labelled based on their enclosed crystal structure errors. Despite the comparatively large reduction in the total data quantity, the final dataset of unique error-labelled crystal structures encompasses a significant portion of the periodic table as depicted in Fig. 2. A broad range of chemical compositions are represented including 78 unique elements appearing in at least one structure and 42 elements existing in more than 100 structures. Ultimately, this is expected to incorporate a majority of the diverse chemistry currently seen in experimental MOFs, and ideally allow for the production of generalizable crystal structure error classification models.

Data processing & graph representation

Given a crystallographic information file representing a MOF crystal structure, an undirected graph $G = (V, E)$ is generated with each vertex V representing an atom and each edge E representing a chemical bond. Chemical bonding tables are inconsistently present in these structural files; thus, graph edges are calculated *via* the Isayev nearest neighbors algorithm[35] implemented in the pymatgen package.[36] This algorithm identifies bonded neighbors according to the presence of a shared Voronoi facet between atoms and interatomic distance of less than the sum of their covalent radii and an additional bond tolerance value. As described in a previous application of GNNs to MOFs,[37] we opted to increase this tolerance value to 0.5 Å to ensure that all bonds, including the highly variable metal–ligand bonding and those crossing a periodic boundary, are

Table 1 Classes of node features employed to describe individual atoms within the crystal structure graph representations

Node feature category	Quantity	Description
Atomic	8	Physical properties of the atom *e.g.*, Atomic number, atomic weight, van der Waals radius, first ionization energy, Cordero covalent radius, Rahm atomic radius, static dipole polarizability & Ghosh electronegativity
Local chemical environment	168	Radial and angular functions encoding the atom's local environment *e.g.*, atomic property weighted distribution functions, weighted average atomic properties, and weighted atom-centered symmetry functions
MOSAEC	3	Atomic formal charges or oxidation state calculated under three distinct charge distribution routines (*i.e.*, spanning from fully local to fully global charge distribution to metal atoms) implemented in MOSAEC

properly identified. Each vertex is assigned a vector containing node features which aim to describe the corresponding atom's physical properties, as well as to encode information regarding its chemical environment within the unit cell. The three primary categories of descriptor that were employed as node features in this work include: (i) atomic property-based, (ii) local chemical environment-based, and (iii) formal charge and oxidation state-based descriptors. Concise descriptions concerning each of the three descriptor categories are provided in Table 1, including their underlying properties and total feature vector count. The first two were calculated as described in our previous work applying graph attention networks to partial atomic charge prediction.[37] The physical properties of each atomic node were sourced from databases contained in the Mendeleev[38] Python library, while the local chemical environment description were computed using an internal code which analyzes atomic pair distances to compute the various, well-known mathematical representations of chemical environments described in the SI (Section S2).The final feature category was calculated *via* an automated oxidation state and formal charge calculation tool known as MOSAEC.[17] This method has been previously validated to produce highly accurate oxidation states for thousands of diverse MOF crystal structures, which represents a significant innovation that—to the best of our knowledge—no previous technique is capable of performing in both an automated and generalizable manner. Crucially, a comprehensive validation of the oxidation states and formal charges has not been completed on an atom-by-atom basis as prior efforts focused on solely the metal oxidation states; therefore, discrepancies may remain between values assigned manually by an expert chemist and the automated MOSAEC method. These metal oxidation states proved to be an invaluable tool in the detection of structural errors in both manual and automated investigations, thus it was anticipated that they would also be an equally valuable addition to the atomic node embeddings. Various combinations of these descriptor categories were applied in the training of distinct GNN models to assess their relative efficacy towards error type classification tasks, which is discussed further in the following sections.

Graph neural network architecture

The SETC model aspires to perform graph-level classification of structural errors based on the output of a GNN trained on the error-labelled dataset. For this purpose, we tested several GNN architectures such as graph convolution (GCN), graph attention (GAT), graph isomorphism (GIN), and GraphSAGE networks. The outcomes of this test are further outlined in the SI (Table S3), but in summary, the GAT models were determined to be better suited to this error classification problem. A demonstrative overview of a GAT model architecture and end-to-end workflow from crystal structure graph input to error label output is provided in Fig. 3. The graph representing a n-atom crystal structure possesses initial d-dimensional node feature vectors $V \in \mathbb{R}^{n \times d}$ which are first standardized using the mean and standard deviations of each feature in the training data. These scaled embeddings are then input into a N-layer GAT to generate hidden node embeddings of size h. These hidden embeddings are subsequently mean-pooled and passed to a readout layer consisting of sequential linear transformations and a final sigmoid activation to produce the desired m-dimensional vector containing probabilities of each outlined structural error type label. Additional details regarding the implementation of this GAT model architecture are available in our corresponding work utilizing GATs to predict DFT-quality partial atomic charges in periodic MOF structures.[37]

Hyperparameter optimization & training

All graph manipulation and neural network functions were implemented using the PyTorch Geometric and PyTorch libraries.[39,40] The Adam optimization algorithm[41] was employed to train the model for a total of 64 epochs with the objective of minimizing the binary cross-entropy loss between the predicted and ground truth crystal structure error labels. A ten-fold cross-validation (CV) routine—comprising an 80 : 10 : 10 ratio of the training, validation, and test sets, respectively—with stratification according to error label was applied during model training and evaluation, hence all reported values represent the average performance across all data splits. The Optuna[42] optimization framework was applied during hyperparameter tuning using the default Tree-Structured Parzen Estimator (TPE) sampling algorithm for 100 trials to maximize each models' classification accuracy on the validation set. Hyperparameters were obtained from the initial data split of each model only and they were assumed to be transferable to all successive folds. A summary of the model hyperparameters which were subject to optimization is presented in Table S2.

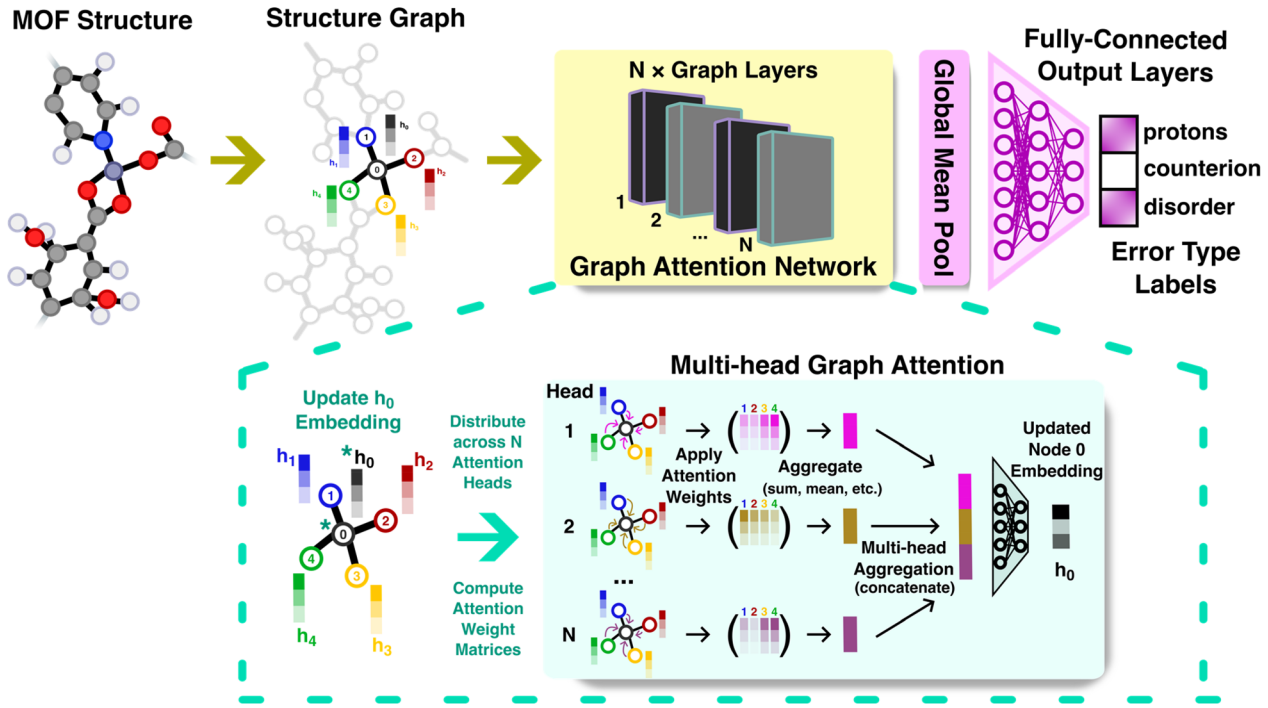

Fig. 3 Overview of the SETC graph attention (GAT) model architecture, including illustrations spanning from the crystal structure input to the multi-head graph attention mechanism to the final error type label outputs.

Graph explainability & feature importance evaluation

Various analyses of the trained error type classification model outputs were performed to gain insight into how SETC differentiated between erroneous and pristine MOF structure graphs. Investigations into the hidden layer embeddings, both the global-pooled structure graph embedding and individual atomic node embeddings, were accomplished through dimensionality reduction utilizing the uniform manifold approximation and projection[43] (UMAP) algorithm. Each of these experiments employed the entire final hidden layer embedding vector to determine how well reduced embedding space clustered with respect to their ground-truth error labels. The GNNexplainer[44] graph explainability technique was also utilized to assess the relative influence of each node feature towards the classification models' predictions. This approach applies a gradient-based optimization of node feature masks and edge masks on the studied graph to identify the edge and feature subsets that best preserve the trained GNN model's prediction. Several GNN explainability algorithms which identify influential subgraphs were tested, but ultimately the SubgraphX[45] method was found to provide the most robust and coherent explanations which is in accordance with a recent survey[46] on graph explainability techniques. This method incorporates a Monte Carlo tree search to rapidly sample combinations of subgraphs and rank their relative contribution towards the prediction using a Shapley value-inspired scoring function. The resulting explanations produced a simplified subgraph and node feature importance values which were subsequently inspected to evaluate the classification models' decision-making processes and whether the identified substructures matched chemical intuition regarding error identification. Implementations of both the GNNExplainer and SubgraphX techniques within the GraphXAI[46] python package were applied to complete this analysis. The overarching aim of the above analyses was determining if the underlying GNN architecture had effectively learned or approximated the fundamental chemical information, such as conventional bonding relationships or proper formal charge and oxidation state values, that a chemist would engage to judge a crystal structure's quality and errors manually.

Results & discussion

Dataset error distribution

The distribution of each error category across the dataset—including both the CoRE-2019 and "augmented" CSD crystal structures following duplicate removal—and their co-occurrence within erroneous structures are presented in Fig. 4 while a full accounting of the contents of each dataset is presented in Table S1. At least one kind of error—herein described as "any" relating of whether any of the three error types are present—was observed in 55.86% of the constituent crystal structures, while the most common single error type relating to charge balancing issues was seen in a total of 33.38% of the dataset. These findings are consistent with the rates of error occurrence presented in our previous inspections of SBU libraries and MOF databases,[17,22] which found that upwards of half of all MOFs in several databases possessed problematic metal oxidation states and 36.7% of all extracted inorganic SBUs contained incorrect charge assignments. The error type

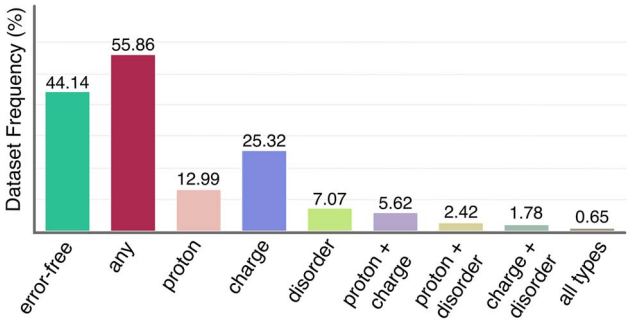

Fig. 4 Visualization of the frequency of each combination of error labels across all crystal structures present in the final dataset comprising the CoRE-2019 and "augmented" datasets post-duplicate crystal structure elimination.

distribution in the initial, investigated sample of 15.6k CoRE-2019 MOFs (Fig. S2) noted similar rates of proton, charge, and general error incidence despite posting a significantly lower total frequency of disordered structures (4.87%) compared to the final dataset (Fig. 4) post-augmentation with disordered crystal structure data. As one might suspect, earlier attempts at training classification models with such low disorder frequency produced limited precision in the disorder-related tasks; thus, dataset augmentation targeting the addition of disordered crystal structures was initiated and successfully increased disorder error frequency by *ca.* 7% while only slightly changing the distribution with respect to the other error categories. Further complicating the error label landscape is the relatively common observation of multiple categories of errors within a single crystal structure as shown by the rightmost columns in Fig. 4. The proton-charge error pairing is detected in 5.62% of all crystal structures in the dataset, while the proton-charge and charge-disorder pairings represent 2.42% and 1.78%, respectively. Approximately 0.65% were also found to contain all three error label categories studied in this work; two illustrative examples of which are highlighted in Fig. S3. This relatively frequent co-occurrence of error types significantly complicated the manual labelling process as the assignment of borderline cases may be ambiguous without details from the original publications' crystallographic analysis; thus, capturing the interplay between the error categories will be integral to achieving highly accurate ML error type classification.

Feature selection

To assess the optimal combination of node features in the graph-level error label classification task, several combinations of the three classes of node features (Table 1) were employed to prepare distinct graph datasets. In addition to the simple pairing of these feature groups, feature reduction using large positive or negative Pearson correlation coefficients ($|r| > 0.5$) as the discriminating variable was examined. Details of this practice are provided in the SI (Fig. S6) and ultimately yielded a "reduced" dataset retaining 74 total features across the three groups in its node embedding. Finally, a greatly reduced combination preserving only 4 node features was proposed through handpicking important descriptors according to chemical intuition. This group of features, herein dubbed "chemist", contains only the atomic number (Z), oxidation states and formal charges (*i.e.* the MOSAEC features) as these concepts—alongside the chemical bonding information contained within graph edges—encompass those employed during the manual inspection process that created the error-labelled dataset. Manually calculating oxidation states is a relatively simple, yet time-consuming process for the average chemist; however, the development of an automated approach to computing these values in a high-throughput manner is incredibly complex due to the diversity of chemical moieties which it must cover. The MOSAEC[17] algorithm represents a significant innovation in this regard by facilitating the incorporation of oxidation states and formal charges as graph node features in an automated and accurate fashion. Comparisons of the performance of binary relevance models trained using each of the seven tested node feature datasets in error label classification are summarized in Table 2. Generally, across all error categories, datasets containing the oxidation state and formal charges as node features achieved the greatest ROC-AUC scores compared to those without this information. Indeed, only considering these oxidation state and formal charge node features observed co-leadership in the general, *i.e.* presence of 'any' error type, error identification task (AUC = 0.967). This discovery is somewhat unsurprising as it concurs with our previous reports on the reliability of using metal oxidation states as a metric to automatically flag problematic crystal structures. Moreover, superior performance in two of the specific error classification tasks, proton and charge error labels, was shared by the chemist and atomic-MOSAEC models, respectively. The disorder errors were most effectively labeled in representations solely possessing simple atomic features, closely followed once again by the simple chemist representation. Each of these instances represent some combination of atomic and oxidation state-based features, thereby indicating that a high degree of classification efficiency may be achieved with relatively interpretable descriptors. We originally anticipated that the features encoding local chemical environment information (*i.e.* radial and angular functions) would be vital in mapping the complex spatial relationships necessary when diagnosing crystallographic disorder; however, models trained on graphs with such features performed worse on average than analogous models which omitted them. It is possible that the sizable quantity of these descriptors introduced redundancy and noise during model training without providing any additional benefits beyond what could be deduced from the chemical bonding information stored in the edge connectivity. The 'reduced' dataset observed similarly poor performance which may be reasonably forecasted as numerous local environment features were kept due to their low correlation coefficients. In summary, the results presented in Table 2 indicate that combinations of atomic properties and oxidation state/formal charge-based features supply sufficient knowledge to generate GAT models with high predictive accuracy towards the graph-level classification of structural errors. Subsequent analyses will focus primarily on those high-performing features, namely

Table 2 Classification performance of the SETC model employing various combinations of node features in the input structure representation

	Error label category			
	Any	Proton	Charge	Disorder
Selected node features	ROC-AUC[a]	ROC-AUC[a]	ROC-AUC[a]	ROC-AUC[a]
Atomic	0.928 ± 0.008	0.832 ± 0.017	0.826 ± 0.017	**0.940 ±0.014**
MOSAEC	**0.967 ±0.005**	0.935 ± 0.008	0.902 ± 0.011	0.923 ± 0.009
Local	0.903 ± 0.009	0.870 ± 0.015	0.888 ± 0.011	0.864 ± 0.020
Atomic & MOSAEC	**0.968 ±0.005**	**0.947 ±0.009**	0.904 ± 0.011	0.901 ± 0.014
Atomic & local	0.888 ± 0.009	0.876 ± 0.020	0.905 ± 0.009	0.916 ± 0.017
MOSAEC & local	0.909 ± 0.007	0.914 ± 0.013	0.861 ± 0.010	0.853 ± 0.016
All	0.923 ± 0.011	0.906 ± 0.010	0.869 ± 0.006	0.911 ± 0.016
Reduced	0.890 ± 0.009	0.914 ± 0.014	0.864 ± 0.011	0.896 ± 0.017
chemist[b]	**0.966 ±0.004**	**0.949 ±0.010**	**0.925 ±0.008**	0.932 ± 0.024

[a] Values are reported as means and standard deviations across ten cross-validation folds. [b] Includes the atomic number (Z) and oxidation sates/formal charges calculated *via* MOSAEC.

studying models trained on the 'chemist' and 'atomic-MOSAEC' datasets as they achieved the foremost aggregate performance across all classification tasks.

Classification approach

Two distinct classification methods, consisting of either binary relevance models or multi-label model output, were contrasted to evaluate whether any advantage could be gained from a unified model learning the correlation between error labels. Table 3 compiles each approach's classification performance for models trained on the 'chemist' graph representations in terms of the ROC-AUC, F1-score, and accuracy for each of the three error label categories, as well as the accuracy and Hamming loss associated with predicting the final output vector containing all labels. In this instance, binary relevance classification outperformed models designed to learn the three distinct error label categories simultaneously. Combining separate models observed a more than 6% advantage in classification accuracy of the overall ground-truth error type vector, and each individual error label noticed similar enhancements in their accuracy culminating in a change of 3.2% in their Hamming loss. The most apparent difference occurred in the disorder error classification models which observed 3.1% improvement relative to the multi-label approach. These results would initially suggest that any information gained by the multi-label models from the correlation between error types did not provide significant benefits to the classification performance; however, this phenomenon was not observed across all node feature combinations as outlined in Table S4. Largely, in the case of the second-best graph representations utilizing atomic and MOSAEC features, the multi-label classification method achieved superior performance in predicting the overall complement of error type labels compared to training three separate binary-relevance models. The multi-label model maintained better performance across all validation metrics for the charge and disorder error labels, however binary models for proton errors observed improved classification ability. This instance of the multi-label approach predicted the exact vector of error labels in 75.1% of structures and a Hamming loss of only 10.5% reflecting low failure in the predictions of the individual error labels components. This marked a nearly 1.3% increase in accuracy relative to the merged binary relevance model outputs. In the case of atomic-MOSAEC node representations, there may be beneficial information gained from the correlation between the distinct error categories that is lost in the case of the binary relevance models. While the binary relevance models using the 'chemist' node features surpassed all others in general, this disparity in the classification approach trends indicates that the utility of shared training of the error labels may be heavily dependent on the quantity and category of node features supplied to the model as we discovered no particular trend which held true across all experiments. The underlying relationship between occurrences of proton, charge, and disorder structural errors was not straightforward from the findings of the manual investigation process, once more demonstrating the potential of GAT neural network-based approaches at disentangling complex connections from crystal structure data.

Table 3 Performance of the binary relevance and multi-label approaches towards the classification of individual and overall error labels using chemist node features

		Classification method	
Error label	Classifier score	Binary relevance	Multi-label
Proton	ROC-AUC	0.949 ± 0.010	0.931 ± 0.010
	F1	0.854 ± 0.017	0.747 ± 0.021
	Accuracy	0.938 ± 0.007	0.894 ± 0.009
Charge	ROC-AUC	0.925 ± 0.008	0.908 ± 0.010
	F1	0.786 ± 0.020	0.749 ± 0.023
	Accuracy	0.868 ± 0.011	0.843 ± 0.013
Disorder	ROC-AUC	0.932 ± 0.024	0.896 ± 0.010
	F1	0.829 ± 0.038	0.628 ± 0.036
	Accuracy	0.961 ± 0.008	0.930 ± 0.006
Overall	Hamming loss	0.079	0.111
	Accuracy	0.797	0.733

Comparison to global features

Given that this method of crystal structure error labelling and classification is relatively nascent within chemistry ML applications, one may speculate whether the graph representation selected in this work is as advantageous as asserted in the above sections. While graph structures provide the most natural representation of chemical structure data, global feature-based methods—which describe each crystal structure as a series of chemical and physical properties calculated for the structure in aggregate—are conventionally utilized in property prediction and classification tasks on MOF datasets. Therefore, a reasonable question becomes whether ML models based on global features are capable of differentiating erroneous crystal structures with a similar degree of efficiency as the graph-based models. To assess the effectiveness of global features in error type classification, we select four categories of features (Table S5) frequently employed in ML regression and classification models studying MOFs and other periodic materials: (i) geometric, (ii) persistent homology, (iii) revised autocorrelation (RAC), and (iv) atomic property-weighted radial distribution function (RDF) descriptors. These features have proven successful in various ML tasks, and they cover a broad range of physical properties describing the pore environment and chemical properties (*i.e.*, nuclear charge and mass, Pauling electronegativity, polarizability, *etc.*) present within each structure's unique local atomic environments. Additionally, several machine learning algorithms regularly applied to chemical classification tasks were studied to fairly gauge the utility of this global approach: (i) random forest (RF), (ii) gradient-boosted decision tree (GBDT), (iii) support vector machine (SVM), and (iv) neural networks (NN). Comprehensive details of the featurization, hyperparameter tuning, and model training procedures for these global feature-based models are provided in the SI (Section S3), but importantly, care was taken to closely replicate the graph-based models' training procedures. A summary of the performance of the leading graph-based and global feature-based models on the targeted error label classification tasks is presented in Fig. 5. The SETC graph-based model outperformed the global features in identifying the presence of any generic error type quite significantly by a margin of greater than 7% accuracy. A less glaring improvement was also discovered in the proton error classification task with the top graph-based model only achieving a *ca.* 2.5% enhancement over the best global counterparts. In contrast, the charge and disorder errors observed more similar performances with relative differences in accuracies of less than 1% across all structural representations for their best performing models. Notably, NN and GBDT models outperformed GNN-based methods in the charge error classification task, but the magnitude of these performance improvements falls within the reported fold-wise variance in classification accuracy. This finding may indicate that, compared to the structure graph representations, certain global features may possess a higher sensitivity to the characteristic pore environment changes induced by charge errors associated with the omission of non-coordinating counterions. Interestingly, investigations into

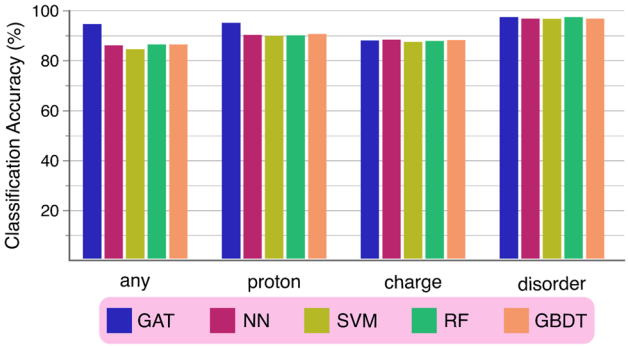

Fig. 5 Error type classification performance of graph-based and global feature-based models compared according to accuracy. The top-performing model achieved for each algorithm is presented for each error category.

each individual global features' effect on classifier performance (Tables S7–S10) indicated that the RAC descriptors were most highly correlated with improved classification performance across all ML algorithms and feature combinations, appearing in nearly all of the best performing models. Notably, this class of features is one of the few that implicitly consider chemical connectivity akin to the information contained within a chemical structure graph.[47] Specifically, when calculating the atom-wise contributions to their RAC descriptor, the function's scope is limited to only other atoms within a certain chemical bond-walk distance (*i.e.* number of bonds separating atom *i* and *j*) of the considered atom, thereby injecting some amount of connectivity information into the otherwise simple, scalar features. These observations further substantiate our hypothesis regarding the importance of considering bond connectivity when assessing crystal structure errors, and how graph representations—and to a lesser extent, connectivity-minded features such as RACs—mimic the approach that an expert chemist would apply when visually assessing a crystal structure's validity.

Outside of the abovementioned stellar performance of the RAC features, the remaining three global feature categories were much less suitable to distinguish between high-quality and erroneous structures. Accuracies ranging between 64.2–65.1%, 75.3–77.2%, and 76.9–77.7% were observed for the geometric, homology, and RDF features, respectively, towards the generic error determination tasks. This level of classification accuracy suggest that these global features do not contain the requisite sensitivity to minor structure errors such as single proton omissions, and as such likely cannot properly learn the necessary structure–error relationships. This outcome is perhaps unsurprising as these features represent aggregate sums over the crystal structure as a whole and are thus inherently less sensitive to errors which may often only affect a small subdomain of the crystal structure. In instances where significant distortions of the correct crystal structure occur, such as crystallographic disorder which can lead to massive quantities of duplicated or overlapping atoms, these other three features—and the persistent homology, in particular—are sufficiently

altered and result in significant improvements in the disorder error label classification tasks. In truth, a single category of feature remained insufficient in even such drastic error cases and ultimately, we found that combinations of several global feature types were typically required. This added complexity in the featurization process represent substantial overhead in the overall global feature-based error classification workflow, particularly considering that competitive accuracy can be achieving with even the simplest graph representations containing simple, rapidly-assigned atomic and oxidation state (MOSAEC) node features.

The complicated nature regarding how crystal structures manifest in the global features is further clarified upon performing an analysis of the dimensionality-reduced space of each global feature category as presented in Fig. S7–S10. If structural errors resulted in large deviations in the computed features' values, one would expect a clear boundary between the erroneous and correct crystal structures within the reduced global feature space; however, datapoints representing structural errors are interspersed throughout the reduced feature coordinate with no clear separation relative to the chemically accurate crystal structures in almost all feature categories. A minor degree of clustering is discovered in the persistent homology and RAC feature spaces with respect to the charge and disorder error labels. This may partially explain the improved classification accuracy observed across the global feature-based models in determining these error labels; however, overall, this inspection of the reduced feature space suggests that a complex relationship exists between the global features and structural error labels which may not be easily interpreted or modeled *via* conventional ML classification approaches. In light of this lack of correlation between the error labels and global features and their relative inability to meet the graph-based methods' classification accuracy, we conclude that the proposed SETC GAT model presents the best way forward towards applying ML error identification tools in MOF dataset curation and future repair workflows.

Hidden embedding and subgraph feature analysis

Various aspects of the final trained models' performance, including the final GAT hidden layer embeddings and the impact of subgraph structure and node features on the output, were probed to deepen our understanding of the mechanism by which crystal structure errors were diagnosed. In a generalizable error classification model, the key structure graph and node features influencing the resulting label would ideally reflect the concepts utilized by chemists in analogous tasks. Initially, the decision boundaries between the positive and negative error labels within the top-performing models were interrogated through analysis of the reduced GAT hidden embeddings. Fig. 6 highlights these boundaries established for all error label categories within the overall pooled graph embeddings for the best performing SETC models, as well as examples for an element-specific node embedding space. Distinct boundaries between the regions representing pristine and erroneous crystal structures can be clearly observed in the UMAP-reduced dimension in all instances of the pooled graph embeddings (Fig. 6a) which one may have expected given the level of classification accuracy achieved by these models. As crystal structure errors were observed to range in severity (*i.e.* affecting a few atom sites up to the entirety of the structure) during manual dataset validation, each crystal structure graph may simultaneously consist of nodes representative of both erroneous and error-free chemical environments. It is, therefore, unsurprising that the boundary between the two error

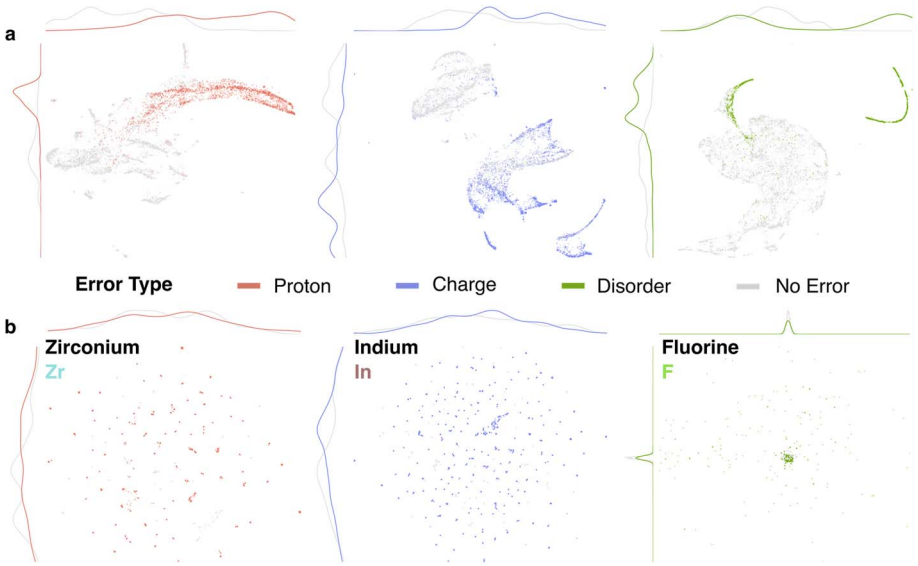

Fig. 6 Analysis of the intermediate GAT embedding vectors extracted prior to the final readout layers for the top-performing SETC models. Visualization of both the (a) pooled graph embeddings and (b) individual atomic node embeddings are achieved using the UMAP dimensionality reduction technique. Each point is colored according to the error status of their associated crystal structure graph with gray representing the absence of the error type under consideration in each plot.

classes in the individual atomic node embeddings becomes more ambiguous in Fig. 6b. However, intriguingly, despite the considerable overlap seen between the two classes for many of the individual node embeddings, definite regions of nodes belonging to only erroneous crystal structures can be discovered for each of the error categories when considering specific elements. For example, the nodes associated with zirconium (Zr) atoms experience shifts in the distribution and clustering of their hidden embeddings produced in the classification of protons errors. This discovery concurs with the knowledge that the most frequently occurring Zr species within the error-labelled dataset, the UiO-66-type Zr_6 inorganic node, frequently experienced hydrogen atom omissions at the μ_3-OH positions. Similarly, the indium (In)-containing crystal structures that frequently noted counterion omissions and F-containing crystal structures which often demonstrated rotational disorder also demonstrated some clustering between certain atomic node embeddings belonging to erroneous structures within the reduced space. However, these relationships do not follow any straightforward trend as can be clearly seen in Fig. S11 which again visualizes the atomic embedding spaces according to their error label for additional elements. For example, the C atomic embeddings (Fig. S11a) show very little correlation between their error state and reduced embedding vector space despite these atoms commonly existing in close proximity to or at the exact location of many proton omission and crystallographic disorder errors. Likewise, the Ag and O atom embeddings which were regularly affected by charge and proton errors, respectively, demonstrate minimal grouping by error status. This indicates that the relationship between these hidden node embeddings and the error state is sophisticated and may not be easily deconvoluted to extract any chemical insight or make determinations regarding the quality of the crystal structure at the studied atomic node position. This intermittent observation of segregation in the specific atomic embeddings suggests furthermore that the node-level error classification task may be possible and could be pursued in the future to generate the targeted crystal structure repair strategies necessary to restore erroneous structures; however, the absence of any clear trends for all atom identities in a given error category hints that developing such a procedure may not be straightforward. Meticulously labelling all erroneous node sites in each crystal structure in a similar fashion as implemented in this work at the graph-level would represent an exponential jump in the quantity of labour necessary to establish a labelled dataset for supervised learning, thus alternative ML learning paradigms such as semi-supervised or unsupervised learning ought to be explored. Finally, these analyses suggests that the mechanism by which SETC identifies errors in crystal structural graphs is not through the pinpointing of specific problematic atomic node positions, but rather through examination of larger, aggregate properties of the crystal structure graphs. This differs somewhat from the manner in which a chemist would manually investigate proton and disorder errors as one would typically seek out particular atoms which violate the accepted principles of chemical bonding. However, this concept does somewhat match the manual classification approach for charge errors which generally need a more global view of the charge balance across the entire crystal structure as non-coordinating charges often play a major role.

Investigations into the relative importance of each node feature and subgraph structures towards the final error label prediction were conducted to again discover any corollaries to the process employed by chemists during crystal structure inspections. Interpretable explanations of the trained models' outputs for all MOF crystal structures were generated using the GNNExplainer and SubgraphX approaches, and the mean feature importance were analyzed to determine the most influential factors. A breakdown of these graph-level explanations for the 'chemist' node representation is shown for three representative examples in Fig. 7, while additional examples are presented in SI (Fig. S12–S15). The computed top-ranking subgraphs match one's expectations regarding the problematic substructure within the crystal structure graph in many instances. For example, Fig. 7a shows that the carbon and nitrogen sites in the 1,8-diaminooctane linker of SOBZOY_clean that are missing their implied hydrogen atoms were differentiated as the most important contributors towards that crystal structure's GAT-predicted error label. Similarly, in the case of disorder errors such as the CUFQUQ_P1 structure displayed in Fig. 7c, the atom sites in the pyrazine linkers which were disordered by symmetry were evaluated to be important in predictions made on that crystal structure graph. In contrast to these two error cases where assigning the erroneous node sites is relatively straightforward, charge-related errors are more complex as the omission of charge-balancing counterions may not be easily attributed to any individual atom and thusly they are more often considered as an aggregate property of the framework itself. Interestingly, subgraphs involving metals atoms and portions of their surrounding ligand environment are frequently identified as important to the charge error prediction. Given that the MOSAEC oxidation state calculation assumes an overall neutrality of the structure graph, the metal sites generally assume the responsibility of compensating for the consequences of excess framework charge implied by charge errors; therefore, it is perhaps logical that the subgraphs including these affected metal sites—such as the overly oxidized In^{4+} atom highlighted in Fig. 7b—would be characterized as relevant to the error classification. Intriguingly, in crystal structures where the subgraph analysis failed to generate an explanation coherent with chemical intuition (Fig. S12c), subgraphs containing metal sites in unlikely oxidation state were generally chosen over the more directly culpable nodes missing hydrogen atoms or possessing disorder. Once again, this finding indicates that the presence of irregular metal oxidation states endures as an effective indicator of crystal structure errors which may explain their prevalence as features in the top-performing models and further as nodes belonging to influential subgraph structures. The node feature explanations of the predicted outputs in each of the examples in Fig. 7 similarly suggests that the metal oxidation state features play a marginally more important role, but overall this effect appears minor. A comprehensive analysis of the node features is presented in Fig. S15 including investigations into other node

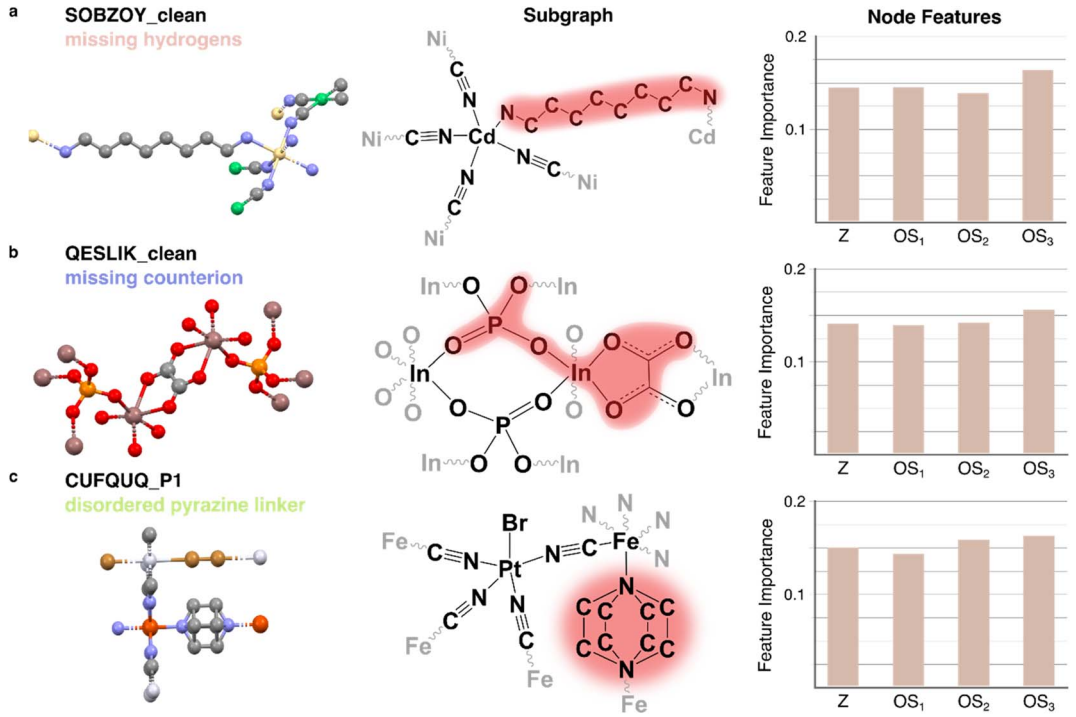

Fig. 7 Visualized explanations of the leading SETC error classification models for three representative examples of (a) proton, (b) charge, and (c) disorder crystal structure errors. Skeletal representations of each crystal structure are depicted with the substructure highlighted in red representing the most important subgraph.

feature categories and aggregate evaluations on the entirety of the dataset rather than individual structures. While individual structures observed varying influence from particular features, no obvious trend in feature importance is seen across graph representations. Additionally, the importance score differences between features of greater and lesser importance were minimal when considering the dataset as a whole, thereby we conclude that no specific feature dominates the models' prediction. Overall, this study of explainability in the SETC models indicates that chemically relevant subgraphs are preferentially utilized during error classification, ultimately signifying that general knowledge of acceptable chemical structure motifs was acquired.

Model generalizability

The ability to diagnose structural errors in various classes of materials outside of the original MOF dataset, such as molecules and transition metal complexes, was also evaluated as a final means to determine if these classifiers truly gained general chemical knowledge regarding acceptable bonding and spatial relationships in 3D chemical representations. Random samples of 100 3D structures each, sourced from several unseen databases used in computational studies of molecules, metal complexes, covalent-organic frameworks (COFs), and MOFs, were inspected in the same manner as the above-described error label dataset curation procedure. A summary of the error rates identified in each of the sampled unseen datasets is provided in Table S11 with further details available in the SI files. The frequency of errors in the molecular database samples were typically lower than either the metal complex or MOF/COF samples, which may have been anticipated from the relative maturity of molecular databases and of the tools used to recognize and repair problematic organic moieties. It is important to note that since SETC aims to investigate errors in the data supplied in 3D structural representations, only such data (*i.e.*, atomic identity and coordinates) was considered irrespective of any other available metadata. In the context of molecular (*e.g.*, ZINC[11]) and transition metal complex (*e.g.*, tmQM[48]) databases which—unlike many MOF databases—correctly label charged species, a positive charge error designation will be predicted in such structures by SETC despite the correct charge label being supplied to users. This is the expected behaviour as this model only considers the data available from the crystallographic information and not any additional database metadata; thus, these instances may be regarded as errors when users improperly utilize ionic crystal structures *i.e.* assumed them to be neutral. An analysis of the generalizability of the optimized error classification models towards the unseen samples of MOFs, COFs, molecules, and transition metal complexes is summarized in Fig. 8 which displays radar plots for each subgroup displaying their observed classification accuracy towards each error category, while Table S12 details the raw classification accuracy values. The manifestations of structural errors in molecules, metal complexes, and COFs shared many similarities to those previously highlighted in MOFs, however, a greater diversity in the chemical moieties and metal coordination environments was noted in the unseen

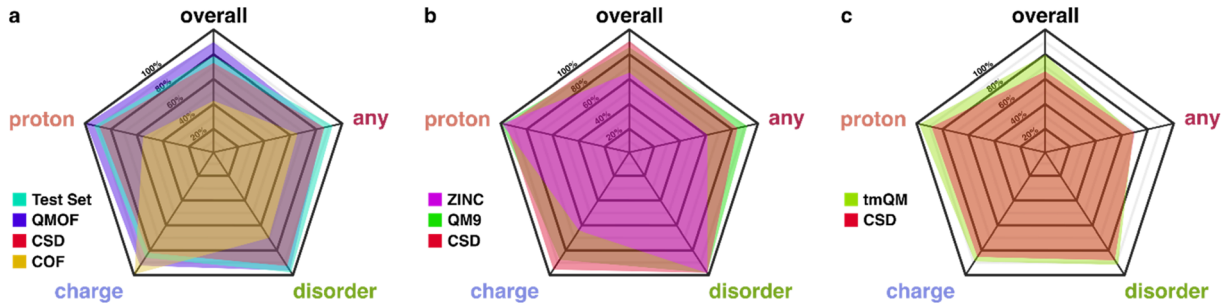

Fig. 8 Analysis of the generalizability of the SETC models employing the chemist node features towards diverse classes of materials and databases unseen during model training. Radar plots summarizing the classification accuracy observed for each of the 100 crystal structure samples of (a) MOF/COFs, (b) molecules, and (c) metal complexes in all error labelling tasks. The 'overall' error classification task represents the accuracy in obtaining the complete error label containing the three major error categories (proton, charge, and disorder), while the 'any' label represents determining the presence of any of the three errors.

samples. Overall, the binary relevance models classifying the individual proton, charge, and disorder errors performed remarkably for nearly all materials despite never encountering such small molecules and complexes during training. A much broader array of success is noted in the classification of the 'any' (*i.e.* presence/absence) and overall error labels across the unseen material groups. Superior generalizability was observed towards the samples containing chemical substructures with a greater degree of similarity to the original error-labelled dataset composed of principally MOFs. Error labels of MOFs contained in the closely related CSD MOF Collection and QMOF samples were largely in the same ranges as those found during the model evaluation on the test set structures. The models possessed dampened efficacy in identifying errors in the next nearest material subgroup, metal complexes, which observed accuracy losses between 10 and 20% in all error categories relative to the performance in MOFs. While these reductions are significant, SETC classified transition metal complex errors at accuracies which still greatly surpass a random guess, thereby suggesting that general chemical knowledge has been transferred from the models' information obtained during training on solely MOFs. Chemically speaking, this matches intuition as one would expect the structural and electronic properties pertaining to metal coordination in MOFs and metal complexes to be largely shared with some variance in the distributions of their ligand types, binding modes, acceptable oxidation states, *etc.* Remarkably, the remaining unseen samples, small molecules and COFs, which are most dissimilar to the original dataset did not universally observe declines in classification accuracy. In contrast with the previous examples, these databases contain predominantly organic chemical structures with significant functional group diversity relative to that which is typically observed in typical MOF linker chemistry; therefore, it was assumed that the loss of the metal nodes' information would lead to considerable performance losses. In fact, on the contrary, SETC observed similar or greater classification accuracies in the case of the CSD and QM9 databases samples across all metrics. The lowered performance on the ZINC20 database largely arises from the drop in charge error classification accuracy, which itself may be explained by the fact that it was sole molecular sample to possess apparent charge errors. A more substantial loss of generalizability is observed in the CURATED COF[49] sample despite their likeness to organic molecules. The disparity in performance on these non-metal-containing structure graphs may perhaps indicate that the metal node environments are highly consequential to the models' output in many situations. This hypothesis aligns with our previous determination of metal oxidation states' utility in flagging erroneous crystal structures;[17] however, the observation that proton and disorder errors continue to be well characterized in these molecular samples indicates that structural relationships learned from the limited set of MOF linker molecules are sufficient to assess structure graph validity for varied datasets of organic moieties. While these models may not provide sufficient accuracy to be reliably applied to screen databases composed of mainly organic molecules and COFs, their ability to properly assess protonation state and crystallographic disorder implies that these models are capable of handling out-of-distribution organic structures and that efforts to develop fine-tuned models specifically for organic crystals such as COFs are plausible. Finally, in summation, the efficient classification of the three major error labels in the unseen datasets suggested that the SETC models acquired a general understanding of what constitutes satisfactory bonding and compositions in chemical structure graphs. These initial findings signal that an all-encompassing, generalizable crystal structure inspection tool may be realized given sufficient understanding of structural errors and the availability of a diverse set of error labelled data from numerous branches of chemistry.

Conclusions

Concerning reports of prevalent structural errors in materials databases applied in screening and ML efforts necessitates enhanced scrutiny into the quality of data enclosed in such studies. In consideration of these concerns, we implemented an error investigation protocol applying expert chemist knowledge and rigorous literature examination to establish a dataset of MOF crystal structures labelled by their constituent structural

errors. We subsequently utilized this unprecedented structural dataset comprising over 11k unique MOFs labelled by hand according to their error type in the development of various ML classification models capable of identifying the three major categories of structural errors—namely, proton, charge, and disorder errors—with considerable classification accuracies upwards of 86% in the top-performing models. This work proposed that graph representations of periodic crystal structures should contain all the necessary information that a chemist would apply when manually assessing a crystal structure's error state. This hypothesis was confirmed by experiments which demonstrated that graph attention neural networks employing straightforward node features, such as atomic number and oxidation states, produced leading performance on average when compared with other features, GNN architectures, and global feature-based classification models. Particularly in the cases of subtle structural errors such as proton omissions, the GAT-based models surpassed the alternative by a substantial margin. However, common global features employed in prior ML investigations of MOFs (*i.e.* geometric, persistent homology, RAC, and RDF descriptors) proved capable of achieving similar performance to GATs in classifying structural errors impacting large domains of the crystal structures such as charge and disorder errors. Future studies could focus on adapting these global features to be more sensitive to even minor structural differences, potentially through incorporation of the concepts of oxidation state and formal charge which have proven to be effective in this and our previous works on error identification and classification. Additionally, there exists potential to improve the SETC error type classification workflow beyond the simple, connectivity-based graph representations towards more sophisticated GNN model architectures and graph representations that innately capture the crystal structure's geometric information. While this study focused specifically on the issue of structural error frequently found in MOF computational databases, one may expect analogous problems in other material classes when databases adapt their crystal structures from imperfect crystallographic data and postprocessing techniques. Despite being solely trained on MOF crystal structures, this GAT approach effectively classifies errors in a broad range of unseen materials classes, such as organic molecules and metal complexes. Particularly in the case of proton and disorder errors, classification accuracy nearing or greater than 90% was observed for almost all of the unseen molecular, metal complex, MOF, and COF samples. This observation, combined with the GAT explainability investigation that proposed a correlation between the chemist-derived and model-derived subgraph importance assessments, indicated that the error classification models developed a general knowledge of proper bonding relationships in crystal structure graphs. We hope this study further highlights the importance of carefully considering dataset quality in structural datasets and demonstrates how combining state-of-the-art methods with expert domain knowledge can offer new avenues of enhancing the reliability of future ML and AI endeavours in chemical applications.

Author contributions

M. G. and T. K. W. conceptualized the project. A. W., C. P., J. L., and M. G. contributed to the core software development. A. W., M. G., and R. A. M. contributed to data curation. M. G. performed the experimental investigation, visualization, and manuscript writing – original draft. All authors participated in writing – reviewing & editing. C. P. and T. K. W. provided project supervision. T. K. W. acquired project funding.

Conflicts of interest

Patent applications were filed for the method of assigning metal oxidation states and method of processing chemical compounds for computation cited in this work. There are no other conflicts of interest to declare.

Data availability

The latest version of the SETC code discussed in this work is freely available at **https://github.com/uowoolab/SETC-GAT**.

Supplementary information: Additional figures and tables relating to the dataset creation, crystal structure error label data, and SETC model training, optimization, classification performance evaluations, and other evaluations such as model explainability and generalizability (PDF). Comprehensive archive of crystal structure error label data for the MOF datasets, and all samples applied during the generalizability analysis, as well as the model outputs for each of the described evaluation efforts (ZIP). See DOI: **https://doi.org/10.1039/d5sc05477j**.

Acknowledgements

Financial support from the Natural Sciences and Engineering Research Council of Canada (DISCOVERY Grant), the University of Ottawa, MITACS (ACCELERATE), and TotalEnergies is greatly appreciated, as well as the computing resources provided by the Canada Foundation for Innovation, TotalEnergies and the Digital Research Alliance of Canada.

References

1 OpenAI, *GPT-4 Technical Report*, 2023, pp. 1–100.
2 A. Dubey *et al.*, *The Llama 3 Herd of Models*, 2024, pp. 1–92.
3 Y. Kang, H. Park, B. Smit and J. Kim, A multi-modal pre-training transformer for universal transfer learning in metal–organic frameworks, *Nat. Mach. Intell.*, 2023, **5**, 309–318.
4 B. P. Munson, *et al.*, De novo generation of multi-target compounds using deep generative chemistry, *Nat. Commun.*, 2024, **15**, 3636.
5 R. Irwin, S. Dimitriadis, J. He and E. J. Bjerrum, Chemformer: a pre-trained transformer for computational chemistry, *Mach. Learn.: Sci. Technol.*, 2022, **3**, 015022.
6 C. Zeni *et al.*, *MatterGen: a Generative Model for Inorganic Materials Design*, 2023, pp. 1–56.

7 B. Zdrazil, *et al.*, The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods, *Nucleic Acids Res.*, 2024, **52**, D1180–D1192.
8 Z. Wu, *et al.*, MoleculeNet: a benchmark for molecular machine learning, *Chem. Sci.*, 2018, **9**, 513–530.
9 O. Mahmood, E. Mansimov, R. Bonneau and K. Cho, Masked graph modeling for molecule generation, *Nat. Commun.*, 2021, **12**, 3156.
10 B. Feng, *et al.*, A bioactivity foundation model using pairwise meta-learning, *Nat. Mach. Intell.*, 2024, **6**, 962–974.
11 J. J. Irwin, *et al.*, ZINC20—A Free Ultralarge-Scale Chemical Database for Ligand Discovery, *J. Chem. Inf. Model.*, 2020, **60**, 6065–6073.
12 R. Ramakrishnan, P. O. Dral, M. Rupp and O. A. von Lilienfeld, Quantum chemistry structures and properties of 134 kilo molecules, *Sci. Data*, 2014, **1**, 140022.
13 S. Kim, *et al.*, PubChem 2023 update, *Nucleic Acids Res.*, 2023, **51**, D1373–D1380.
14 A. Jain, *et al.*, Commentary: The Materials Project: A materials genome approach to accelerating materials innovation, *APL Mater.*, 2013, **1**, 011002.
15 D. Zagorac, H. Müller, S. Ruehl, J. Zagorac and S. Rehme, Recent developments in the Inorganic Crystal Structure Database: theoretical crystal structure data and related features, *J. Appl. Crystallogr.*, 2019, **52**, 918–925.
16 C. R. Groom, I. J. Bruno, M. P. Lightfoot and S. C. Ward, The Cambridge Structural Database, *Acta Crystallogr., Sect. B:Struct. Sci., Cryst. Eng. Mater.*, 2016, **72**, 171–179.
17 A. J. White, M. Gibaldi, J. Burner, R. A. Mayo and T. K. Woo, High Structural Error Rates in "Computation-Ready" MOF Databases Discovered by Checking Metal Oxidation States, *J. Am. Chem. Soc.*, 2025, **147**, 17579–17583.
18 J. Park, Y. Lee and J. Kim, Multi-modal conditioning for metal-organic frameworks generation using 3D modeling techniques, *ChemRxiv*, 2024, 1–34, DOI: **10.26434/chemrxiv-2024-w8fps**.
19 Y. G. Chung, *et al.*, In silico discovery of metal-organic frameworks for precombustion CO 2 capture using a genetic algorithm, *Sci. Adv.*, 2016, **2**, e1600909.
20 Z. Yao, *et al.*, Inverse design of nanoporous crystalline reticular materials with deep generative models, *Nat. Mach. Intell.*, 2021, **3**, 76–86.
21 H. Park, S. Majumdar, X. Zhang, J. Kim and B. Smit, Inverse design of metal–organic frameworks for direct air capture of CO2 *via* deep reinforcement learning, *Digital Discovery*, 2024, **3**, 728–741.
22 M. Gibaldi, O. Kwon, A. White, J. Burner and T. K. Woo, The HEALED SBU Library of Chemically Realistic Building Blocks for Construction of Hypothetical Metal–Organic Frameworks, *ACS Appl. Mater. Interfaces*, 2022, **14**, 43372–43386.
23 A. S. Rosen, *et al.*, Machine learning the quantum-chemical properties of metal–organic frameworks for accelerated materials discovery, *Matter*, 2021, **4**, 1578–1597.
24 A. S. Rosen, *et al.*, High-throughput predictions of metal–organic framework electronic properties: theoretical challenges, graph neural networks, and data exploration, *npj Comput. Mater.*, 2022, **8**, 112.
25 J. Burner, *et al.*, ARC–MOF: A Diverse Database of Metal-Organic Frameworks with DFT-Derived Partial Atomic Charges and Descriptors for Machine Learning, *Chem. Mater.*, 2023, **35**, 900–916.
26 Y. G. Chung, *et al.*, Computation-Ready, Experimental Metal–Organic Frameworks: A Tool To Enable High-Throughput Screening of Nanoporous Crystals, *Chem. Mater.*, 2014, **26**, 6185–6192.
27 Y. G. Chung, *et al.*, Advances, Updates, and Analytics for the Computation-Ready, Experimental Metal–Organic Framework Database: CoRE MOF 2019, *J. Chem. Eng. Data*, 2019, **64**, 5985–5998.
28 A. Li, *et al.*, The launch of a freely accessible MOF CIF collection from the CSD, *Matter*, 2021, **4**, 1105–1106.
29 P. Müller, R. Herbst-Irmer, A. L. Spek, T. R. Schneider and M. R. Sawaya, *Crystal Structure Refinement: A Crystallographer's Guide to SHELXL. Crystal Structure Refinement: A Crystallographer's Guide to SHELXL Vol. 9780198570*, Oxford University Press, 2010.
30 P. Z. Moghadam, *et al.*, Development of a Cambridge Structural Database Subset: A Collection of Metal–Organic Frameworks for Past, Present, and Future, *Chem. Mater.*, 2017, **29**, 2618–2625.
31 M. Gibaldi, A. Kapeliukha, A. White and T. K. Woo, Incorporation of Ligand Charge and Metal Oxidation State Considerations into the Computational Solvent Removal and Activation of Experimental Crystal Structures Preceding Molecular Simulation, *J. Chem. Inf. Model.*, 2025, **65**, 275–287.
32 R. de Gelder, R. Wehrens and J. A. Hageman, A generalized expression for the similarity of spectra: application to powder diffraction pattern classification, *J. Comput. Chem.*, 2001, **22**, 273–289.
33 A. Otero-de-la-Roza, E. R. Johnson and V. Luaña, Critic2: A program for real-space analysis of quantum chemical interactions in solids, *Comput. Phys. Commun.*, 2014, **185**, 1007–1018.
34 J. Riebesell, H. Yang, R. Goodall and S. G. Baird, *Pymatviz: visualization toolkit for materials informatics*, 2022, DOI: **10.5281/zenodo.7486816**.
35 O. Isayev, *et al.*, Universal fragment descriptors for predicting properties of inorganic crystals, *Nat. Commun.*, 2017, **8**, 1–12.
36 S. P. Ong, *et al.*, Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis, *Comput. Mater. Sci.*, 2013, **68**, 314–319.
37 J. Luo, *et al.*, MEPO-ML: a robust graph attention network model for rapid generation of partial atomic charges in metal-organic frameworks, *npj Comput. Mater.*, 2024, **10**, 224.
38 Ł. M. Mentel, *A Python Resource For Properties Of Chemical Elements, Ions And Isotopes*.
39 M. Fey and J. E. Lenssen, *Fast Graph Representation Learning with PyTorch Geometric*, 2019, pp. 1–9.

40 A. Paszke, *et al.*, PyTorch: An imperative style, high-performance deep learning library, *Adv. Neural Inf. Process. Syst.*, 2019, **32**, 1–12.
41 D. P. Kingma and J. Ba, Adam: A Method for Stochastic Optimization, *3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc*, 2014, pp. 1–15.
42 T. Akiba, S. Sano, T. Yanase, T. Ohta and M. Koyama, Optuna: A Next-generation Hyperparameter Optimization Framework, *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min*, 2019, pp. 2623–2631, DOI: **10.1145/3292500.3330701**.
43 L. McInnes, J. Healy and J. Melville, UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, *arXiv*, 2018, preprint, arXiv:1802.03426, DOI: **10.48550/arxiv.1802.03426**.
44 R. Ying, D. Bourgeois, J. You, M. Zitnik and J. Leskovec, *GNNExplainer: Generating Explanations for Graph Neural Networks*, 2019.
45 H. Yuan, H. Yu, J. Wang, K. Li and S. Ji, On Explainability of Graph Neural Networks *via* Subgraph Explorations, *Proc. Mach. Learn. Res.*, 2021, **139**, 12241–12252.
46 C. Agarwal, O. Queen, H. Lakkaraju and M. Zitnik, Evaluating explainability for graph neural networks, *Sci. Data*, 2023, **10**, 144.
47 S. M. Moosavi, *et al.*, Understanding the diversity of the metal-organic framework ecosystem, *Nat. Commun.*, 2020, **11**, 4068.
48 D. Balcells and B. B. Skjelstad, tmQM Dataset—Quantum Geometries and Properties of 86k Transition Metal Complexes, *J. Chem. Inf. Model.*, 2020, **60**, 6135–6146.
49 D. Ongari, A. V. Yakutovich, L. Talirz and B. Smit, Building a Consistent and Reproducible Database for Adsorption Evaluation in Covalent-Organic Frameworks, *ACS Cent. Sci.*, 2019, **5**, 1663–1675.