Machine Learning for Organic Cage Property Prediction

Lukas Turcani,[†] Rebecca L. Greenaway,[‡] and Kim E. Jelfs[*,†]

[†]*Department of Chemistry, Imperial College London, South Kensington, London, SW7 2AZ, United Kingdom*

[‡]*Department of Chemistry and Materials Innovation Factory, University of Liverpool, 51 Oxford Street, Liverpool L7 3NY, United Kingdom*

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

Phone: +44 (0) 20759 43438

Abstract

We use machine learning to predict shape persistence and cavity size in porous organic cages. The majority of hypothetical organic cages suffer from a lack of shape persistence and as a result lack intrinsic porosity, rendering them unsuitable for many applications. We have created the largest computational database of these molecules to date, numbering 63,472 cages, formed through a range of reaction chemistries and in multiple topologies. We study our database and identify features which lead to the formation of shape persistent cages. We find that the imine condensation of trialdehydes and diamines in a [4+6] reaction is the most likely to result in shape persistent cages, whereas thiol reactions are most likely to give collapsed cages. Using this database, we develop machine learning models capable of predicting shape persistence with an accuracy of up to 93%, reducing the time taken to predict this property to milliseconds, and removing the need for specialist software. In addition, we develop machine learning models for two other key properties of these molecules, cavity size and symmetry.

We provide open-source implementations of our models, together with the accompanying data sets, and an online tool giving users access to our models to easily obtain predictions for a hypothetical cage prior to a synthesis attempt.

Introduction

Porous organic cages are a class of molecules distinguished by the presence of an internal cavity, made accessible to guest molecules via molecular windows.[1,2] These features provide the potential for organic cages to be used in a number of applications, notably encapsulation,[3] molecular separations[4–7] and catalysis,[8] and has led to the development of an increasingly active area of research. Organic cages are distinguished from other porous materials, such as zeolites and metal-organic frameworks (MOFs), by lacking an extended network of covalent bonds in the solid state. In addition, unlike extended frameworks, cage molecules are often soluble in organic solvents, allowing for solution processing into thin films or membranes, both in the crystalline or amorphous solid-state.[9] The lack of 3-dimensional chemical bonding can allow the solid-state structures to undergo large rearrangements, which has been utilized in the creation of molecular crystals with "on/off" porosity with polymorph switching.[10] However, the additional flexibility also means organic cages are less likely to be shape persistent, which means they collapse and lose porosity as a result of desolvation.[11] Shape persistence itself is often difficult to predict *a priori* without employing computational modelling.[11,12]

In most cases, cages are assembled using reversible dynamic covalent chemistry (DCC) from two multifunctionalized molecular precursors. Cages may exhibit two distinct, but not mutually exclusive, forms of porosity, termed intrinsic and extrinsic. Intrinsic porosity refers to voids found within molecules, such as the cavities of cages, while extrinsic porosity refers to voids which result from the inefficient packing of multiple molecules. To date, the largest synthesised organic cage possesses an internal cavity diameter of 2.3 nm and Brunauer-Emmett-Teller (BET) surface area of 3758 m^2 g^{-1}.[13] A key bottleneck to the prac-

tical application of cages is the discovery of viable candidate molecules. Currently, only a few hundred intrinsically porous molecules are known.[14] Typically, the development time associated with a single cage spans multiple months or years, including time needed for the identification of a suitable candidate molecule, synthesis and characterisation. In recognition of this obstacle, approaches such as high-throughput computational and synthetic robotic screening have recently been put to use in this field.[12] Nevertheless, important challenges remain. While computational modelling requires significantly less time than experimental approaches, modelling a large number of molecules is still time-consuming using forcefield approaches and intractable with quantum mechanical methods, particularly as these are large molecules, typically with >100 atoms.

In order to combat excessive computational times, use of machine learning (ML) techniques for molecular property prediction has been growing in popularity. The origins of this growth are manyfold. Though application of ML to chemistry has a long history, ML techniques have recently been popularized by a number of high-profile success stories in the broader scientific literature, such as AlphaGo,[15] and in commercial applications such as personal assistants. Naturally, the success of ML techniques in these areas has facilitated their uptake in others. The use of ML approaches in materials science in particular has seen much progress, from text-mining literature[16] to synthesis[17,18] and chemical design,[19] and been subject to a recent review.[20] ML has previously been applied in the discovery of new extrinsically porous molecular crystals, resulting in models which predict the porosity of crystals with up to 70% accuracy.[14] The advent of high-performance hardware has not only allowed the training of ever more complex ML models, but has increasingly allowed the creation of large computational data sets which provide fertile ground to ML algorithms. Examples of such initiatives are the Harvard Clean Energy Project,[21] NOMAD,[22] and the Materials Project,[23] among many others.[20] As a result, ML algorithms have been applied in a a broad range of prediction tasks including, but not limited to, energy,[24] prediction of MOFs,[25,26] toxicity,[27] and partial charges.[28]

Despite these advances, ML has yet to be applied to the property prediction of organic cages, despite the considerable computational expense associated with modelling them and the fact that the majority of hypothetical cages will lack shape persistency. A significant barrier to such an attempt has been the lack of a sufficiently large computational or experimental cage database. In this work, we attempt to address this firstly through the development of a large computational cage database consisting of over 60,000 cage molecules and secondly by training random forest models for the prediction of cage collapse, cavity size and window symmetry. All models rely only on the molecular graphs of cage precursors, bypassing the need for any structure determination of the cage. As a result, our cage collapse model can reduce the time taken to determine if a cage is collapsed from minutes to milliseconds, allowing for significant improvements in high-throughput screening. In order to provide easy access to our models we provide an online tool, available at `https://ismycageporous.ngrok.io`, which predicts if user-uploaded cage precursors are likely to form a shape persistent cage. The source code for our models is available at `github.com/lukasturcani/cage_prediction`. Finally, we analyse the cage precursors in our database to determine which structural features are conducive to the formation of shape persistent cages, informing future design.

Methods

Data sets

A data set of organic cages was assembled from a series of di-, tri- and tetra-topic precursors, designed by a synthetic chemist with many years of cage synthesis experience to be suitable for the generation of cages. In total, 118 di-topic, 51 tri-topic, and 20 tetra-topic precursor cores were included, each with locations of functional groups marked. These are available as SMILES strings, along with a script performing the functional group substitution and 3D embedding, in the supporting information. Throughout, we refer to the precursor with more reactive functional groups as the "building block", and the precursor with fewer functional

groups as the "linker". Each precursor backbone was generated with each of the different functional groups to include cages that could be generated through a variety of reactions. This process is summarized in Figure 1.

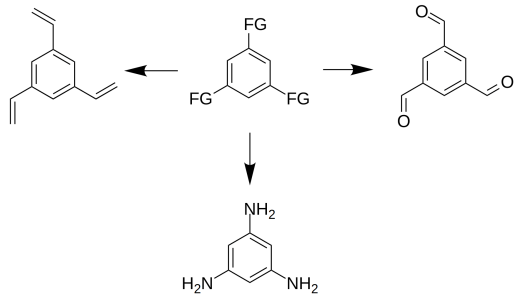

Figure 1: An example showing the creation of multiple precursors from a backbone with functional group (FG) positions marked.

As the majority of intrinsically porous organic cages reported to date have used DCC, in particular imine condensation reactions, we have focused on DCC reactions. The functional groups used in this study are aldehydes, alkynes, amines, carboxylic acids, alkenes, and thiols, which are combined using the imine/amide condensation, alkyne/alkene metathesis and disulfide formation reactions. For each pair of functional groups capable of undergoing a reaction, every possible pairing of precursors was used to generate the corresponding cage. These pairings are listed in Table 1. In addition to this, cages were generated in a number of topologies. We use the topology nomenclature defined previously by Santolini *et al.*, where a topology such as $\mathbf{Tri^4Di^6}$ refers to a cage formed from 4 tri-functionalized and 6 di-functionalized precursors.[29] The topologies considered for this study include $\mathbf{Tri^4Di^6}$, $\mathbf{Tri^8Di^{12}}$, $\mathbf{Tet^6Di^{12}}$, $\mathbf{Tet^6Tri^8}$, and $\mathbf{Tri^4Tri^4}$, which are shown in Figure 2. This led to a total of 63,472 organic cages. This is the largest number of organic cages studied to date and compares to the few hundred that have been synthetically reported so far.[2] The data set is available for download at `https://doi.org/10.14469/hpc/4618`.

The 3-dimensional structure of each precursor was determined by embedding 100 conformations using the RDKit ETKDG method.[30] The energy of each conformation was evaluated using UFF,[31] and the lowest energy conformation was chosen as the final structure. The

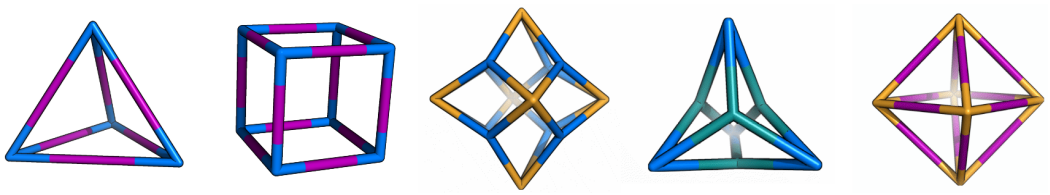

Figure 2: All cage topologies considered for this study. Left to right: $\mathbf{Tri^4Di^6}$, $\mathbf{Tri^8Di^{12}}$, $\mathbf{Tet^6Tri^8}$, $\mathbf{Tri^4Tri^4}$ and $\mathbf{Tet^6Di^{12}}$. Di-topic components are shown in purple, tri-topic in blue (or cyan) and tetra-topic in orange.

cage structures were then generated using our supramolecular toolkit (*stk*), an open-source Python library which allows for the assembly of complex supramolecular structures such as porous organic cages in various topologies.[32] Structures produced by *stk* were optimized in MacroModel[33] using a three-stage approach. Firstly, only the bonds added to the precursors during cage assembly were geometry optimized using the OPLS3 forcefield,[34] while keeping all other intramolecular distances fixed. Secondly, the entire molecule was optimized and finally, a molecular dynamics (MD) run was performed to search for low energy conformations. Each cage underwent an MD simulation for 2 ns after a 100 ps equilibration, with a timestep of 1 fs and at a temperature of 700 K. At 50 points during the MD trajectory, the cage structure was geometry optimized and the lowest energy sampled conformation was then used for further analysis. The convergence criteria for all geometry optimizations was a maximum of 2500 iterations and a gradient of 0.05.

Potential limitations of the forcefield description

In this work, we exclusively use the OPLS3 calculations described above to give the definitive structures of the molecules for ongoing classification and then machine learning of cage shape persistency, window sizes and cavity diameters. This assumes two key points: (i) that the forcefield can appropriately describe the cage chemistry involved in this diverse training set and (ii) that the MD procedure we use is sufficient to discover the relevant low energy conformations of each cage. OPLS3 was designed to provide broad coverage of small molecules,[34] thus is suitable for use as a transferable forcefield. We have extensive experi-

Table 1: Reactions and precursors used to generate cages. 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).

Building block	Linker	Topologies	Reaction	No. cages
aldehyde 3	amine 2	**Tri^{4}Di^6**, **Tri^{8}Di^{12}**	imine condensation	12036
amine 3	aldehyde 2	**Tri^{4}Di^6**, **Tri^{8}Di^{12}**	imine condensation	12036
aldehyde 4	amine 2	**Tet^{6}Di^{12}**	imine condensation	2360
amine 4	aldehyde 2	**Tet^{6}Di^{12}**	imine condensation	2360
aldehyde 4	amine 3	**Tet^{6}Tri^8**	imine condensation	1020
amine 4	aldehyde 3	**Tet^{6}Tri^8**	imine condensation	1020
amine 3	aldehyde 3	**Tri^{4}Tri^4**	imine condensation	2601
alkene 3	alkene 2	**Tri^{4}Di^6**	alkene metathesis	6018
alkyne 3	alkyne 2	**Tri^{4}Di^6**	alkyne metathesis	6018
[a]thiol 3	thiol 2	**Tri^{4}Di^6**	disulfide formation	5967
carboxylic acid 3	amine 2	**Tri^{4}Di^6**	amide condensation	6018
amine 3	carboxylic acid 2	**Tri^{4}Di^6**	amide condensation	6018

[a]One ditopic thiol precursor is excluded as the precursor backbone already included two thiol groups.

ence in previous work of the use of OPLS3 for the exploration of cage conformation, shape persistence, flexibility and host-guest chemistry, including in high-throughput studies,[29,35–39] which reassures us that OPLS3 is suitable for the task of both structure and energetics of the molecules involved here. However, we note that our previous focus has been on imine cages, and therefore there is the possibility that more diverse cage chemistries are not as thoroughly validated. Our manual inspection of molecules suggested plausible geometries, but of course it was not possible to validate all 63,472 cages, so we must accept that a poor forcefield description is a possible source of error for our data.

Structural characterisation

The internal cavity size for each cage was calculated by translating the centroid of the cage onto the origin. An example cavity can be seen on the left in Figure 3. The cavity size was measured as the distance from the origin to the nearest atom, with the van der Waals radius of the atom subtracted. This functionality was provided by *stk*. Next, we calculated the difference in the window diameters. An example cage window is highlighted on the

right of Figure 3. We wished to calculate the difference in the window diameters in a single cage, as we consider it to be a useful proxy for the degree of symmetry, and often shape-persistence in a cage. The first criterion is assumed to correlate with ease of synthesis, as when highly symmetrical precursors are used, high symmetry in the resultant assembly suggests a good correspondence of the precursors to the topology and a low strain in the assembly. A shape persistent, open, cage built from high symmetry precursors will typically form a high symmetry assembly with all windows being of similar size. Shape persistency is a primary property of interest in organic cages.

To calculate the difference in the window diameters, the diameters of all windows were calculated using pywindow.[40] If pywindow failed to detect the expected number of windows for a given topology, for example four windows in a **Tri^{4}Di^6** topology, the cage was not used for training window difference models. For cages where all windows were successfully detected, all possible pairs of windows were compared. The difference in diameters between the two windows in each pair was calculated and the resulting differences averaged to produce the final measure of "window difference". This calculation was performed in *stk* using the Cage.window_difference() method. Finally, in the supporting information we show results of predicting the standard deviation in the window sizes. We calculated the standard deviation in the window sizes as a simpler measure of window size differences, but found no notable change in performance.

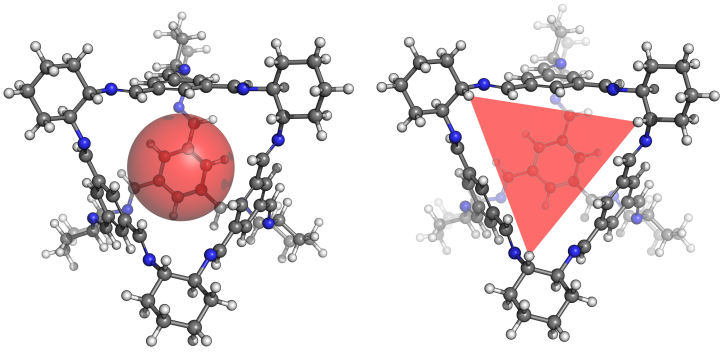

Figure 3: A porous organic cage with highlighted cavity (left) and window (right) in red.

Assignment of collapse labels

To model cage collapse, each cage needed to be labelled as either "collapsed", "not collapsed" (*i.e.* shape persistent) or "undetermined". Cages in the final category were not used when training cage collapse models. An example of a shape persistent cage can be seen in Figure 3 and of a collapsed cage in Figure 4. A number of approaches were trialled to accomplish the labelling. Initially, we developed a graphical user interface, allowing us to manually assign labels, as shown in Figure S1. This software is freely available at `github.com/lukasturcani/molder` and allows users to manually label molecules over the web. The molecules to label and which labels to assign can be freely modified depending on the task. However, such an approach soon proved to be too time consuming and unscalable to data sets with tens of thousands of molecules. As a result, we instead developed an automated approach to cage labelling.

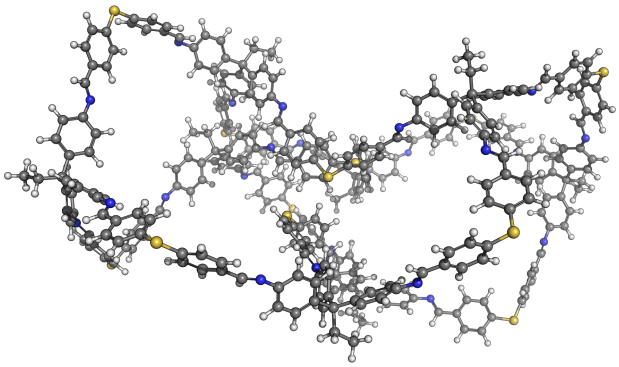

Figure 4: An example of a collapsed organic cage that does not have symmetric windows or a single, central cavity.

To automate cage labelling, we used pywindow to detect all the windows present in a cage. This is because in all cases where a cage is collapsed, windows which should be topologically identical will take on different sizes or some windows may be removed completely. Naive collapse detection methods relying only on cage volume and cavity size frequently lead to incorrect labels, as collapsed cages may flatten in such a way that they retain a central cavity and a large diameter. As we use *stk* to assemble cages of a given topology, we know how

many windows to expect in shape persistent cages of a given topology. If pywindow does not find the expected number of windows, we labelled the cage as collapsed. By manually examining thousands of cages labelled with this approach, we concluded that nearly all cages labelled "collapsed" are labelled correctly. However, not all collapsed cages are detected with this approach. For cages where pywindow detects the expected number of windows for a given topology, we apply the following criterion:

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

If $\alpha < 0.035$ and cavity size is greater than 1 Å, the cage is labelled as "not collapsed", else it is labelled "undetermined". This equation was derived through trial-and-error in order to maximize the number of cages correctly labelled as "not collapsed". While the majority of cages are correctly labelled, mislabelling does occur and the "not collapsed" labels can be considered slightly noisy as a result. In addition, not all shape persistent cages are detected using this method. There was a high imbalance in the data set for a reaction of di-topic thiols with tri-topic thiols; manual inspection of this data set revealed that the 99 cages labelled "not collapsed" were in fact collapsed. Due to lacking any cages in the "not collapsed" class, this data set was not used to train any of the following models.

Input featurization.

The base input for all models consisted of an extended-connectivity fingerprint (ECFP), as implemented by RDKit.[41] An ECFP is a vector, of size given by the number of fingerprints bits, which indicates the presence of certain substructures in the molecule. The size of the substructures represented by the fingerprint is determined by the fingerprint radius. After testing an initial collapse prediction model using a variety of fingerprint radii and bit sizes, a radius of 8 and bit size of 512 was found to give the smallest fingerprint with the highest accuracy score and was used for all successive models. The results of this featurization anal-

ysis can be seen in the supporting information. The input for a given cage was constructed by concatenating the ECFP of radius 8 and bit size 512 of each precursor, leading to a total bit size of 1024. Each element of the fingerprint indicates how many times it was activated by some molecular substructure. This is in contrast to the more common approach of using either 1 or 0 to indicate if a given element was activated by a molecular substructure at least once. A count-based fingerprint should provide more information to our models and be more suitable for regression modelling, by allowing for a more continuous input space. For models trained or evaluated on cages of multiple topologies, the final fingerprint is extended with a 1-hot vector indicating the topology of the cage. This featurization approach intentionally excludes the fingerprint of the cage molecule itself. It means that the trained models can be used for predictions without having to construct or geometry optimize cage molecules. Only the precursors themselves are necessary to predict the properties of a cage. This means less time is required to generate a prediction and no molecular modelling expertise or software is required, at the cost of some prediction accuracy.

Evaluation metrics.

In order to avoid excessive complexity, we organise our models into two types. Cross-reaction models are trained and evaluated using cages of the same topology, $\mathbf{Tri^4Di^6}$, but assembled from precursors with different functional groups. Cross-topology models are trained with cages of different topologies, but formed via the imine condensation reaction. Cross-reaction models allow us to evaluate how transferable models trained on cages generated with one set of reactions are to cages generated with a different set of reactions. Cross-topology models allow us to evaluate how additional training examples, when consisting of cages with different topologies but formed by the same reaction, affect performance.

Cross-reaction models are organized by data set, indicated by a row with a $\mathbf{Tri^4Di^6}$ topology in Table 1. For each of these data sets, three distinct evaluations take place. Firstly, we evaluate the performance using 5-fold cross-validation (CV) using cages from

that data set only. We denote this as "train & test" in the results later. Secondly, we train a model with all cages in the data set and use all cages in the remaining data sets of $\mathbf{Tri^4Di^6}$ topology as the test set. Thirdly, we train a model using all $\mathbf{Tri^4Di^6}$ cages, with the exception of the chosen data set, which is used as the test set. The latter two evaluation schemes allow us to determine how transferable our models are to cages generated from precursors with functionalities different to those in the training set and correspond to the "train" and "test" columns respectively in the results. For each task discussed in this paper, we train a single cross-topology model. These models are trained and evaluated with the data set assembled by combining all cages generated with the imine condensation reaction. We use 5-fold CV to train and evaluate these models.

For the regression models of cavity size and window size difference, performance was evaluated with two metrics, mean-absolute-error (MAE) and the correlation coefficient, R^2, which shows the percentage of variance in the training data accounted for by our models. These metrics are given by:

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_{\mathrm{true}}^{(i)} - y_{\mathrm{predicted}}^{(i)}|}{n} \tag{2}$$

$$\mathrm{R}^2 = 1 - \frac{\sum_{i=1}^{n} (y_{\mathrm{true}}^{(i)} - y_{\mathrm{predicted}}^{(i)})^2}{\sum_{i=1}^{n} (y_{\mathrm{true}}^{(i)} - \bar{y})^2} \tag{3}$$

respectively. Here, n is the total number of examples in the evaluation set, $y_{\mathrm{true}}^{(i)}$ is the ith example's true value, $y_{\mathrm{predicted}}^{(i)}$ is the ith example's predicted value, and $\bar{y}$ is the mean value of $y_{\mathrm{true}}^{(i)}$. The classification model for cage collapse was evaluated using prediction accuracy and the precision and recall of each class, defined as:

$$\mathrm{accuracy} = \frac{\sum_{i=1}^{n} I(y_{\mathrm{true}}^{(i)} = y_{\mathrm{predicted}}^{(i)})}{n} \tag{4}$$

$$\mathrm{precision} = \frac{\mathrm{true\ positives}}{\mathrm{true\ positives} + \mathrm{false\ positives}} \tag{5}$$

$$\mathrm{recall} = \frac{\mathrm{true\ positives}}{\mathrm{true\ positives} + \mathrm{false\ negatives}} \tag{6}$$

respectively, where $I(x)$ is a function returning 1 if the statement x is true and 0 if x is false. In addition, all CV folds were stratified and during training, examples were weighted to counter any class imbalance present in the training set.

Implementation.

All models were implemented in Python 3.6 using the scikit-learn library[42] version 0.19.1 and the random forest algorithm. The random forest algorithm works by creating an ensemble of decision trees. Decision trees attempt to split examples in the training set into subsets until either all examples in the subset belong to the same class, or each subset holds a minimum number of examples. Each decision tree in the forest attempts to split a bootstrap sample of the entire training set and uses a random subset of input features as the basis for splits. This sampling of examples and features results in a decorrelation between trees, trading a large decrease in variance of the forest for a smaller increase in bias and increasing quality of predictions overall. The prediction of a forest is given by a majority vote across trees in classification tasks and by taking the average across predictions in regression tasks. Random forests are ubiquitous in ML for chemistry, used to model quantitative structure-activity relationships (QSARs) and comparing favourably to other approaches such as support vector machines (SVMs) in terms of both performance and interpretability.

We used random forests for both regression and classification tasks for a number of reasons. Firstly, preliminary tests showed that the choice of learning algorithm between random forests, support vector machines (SVMs), and neural networks did not significantly affect prediction accuracy. Secondly, compared to the other two models, random forests have either few or easily interpretable hyperparameters. Finally, in many cases, random forests have shorter training times, though this is subject to the choice of hyperparameters. All of our code is available at `github.com/lukasturcani/cage_prediction`.

Parameters and hyperparameters.

Apart from the hyperparameters of fingerprint radius and size that concern the input featurization (discussed in detail in the supporting information), random forests come with their own set hyperparameters. We use 100 trees in each model, as this provides a reasonable compromise between training time and model accuracy. Increasing this hyperparameter increases the model accuracy, subject to diminishing returns. Our models use the RandomForestClassifier and RandomForestRegressor classes defined by scikit-learn, and we use default values for all parameters, with the exception of "class_weight", which we set to "balanced" when training classifiers. This means that during training, examples are weighted so as to counteract class imbalance. Gini impurity was used to measure the quality of the splits for classification tasks and mean-squared-error for regression tasks. Trees in the forest were grown until all examples in a leaf node were from the same class or only 2 examples remained in the leaf node.

Results and discussion

Shape persistency prediction

A prediction model for shape persistency allows us to predict if a cage will collapse or not without performing a structural optimization step. This reduces the time taken to establish if a cage is shape persistent from minutes to milliseconds, making high-throughput screening possible. By filtering out cages which are highly likely to be collapsed, a full structural optimization and characterisation only needs to be applied to molecules more likely to be porous. This drastically reduces the number of molecules on which a structural optimization needs to be performed. Further, a prediction model is accessible to those without custom software for molecular modelling, such as experimental chemists, allowing a check for shape persistency to be a routine part of synthesis planning.

In order for the shape persistency prediction models to be effective, a high precision and recall for identifying collapsed cages is necessary, as we prioritise avoiding cages which are almost certainly non-porous. Precision and recall for cages labelled "not collapsed" (*i.e.* shape persistent) is less important, as this label does not affect whether a cage undergoes an optimization or not. Despite our models being relatively naive, Table 2 shows it is possible to achieve very high accuracy, precision, and recall scores, especially when cages in the training and test sets are formed through the same reaction. Under such circumstances, the lowest precision and recall scores for the "collapsed" class are both 0.87, while the highest are 0.90 and 0.96, respectively. As a result, we foresee these models having practical applications in computational high-throughput screening of cages and for experimental chemists planning cage syntheses. In Figure 5 we see that some models may benefit from an increased data set size, though any gains in accuracy are likely to be quite small.

The worst performing scores in Table 2 are the anomalously low recall scores for shape persistent cages formed through the condensation reaction between a carboxylic acid and an amine molecule. These can be seen in the rows "carboxylic acid 3 amine 2" and "amine 3 carboxylic acid 2", with recall scores of 0.59 and 0.72, respectively. The low scores for these data sets can be linked to them having the lowest proportion of cages labelled "shape persistent" with 18% and 23%, respectively, as shown in Table S1. We conclude that due the lack of shape persistent cages in the training sets of these models, they are unable to generalise well and as a result are poor at identifying shape persistent cages in the test set.

In general, across all models, we see better performance at identifying collapsed rather than shape persistent cages, as indicated by the precision and recall scores for these respective classes. We identify two factors which contribute to this result. Firstly, labels for shape persistent cages are noisy, while those of collapsed cages are not. This is a reflection of the fact that the criterion to label a cage collapsed is relatively simple (the expected number of windows is not found), while the criteria for a shape persistent cage are more complex. Secondly, the features which lead to cage collapse may be easier to identify. For example,

Table 2: Shape persistency prediction using 100 tree random forest models. The results are for when a model was trained on a single data set, given by the row.

Building block	Linker	Accuracy	Precision (shape persistent)	Recall (shape persistent)	Precision (collapsed)	Recall (collapsed)
aldehyde3	amine2	0.88	0.88	0.88	0.87	0.88
amine3	aldehyde2	0.88	0.87	0.87	0.88	0.89
alkene3	alkene2	0.86	0.84	0.84	0.87	0.87
alkyne3	alkyne2	0.92	0.94	0.93	0.89	0.91
carboxylic acid3	amine2	0.88	0.80	0.59	0.89	0.96
amine3	carboxylic acid2	0.88	0.82	0.72	0.90	0.94

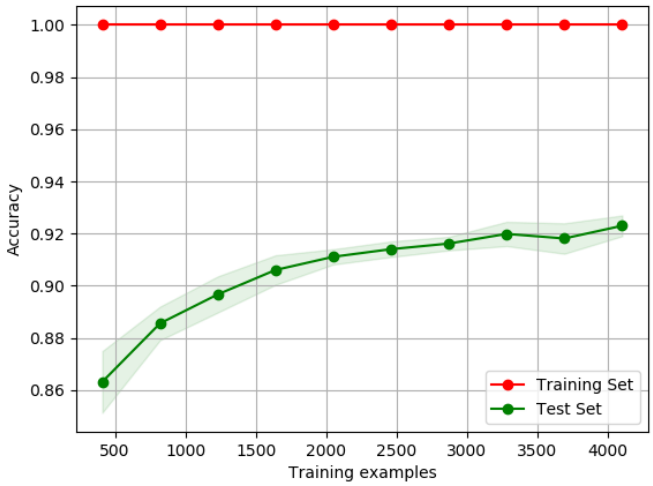

Figure 5: The accuracy of collapse prediction for the "alkyne 3 alkyne 2" data set at different data set sizes. The model was trained and tested using alkyne cages only. The green and red points represent mean cross-validation accuracy for the test and training sets, respectively, while the shaded areas represent the standard deviation across the cross-validation folds.

while all excessively long and flexible linkers will produce collapsed cages, not all short and rigid linkers will produce shape persistent ones.

When models are trained and tested on cages generated through different reactions, shown in Tables 3 and 4, drops in performance are significant. Table 3 shows the results of models trained only on the data set specified by the row and tested on the data sets shown in the remaining rows. Conversely, Table 4 shows the results of models tested on the data set specified by the row and trained on the data sets listed in the other rows. While these models have some predictive capacity, prediction should be made using models trained on cages generated from the same reaction where possible. The average accuracy of these cross-reaction models is only 0.65, compared to 0.88 of models in Table 2.

Table 3: Cage collapse prediction using 100 tree random forest models. The results are for when a model was trained on a single data set, given by the row. The remaining rows are the data sets used as the test set.

Building block	Linker	Accuracy	Precision (shape persistent)	Recall (shape persistent)	Precision (collapsed)	Recall (collapsed)
aldehyde3	amine2	0.67	0.50	0.56	0.77	0.73
amine3	aldehyde2	0.66	0.48	0.35	0.72	0.81
alkene3	alkene2	0.66	0.49	0.45	0.73	0.77
alkyne3	alkyne2	0.62	0.36	0.32	0.71	0.75
carboxylic acid3	amine2	0.65	0.80	0.08	0.64	0.99
amine3	carboxylic acid2	0.67	0.76	0.13	0.66	0.98

Table 4: Cage collapse prediction using 100 tree random forest models. The results are for when a model was tested on a single data set, given by the row. The data sets given by the other rows were used as the training set.

Building block	Linker	Accuracy	Precision (shape persistent)	Recall (shape persistent)	Precision (collapsed)	Recall (collapsed)
aldehyde3	amine2	0.61	0.96	0.23	0.56	0.99
amine3	aldehyde2	0.72	0.88	0.47	0.66	0.94
alkene3	alkene2	0.63	0.82	0.22	0.61	0.96
alkyne3	alkyne2	0.41	0.97	0.04	0.40	1.00
carboxylic acid3	amine2	0.71	0.42	0.88	0.95	0.66
amine3	carboxylic acid2	0.73	0.50	0.88	0.94	0.67

A close analysis of the precision and recall scores in Tables 3 and 4 shows that trained models have a strong tendency to overwhelmingly predict a single class, demonstrated by a combination of a high recall and low precision for the given class. An example of this can be seen for the trialdehyde and diamine cages in Table 4 (first row). Here, the model was trained on cages from all data sets with the exception of the first row, which was used as the test set. In this case, the trained model predicts that virtually all cages in the test set are collapsed, as shown by a recall score of 0.99 and precision score of 0.56 for collapsed cages. In contrast, the model correctly identifies only 23% of shape persistent cages in the test set. However, it achieves a precision score of 0.96 for these cages. This implies that the model learnt specific features which prevent collapse in the training data set, and that these features are not found in the majority of porous trialdehyde and diamine cages. However, when such features are found, they are strong indicators that a cage will be shape persistent. The remaining features which indicate a trialdehyde and diamine cage is shape persistent must be unique to this reaction pairing, and this accounts for the difference between the recall score of 0.88 in Table 2 and 0.22 in this model. This is an interesting result, as it suggests that during the design of a porous organic cage, the chemical features that are required to make sure the cage is shape persistent are highly dependent on the reaction chemistry. This is likely contrary to the typical current approach for designing shape persistent cages, whereby you would assume you want the same features in the precursor cores regardless of the reaction chemistry. Overall, Figure 6 shows that our models do not strongly depend on a particular set of features, with feature importance being relatively similar among the top 20 features. In addition, the importance of each feature varies widely from tree to tree in the forest as indicated by the confidence intervals of the feature importances. Because of this, it is difficult to draw conclusions about the internal decision making of our models, however it does imply that the collapse prediction cannot be reduced to the simple presence or absence of a few molecular substructures. This is to be expected as chemists often struggle to design and predict shape-persistence in cages.

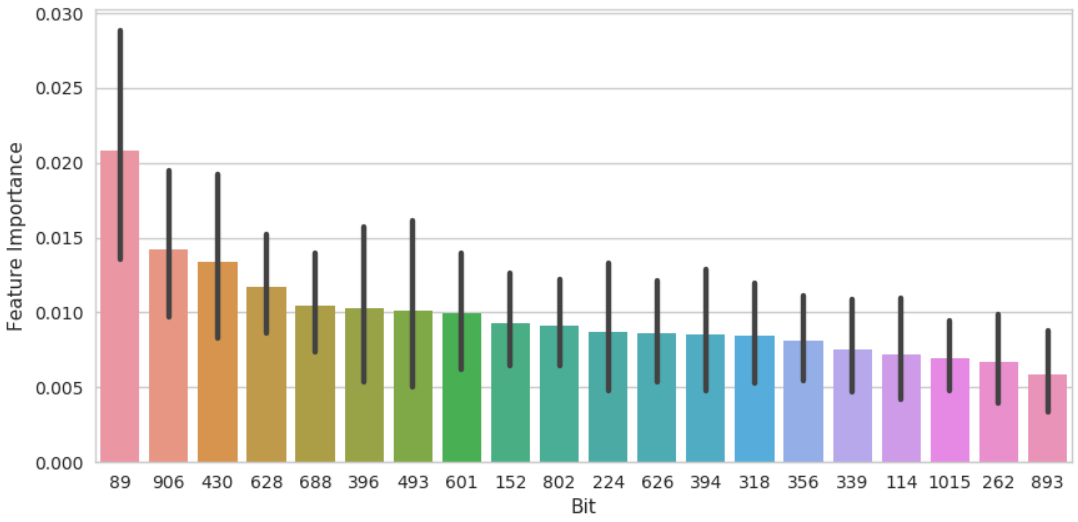

Figure 6: Feature importance for the 20 most important fingerprint bits on a random forest trained on cages generated through alkyne metathesis. Black bars represent the 95% confidence intervals across all the trees in the forest.

Finally, the cross-topology model for the collapse prediction task achieved an accuracy of 0.86, precision scores of 0.82 and 0.89, and recall scores of 0.79 and 0.90 for shape persistent and collapsed cages respectively. As a result, this model does not show a strong benefit from additional training examples corresponding to cages with different topologies. However, the model does show a slight improvement in both the precision and recall of collapsed cages, compared to the trialdehyde and diamine model in Table 2. This result indicates that a topologically diverse training set may yield some benefit to model performance. Due to the strong performance of models trained on single reaction, compared to cross-reaction and cross-topology models, our online prediction tool only provides the trained single reaction models, though cross-reaction and cross-topology models can be trained using our open source code.

Cavity size prediction

The results of 18 different random forest models used for predicting cavity sizes of organic cages are presented in Table 5 for all cross-reaction models. In order to predict cavity size,

only shape persistent cages were used. This is because collapsed cages lack a well defined, central cavity. A clear trend emerges, in that all models perform best when trained and evaluated on cages generated from the same reaction chemistry, as shown in the "Test & Train" column. The average MAE of the models is 1.5 Å. While such an error is unacceptable for the screening of small cages (<8 Å cavity diameter), it can be used to screen cages with medium to large cavities. The best performing model is for cages generated through alkyne metathesis, with a MAE of 1.13 Å, and we attribute this to the rigid triple bond formed by this reaction, as structures with fewer degrees of freedom are likely to exhibit less structural variance. Indeed, sampling more than one conformation from the original MD simulations of the cage is likely to find a range of cavity sizes. The relatively rigid porous imine cage, **CC3**,[43] was found to have a pore diameter range of approximately 1.5 Å, from a mean value of 5.2 Å during MD simulations.[40] When using the alkyne cage model as a representative example, in Figure 7 we can see that cavity size prediction models may benefit from additional training examples, as there is a steady decrease in MAE when using 100% of the data set, up from using only 90% of the data set for training.

Table 5: Cavity size prediction using 100 tree random forest models. "Test & Train" column indicates that the data set specified in the row was split using 5-fold cross validation in the test and train data sets. The "Train" column indicates that the data set specified in the row was used only as the training set and that the remaining data sets were used as the test set. The "Test" column indicates that the data set specified in the row was used only as a test set and the model was trained on the remaining data sets.

Building block	Linker	Test & Train		Train		Test	
		MAE / Å	R^2	MAE / Å	R^2	MAE / Å	R^2
aldehyde3	amine2	1.33	0.89	2.52	0.75	3.85	0.49
amine3	aldehyde2	1.45	0.89	2.25	0.81	4.92	0.17
alkene3	alkene2	1.62	0.85	2.24	0.79	5.31	0.16
alkyne3	alkyne2	1.13	0.92	3.87	0.64	3.89	0.42
carboxylic acid3	amine2	1.74	0.86	2.17	0.80	6.44	-0.20
amine3	carboxylic acid2	1.74	0.85	1.98	0.84	4.82	0.28

We also trained a cross-topology model, with cages produced through the imine condensation reaction but consisting of five different topologies. This model was evaluated using

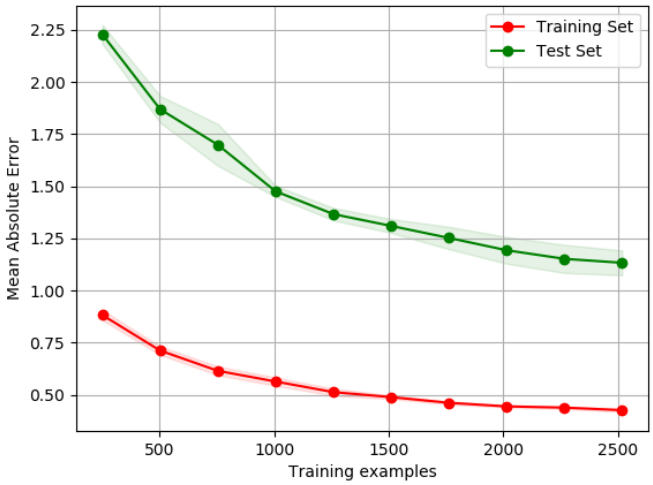

Figure 7: The mean absolute error for the "alkyne 3 alkyne 2" data set at different data set sizes. The model was trained and tested using alkyne cages only. The green and red points represent mean cross-validation MAE for the test and training sets, respectively, while the shaded areas represent the standard deviation across the cross-validation folds.

5-fold CV to give a MAE of 2.28 Å and an R^2 of 0.81. The model performs worse than both amine & aldehyde models shown by the rows "aldehyde 3 amine 2" and "amine 3 aldehyde 2" in Table 5, both of which are trained on cages also used in the cross-topology model. This demonstrates that additional training examples, consisting of cages with different topologies, do not produce a clear benefit to model performance. Previous findings have shown that the same precursors when changing from a $\mathbf{Tri^4Di^6}$ topology to a $\mathbf{Tri^8Di^{12}}$ topology, can result in the complete loss of shape persistency, and hence dramatic decrease or complete loss of cavity size.[11,29] In this context, it is not surprising that our model struggles with this prediction.

Figure 8 (left) shows the predicted versus reference cavity size for the cross-topology model. The model's predicitions fall further from the target value as cavity size increases. This can be seen more clearly in Figure S5. Due to this, the model displays a clear heteroscedasticity, as the variance of residuals increases with cavity size. The origin of this is likely an increase in the variance of cavity sizes as cage size increases, which follows from the fact that larger cages are more flexible. However, a secondary cause is likely that our input

featurization does not capture all the factors that influence cavity size.

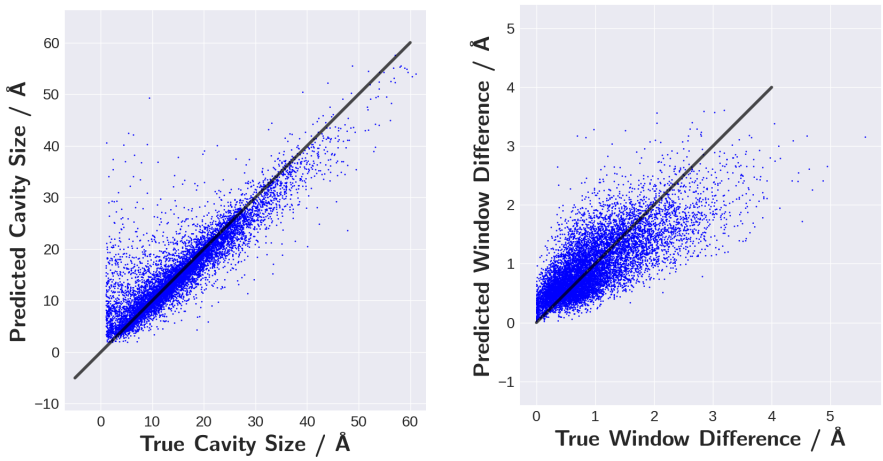

Figure 8: (left) The true cavity size versus the predicted cavity size. (right) The true window size difference versus the predicted window size difference. The black lines show where predictions match true cavity size.

When training and evaluating with cages generated through different reactions, shown by both the "Train" and the "Test" columns in Table 5, prediction accuracy falls drastically, leading to an approximate doubling of the MAE. In the case of the "Train" column, this loss in performance occurs despite the training set being significantly larger. This highlights the need to ensure cages found in the training set and cages which are to be screened, are formed using the same reaction. In addition, this result shows that structural features which lead to shape persistence differ between cages produced through different reactions, as decisions learned by the forests are not entirely transferable between data sets.

Window size difference prediction

We developed the window size difference models to provide a measure for the asymmetry of a cage, as shape persistent molecules often exhibit high degrees of symmetry. An alternative approach to measure the asymmetry by predicting the standard deviation of window sizes was found to produce similar results and is discussed in the supporting information. As with cavity size prediction, we restrict our models to training on shape persistent cages as these are the only ones with well defined windows. Collapsed cages may lack any windows

whatsoever. The results are shown in Table 6. The mean window size difference across shape persistent, $\mathbf{Tri^4Di^6}$ cages in the database is 0.56 Å, with a standard deviation of 0.36 Å. Our results show that once again, cages generated by alkyne metathesis produce the most reliable models, with a MAE of 0.2 Å, compared to a MAE of 0.24 Å across the models on average. Considering the magnitude of this error in relation to the standard deviation of the data, these models are unsuitable for applications which require ranking molecules in terms of their asymmetry. On the other hand, they may be suitable for screening out cages on the basis of some threshold window difference value. Figure 9 shows that in order to increase the accuracy of these models, more training examples may produce some benefit.

Table 6: Window difference prediction using 100 tree random forest models. "Test & Train" column indicates that the data set specified in the row was split using 5-fold cross validation in the test and train data sets. The "Train" column indicates that the data set was used only as the training set and that the remaining data sets were used as the test set. The "Test" column indicates that the data set was used only as a test set and the model was trained on the remaining data sets.

Building block	Linker	Test & Train		Train		Test	
		MAE / Å	R^2	MAE / Å	R^2	MAE / Å	R^2
aldehyde3	amine2	0.24	0.27	0.28	0.05	0.29	0.10
amine3	aldehyde2	0.23	0.25	0.25	0.16	0.29	0.11
alkene3	alkene2	0.25	0.17	0.25	0.18	0.29	0.08
alkyne3	alkyne2	0.20	0.41	0.32	-0.14	0.29	-0.04
carboxylic acid3	amine2	0.26	0.21	0.31	-0.14	0.29	0.05
amine3	carboxylic acid2	0.26	0.23	0.29	0.10	0.27	0.17

The cross-topology model for this task had a MAE of 0.34 Å and R^2 of 0.56. This again shows that additional training examples from cages with a different topology, despite similar chemistry, do not produce improved model performance. The plot of true versus predicted window difference can be seen in Figure 8 (right). This model is also heteroscedastic (see Figure S5) and we attribute this to the same causes as our cavity size models. Testing and training on cages formed by different reactions produces extremely weak models, with some R^2 values being negative. This shows the trained models perform worse than predicting some constant value for each cage. As a result, we conclude that training and predicting cages

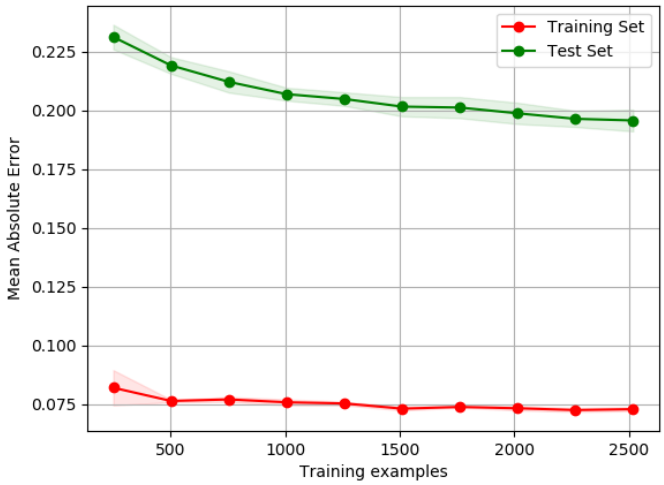

Figure 9: The mean absolute error for the "alkyne 3 alkyne 2" data set at different data set sizes. The model was trained and tested using alkyne cages only. The green and red points represent mean cross-validation MAE for the test and training sets, respectively, while the shaded areas represent the standard deviation across the cross-validation folds.

formed via the same reactions is a requirement of these models, as cages formed through different chemistries do not provide transferable information.

Analysis of 63,472 organic cages

The creation of a large, labelled cage data set presents opportunities for data mining, and thus to learn which chemical features infer which properties in a cage. We begin by analysing the characteristics of cages in the various data sets and topologies. Figure 10 shows the proportion of collapsed, shape persistent, and undetermined cages for each reaction. A number of trends emerge. Firstly, thiol cages are a significant outlier in that the entire data set consists of collapsed cages. The cages in this data set labelled "shape persistent" were mislabelled, as noted earlier. Cages with a $\mathbf{Tri}^{8}\mathbf{Di}^{12}$ topology have a higher proportion of collapsed cages than other topologies, 71% and 73% collapsed for the two $\mathbf{Tri}^{8}\mathbf{Di}^{12}$ data sets compared to 33 - 66% for the other data sets (excluding thiols). This can be attributed to $\mathbf{Tri}^{8}\mathbf{Di}^{12}$ being the largest topology in our study and as a result the most flexible.

Figure 10 also shows that with the exception of the "alkyne2 alkyne3" (52% shape per-

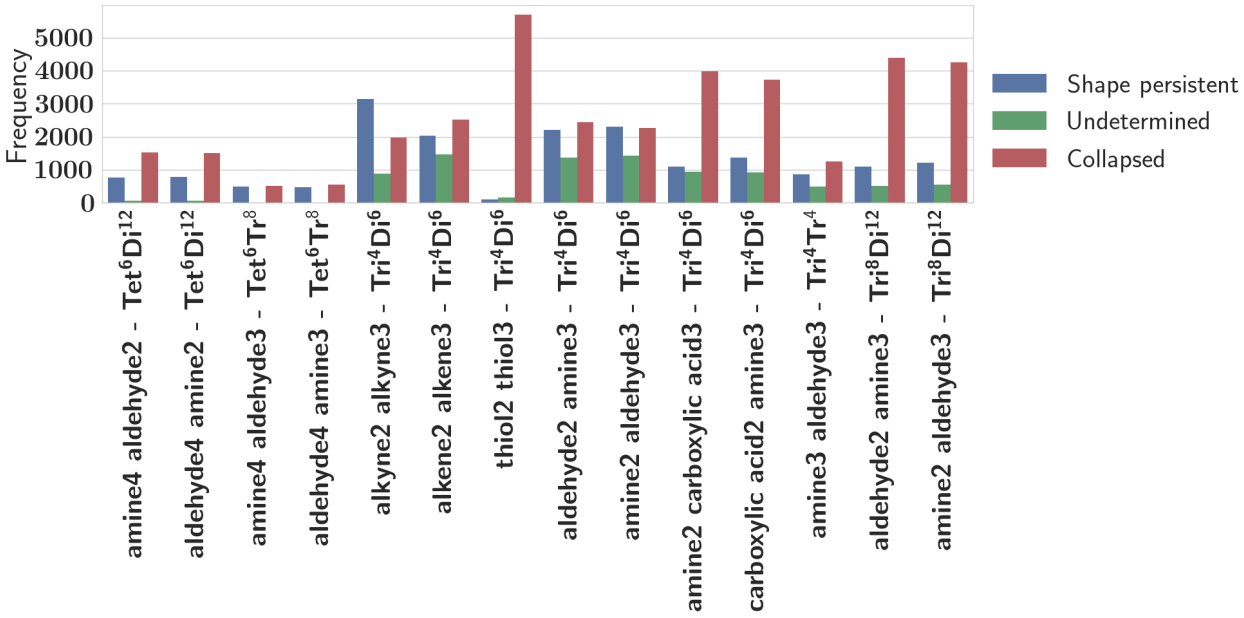

Figure 10: Number of cages labelled "collapsed", "shape persistent" and "undetermined" by our automated labelling approach, show across precursor combination and topology.

sistent, 33% collapsed) and "amine2 aldehyde3 - **Tri**^{4}**Di**^{6}" (49% shape persistent, 38% collapsed) data sets, all data sets have more collapsed than shape persistent cages. This highlights the relatively rarity of shape persistent molecules, as despite the precursors being explicitly designed to maximize the likelihood of their construction, they are still a minority (overall only 28% are shape persistent). While the abundance of shape persistent examples in the alkyne data set can be attributed to the rigid triple bond, the reason why cages in the "amine2 aldehyde3 - **Tri**^{4}**Di**^{6}" data set tend toward shape persistence is less clear. It can be seen that simply reversing the functional groups between building blocks and linkers, as given by the "aldehyde2 amine3 - **Tri**^{4}**Di**^{6}" data set, results in a prevalence of collapsed cages, despite the formation reaction and topology being identical. It is interesting that the "aldehyde2amine3 - **Tri**^{4}**Di**^{6}" data set is so successful at shape persistency, given that a large number of experimentally reported porous organic cages with shape persistent cavities use this topology.[2]

Figure 11 shows how cavity size varies between the data sets. An interesting observation is that imine **Tet**^{6}**Tri**^{8} cages have the largest cavities on average, despite having fewer shape

persistent molecules than the "alkyne2 alkyne3" data set. Cages formed by the amide condensation reaction have the smallest cavities on average, which follows from the fact that these cages also have a high proportion of collapsed individuals. In general, most data sets have a mean cavity size of approximately 6 Å, while the most common cavity size is 9 Å to 10 Å (Figure S4). In Figure 12, the change in window difference can be seen across data sets. The worst performing data sets in this area belong to the $\mathbf{Tri}^{8}\mathbf{Di}^{12}$ topologies, which can once again be attribute to the size and flexibility of cages with this topology. On average, cages in the "alkyne2 alkyne3" data set had the lowest window difference and therefore the highest symmetry, as would be expected from our previous discussions regarding the structural properties of this data set. In general, a comparison of window size difference against the cavity size shows that there is no correlation between the two properties.

We conclude our analysis with a general set of recommendations when searching for new shape persistent cages. Based on our analysis, alkyne cages appear to be the most likely to produce positive results, followed by imine cages and alkene cages. Thiol and amide cages on the other hand should be avoided. When targeting cages with large cavity sizes, imine cages perform best on average, followed by alkyne cages. While $\mathbf{Tri}^{8}\mathbf{Di}^{12}$ cages usually lack shape persistence, and as a result generally do not have large cavity sizes, the largest cages in our entire data set are all of this topology. This means it may be the most suitable topology to target when a single molecule with a record-breaking size is desired. For reliably large cavity sizes, the $\mathbf{Tet}^{6}\mathbf{Tri}^{8}$ topology is recommended.

Following our analysis, we wished to identify some of the most promising organic cages in our data set. The structures of these cages can be seen in Figure 13. We selected the candidates based on a number of factors, such as size, reaction chemistry, symmetry and structural diversity of precursors. Among these cages are some of the largest in our data set, ranging from 51 - 60 Å, showing the $\mathbf{Tri}^{8}\mathbf{Di}^{12}$ topology. We also show large cages (16 Å cavity) with high symmetry. Finally, Figure 13 shows three of the largest alkyne cages. While all of these cages are among the best from a purely property based perspective, this

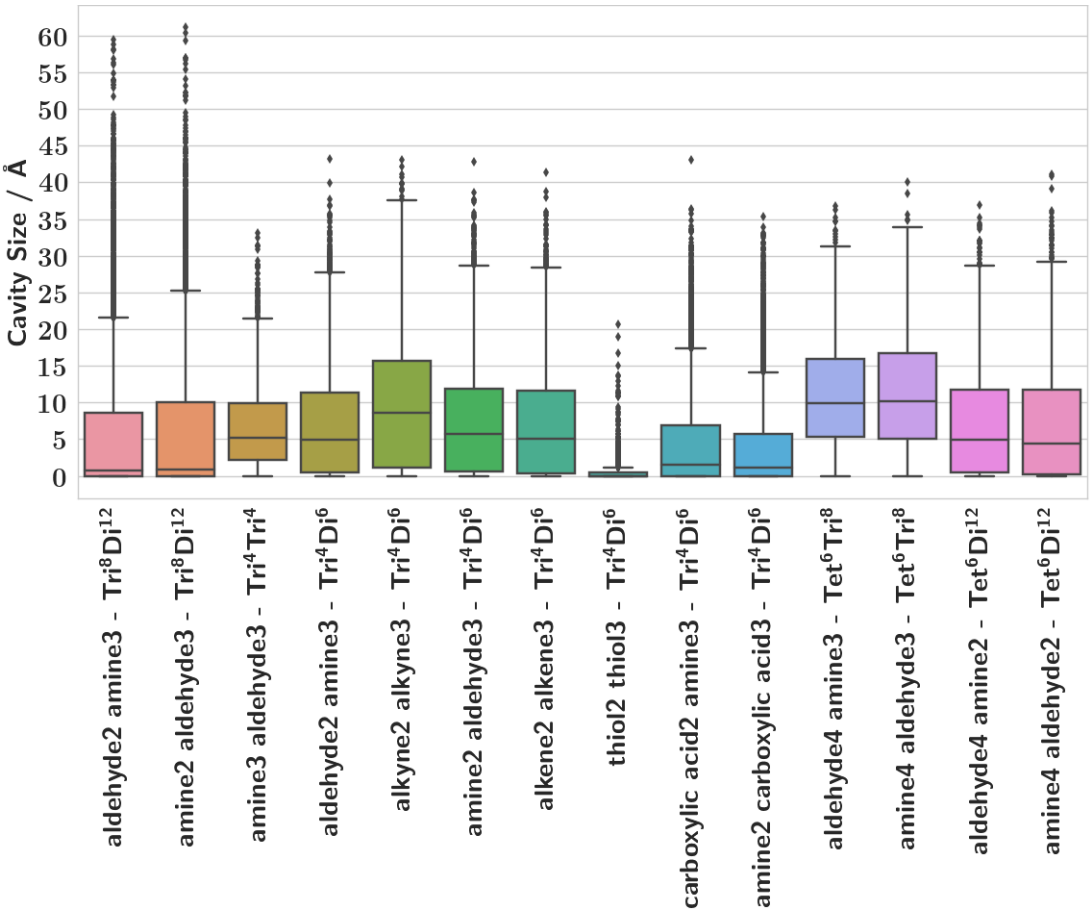

Figure 11: A box plot of cavity size of cages by precursor combination and topology. The colored box represents the values from the 25th to the 75th percentile. The line within the colored box represents the median of the data. The lines extending from the box represent data that lies within 1.5 inter-quartile ranges of the lower and upper quartile and the remaining points represent the left-over data points.

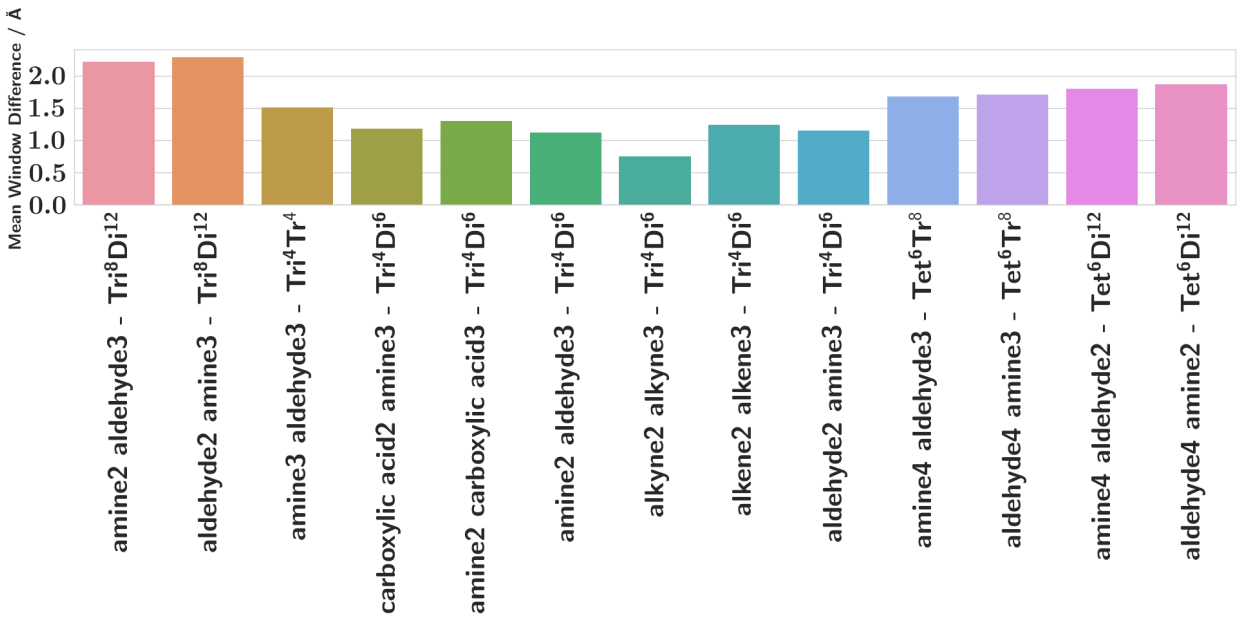

Figure 12: Mean window difference of cages by precursor combination and topology.

does not address the synthesis of these molecules, which may not be possible in some cases. The precursors for these cages are shown in Figure S7.

We have extracted the ten building blocks and linkers most commonly found in cages belonging to the "collapsed" and "shape persistent" classes. Their structures can be seen in Figures S8-12. Analysis of these precursor structures allows us to draw conclusions about molecular features conducive to forming shape-persistent cages. In turn, this analysis can be used to inform design of cage precursors in the future. We discuss the prevalence of four structural features between building blocks and linkers belonging to collapsed and shape persistent cages. In addition to this, the supporting information shows an additional nine structural features (Figures S12-S20). Figure 14A shows the number of rotatable bonds among the best and worst precursor molecules. The number of rotatable bonds in a molecule relates directly to its rigidity and the results in Figure 14A conform to our expectations. Good building blocks and linkers have, on average, significantly fewer rotatable bonds than bad ones.

One of the most apparent differences between building blocks and linkers commonly found in collapsed versus shape persistent cages is the presence of bridgehead atoms, as shown in

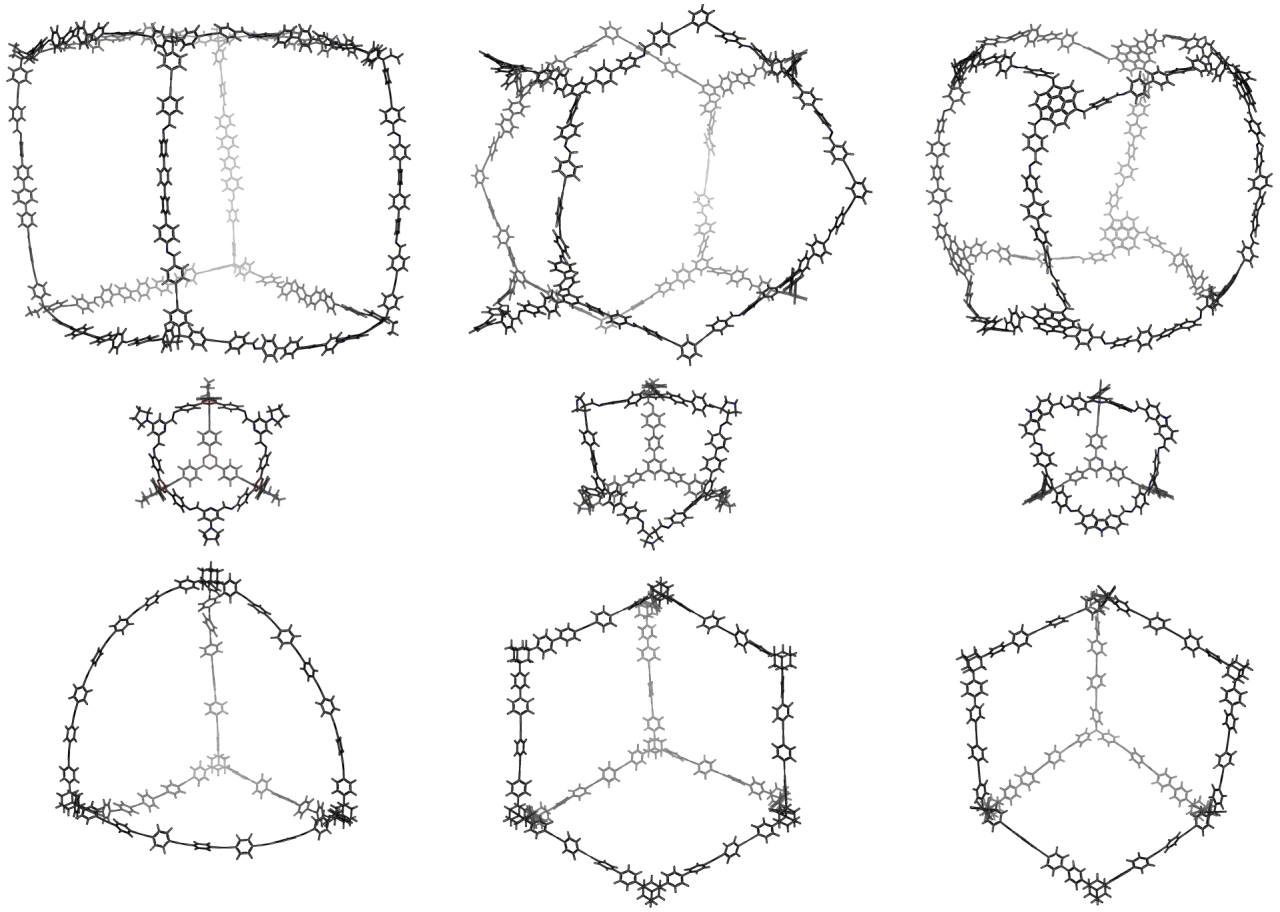

Figure 13: Structures of some the "best" cages found in our data set. Top row shows some of the largest cages, with cavities ranging from 51 - 60 Å. Middle row shows cages with a cavity size of 16 Å and high symmetry (low window difference). Bottom row shows alkyne metathesis generated cages with large (>40 Å cavities). Atoms colors are gray, blue, red, white and pink for carbon, nitrogen, oxygen, hydrogen and boron, respectively.

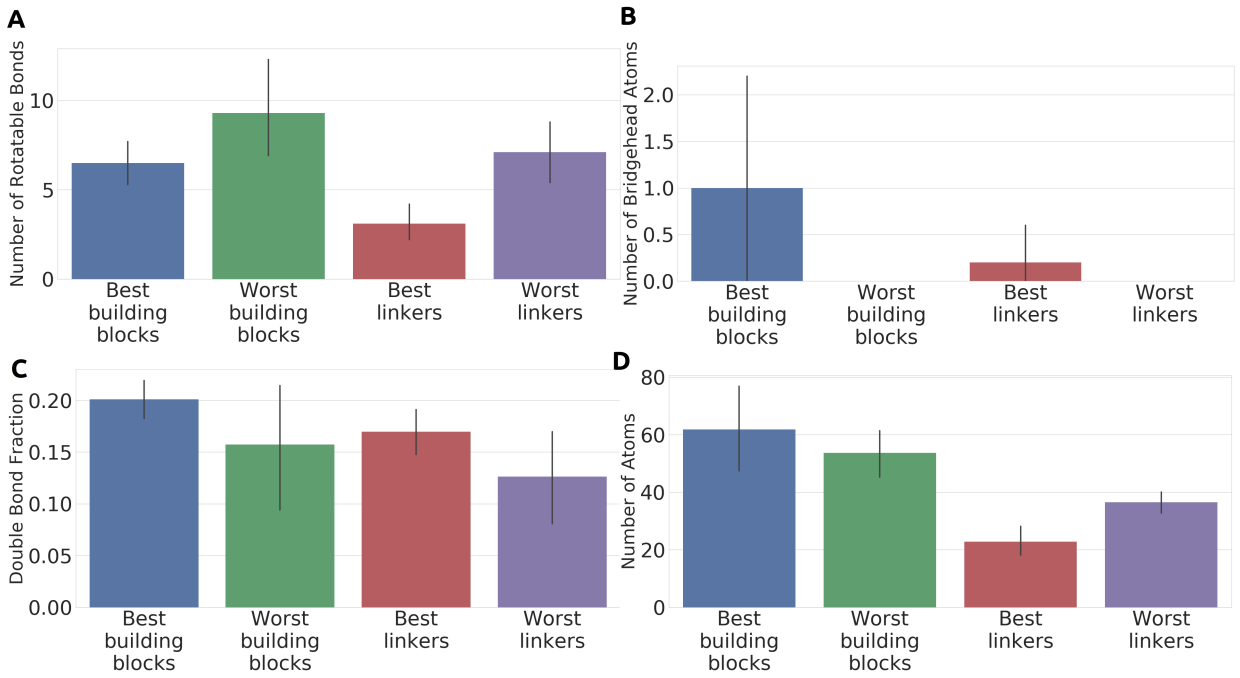

Figure 14: Structural features of 40 precursors, consisting of the 10 best and 10 worst cage building blocks and linkers from the total of 1054. Best building blocks and linkers are those found most frequently in shape persistent cages, whilst worst building blocks and linkers are those found most often in collapsed cages. The bars indicate mean values of the given properties, while the black lines show 95% confidence intervals. (A) Number of rotatable bonds; (B) number of bridgehead atoms; (C) double bond fraction and (D) number of atoms.

Figure 14B. On average, each of the 10 most successful building blocks had at least one bridgehead atom, while none were present in any of their poorly performing counterparts. We attribute this to the fact that bridgehead atoms greatly restrict the range of movement possible within the ring being bridged. This structural rigidity inhibits the motion necessary for cage collapse. Similarly, bridgeheads are found in two of the most successful linkers, while not being found at all in the worst performing ones. Compared to building blocks, bridgehead atoms may be rare in linkers because their smaller size confers sufficient rigidity already and other factors, such as the number of double bonds in the linker, may be more important. A graphical guide identifying bridehead atoms is shown in Figure S21.

Figure 14C shows that, on average, good linkers that lead to shape persistent cages have a larger proportion of double bonds compared to bad ones. Similarly to bridgehead atoms, these are key to conferring rigidity to a molecule. This result is also mirrored in building block molecules. Most noticeably in this figure however, is that the distribution of values is much narrower in good building blocks and linkers compared to the bad ones. While the double bond fraction present in good building blocks and linkers is relatively consistent, this is not the case for the "worst" molecules. This highlights the fact that while the presence of a double bond fraction of approximately 0.20 for building blocks and 0.17 for linkers does not guarantee a good cage precursor, large deviations from these values are likely to make it unsuitable for cage construction.

Finally, in Figure 14D, an examination of the number of atoms in a molecule, a proxy for its size, reveals a somewhat surprising result, good building blocks are on average larger than bad building blocks. This is unexpected, as larger molecules generally have more degrees of freedom and are therefore less rigid, leading to cage collapse. This trend is apparent when comparing the linkers in Figure 14D. Good linkers have significantly fewer atoms on average than bad ones. In the case of building blocks, specifically those found in our data set, larger sizes may be associated with other features such as rings and bridgeheads, which may increase the suitability of the building block toward constructing shape persistent cages,

despite leading to an increase in the number of atoms.

Online tool for using trained models

In order to allow easy access to our prediction models, we provide our online tool at `https://ismycageporous.ngrok.io`, which allows users to input the SMILES strings of a cage building block and linker and select one of the reaction specific models, described in Table 2. The tool will then return a prediction of either "collapsed" or "shape persistent" to the user. The models available on the site have been trained using the entire data set, including the test set. The source code for this tool is available at `github.com/lukasturcani/cage_prediction`.

Conclusions

We have presented a large data set of 63,472 organic cages, constructed through a number of reactions and in a variety of topologies. The cage molecules were all computationally assembled and their structure optimized. Analysis of the cages revealed that cages produced by alkyne metathesis tended to outperform other cages in terms of cavity size, shape persistence and symmetry. On the other hand, cages with a $\mathbf{Tri}^{8}\mathbf{Di}^{12}$ topology perform noticeably worse on these criteria. Despite this, the largest shape persistent cages have a $\mathbf{Tri}^{8}\mathbf{Di}^{12}$ topology. We find that cages formed by the disulfide formation reaction are least likely to be shape persistent and amide cages also perform poorly in most criteria. By analysis of the cage precursors we show that structural features such as bridgehead and double bonds are conducive to forming shape persistent cages, while increasing the total number of atoms and rotatable bonds encourages collapse. In addition, we show that for precursors which form shape persistent cages, these features tend to fall within a narrow range.

After categorization of the molecules as either collapsed or shape persistent, we trained random forest models to predict three properties of interest: shape persistence, cavity size,

and window size difference. A number of validation methods have been described, considering the transferability of models trained on cages generated by different chemical reactions. In all cases, we found our models require a training set composed of cages assembled through the same reaction chemistry as the test set. Analysis of the performance of our models across cages generated through different formation reactions indicates that structural features which lead to cage collapse differ with formation reaction.

The regression models, at this stage, suffer from a high error making them unsuitable for tasks requiring fine discrimination between individual cages, however, they may still find application in screening procedures requiring comparison only to some threshold, desirable value. On the other hand, the classification model used to identify collapsed cages provides good discrimination, as shown by high precision and recall scores. We anticipate this model will drastically reduce wasted computation spent on uninteresting, collapsed cages, allowing computational resources to be more usefully allocated to the modelling of their porous counterparts. We provide an online tool which provides predictions for user submitted cages. This can be accessed on `https://ismycageporous.ngrok.io` and the source code for our model is available at `github.com/lukasturcani/cage_prediction`.

Acknowledgement

We acknowledge a Royal Society University Research Fellowship (K.E.J.) and the EPSRC (EP/M017257/1, EP/N004884/1 and EP/R005710/1) and ERC through grant agreement number 758370 (ERC-StG-PE5-CoMMaD) and 321156 (RobOT) for funding. The authors thank Ioannis Karamanlakis for many helpful discussions and assistance with the interpretation of results and Enrico Berardo for helpful discussions.

Supporting Information Available

We provide as supporting information additional methodological details, figures and a database of SMILES strings of all precursors, with a script that can convert these to structures.

References

(1) Hasell, T.; Cooper, A. I. Porous organic cages: soluble, modular and molecular pores. *Nat. Rev. Mater.* **2016**, *1*, 16053.

(2) Beuerle, F.; Gole, B. Covalent Organic Frameworks and Cage Compounds: Design and Applications of Polymeric and Discrete Organic Scaffolds. *Angew. Chem., Int. Ed. 57*, 4850–4878.

(3) Yoshizawa, M.; Yoshizawa, M.; Klosterman, J. K.; Klosterman, J. K.; Fujita, M.; Fujita, M.; Fujita, M. Functional Molecular Flasks: New Properties and Reactions within Discrete, Self-Assembled Hosts. *Angew. Chem., Int. Ed.* **2009**, *48*, 3418–3438.

(4) Kewley, A.; Stephenson, A.; Chen, L.; Briggs, M. E.; Hasell, T.; Cooper, A. Porous Organic Cages for Gas Chromatography Separations. *Chem. Mater.* **2015**, *27*, 3207–3210.

(5) Mitra, T.; Jelfs, K. E.; Schmidtmann, M.; Ahmed, A.; Chong, S. Y.; Adams, D. J.; Cooper, A. Molecular shape sorting using molecular organic cages. *Nat. Chem.* **2013**, *5*, 276–281.

(6) Hasell, T.; Miklitz, M.; Stephenson, A.; Little, M. A.; Chong, S. Y.; Clowes, R.; Chen, L.; Holden, D.; Tribello, G. A.; Jelfs, K. E.; Cooper, A. I. Porous Organic Cages for Sulfur Hexafluoride Separation. *J. Am. Chem. Soc.* **2016**, *138*, 1653–1659, PMID: 26757885.

(7) Chen, L. et al. Separation of rare gases and chiral molecules by selective binding in porous organic cages. *Nat. Mat.* **2014**, *13*, 954–960, Article.

(8) Lee, T.-C.; Kalenius, E.; Lazar, A. I.; Assaf, K. I.; Kuhnert, N.; Grün, C. H.; Jänis, J.; Scherman, O. A.; Nau, W. M. Chemistry inside molecular containers in thegas phase. *Nat. Chem.* **2013**, *5*, 1–7.

(9) Song, Q.; Jiang, S.; Hasell, T.; Liu, M.; Sun, S.; Cheetham, A. K.; Sivaniah, E.; Cooper, A. I. Porous Organic Cage Thin Films and Molecular-Sieving Membranes. *Adv. Mater. 28*, 2629–2637.

(10) Jones, J. T. A.; Holden, D.; Mitra, T.; Hasell, T.; Adams, D. J.; Jelfs, K. E.; Trewin, A.; Willock, D. J.; Day, G. M.; Bacsa, J.; Steiner, A.; Cooper, A. I. OnâĂŞOff Porosity Switching in a Molecular Organic Solid. *Angew. Chem., Int. Ed. 50*, 749–753.

(11) Jelfs, K. E.; Wu, X.; Schmidtmann, M.; Jones, J. T. A.; Warren, J. E.; Adams, D. J.; Cooper, A. I. Large Self-Assembled Chiral Organic Cages: Synthesis, Structure, and Shape Persistence. *Angew. Chem., Int. Ed. 50*, 10653–10656.

(12) Greenaway, R. L.; Santolini, V.; Bennison, M. J.; Alston, B. M.; Pugh, C. J.; Little, M. A.; Miklitz, M.; Eden-Rump, E. G. B.; Clowes, R.; Shakil, A.; Cuthbertson, H. J.; Armstrong, H.; Briggs, M. E.; Jelfs, K. E.; Cooper, A. I. High-throughput discovery of organic cages and catenanes using computational screening fused with robotic synthesis. *Nat. Commun.* **2018**, *9*, 2849.

(13) Zhang, G.; Presly, O.; White, F.; Oppel, I. M.; Mastalerz, M. A Permanent Mesoporous Organic Cage with an Exceptionally High Surface Area. *Angew. Chem., Int. Ed. 53*, 1516–1520.

(14) Evans, J. D.; Huang, D. M.; Haranczyk, M.; Thornton, A. W.; Sumby, C. J.; Doonan, C. J. Computational identification of organic porous molecular crystals. *CrystEngComm* **2016**, *18*, 4133–4141.

(15) Silver, D. et al. Mastering the game of Go without human knowledge. *Nature* **2017**, *550*, 354–359, Article.

(16) Kim, E.; Huang, K.; Saunders, A.; McCallum, A.; Ceder, G.; Olivetti, E. Materials Synthesis Insights from Scientific Literature via Text Extraction and Machine Learning. *Chem. Mater.* **2017**, *29*, 9436–9444.

(17) Segler, M. H. S.; Preuss, M.; Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **2018**, *555*, 604–610, Article.

(18) Wei, J. N.; Duvenaud, D.; Aspuru-Guzik, A. Neural Networks for the Prediction of Organic Chemistry Reactions. *ACS Cent. Sci.* **2016**, *2*, 725–732, PMID: 27800555.

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

(20) Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine learning for molecular and materials science. *Nature* **2018**, *559*, 547–555.

(21) Hachmann, J.; Olivares-Amaya, R.; Atahan-Evrenk, S.; Amador-Bedolla, C.; Sánchez-Carrera, R. S.; Gold-Parker, A.; Vogt, L.; Brockway, A. M.; Aspuru-Guzik, A. The Harvard Clean Energy Project: Large-Scale Computational Screening and Design of Organic Photovoltaics on the World Community Grid. *J. Phys. Chem. Lett.* **2011**, *2*, 2241–2251.

(22) NOMAD. `https://nomad-coe.eu/`, (accessed August 10, 2018).

(23) Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; Persson, K. A. Commentary: The Materials Project:

A materials genome approach to accelerating materials innovation. *APL Mater.* **2013**, *1*, 011002.

(24) Pereira, F.; Xiao, K.; Latino, D. A. R. S.; Wu, C.; Zhang, Q.; Aires-de Sousa, J. Machine Learning Methods to Predict Density Functional Theory B3LYP Energies of HOMO and LUMO Orbitals. *J. Chem. Inf. Model* **2017**, *57*, 11–21, PMID: 28033004.

(25) He, Y.; Cubuk, E. D.; Allendorf, M. D.; Reed, E. J. Metallic MetalâĂŞOrganic Frameworks Predicted by the Combination of Machine Learning Methods and Ab Initio Calculations. *J. Phys. Chem. Lett.* **2018**, *9*, 4562–4569, PMID: 30052453.

(26) Evans, J. D.; Coudert, F.-X. Predicting the Mechanical Properties of Zeolite Frameworks by Machine Learning. *Chemistry of Materials* **2017**, *29*, 7833–7839.

(27) Luechtefeld, T.; Marsh, D.; Rowlands, C.; Hartung, T. Machine Learning of Toxicological Big Data Enables Read-Across Structure Activity Relationships (RASAR) Outperforming Animal Test Reproducibility. *Toxicol. Sci.* **2018**, kfy152.

(28) Bleiziffer, P.; Schaller, K.; Riniker, S. Machine Learning of Partial Charges Derived from High-Quality Quantum-Mechanical Calculations. *J. Chem. Inf. Model* **2018**, *58*, 579–590, PMID: 29461814.

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

(30) Riniker, S.; Landrum, G. A. Better Informed Distance Geometry: Using What We Know To Improve Conformation Generation. *J. Chem. Inf. Model* **2015**, *55*, 2562–2574, PMID: 26575315.

(31) Rappe, A. K.; Casewit, C. J.; Colwell, K. S.; Goddard, W. A.; Skiff, W. M. UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *J. Am. Chem. Soc* **1992**, *114*, 10024–10035.

(32) Turcani, L.; Berardo, E.; Jelfs, K. E. STK: A Python Toolkit for Supramolecular Assembly. *J. Comp. Chem* **2018**, DOI: 10.1002/jcc.25377.

(33) Schrödinger, L. N. Y. Schrödinger Release 2018-1: MacroModel. 2018.

(34) Harder, E. et al. OPLS3: A Force Field Providing Broad Coverage of Drug-like Small Molecules and Proteins. *J. Chem. Theory Comput.* **2016**, *12*, 281–296, PMID: 26584231.

(35) Santolini, V.; Tribello, G. A.; Jelfs, K. E. Predicting solvent effects on the structure of porous organic molecules. *Chemical Communications* **2015**, *51*, 15542–15545.

(36) Stackhouse, C.; Santolini, V.; Greenaway, R.; Little, M. A.; Briggs, M. E.; Jelfs, K. E.; Cooper, A. Cage Doubling: Solvent-Mediated Re-equilibration of a [3+6] Prismatic Organic Cage to a Large [6+12] Truncated Tetrahedron. *Crystal Growth & Design* **2018**, acs.cgd.7b01422–17.

(37) Greenaway, R. L.; Santolini, V.; Bennison, M. J.; Alston, B. M.; Pugh, C. J.; Little, M. A.; Miklitz, M.; Eden-Rump, E. G. B.; Clowes, R.; Shakil, A.; Cuthbertson, H. J.; Armstrong, H.; Briggs, M. E.; Jelfs, K. E.; Cooper, A. High-throughput discovery of organic cages and catenanes using computational screening fused with robotic synthesis. *Nature Communications* **2018**, *9*, 1–11.

(38) Berardo, E.; Turcani, L.; Miklitz, M.; Jelfs, K. E. An Evolutionary Algorithm for the Discovery of Porous Organic Cages. *Chemical Science* **2018**,

(39) Miklitz, M.; Jiang, S.; Clowes, R.; Briggs, M. E.; Cooper, A.; Jelfs, K. E. Computational Screening of Porous Organic Molecules for Xenon/Krypton Separation. *The Journal of Physical Chemistry C* **2017**, *121*, 15211–15222.

(40) Miklitz, M.; Jelfs, K. E. pywindow: Automated Structural Analysis of Molecular Pores. *ChemRxiv* **2018**, DOI: 10.26434/chemrxiv.6850109.v1.

(41) Landrum, G. A. RDKit: Open-source cheminformatics. `http://www.rdkit.org`, (accessed August 10, 2018).

(42) Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

(43) Tozawa, T. et al. Porous organic cages. *Nat. Mater.* **2009**, *8*, 973–978, Article.

Graphical TOC Entry