Explainable Graph Neural Networks for Organic Cages

Qi Yuan, Filip T. Szczypiński, and Kim E. Jelfs*

Department of Chemistry, Molecular Sciences Research Hub, White City Campus, Imperial College London, Wood Lane, London, UK

*E-mail address: k.jelfs@imperial.ac.uk

Abstract

The development of accurate and explicable machine learning models to predict the properties of topologically complex systems is a challenge in material science. Porous organic cages, a class of polycyclic molecular materials, have potential application in molecular separations, catalysis and encapsulation. For most applications of porous organic cages, having a permanent internal cavity in the absence of solvent, a property termed “shape persistency” is critical. Here, we report the development of Graph Neural Networks (GNNs) to predict the shape persistence of organic cages. Graph neural networks are a class of neural networks where the data, in our case that of organic cages, are represented by graphs. The performance of the GNN models was measured against a previously reported computational database of organic cages formed through a range of [4+6] reactions with a variety of reaction chemistries. The reported GNNs have an improved prediction accuracy and transferability compared to random forest predictions. Apart from the improvement in predictive power, we explored the explicability of the GNNs by computing the integrated gradient of the GNN input. The contribution of monomers and molecular fragments to the shape persistence of the organic cages could be quantitatively evaluated with integrated gradient. With the added explicability of the GNNs, it is possible not only to accurately predict the property of organic materials, but also to interpret the predictions of the deep learning models and provide structural insights to the discovery of future materials.

1. Introduction

Porous organic cages are a class of molecules with an internal cavity that is made accessible to guest molecules via at least two molecular windows[1,2]. The cavity of porous organic cages offers potential applications including encapsulation[3], molecular separation[4–7], and catalysis[8]. Organic cages are distinguished from other porous materials such as zeolites and metal-organic frameworks (MOFs) due to the absence of an extended network of bonds in the solid state. In addition, organic cage molecules are usually soluble in organic solvents, allowing for solution processing into thin films or membranes both in the crystalline and amorphous solid state[9]. The lack of three-dimensional chemical bonding can allow the solid-state structures to undergo large rearrangements, which has been used in the creation of molecular crystals with "on/off" porosity with polymorph switching[10]. However, such flexibility also means that organic cages are more likely to collapse and lose porosity as a result of desolvation[11], which is known as a lack of "shape persistence". The shape persistence of organic cages is difficult to predict without employing computational modelling[11,12]. High-throughput computational screening has been used in combination with robotic synthesis for the discovery of novel organic cages[12]. The cost of computational screening of organic cages is significantly cheaper than for experimental measurements, however modelling larger systems is still time consuming, especially for organic cages that often have several hundred atoms.

Machine learning (ML) has many potential uses within material discovery, including to reduce the cost of property calculation compared to carrying out computational simulations (especially via quantum mechanical methods), to focus experimental synthesis and measurement effort on the most promising materials reducing wasted laboratory effort[13,14], as well as to help facilitate the exploration of larger chemical space[15–17]. Apart from the widely reported ML models for molecular discovery, especially drug discovery, the applications of ML to porous materials such as MOFs have gained significant interest[18]. Various structural and geometrical descriptors for MOFs have been developed for the prediction of their gas sorption[19], and open-source databases recording the structures with experimental and/or computational properties have been published and the diversity of the chemical space examined[20]. The development and application of ML to modelling the properties of organic cages, on the other hand, is less reported.

We have previously developed a computational database of >60,000 organic cages formed through a range of reaction chemistries via a [4 + 6] reaction of four tritopic and six ditopic building blocks and studied their behaviour using molecular dynamics calculations.[21] We then modelled the computed shape persistence of these cages using random forest models and found them to be very effective when applied to systems with the same reaction chemistry, for example a random forest model trained on cages formed from imine chemistry was effective at predicting shape persistency in other imine cages[21]. However, the random forest model did not translate well between cages formed from different reaction mechanisms: an imine-trained random forest model was not as effective at predicting the shape persistence of a cage formed by alkyne metathesis chemistry. This was not surprising given that experimentally, extremely small changes to the synthesis, for example adding a single CH_2 group to one cage precursor, could completely invert the shape persistency behaviour.[11] The prediction result of the random forest models could not be attributed to specific monomers of the cage or fragments of the monomers, because the feature importance analysis did not show a strong preference to any specific molecular features[21]. In addition, the Morgan molecular fingerprint, a vector indicating the presence of specific substructures within a molecule, was adapted as the input feature to the random forest model. Recently developed graph neural networks (GNNs), which encode molecular information into neural graph fingerprints with machine-learned continuous numeric vector representation have exhibited improved predictive performance on various tasks including chemical reactivity[22], compound protein interaction[23] and partial charge assignment[24], because of the flexibility of such fingerprints, especially when a larger dataset is available[25].

An additional benefit of prediction via GNNs is that it is possible to identify key building blocks or molecular fragments contributing to the models' predictions through calculating attribution scores of the input features. Sundararajan *et al.* developed the integrated gradients to compute the contribution of input features for ML tasks and highlighted a case study of explaining molecular binding mechanisms using integrated gradients[26]. McCloskey *et al.* calculated the attribution score of fragments of molecules with a hypothesized binding mechanism and proposed a sanity check to determine whether a hypothesized mechanism can be learned[27]. The explicability of ML models for predictive tasks in material and molecular discovery has gained increasing research interest, since explainable models can not only provide insight for the monomers and fragments that contribute exclusively to the prediction

to help future discovery, but also suggest possible pitfalls of the models where predictions are accurate, but the underlying chemical mechanism has not been learnt.

In this study, we developed GNN models to predict the shape persistence of organic cages formed via different [4+6] reaction chemistries: imine condensation, amide condensation, and alkene/alkyne metathesis. Graph representations of the organic cages were developed and neural fingerprints for cages were trained using the GNN architecture. The shape persistence of the organic cages was accurately predicted using the GNN model, with significant improvement of generalisability towards unseen monomers compared to prior work with random forest models. In addition, to obtain explicability of the prediction of the GNNs, the integrated gradient was implemented and computed for precursors of the organic cages and fragments of the precursors. It was therefore possible to quantify the contribution of precursors as well as fragments to the shape persistence of organic cages and provide insight for the design of future precursors for organic cages.

2. Methods

2.1 Dataset

The dataset for organic cages used here was reported in our previous work[21]. In brief, the synthetically viable library of di- and tri- precursors were generated based on synthetic experience, and 118 di-topic and 51 tri-topic precursor cores were included, each with locations of functional groups marked. In this work, the precursors with the greater number of reactive functional groups are referred to as the "building block", and the precursors with fewer functional groups are referred as the "linker". Each precursor backbone was expanded with different functional groups to include organic cages synthesized with different reaction chemistry. The functional groups included were aldehydes, alkynes, amines, carboxylic acids, alkenes, which are combined using imine or amide condensation, alkyne or alkene metathesis, and disulfide formation reactions. The topologies of organic cages were defined previously by Santolini *et al.*[28]. Here, we used only the **Tri^{4}Di^{6}** cages assembled from the four tritopic precursors and six ditopic precursors in a [4+6] reaction (example cage in **Tri^{4}Di^{6}** topology is shown in Figure 1), and the previously reported random forest models were used as a benchmark for our work. For each pair of functional groups capable of undergoing a reaction, every possible pair of precursors was used to generate a cage. For each reaction, 6018 distinct

precursor pairs were generated, resulting in a total of 36,108 cages. A summary of the precursor pairing for the **Tri^{4}Di^6** cages is shown in Table 1.

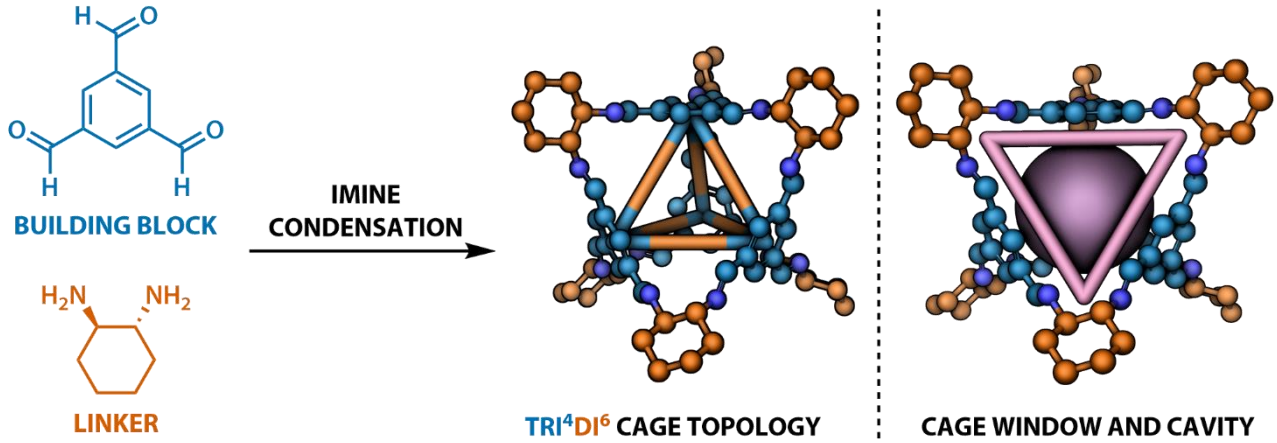

*Figure 1: The **Tri^{4}Di^6** tetrahedral topology of the organic cages considered in this study. The tritopic precursor ("building block") is shown in blue and the ditopic precursor in orange ("linker"). The resulting cavity and one of the four windows are highlighted in purple (right).*

*Table 1 Reaction and precursor information for the **Tri^{4}Di^6** cages in this study.*

Group name	Building block	Linker	Reaction	No. cages
aldehyde3amine2	aldehyde 3	amine 2	imine condensation	6018
amine3aldehyde2	amine 3	aldehyde 2	imine condensation	6018
alkene3alkene2	alkene 3	alkene 2	alkene metathesis	6018
alkyne3alkyne2	alkyne 3	alkyne 2	alkyne metathesis	6018
carboxylicacid3amine2	carboxylic acid 3	amine 2	amide condensation	6018
amine3carboxylicacid2	amine 3	carboxylic acid 2	amide condensation	6018

The number following a functional group name indicates the number of functional groups present in the precursor (i.e., a "2" means that it is di-topic).

2.2 Dataset labelling

The same computational labels of shape persistency for the [4+6] organic cages published previously[21] were used in this work, where the shape persistency of the cages were calculated from the geometrically optimized structures. Specifically, the geometry of the organic cages formed using the precursors were previously optimized using molecular dynamics (MD) simulations. The cavity

size, window diameter and number of windows of the MD optimized cages were calculated using pywindow[29], and the cages were labelled as either "collapsed", "not collapsed" (i.e., shape persistent), or "undetermined" using the above parameters. If the cages did not contain the expected 4 windows for a tetrahedral topology, the cage was labelled as collapsed. For cages with the expected number of windows detected by pywindow, the following empirical criterion was applied:

$$\alpha = \frac{4 \times average\ difference\ in\ window\ diameter}{maximum\ window\ diameter \times expected\ no.of\ windows} \quad (1)$$

If α < 0.035 and the cavity size was greater than 1 Å, the cage was labelled as "not collapsed", else it was labelled "undetermined". Only the "collapsed" and "not collapsed" cages were used to train the ML models in this study. A summary of cage collapse labels for different chemical reactions in this study are provided in Table S1. The cavity size for each cage was calculated by translating the centroid of the cage onto the origin. An example cavity with the corresponding window can be seen on the right of Figure 1.

2.3 Representation of cages

Building blocks and linkers of the organic cages were encoded using the graph neural network (GNN), where representation of each atom in the molecule was obtained by aggregating the information of the atom and its neighbours. The design of the GNN layer for encoding the building blocks and linkers is shown in Figure 2. Each non-hydrogen atom X in the molecule was represented using a numeric vector in the form of $X_i = (V_{atom}, V_{neighbour}, V_{2nd\ neighbour})$. V_{atom} contains information including atomic symbol, number of neighbour non-hydrogen atoms, implicit and explicit valence, and whether the atom is aromatic. $V_{neighbour}$ is the weighted sum of the atomic vector of the atom and its neighbours, while $V_{2nd\ neighbour}$ contains the weighted sum of the atomic vector and its neighbours up to the second order. The representation of the building block or linker molecule was obtained be summing up all the atomic vectors X_i in the molecule.

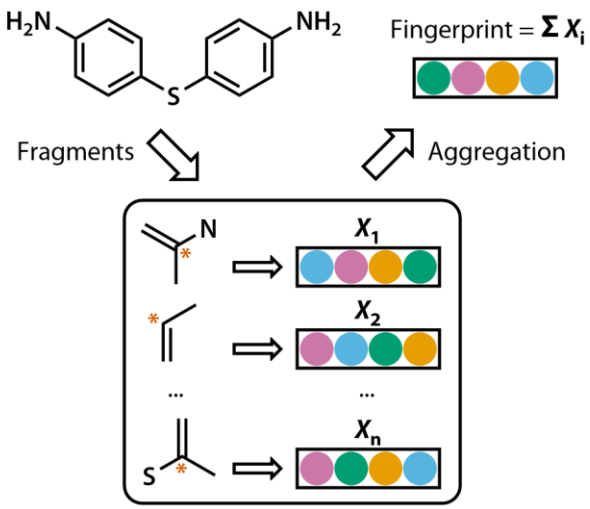

Figure 2 GNN encoding of the molecular features of the building blocks and linkers of organic cages in this study.

Similar to our previous work[21], the neural fingerprints for the organic cages in this study were obtained by concatenating the molecular vectors of the building blocks and linkers. Such neural fingerprints were then processed by a multi-layer neural network followed by a prediction layer (see Figure 3). The architecture of the prediction layer is determined by the predictive task in this study, for the classification tasks such as predicting the organic cage shape persistence, the output layer has two neurons, each of which was interpreted the organic cage being "collapsed" or "not collapsed", the output of the two neurons are noted as $z_i (i = 1,2)$, $z_i$would be processed using the softmax function:

$$\sigma(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{2} e^{z_j}} \quad (2)$$

The neuron with the larger softmax output $\sigma(z_i)$ would be treated as the "predicted" label.

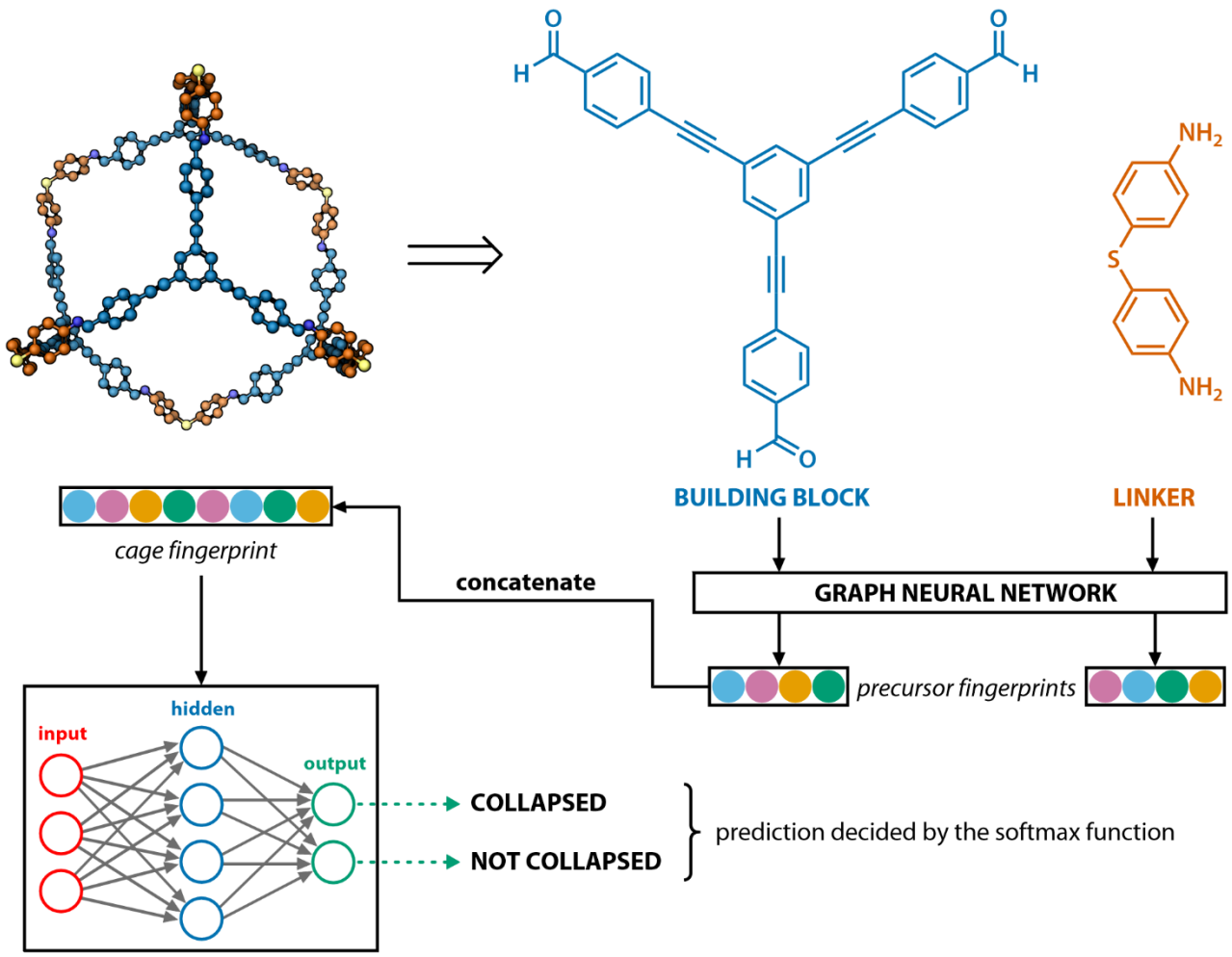

Figure 3 Architecture of the GNN in this study: Monomers (building blocks and linkers) of the organic cages were encoded to numeric vectors using a graph neural network (see Figure 2), the vectors were then concatenated and processed by a multi layer neural network to output a shape persistence prediction. The prediction by the two neurons in the output layers was processed using the softmax function to obtain the final classification.

2.4 Training and evaluating the GNN models

In this study, we focused primarily on the classification GNN model, where the building block and linker of the organic cages were represented using GNN encoding, and the encoded vectors for the building block and linker molecules were concatenated so as to form a feature vector of the organic cage. The feature vector is then processed through a multi-layer neural network to predict the shape persistence of the organic cages. To examine the predictive power as well as the generalizability of the GNN models, two types of prediction tasks were employed. For the All-vs-One task, cross-reaction prediction was performed: the "collapsed" and "not collapsed" data in all but one rows in Table 1 were used as the training set, and data in the remaining row were used as the test set. All rows in Table 1 were used iteratively for the All-vs-One task. For the All-vs-All task, on the other hand,

the data for "collapsed" and "not collapsed" cages in Table 1 were randomly split to the training (80%) and test (20%) set. Performance of the All-vs-One model is an indicator of how transferrable the GNN model is towards cages generated via different reaction chemistries.

The performance of the GNN model on the classification task of "collapsed" and "not collapsed" cages was evaluated using the accuracy, precision and recall scores on the test sets, defined as follows:

$$Accuracy = \frac{True\ positive + True\ negative}{Size\ of\ test\ set} \quad (3)$$

$$Precision = \frac{True\ positive}{True\ positive + False\ positive} \quad (4)$$

$$Recall = \frac{True\ positive}{True\ positive + False\ negative} \quad (5)$$

In this study, the "collapsed" organic cages were regarded as "positive" in our predictions. "True positive" represents the data where cages were "collapsed" from both the GNN model prediction and as labelled in the database; "False positive" represents the data where cages were "collapsed" according to the GNN model prediction but "not collapsed" as labelled in the database; "True negative" represents the data where cages were "not collapsed" from both the GNN prediction and as labelled in the database; "False negative" represents the cages that were "not collapsed" according to the GNN prediction but were "collapsed" as labelled in the database.

2.5 Explicability of the GNN models

The explicability of the GNN model predictions was analysed by calculating the attribution score of the input features, which is the atomic input vectors to the GNN in this study. By calculating the attribution score, we aim to analyse which building block or linker molecules contribute more to the collapse of an organic cage, and which fragments in these molecules contribute more to the building block or linker being a "collapsed" component of the cage. An example of per-atom contribution to the prediction is shown in Figure 4, where the fragments with a positive attribution score (likely to contribute to pore collapse) are shown in red, and fragments with a negative attribution score (not likely to contribute to pore collapse) are shown in blue.

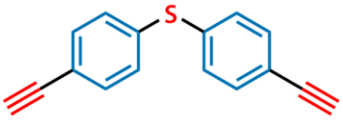

Figure 4 Example visualization of per-atom contribution to the model prediction. Fragments with positive contributions (likely to contribute to pore collapse) are shown in red, while fragments with negative contributions (not likely to contribute to pore collapse) are shown in blue.

The attribution scores in this study were calculated and represented using integrated gradients. The formal definition for attribution scores, as well as the axiomatic justification of the integrated gradients satisfying certain properties is provided by Sundararajan *et al.*[26]. To explain briefly here, let function $F: R^n \rightarrow [0,1]$ represent a deep neural network. Given an input feature x and some baseline feature x', the integrated gradient of x along the i^{th} dimension of x was defined as follows:

$$a_i = (x_i - x_i') \times \int_{t=0}^{1} \frac{\partial F\left(x' + t \times (x - x')\right)}{\partial x_i} dt \qquad (6)$$

where $\frac{\partial F(x)}{\partial x_i}$ is the gradient of F along the i^{th} dimension of x. In this study, the input x was the numeric vector for the organic cages, which is the concatenation of the feature vectors of building block and linker molecules, and F is the probability of the organic cage being "collapsed" as predicted by the GNN. This definition of integrated gradient is justified by the axiomatic result that it satisfies several desirable properties of an attribution method[26].

The integrated gradient attribution was defined relative to a baseline, and the selection of the baseline is essential to causal analysis of ML models[30]. A robust baseline input should give uninformative predictions; for example, for a classification task, the ML model should give the probability of approximately 0.5 for the baseline input. Here, we used the input of zero vectors as the baseline molecule and augmented the training set using the baseline cages to achieve uninformative predictions for the baseline cages. The detailed implementation is provided in Section S2. Once the integrated gradients for all input atoms of the organic cage were calculated, the contribution of the building block and linker of the cage was calculated by summing up the integrated gradients for the atoms corresponding to the building block and linker, respectively. The attribution of fragments in the building block and linker molecules were visualized using the atomic integrated gradients in the molecules.

The GNN model, as well as the computation of integrated gradients were implemented in Python 3.7.5 combined with PyTorch 1.1.0; the source code is provided at github.com/qyuan7/Cage_GNN.

3. Results and discussion

3.1 Predictive performance of the GNN for organic cage shape persistence

A comparison of the predictive performance of the previously reported random forest model (used here as a benchmark) and the GNN model on the All-vs-All task is shown in Table 2, where the data

for all cages in this study were randomly split to training and test sets. It can be seen that the GNN model and the random forest benchmark have comparable performance for the All-vs-All task, with the GNN model slightly outperforming the random forest model based upon the accuracy and precision metrics. The reason for the almost equally good performance of the GNN and random forest models on the All-vs-All task could originate from the dataset in this study. The building block and linker molecules in this study were built by changing the functional groups on a fixed set of precursor cores, and each row of organic cages in Table 1 was generated from only 118 unique di-topic precursors and 51 unique tri-topic precursors. For the All-vs-All task, the dataset of all the organic cages in Table 1 was randomly split between the training and test sets, and the same precursor would possibly be present in both the training and test sets. In addition, for both the GNN and random forest models, the organic cages were represented by concatenating the molecular vectors of the precursors. It is therefore possible for both models to learn the "possibility" of a certain precursor belonging to a collapsed cage from the training set and to then make predictions on the test set. In this sense, both the GNN and the random forest learned the probability distribution of certain precursors being collapsed, and it is not clear how much contribution to shape persistency of the precursors and organic cages were learnt from the All-vs-All task. Therefore, the advantage of the neural fingerprints learnt from the GNN model of being more flexible is minimized in the All-vs-All task.

Table 2 Shape persistence prediction of the GNN and random forest models on the All-vs-All task. The models with better performance for each metric are highlighted in bold.

	GNN	Random forest
Accuracy	**0.89**	0.88
Precision	**0.90**	0.89
Recall	0.90	**0.91**

The All-vs-One task, where data for cages in all but one row in Table 1 was used as the training set and the remaining row is used as the test set, is more challenging compared to the All-vs-All task, as most of the precursors in the test set were not included in the training set (except for the amine2 linkers and amine3 building blocks, which were used by two rows in Table 1). The All-vs-One task provides better evaluation of the transferability of the ML models towards different families of precursors with different functional groups, which carries more application significance for the design of future organic cages. The accuracy scores for the GNN and random forest models are shown in Table 3, and the corresponding precision and recall scores are provided in Table S2.

For the All-vs-One task, the GNN model consistently outperformed the random forest model and by a larger margin compared to the All-vs-All task. The biggest improvement in the predictive performance of the GNN model compared to the random forest benchmark was for the alkene3alkene2 cages (alkene metathesis of a tri-alkene and di-alkene) and the alkyne3alkyne2 cages (alkyne metathesis of a tri-alkyne and di-alkyne). As shown in Table 1, the building blocks for alkene3alkene2 and alkyne3alkyne2 cages were not used for the other cages, and the benchmark random forest model failed to give reasonably accurate predictions on the shape persistence of the alkene3alkene2 and alkyne3alkyne2 cages, thus the transferability for the benchmark random forest model is poor to building blocks that were not used in the training sets. The GNN model, on the other hand, was equally accurate for the predictions of the alkene3alkene2 and alkyne3alkyne2 cages compared to the other groups of cages. The consistent improvement in predictive power of the GNN model compared to the random forest model indicates that the GNN model has better transferability to novel precursors for cages and different reaction types. In addition, the improved performance of the GNN model for the alkene3alkene2 and alkyne2alkyne2 cages suggests that the GNN model has learnt some structural features of the precursors that led to collapse from the training process, providing the model with some "chemical intuition", which can be investigated further by trying to explain and interpret the predictions of the GNN model using the integrated gradients.

Table 3 Shape persistence prediction of the GNN and random forest models on the All-vs-One task[a]. Model with better performance for each task is highlighted in bold.

Building block	Linker	Test accuracy (Random forest)	Test accuracy (GNN)
aldehyde 3	amine 2	0.61	**0.72**
amine 3	aldehyde 2	0.72	**0.73**
alkene 3	alkene 2	0.63	**0.81**
alkyne 3	alkyne 2	0.41	**0.77**
carboxylic acid 3	amine 2	0.71	**0.76**
amine 3	carboxylic acid 2	0.73	**0.79**

The results are for when a model was tested on a single data set within a row, i.e. with cages formed by a single reaction chemistry type. The data sets in the other rows were used as the training set. For example, for the test of aldehyde3amine2 cages (row 1), all the precursor pairs in the other rows were used as the training set (amine3aldehyde2, alkene3alkene2, etc), only the aldehyde3amine2 cages were used as the test set.

3.2 Explicability of the GNN predictions

To interpret the predictions of the GNN models for the All-vs-One task, we computed the integrated gradients of input vectors to the GNN, which were further summed up to get the integrated gradients of cage precursors and fragments in the test sets. Before analysing the results, we have validated our calculations by checking the *completeness* of the integrated gradient in this study: the attributions of the input features (cage atoms) should add up to the difference between the output of F at the input x and the baseline x' for equation (4)[26]. The probability of the organic cages and the corresponding baseline cages being "collapsed" were computed from the GNN models, the *completeness* of the attribution model requires that the difference between the two probability values ΔP should be equal to the integrated gradient of the input features for the organic cage Σ_{ig}. The distribution of the difference between the ΔP and Σ_{ig} values for all the cages in the test sets of the All-vs-One task is shown in Figure 5 (a). The distribution is centred around 0, with a mean value of 0.008 and standard deviation of 0.013, indicating that the integrated gradient computed in this study meets the requirement of *completeness* for an attribution model.

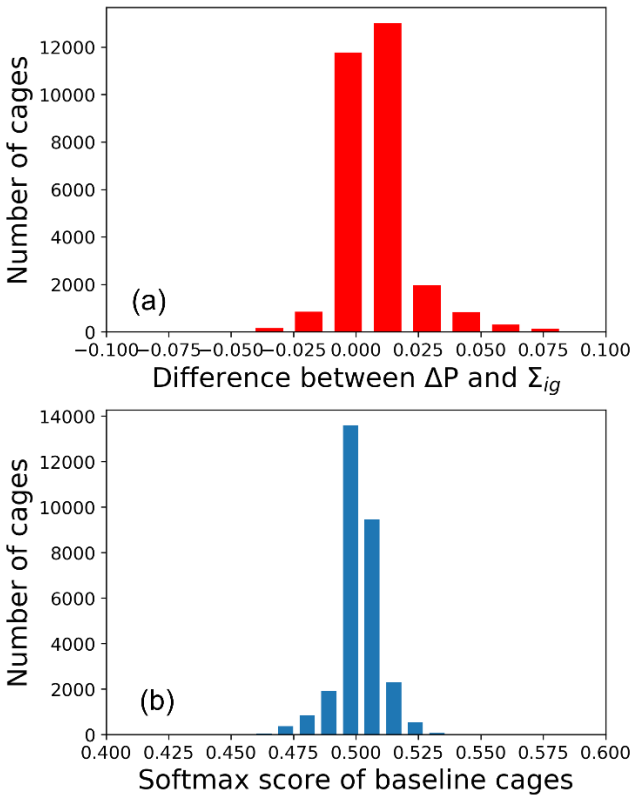

Figure 5 Validation of the integrated gradient calculations in this study: (a) distribution of the difference between the ΔP and Σ_{ig} values for all the cages in the test sets of the All-vs-One task; (b) distribution of the predicted softmax score of baseline cages.

In this study, the integrated gradient of the cage input feature x for an atom is defined relative to the baseline input x' in equation (4), thus it is important that the GNN model F should give uninformative predictions to the baseline input. For the classification task in this study, the baseline input should render a probability close to 0.5, indicating the baseline cage composed of vector of zeros should have neutralized probability of being "collapsed". When calculating the integrated gradients, we used the data augmentation technique on the training set, as described in Section S2 of the supporting information and the work by McCloskey *et al.*[27]. The distribution of the predicted softmax scores for the "collapse" neuron in the output layer (which can be interpreted as the probability of the cage being collapsed) on the baseline cages used in training the GNN model for calculating the integrated gradient is shown in Figure 5 (b). The softmax score of the baseline cages centres around 0.5 with a mean value of 0.501 and standard deviation of 0.008. This result indicates that the GNN model gives neutral predictions to the baseline cages, and for a cage with softmax score larger than 0.5 for the "collapse" neuron in the output layer that is classified to be "collapsed", the majority of the attribution to the increased softmax score can be ascribed to the molecular features of the building block and linker molecules of the cage.

3.3 Explicability of the GNN models – Precursors with the highest integrated gradients

With the validation of the integrated gradient calculations completed, it was possible to calculate the attributions of the cage building blocks and linkers and identify the precursors with a high integrated gradient contribution for the "collapsed" predictions. If some precursors have high integrated gradient scores in "collapsed" cages, it is possible that such precursors can be regarded as the "collapse-directing precursors" that should be avoided in the design of novel organic cages. However, if the precursors' integrated gradient attribution scores had no strong correlation with the shape persistence, then the structural features of the "collapsed" precursors were not learnt.

We calculated the integrated gradients of the precursors in the test sets for the All-vs-One tasks and ranked the building blocks and linkers according to their integrated gradient attribution scores. The top 5 building blocks **BB1-5** for the cages generated from the aldehyde3amine2 (imine reaction of trialdehydes and diamine) cages with the largest integrated gradient are shown in Figure 6. The percentage of aldehyde3amine2 cages containing the building blocks that were "collapsed" in the All-vs-One test set are also shown. It can be seen that almost all the building blocks in Figure 6 have a probability of larger than 90% of being "collapsed", indicating that cages with these building blocks have a great chance of being "collapsed" and that these building blocks should be avoided in the design of future organic cages for the sake of shape persistence. The top 5 linker molecules **L1-5** for

the aldehyde3amine2 cages with the largest integrated gradient attribution to "collapsed" cages are shown in **Figure 7**, with the percentage of "collapsed" aldehyde3amine2 cages containing the linker molecules shown. Apart from **L1**, the cages in the test set containing these linkers have a high probability of being "collapsed". The integrated gradients of the building block and linker molecules can thus serve as an indicator for the organic cages being "collapsed" – using building block/linker molecules with high integrated gradient attributions means there is a high probability of collapsed cages. It might be tempting to assume that precursors with smallest gradient attributions would indicate "uncollapsed" cages. The "bottom 5" building blocks and linkers of the aldehyde3amine2 cages are shown in Figures S11 and S12. Cages with such building blocks and linkers still have a considerable possibility of being "collapsed", thus the integrated gradient has only limited effect of identifying "stable" precursors, and we therefore focus on the "collapsible" precursors in this study.

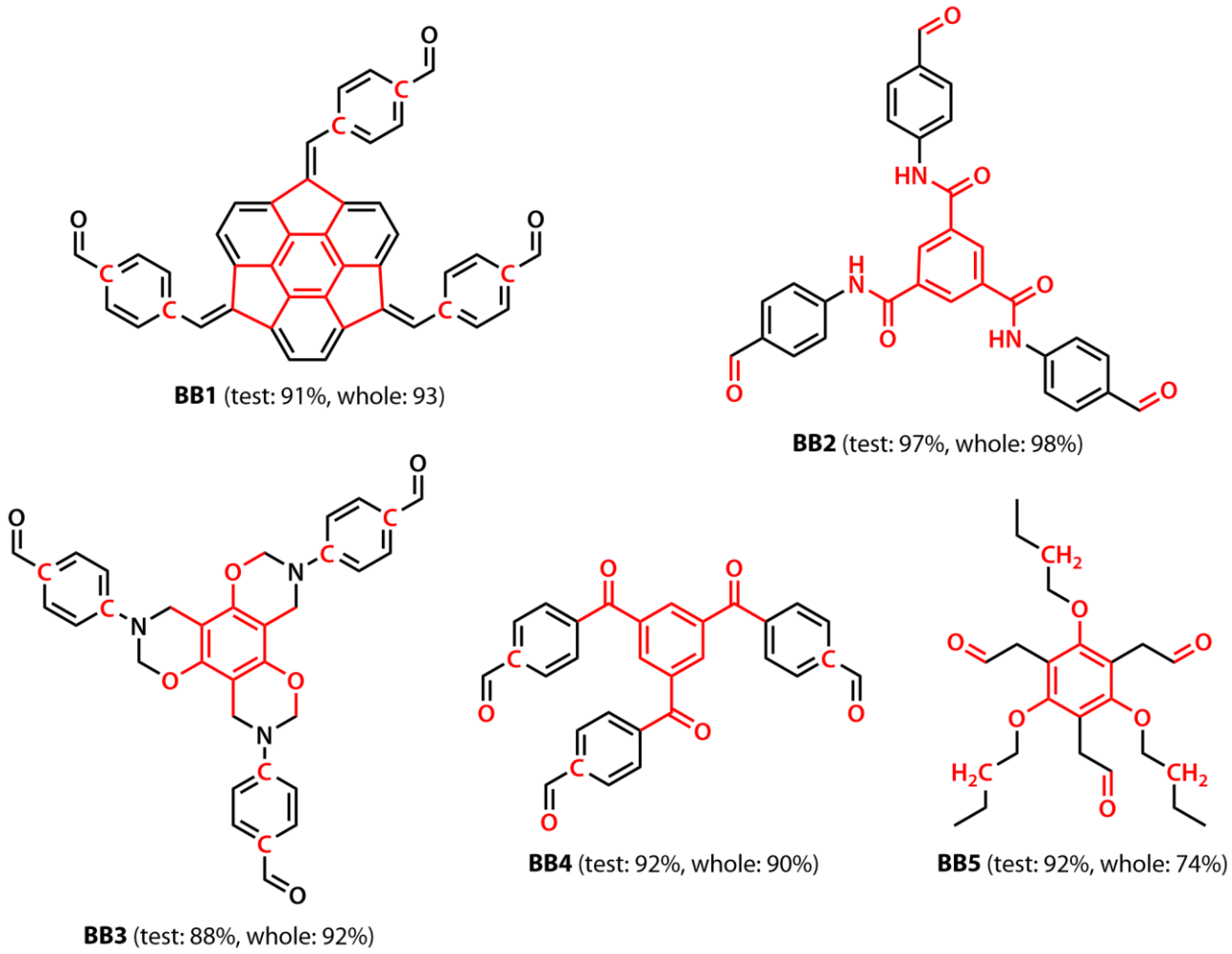

Figure 6 The top 5 building blocks with the largest overall integrated gradient attributions for the aldehyde3amine2 cages. Atoms with integrated gradients greater than 0.01 are highlighted in red.

*Percentages of cages containing each building block identified as "collapsed" in the **test** set and the highlighted backbones in the **whole** database are shown. The building blocks can be regarded as "collapsible" precursors which tend to form cages with high probability of collapse. The highlighted fragments in the building blocks are those contributing most to the integrated gradient of the corresponding building block.*

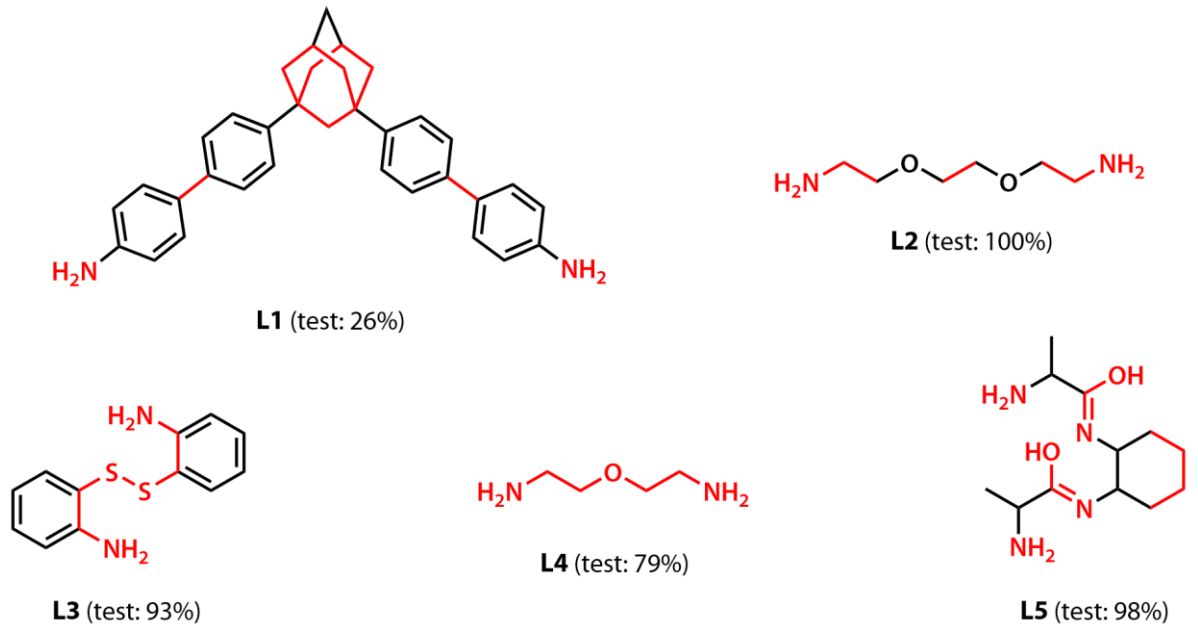

Figure 7 The top 5 linkers with the largest integrated gradient attributions for the aldehyde3amine2 cages. Atoms with integrated gradients greater than 0.01 are highlighted in red. Percentages of cages containing each building block identified as "collapsed" in the test set are shown. The linkers can be regarded as "collapsible" precursors which tend to form cages with high probability of collapse. The highlighted fragments in the linkers are those contributing most to the integrated gradient of the corresponding building block.

The top building blocks and linkers for the other groups of organic cages with the largest integrated gradient together with the probability of a cage being "collapsed" with such precursors are provided in Section S4 of the supporting information. For the carboxylicacid3amine2 cages (amide condensation of a tricarboxylic acid and diamine), the integrated gradient attributions of the top building blocks had poor correlation to the cage shape persistence, which could be because the carboxylic acid functional group was used less in the database (Table 1), and the GNN model therefore had poorer transferability to the cages with the tricarboxylic acid building blocks. Further improvement of the GNN model for the cages formed via amide condensation reaction would require a larger dataset labelled as per the current dataset. The relationship of cage shape persistency and the average integrated gradient attribution scores for the building block/linker molecules in the All-vs-One task is shown in Figures S11 and S12. Qualitative agreement of cage collapse and high integrated

gradient scores can be found for cages formed via imine condensation, alkene metathesis and alkyne metathesis, which could provide initial insight into the shape persistence of organic cages formed via such reaction chemistries (see Figures S13 and S14).

If specific precursor fragments could be identified as "collapsible" from the above analysis, then such a fragment could be usefully avoided in the design of novel precursors. Atoms in the cage components with integrated gradient attribution score greater than 0.01 are highlighted in red in Figure 6 andFigure 7. The majority of the integrated gradient attribution is located in the central core of the building blocks; and such fragments could contribute to the poor shape persistence of the corresponding cages. It is thus possible to identify molecular fragments/center cores that have high attribution to the collapse of organic cages and to therefore avoid/alter such fragments when selecting precursors for cage synthesis. In order to validate whether the identified central cores correlate with the shape persistence of all the organic cages in this study, we performed a sub-structure match of the cores across all the cages in this study and calculated the probability of a cage with precursors containing the backbones being "collapsed", which is also shown in Figure 6.

Meanwhile, the linkers in Figure 7 (apart from the outlier **L1**) contain more saturated carbon chains and hence more internal degrees of freedom. Furthermore, the amine part of the imine bond (resulting from the condensation to give cages in the aldehyde3amine2 set) contains one more flexible methylene unit compared to the aldehyde contribution. As a result, the fragments with high integrated gradients for linkers **L2-5** span over both the linker backbone and the functional group, making it difficult to attribute the GNN prediction to any particular motif within those molecules, and therefore substructure matching of the linker molecules was not performed.

4. Conclusions

We have developed graph neural network (GNN) models to predict the shape persistence of organic cages computationally generated via a range of reactions. The GNN model has better performance compared to our previously published random forest models[21], especially for cross-reaction prediction tasks. Apart from the improved predictive performance, we evaluated the explicability of the GNN models by computing the precursor-wise and atom-wise integrated gradients. We have shown that integrated gradients can be used to learn structural features of the precursors that contribute to the collapse of organic cages, which could help exclude precursors that are more likely to result in collapsed cages. For the generally more rigid building blocks, the core backbones appear to be of greatest importance for collapse prediction, while for the smaller and more flexible linker

molecules, the collapsibility appears to originate from saturated aliphatic chains and the corresponding increased degrees of freedom, as would be expected.

The computational study of supramolecular systems such as organic cages is time consuming using physical simulations, and the development of ML techniques has the potential to provide data-driven solutions that might accelerate the evaluation of supramolecular systems. However, in many cases the ML models are regarded as powerful black-boxes, providing limited insight to further the materials discovery process further. In this study, we aimed to develop an explainable GNN model both to ensure the transferability of our model and to provide guidance of further material discovery.

Conflicts of interest

There are no conflicts of interest to declare.

Acknowledgement

We acknowledge funding from the European Research Council under FP7 (CoMMaD, ERC Grant No. 758370), K. E. J. thanks the Royal Society for a University Research Fellowship and we thank the Leverhulme Trust for a Leverhulme Trust Research Project Grant. We thank Dr. Lukas Turcani for useful discussions.

References

[1] T. Hasell, A.I. Cooper, Porous organic cages: soluble, modular and molecular pores, Nat. Rev. Mater. 1 (2016) 1–14.

[2] F. Beuerle, B. Gole, Covalent organic frameworks and cage compounds: design and applications of polymeric and discrete organic scaffolds, Angew. Chemie Int. Ed. 57 (2018) 4850–4878.

[3] M. Yoshizawa, J.K. Klosterman, M. Fujita, Functional molecular flasks: new properties and reactions within discrete, self-assembled hosts, Angew. Chemie Int. Ed. 48 (2009) 3418–3438.

[4] A. Kewley, A. Stephenson, L. Chen, M.E. Briggs, T. Hasell, A.I. Cooper, Porous organic cages for gas chromatography separations, Chem. Mater. 27 (2015) 3207–3210.

[5] T. Mitra, K.E. Jelfs, M. Schmidtmann, A. Ahmed, S.Y. Chong, D.J. Adams, A.I. Cooper, Molecular shape sorting using molecular organic cages, Nat. Chem. 5 (2013) 276.

[6] T. Hasell, M. Miklitz, A. Stephenson, M.A. Little, S.Y. Chong, R. Clowes, L. Chen, D. Holden, G.A. Tribello, K.E. Jelfs, others, Porous organic cages for sulfur hexafluoride separation, J. Am. Chem. Soc. 138 (2016) 1653–1659.

[7] L. Chen, P.S. Reiss, S.Y. Chong, D. Holden, K.E. Jelfs, T. Hasell, M.A. Little, A. Kewley, M.E. Briggs, A. Stephenson, others, Separation of rare gases and chiral molecules by selective binding in porous organic cages, Nat. Mater. 13 (2014) 954–960.

[8] T.-C. Lee, E. Kalenius, A.I. Lazar, K.I. Assaf, N. Kuhnert, C.H. Grün, J. Jänis, O.A. Scherman, W.M. Nau, Chemistry inside molecular containers in the gas phase, Nat. Chem. 5 (2013) 376–382.

[9] Q. Song, S. Jiang, T. Hasell, M. Liu, S. Sun, A.K. Cheetham, E. Sivaniah, A.I. Cooper, Porous organic cage thin films and molecular-sieving membranes, Adv. Mater. 28 (2016) 2629–2637.

[10] J.T.A. Jones, D. Holden, T. Mitra, T. Hasell, D.J. Adams, K.E. Jelfs, A. Trewin, D.J. Willock, G.M. Day, J. Bacsa, others, On--off porosity switching in a molecular organic solid, Angew. Chemie Int. Ed. 50 (2011) 749–753.

[11] K.E. Jelfs, X. Wu, M. Schmidtmann, J.T.A. Jones, J.E. Warren, D.J. Adams, A.I. Cooper, Large self-assembled chiral organic cages: synthesis, structure, and shape persistence, Angew. Chemie. 123 (2011) 10841–10844.

[12] R.L. Greenaway, V. Santolini, M.J. Bennison, B.M. Alston, C.J. Pugh, M.A. Little, M. Miklitz, E.G.B. Eden-Rump, R. Clowes, A. Shakil, others, High-throughput discovery of

organic cages and catenanes using computational screening fused with robotic synthesis, Nat. Commun. 9 (2018) 1–11.

[13] J. Fine, J. Kuan-Yu Liu, A. Beck, K.Z. Alzarieni, X. Ma, V.M. Boulos, H.I. Kenttämaa, G. Chopra, Graph-based machine learning interprets and predicts diagnostic isomer-selective ion–molecule reactions in tandem mass spectrometry, Chem. Sci. 11 (2020) 11849–11858. https://doi.org/10.1039/D0SC02530E.

[14] J. Jang, G.H. Gu, J. Noh, J. Kim, Y. Jung, Structure-Based Synthesizability Prediction of Crystals Using Partially Supervised Learning, J. Am. Chem. Soc. 142 (2020) 18836–18843. https://doi.org/10.1021/jacs.0c07384.

[15] M.H.S. Segler, T. Kogej, C. Tyrchan, M.P. Waller, Generating focused molecule libraries for drug discovery with recurrent neural networks, ACS Cent. Sci. (2018). https://doi.org/10.1021/acscentsci.7b00512.

[16] B. Sattarov, I.I. Baskin, D. Horvath, G. Marcou, E.J. Bjerrum, A. Varnek, De Novo Molecular Design by Combining Deep Autoencoder Recurrent Neural Networks with Generative Topographic Mapping, J. Chem. Inf. Model. (2019). https://doi.org/10.1021/acs.jcim.8b00751.

[17] Q. Yuan, A. Santana-Bonilla, M.A. Zwijnenburg, K.E. Jelfs, Molecular generation targeting desired electronic properties: Via deep generative models, Nanoscale. 12 (2020) 6744–6758. https://doi.org/10.1039/c9nr10687a.

[18] K.M. Jablonka, D. Ongari, S.M. Moosavi, B. Smit, Big-Data Science in Porous Materials: Materials Genomics and Machine Learning, Chem. Rev. 120 (2020) 8066–8129. https://doi.org/10.1021/acs.chemrev.0c00004.

[19] S. Chong, S. Lee, B. Kim, J. Kim, Applications of machine learning in metal-organic frameworks, Coord. Chem. Rev. 423 (2020) 213487. https://doi.org/https://doi.org/10.1016/j.ccr.2020.213487.

[20] S.M. Moosavi, A. Nandy, K.M. Jablonka, D. Ongari, J.P. Janet, P.G. Boyd, Y. Lee, B. Smit, H.J. Kulik, Understanding the diversity of the metal-organic framework ecosystem, Nat. Commun. 11 (2020) 4068. https://doi.org/10.1038/s41467-020-17755-8.

[21] L. Turcani, R.L. Greenaway, K.E. Jelfs, Machine Learning for Organic Cage Property Prediction, Chem. Mater. 31 (2019) 714–727. https://doi.org/10.1021/acs.chemmater.8b03572.

[22] C.W. Coley, W. Jin, L. Rogers, T.F. Jamison, T.S. Jaakkola, W.H. Green, R. Barzilay, K.F. Jensen, A graph-convolutional neural network model for the prediction of chemical reactivity, Chem. Sci. 10 (2019) 370–377. https://doi.org/10.1039/C8SC04228D.

[23] M. Tsubaki, K. Tomii, J. Sese, Compound–protein interaction prediction with end-to-end learning of neural networks for graphs and sequences, Bioinformatics. 35 (2019) 309–318. https://doi.org/10.1093/bioinformatics/bty535.

[24] A. Raza, A. Sturluson, C.M. Simon, X. Fern, Message Passing Neural Networks for Partial Charge Assignment to Metal–Organic Frameworks, J. Phys. Chem. C. 124 (2020) 19070–19082. https://doi.org/10.1021/acs.jpcc.0c04903.

[25] D.K. Duvenaud, D. Maclaurin, J. Iparraguirre, R. Bombarell, T. Hirzel, A. Aspuru-Guzik, R.P. Adams, Convolutional networks on graphs for learning molecular fingerprints, in: Adv. Neural Inf. Process. Syst., 2015: pp. 2224–2232.

[26] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, ArXiv Prepr. ArXiv1703.01365. (2017).

[27] K. McCloskey, A. Taly, F. Monti, M.P. Brenner, L.J. Colwell, Using attribution to decode binding mechanism in neural network models for chemistry, Proc. Natl. Acad. Sci. 116 (2019) 11624–11629.

[28] V. Santolini, M. Miklitz, E. Berardo, K.E. Jelfs, Topological landscapes of porous organic cages, Nanoscale. 9 (2017) 5280–5298.

[29] M. Miklitz, K.E. Jelfs, pywindow: Automated Structural Analysis of Molecular Pores, J. Chem. Inf. Model. 58 (2018) 2387–2391.

[30] D. Kahneman, D.T. Miller, Norm theory: Comparing reality to its alternatives., Psychol. Rev. 93 (1986) 136.