Materials Precursor Score: Modelling Chemists' Intuition for the Synthetic Accessibility of Porous Organic Cages

Steven Bennett,[†] Filip T. Szczypiński,[†] Lukas Turcani,[†] Michael E. Briggs,[‡] Rebecca L. Greenaway,[†] and Kim E. Jelfs[*,†]

[†]*Department of Chemistry, Imperial College London, Molecular Sciences Research Hub, White City Campus, Wood Lane, London, W12 0BZ, UK*

[‡]*Materials Innovation Factory, University of Liverpool 51 Oxford Street, Liverpool, L7 3NY, UK*

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

Abstract

Computation is increasingly being used to try to accelerate the discovery of new materials. One specific example of this is porous molecular materials, specifically porous organic cages, where the porosity of the materials predominantly comes from the internal cavities of the molecules themselves. The computational discovery of novel structures with useful properties is currently hindered by the difficulty in transitioning from a computational prediction to synthetic realisation. Attempts at experimental validation are often time-consuming, expensive and, frequently, the key bottleneck of material discovery. In this work, we developed a computational screening workflow for porous molecules that includes consideration of the synthetic difficulty of material precursors, aimed at easing the transition between computational prediction and experimental realisation. We trained a machine learning model by first collecting data

on 12,553 molecules categorised either as 'easy-to-synthesise' or 'difficult-to-synthesise' by expert chemists with years of experience in organic synthesis. We used an approach to address the class imbalance present in our dataset, producing a binary classifier able to categorise easy-to-synthesise molecules with few false positives. We then used our model during computational screening for porous organic molecules to bias towards precursors whose easier synthesis requirements would make them promising candidates for experimental realisation and material development. We found that even by limiting precursors to those that are easier-to-synthesise, we are still able to identify cages with favourable, and even some rare, properties.

Introduction

Functional materials underpin the foundations of modern society, but their discovery is a long and challenging process. High-throughput computational screening seeks to guide the process and thus to accelerate novel material discovery.[1–3] A key component in computational materials discovery needs to be a consideration of whether a hypothetical material with promising properties can actually be experimentally obtained.[4] There are many elements to that challenge, including the ability to obtain or synthesise the precursors, finding a successful synthetic method to form the material, and being able to control the solid-state form and assembly, for example, to enable incorporation into a device. For materials built from entirely or primarily organic components, the consideration of whether the precursor building blocks of the material can be easily, and ideally cheaply, synthesised is vital. Without expert chemists guiding some element of the computational screening process, it is challenging to foresee which new candidate materials are synthetically viable and accessible prior to experimental design.

The growth of data-driven tools in chemistry have allowed chemists to ease the transition from computational prediction to experimental realisation of targeted molecules.[5,6] In the pharmaceutical industry, there has been a growing interest in the computational prediction

of synthetic difficulty and the automated prediction of retrosynthetic pathways for organic molecules.[6] An assessment of synthetic difficulty allows medicinal chemists to more efficiently allocate their time and resources, prioritising molecules with greater drug-likenesses and lower synthetic difficulties. The approaches used to calculate synthetic difficulty can be broadly categorised as follows: calculating the structural complexity of the molecule;[7,8] using retrosynthetic analysis to elucidate viable synthetic pathways to a molecule;[9] modelling the intuition of expert chemists;[10] and finally, machine learning (ML) models trained on extensive reaction databases or datasets on easy- and difficult-to-synthesise molecules.[11,12] Existing approaches to calculate the synthetic difficulty of organic molecules include the synthetic accessibility score (SAScore), the synthetic complexity score (SCScore) and the SYnthetic Bayesian Accessibility (SYBA) score.[8,11,12] The SAScore uses a combination of structural complexity and fragment contributions to calculate synthetic difficulty.[8] A similar approach is employed by SYBA, which also uses the frequency fragments appear in public compound datasets to estimate synthetic difficulty.[12] The synthetic complexity score (SCScore) meanwhile uses a neural network to predict the number of reaction steps required to synthesise a molecule, defining synthetically complex molecules as those that require a greater number of reaction steps to synthesise.[11] Each of these approaches is subject to individual limitations, making creation of a useful heuristic for synthetic difficulty challenging. However, both scores are able to provide a continuous score extremely quickly, which can be used to bias towards easier-to-synthesise molecules.[13]

Structural complexity does not imply that a molecule is easy to synthesise, as complexity can be introduced into a molecule with simple reaction transformations. Identifying retrosynthetic routes is often computationally intensive,[9] however, recent open-source developments in computer-aided synthesis planning have been shown to produce viable synthesis pathways and be good estimators of synthetic difficulty.[13,14] Meanwhile, data-driven approaches can struggle to generalise to molecules outside of their training set,[11] and can contain inherent bias due to the lack of failed reactions. Obtaining scores derived from the intuition of ex-

pert chemists is often labour intensive and can be subject to human bias,[15] however, this approach can create a useful model if large amounts of unbiased training data are obtained. Moreover, many of these predictive methods are developed with the primary objective of obtaining synthesisable drug candidates, whose synthesis requirements may not perfectly align to those of a materials discovery programme.

In this work, we focus on predicting the synthetic accessibility of precursors for porous molecular materials, whose properties depend not only on the solid-state packing, but also on the structure of their discrete building blocks. The molecular nature of porous molecular materials means that they are typically soluble in common solvents, inferring advantages particularly in their solution synthesis and processability, for instance allowing relatively easy formation into membranes.[16] Porous materials have been investigated for applications such as gas storage,[17,18] separations,[19,20] catalysis,[21] and sensing.[22,23] Porous organic cages (POCs), an example of which can be seen in Figure 1, are a class of porous molecular materials where the porosity predominantly arises from the internal cavity of the molecule (known as 'intrinsic' porosity) and can form a porous material in the solid-state by directing the packing to create interconnected pore networks, sometimes combined with the 'extrinsic' porosity that arises between the molecules due to inefficient packing.

POCs pack together in the solid-state through intermolecular interactions, and do not have the extended chemical bonding found in porous network materials such as metal-organic frameworks (MOFs) and zeolites. Furthermore, the intrinsic porosity of POCs means they have solution-based applications and can give rise to porous liquids.[24,25] The number of previously reported POCs is in the low hundreds, this is comparatively few compared to the number of discovered MOFs, numbering 70,000 in the Cambridge Structural Database.[26,27] Recently, we reported an approach to explore the vast chemical space of POCs using an evolutionary algorithm.[28,29] Using this approach, we were able to computationally identify new cages with less common properties, including a large internal cavity diameter of 16 Å. However, the vast majority of precursors are unlikely to form a shape-persistent POC with a

permanent internal cavity, making screening all potential precursors extremely inefficient.[30] Many POCs are thus discovered through serendipity or are designed using the intuition of an expert chemist in the POC field.

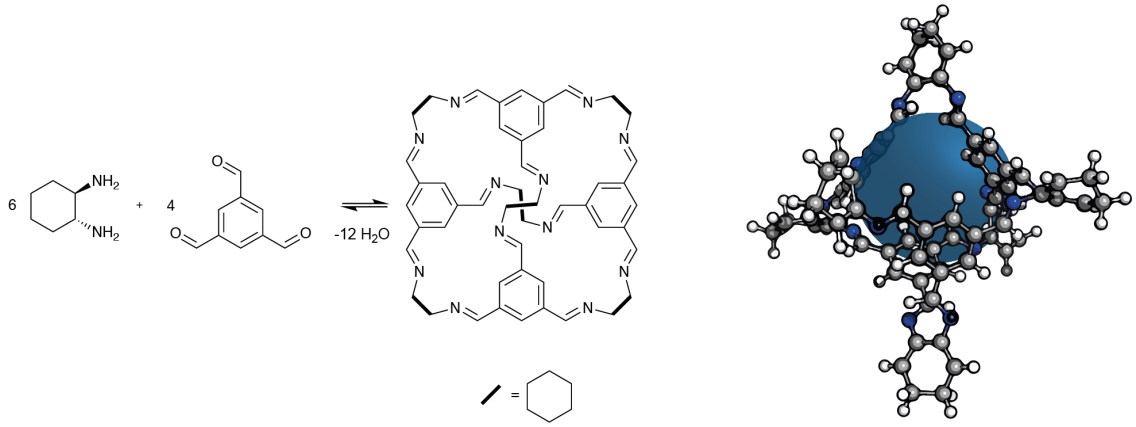

Figure 1: An imine condensation of 1,3,5-triformylbenzene (4 equivalents) with trans-1,2-diaminocyclohexane (6 equivalents) to form covalent cage 3 (**CC3**), a prototypical porous organic cage (POC) with a permanent internal cavity (coloured blue).[31] Carbon atoms are shown in grey, nitrogen atoms in blue, and hydrogen atoms in white

POCs are typically synthesised via dynamic covalent chemistry (DCC), where the reversible reaction allows for error-correction and affords the opportunity to reach either the thermodynamic product, or a desired kinetic product,[32] rather than an oligomer, or polymeric adduct. The computational prediction of a synthetically accessible POC with a desired set of properties is accompanied by a series of challenges. The majority of POCs reported thus far were synthesised through imine bond formation. The precursor components of the POC, for example an amine and an aldehyde for an imine-based molecule, must themselves be synthesised and they must subsequently react in a predictable manner to form the desired POC of the targeted molecular mass and topology. For example, imine condensation between a trialdehyde and a diamine (arguably the most common precursor pair in reported POCs) can result in cages of six different topologies, in addition to unpredictable mixtures of polymers.[33] The challenge of synthesising a POC can extend beyond precursor synthesis; a combination of insoluble products can result in inseparable mixtures, decreased reaction yields or driving the equilibrium away from the desired thermodynamic minimum.[34,35]

The most common topology observed in POCs, denoted as **Tri^{4}Di^6** using the terminology of Santolini *et al.*,[33] results from the condensation of four tritopic molecules with six ditopic molecules into a single cage unit, as shown for the formation of **CC3** in Figure 1. As POCs formed through imine condensation are often the thermodynamically most stable product, we have previously used formation energies to predict the likely topological outcome of a reaction between a precursor pair,[33] and the likely sorting outcome within a given topology if a mixture of precursors is used.[36] The required POC model construction and energy calculations can be automated, for example by our own open-source supramolecular toolkit software (*stk*),[37] which we have previously exploited to computationally aid the experimental discovery of 33 cleanly-formed POCs using a robotic platform.[26] We have in multiple cases ourselves used computation to accelerate the discovery of POCs,[38,39] but time and again, the synthesis of the POC is the most time consuming component in their development, taking months to years compared to weeks for the computational screening.[40] Worse, the synthesis is often not successful at all.

Here, we investigate the best approach to consider the synthetic difficulty of an organic material's precursors in a computational screening workflow, with a focus on POCs. The long term goal is to increase the success rate of experimental materials discovery programmes in relation to the synthetic realisation of computational targets. We develop our own synthetic difficulty prediction model, the Material Precursor Score (MPScore), and compare how this performs relative to the previously reported SAScore and SCScore. Our model reformulates synthetic difficulty prediction as a classification problem modelled using a random forest to answer the following question: *can you make 1 g of this compound in under 5 steps?* In the end, we demonstrate the applicability of our classifier for chemists' intuition in a context that would normally require significant database reduction to human-tractable size. We demonstrate the model's ability to bias against precursors that chemists themselves would avoid in materials synthesis, allowing us to focus our computational resources on POCs with a greater probability of being synthetically realised. We show that even when limiting the

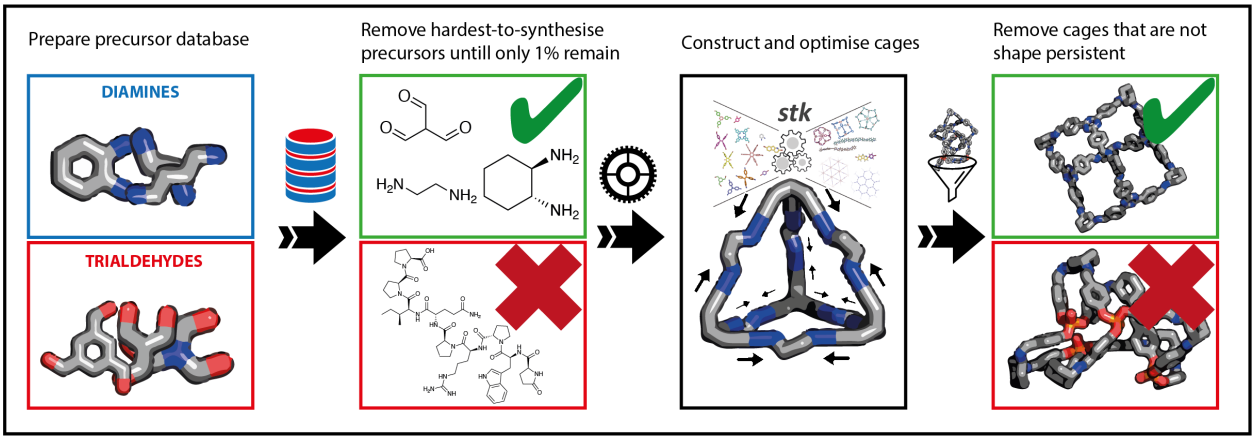

Figure 2: High-throughput computational screening workflow for the discovery of synthetically viable shape-persistent porous organic cages (POCs)

precursors to the easiest-to-synthesise, we are able to identify shape-persistent POCs with unconventional properties, such as large pore diameters.

Methods

Overall, our ambition is to include a synthetic difficulty consideration of organic precursors within our workflow for computational screening to identify functional POCs. We test two existing approaches (SCScore and SAScore) and compare the results to a new synthetic difficulty model we develop here. We test the three approaches in a computational workflow aimed at identifying *synthesisable* functional POCs that are shape-persistent. Shape persistency is a property that POCs must exhibit to achieve porosity, where they remain rigid and have a permanent cavity even in the absence of scaffolding solvent. We calculate shape persistency as part of our workflow here as it is relatively computationally cheap to assess, allowing for easy comparison between different computational screening workflows. We have chosen to target POCs with the $\mathbf{Tri^4Di^6}$ topology, which is the most common topology in previously reported POCs, formed by imine condensations between diamines and trialdehydes. Our computational workflow, which will be described in full detail below, is depicted in Figure 2. The workflow involves screening a precursor database to remove molecules pre-

dicted to have the largest synthetic difficulty, followed by automated cage construction and POC structure prediction, and finally characterisation of the POC pores to identify those with permanent internal cavities. In the below subsections, we first discuss the creation of a labelled training database used to train our synthetic difficulty model. We then evaluate our model using cross-validation, and finally demonstrate the model's utility, selectively screening for easy-to-synthesise precursors as part of a POC screening workflow.

Creating the Synthetic Difficulty Model for Organic Material Precursors

Training Database Construction

First, to train a ML model to classify potential POC precursors as either easy-to-synthesise or difficult-to-synthesise, we needed to generate a diverse training dataset with an approximately equal number in each group. Initially, molecules were extracted from a subset of the proprietary Reaxys[41] and eMolecules[42] databases. This initial subset consisted of molecules with functional groups that are likely to undergo common DCC reactions, frequently used in materials synthesis. This set included di-, tri-, tetra-, penta- and hexa-topic amines and aldehydes, which was extended with molecules that have been used previously in POC synthesis by our experimental collaborators, known to be easy-to-synthesise. Additionally, we included molecules that our experimental collaborators have previously avoided due to their challenging syntheses. To this set of molecules, functional group substitutions were performed to extend the size and diversity of our training set, resulting in a set of molecules with a fairly even distribution of different functional groups. For each molecule in the initial starting set, a SMARTS substitution was performed, exchanging all functional groups in the molecule with those from a pre-defined list, resulting in a dataset consisting of 14,859 molecules. The final database, in addition to the script used to perform functional group substitutions is available on Github.[43] A full breakdown of molecules by their respective

functional groups can be found in Table S1.

The molecules were then assessed by three experimental chemists with at least PhD-level training in synthetic chemistry. Existing measures of synthetic difficulty of organic compounds originate from the drug discovery field and might not capture the scale and simplicity required for materials synthesis. Therefore, we asked experienced synthetic chemists to label molecules relevant to POC synthesis as 'easy-to-synthesise' based on the question *"can you make 1 g of this compound in under 5 steps?"*. Instead of producing a continuous measure of synthetic difficulty, as is the case in both the SAScore and SCScore, we aimed to create a discrete binary classifier, with the goal of identifying easy-to-synthesise precursors for POC synthesis. A binary classification approach was chosen to reduce the challenge of collecting training data, and it is much easier to obtain a consensus on a binary classification.

Figure 3 shows the graphical user interface developed to collect the training labels. Three of our authors labelled the molecules, these are: author RLG, with 12 years of research experience in organic synthesis and 8 years experience synthesising POCs; author FS with more than 4 years research experience in organic chemistry and author MEB with 20 years organic chemistry experience and 8 years experience synthesising POCs. Those labelling the molecules were presented with a two-dimensional representation of a molecule they had not previously scored, and tasked with answering 'yes' or 'no' to the previously mentioned criterion. Each molecule was presented randomly, with equal probability of selection, to reduce the likelihood of systematic scoring occurring if the chemist's opinion was influenced by the preceding molecule. The 'unsure' option was added to avoid appending anything to the database, instead skipping to the next molecule.

Random Forest Model

We aimed to replicate the decision-making process that experimental chemists themselves would use when selecting precursors for material synthesis in our synthetic difficulty model. We chose a random forest (RF) classifier to model the data due to its practical utility, such

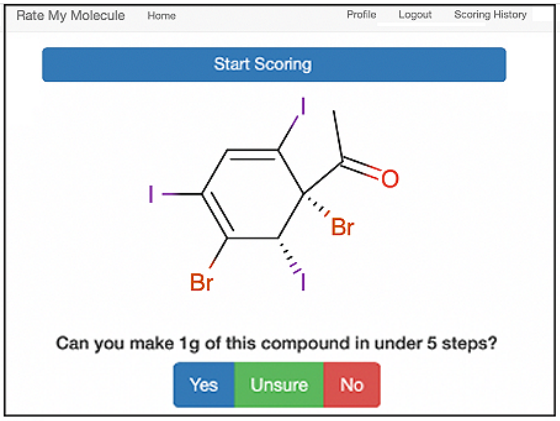

Figure 3: The interface for labelling of molecules by experimental chemists.

as the fact that RF models are quick to train and produce good performance on small to medium datasets. RF models are frequently used in chemistry problems to develop quantitative structure-property relationships for both classification and regression problems, and the models can offer some interpretability.[44,45] We used the `RandomForestClassifier` Python class, as implemented in `scikit-learn` version 0.22.1,[46] to construct the model. We chose "balanced" as the "class-weight" hyperparameter, which reduces the effect of class imbalance by weighing each data point inversely to the class frequency, increasing the importance of classes that appear fewer times in the training dataset. Default values were chosen for all other hyperparameters and 100 decision trees were used in the ensemble.

The extended-connectivity fingerprint (ECFP) was chosen as the vector representation of each molecule for the input vector, as implemented in `RDKit` version 2019.09.1.[47,48] A bit-size of 1024 and radius of 2 was chosen for the final model. ECFP was chosen as it has previously been shown to be one of the best performing fingerprints when similarity searching molecules to identify those with similar bioactive properties.[49] A count-based fingerprint was implemented to encode the number of times a feature appears in the molecule. This count-based approach is thought to provide greater information within the vector encoding of the molecule, and shows improved performance when predicting bioactivity compared to their bit-encoded counterparts.[50]

Cross-validation of the Model

We used five-fold cross-validation to assess the performance and generalisability of the RF model trained on the entire training set. During cross validation, an individual RF model was trained and evaluated using each fold. In five-fold sampling, the dataset is split into five different folds randomly, and the fold that is used as the test set is changed each time. This evaluation procedure allows us to assess how well the model can generalise to unseen data, allowing all the labelled samples to be used as test data at least once. We calculated the average accuracy, precision, recall, and F1 scores across all folds to assess the performance of the models (as discussed later in the results). These scores are defined as:

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Dataset}} \tag{1}$$

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{2}$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{3}$$

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

Following cross-validation, we trained the model on the entire dataset, using the scores achieved during cross-validation as an estimate of the performance of the entire model. We named this final model the Material Precursor score (MPScore), in line with its aim of selecting easy-to-synthesise precursors for materials development. An ideal model would aim to maximise both precision and recall, achieving few false positives and negatives, however, for our intended purpose, we focused on maximising the precision score, as this minimises the number of false positives obtained, such that very few difficult-to-synthesise molecules are incorrectly classed as easy-to-synthesise. This makes sense if we consider the large amount of experimental time that needs to be invested to try to synthesise even a few of the precursors and subsequently investigate their use in cage synthesis. We also calculated

the F1 score, which provides a combined measure of precision and recall in a single metric of model performance.

High-Throughput Virtual Screening

Next, we compared how the three automated synthetic difficulty scores (SAScore, SCScore and our MPScore) perform as filters for easy-to-synthesise precursors in a POC virtual screening workflow. By selecting precursors with the lowest 1% of synthetic difficulty scores using each of the three scoring methods, we hoped to identify shape-persistent POCs that could be readily accessed from the easiest-to-synthesise precursors.

Preparation of the Precursor Database

We wanted a database of potential precursors for POCs in the **Tri^{4}Di^6** topology, which are formed through a [4+6] imine condensation reaction between a tritopic aldehyde and a ditopic primary amine. This single topology was chosen as we previously showed that a significant number of precursors will preferentially form **Tri^{4}Di^6** over the other common **Tri^{8}Di^{12}** topology,[28] and because **Tri^{4}Di^6** molecules are more likely to be shape-persistent (38% of 6,018 cages we investigated in a previous study).[30] To create our database of POC precursor molecules, we used the eMolecules[42] and Reaxys[41] databases, selecting only molecules that contained exactly exactly three aldehydes or two primary aliphatic amines. All functional groups within molecules were identified using an automated detection algorithm, implemented using `RDKit`.[51] The final precursor database comprised 7,190 ditopic amines and 98 tritopic aldehydes, resulting in a possible 704,620 POC combinations that can be formed when restricted to the **Tri^{4}Di^6** topology.

Identification of Synthesisable Precursors

For each combination of diamine and trialdehyde precursors, we calculated the sum of their synthetic difficulty scores with each of the three scoring methods. These sums were then

scaled to values between 0 and 1 to allow for comparison between the models. For the MP-Score, we interpreted the probability that a molecule belonged to the 'difficult-to-synthesise' class as a continuous score, such that a higher value indicates the molecule is more challenging to synthesise. For each synthetic difficulty model, precursors within the first percentile of synthetic difficulty values (in total 21,133 precursor pairs, assuming no duplicates) were investigated for shape-persistence. Duplicate precursor combinations were identified by concatenating the SMILES strings of the diamine and trialdehyde precursors together and finding the overlap between precursor combinations selected by each score. In addition to easy-to-synthesise precursors filtered using synthetic difficulty scores, we selected a further 1% of precursor combinations randomly as a control sample, to investigate whether the three synthetic difficulty scores also bias towards precursors likely to form a shape-persistent cage.

Cage Construction and Conformational Search

The cages were constructed from precursors in an automated approach that utilises our supramolecular toolkit (`stk` - version 18.12.2019).[37] `stk` is a Python library for constructing and optimising complex supramolecular species, by providing precursor molecules and a predefined molecular topology. Structures built in `stk` then underwent a three-step procedure to identify plausible geometries for the lowest energy conformation of each cage. Each step of the process employed the OPLS3 force field,[52] used within Schrödinger's MacroModel software,[53] which has been shown to be able to accurately reproduce geometries of flexible imine cages.[33] Firstly, only bonds created during the `stk` build process were optimised, fixing the geometries of all other atoms. Subsequently, a molecular dynamics (MD) simulation was performed in the NVE ensemble for 2 ns after a 100 ps equilibration, with a time step of 1 fs and a temperature of 700 K. Structures were sampled every 40 ps along the MD trajectory and each of the resulting 50 sampled structures underwent a further geometry optimisation. All geometry optimisations employed the Polak–Ribière Conjugate Gradient algorithm, using a gradient conversion criteria of 0.05 kJ $Å^{-1}$ mol^{-1}. The resulting lowest-energy conformation

was used for further analysis.

Identification of Shape Persistent Cages

Following the optimisation procedure, organic cages that did not remain shape-persistent were removed. `pyWindow` was used to detect and analyse all the windows in the cages. `pyWindow` is a Python package for the analysis of structural properties of molecular pores, shown to be able to accurately reproduce pore sizes of POCs with a $\mathbf{Tri^4Di^6}$ topology.[54] For the POCs in which the expected number of four windows was identified, we calculated a parameter α

$$\alpha = \frac{4 \times \text{ Average Difference in Window Diameter}}{\text{Maximum Window Diameter } \times \text{ Expected Number of Windows}} \tag{5}$$

to classify shape-persistent cages. We developed this equation in our previous work, with the aim of maximising the number of organic cages labelled as shape-persistent using an automated approach.[30] The average window difference in equation 5 is the average difference in window diameter for all possible pairs of windows. If α is less than 0.035 and cavity size is greater than 1 Å, the cage was classified as shape-persistent. Otherwise, it was assigned as undetermined and disregarded. For organic cages that were deemed shape-persistent by this analysis, the central cavity diameter was then calculated using `pyWindow`.

Results and Discussion

Material Precursor Score (MPScore)

Evaluating Chemists Scores

To train our MPScore, we first constructed a diverse database of 14,859 molecules to provide to experimental chemists for labelling, as was discussed in the previous section. Ideally, we would have a very large number of chemists rank all molecules in our database to achieve an

overall consensus. However, the labelling of these molecules is extremely time-consuming, taking approximately 1 hour to assess 180 molecules, and thus it was only possible to obtain 12,553 labelled datapoints, with the largest amount ranked by RLG (10,000), followed by FS (1,858), and MEB (695). As all three chemists work in fields related to organic synthesis, their classifications are of particular value in a model that considers ease-of-synthesis scoring for POCs (and molecular materials in general). Despite the relatively small amounts of labelled data obtained from MEB and FS, we believed model performance would benefit from having multiple labelled datapoints for at least some of the molecules. Averaging the scores assigned by expert chemists has been shown to lead to a better prediction of synthetic difficulty than the opinion of individuals, due to bias originating from personal preference and experiences.[55] The training database labelled by three experienced chemists contained 12,553 data points in total, of which we obtained 2,008 positive easy-to-synthesise labels and 10,545 negative difficult-to-synthesise labels, including overlapping molecules scored by multiple chemists. Table 1 shows the number of molecules scored by each chemist, in addition to the percentage of molecules each chemist labelled easy- and difficult-to-synthesise, and their years of synthetic chemistry experience. RLG labelled the smallest percentage of molecules as easy-to-synthesise (11% of all labelled molecules), followed by MEB (33%) and FS (36%). Despite the widely varying percentages of molecules assigned as easy- and difficult-to-synthesise, the number of molecules scored by multiple chemists was relatively low (11% of all the molecules in the training set), as seen in Table 2. This relatively low overlap in the molecules scored by each chemist suggests that the greater proportion of difficult-to-synthesise labels, especially in the case RLG's labels, is not indicative of systematic labelling of one class over the other.

Although the number of molecules scored by multiple chemists was low (1,667 molecules in total), the chemists agreed with each other 73% of the time on average. As shown in Table 2, for the 42 molecules labelled by three chemists, the chemists agreed 74% of time and for the 1,625 molecules labelled by two chemists, 73% of the time. The 11 molecules that at least one out of the three chemists labelled differently are shown in Figure S1, followed

Table 1: The number of molecules labelled by each chemist, in addition to the number of easy- and difficult-to-synthesise, and their respective percentages. The number of years of synthetic chemistry experience of each chemist is also given.

	RLG	**FS**	**MEB**
Molecules	10,000	1,858	695
Years Experience	12	4	20
Easy-to-synthesise	1,109 (11%)	667 (36%)	232 (33%)
Difficult-to-synthesise	8,891 (89%)	1,191 (64%)	463 (67%)

by the labels assigned by the three chemists in Table S2. We found that disagreement between chemists was relatively large (27% of molecules labelled by two or more chemists), which, as expected, shows chemical intuition can be variable and somewhat subject to prior experiences. Discrepancies in chemist labels were counteracted by providing both positive and negative labels for the same molecule as training data for the RF model, which reduces the importance of that training sample on the overall decision made by the model. Indeed, the disagreement between chemists and the relatively short time frame in which molecules were assessed indicates this classifier is an attempt to model the fast intuition of an expert synthetic chemist when selecting precursors, rather than an in-depth retrosynthetic analysis of molecules.

Table 2: A summary of the molecules labelled by each synthetic materials chemist. Percentages of each label compared to the total number of molecules labelled by each chemist are given in brackets.

	Labelled by Three	**Labelled by Two**	**Labelled by One**
Molecules	42	1,625	10,886
In agreement	31 (74%)	1,179 (73%)	-
In disagreement	11 (26%)	446 (27%)	-

Comparison with Existing Methods

Figure 4 compares the labels expert chemists assigned molecules in the training database to that of the SAScore and SCScore. The correlation between the SAScore and the SCScore for our training database was very weak, with a correlation coefficient of 0.12. The SCScore

aims to capture synthetic complexity as defined by the predicted number of reaction steps required to make a molecule. Therefore, in the SCScore, complex molecules appearing a greater number of times at the start of a reaction pathway would exhibit a lower perceived synthetic difficulty than less complex molecules that appear more frequently near the end of a reaction pathway. By contrast, the SAScore primarily uses a measure of structural complexity to measure synthetic difficulty. The different assumptions included in each scoring method can lead to very different results in quantifying synthetic difficulty, which is what we observe in Figure 4. Complexity can be easily introduced into a molecule using many robust reaction transformations, meaning apparently complex molecules are not necessarily more synthetically challenging to access. Furthermore, the overlapping distributions of synthetic difficulty scores for molecules labelled as easy- and difficult-to-synthesise suggest that these models are not able to readily distinguish between precursors used for materials synthesis in agreement with experienced chemists working in the field. These results highlight the necessity to develop an alternate heuristic model that can identify molecules that synthetic chemists in the field of materials discovery would select themselves.

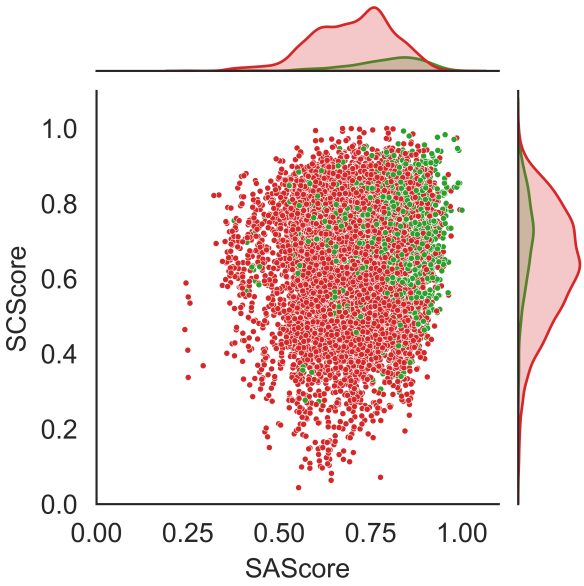

Figure 4: Synthetic difficulty scores of molecules in the training dataset, calculated using the SCScore and SAScore methods, scaled between 0 and 1. Colour coding refers to those labelled as easy-to-synthesise (green) or difficult-to-synthesise (red) by the synthetic chemists. Kernel density estimates of the synthetic difficulty score distributions are shown on the top and side panels.

MPScore: Model Training and Validation

We trained a RF model using the labelled training set discussed above to see if this would give an improved assessment of the synthetic difficulty of material precursors, in particular for POC systems. The developed model was evaluated using five-fold cross validation, and the resulting average scores across all folds and their standard deviations are shown in Table 3, in addition to the sum of false positives, false negatives, true positives and true negatives across all five folds, shown in Table 4.

Ideally, high precision and recall are desirable for any binary classifier, but their impact

Table 3: Evaluation metrics for the MPScore model. Scores are averaged across the five cross-validation folds.

Accuracy	**Precision**		**Recall**		**F1 Score**
	Easy-to-synthesise	Difficult-to-synthesise	Easy-to-synthesise	Difficult-to-synthesise	
0.85 ± 0.04	0.56 ± 0.13	0.91 ± 0.08	0.60 ± 0.17	0.92 ± 0.03	0.53 ± 0.07

Table 4: Sum of all outcomes from each fold of the cross-validation procedure used to estimate the performance of the MPScore, in addition to the total number of predictions made. False positive outcomes refer to molecules labelled as difficult-to-synthesise by the chemist but easy-to-synthesise by our MPScore, whereas false negative outcomes refer to molecules labelled as easy-to-synthesise by the chemist but difficult-to-synthesise by our MPScore.

False Negatives	**False Positives**	**True Positives**	**True Negatives**	**Total**
945	768	1063	9,777	12,553

on the usefulness of a synthetic difficulty score is very different. Low precision (large number of false positives), when a large number of molecules labelled easy-to-synthesise are actually difficult-to-synthesise, wastes the resources of experimental chemists. Low recall (high proportion of false negatives) results in precursor candidates being missed that could have had favourable properties when formulated into the final material, but were incorrectly classed as difficult-to-synthesise and disregarded. However, the latter misclassification is less problematic, as there is no experimental cost associated with missing an easy-to-synthesise molecule. While pre-screening precursors according to their synthetic difficulty with the MPScore may miss candidates with exceptional properties, whose precursors are challenging to synthesise, we instead focus our computational resources on cages that are most likely to be made in a lab, given lab synthesis is the bottleneck we currently face in our computational screening approaches. Therefore, we aimed to develop a model that maximised precision, reducing the number of false positives obtained.

The MPScore model was able to achieve an overall accuracy of 0.85, as seen in Table 3. The mean precision and mean recall for the difficult-to-synthesise label were 0.91 and 0.92 respectively. A lower mean precision and mean recall score is seen for the easy-to-synthesise molecule label of 0.56 and 0.60 respectively. We hypothesise that the lower precision and recall for the easy-to-synthesise label is due to the model's tendency to classify a molecule as difficult-to-synthesise, as suggested by the recall of 0.92 in the difficult-to-synthesise class. This is exemplified by the 9,777 true negatives assigned by the MPScore, which out-weigh the

1,063 true positives the model correctly classified, as shown in Table 4. This can be explained by the imbalanced dataset to train the MPScore, in which the overwhelming majority of molecules were labelled as difficult-to-synthesise. This differs to the typical training dataset used in ML for chemistry, which consists of only positive examples due to the absence of failed experiments or negative results. In such cases, the generation of negative results proved advantageous for generating useful ML models.[56] Later in this sub-section, we show how the threshold the classifier uses to make predictions can be optimised to increase the precision of the model, reducing the negative consequences of class imbalance.

The F1 score, the harmonic mean of the precision and recall scores, provides a combined metric of precision and recall. The F1 score of 0.53 obtained for the MPScore shows the model has lower combined precision and recall scores for the easy-to-synthesise class, which can also be explained by the imbalanced training dataset. A minor contributing factor to the low precision score could be due to some of the features of easy-to-synthesise molecules being similar to those of difficult-to-synthesise, resulting in the model having difficulty distinguishing between the two classes. Figure 5a shows that many fingerprint bit values contributed similar amounts to the classification of the RF model. This shows, as would be expected, that the chemist's opinion was influenced by multiple fragments present within a molecule. Additionally, the confidence intervals in Figure 5a show how the importance of features differ widely across each decision tree in the forest.

One approach to minimise the impact of class imbalance is to adjust the probability threshold of the classifier.[57] Increasing the threshold required for a molecule to belong to the minority class (easy-to-synthesise) reduces the false positive rate, increasing the precision of the classifier. The probability of a sample belonging to a particular class in a RF model is determined by the proportion of decision trees in the ensemble assigning that sample to a particular class. Figure 5b shows the effect of changing the probability cut-off threshold for the difficult-to-synthesise class on the aggregated precision and recall values achieved by the RF models during the cross-validation processes. It can be seen that by increasing the

probability thresholds, the model's precision outperforms the recall. At almost all thresholds, the MPScore outperforms the naive classifier, defined as a model randomly assigning molecules as easy- and difficult-to-synthesise. Choosing a high probability threshold for the MPScore minimised the risk of obtaining false positives (high precision), at the expense of a larger number of false negatives (low recall). Despite the potential to miss easy-to-synthesise molecules, this compromise ensured we can be more certain that the molecules that are selected by the MPScore are not false positives and waste synthetic efforts. One estimate of the size of the chemical space of synthesisable drug-like molecules alone is 10^{11} molecules,[58] meaning the cost of obtaining a false negative and missing a potential easy-to-synthesise candidate is low.

To reduce the probability of obtaining false positive scores for molecules, a final threshold of 0.12 was chosen for our MPScore to select easy-to-synthesise molecules during the precursor screening process, chosen to assign only 1% of all molecules in the POC precursor database as easy-to-synthesise. Molecules with a probability of below 0.12 of belonging to the difficult-to-synthesise were classified as easy-to-synthesise and carried forward in the workflow. At this probability threshold, the precision and recall scores for the easy-to-synthesise test-set class were 0.78 and 0.36 receptively. This means, despite missing 64% of easy-to-synthesise molecules, we can be 78% certain our selected molecule is not a false positive; a necessary trade-off for our materials screening workflow. We trained the final MPScore model used in the screening workflow on the entire dataset of labelled molecules, using the scores obtained from the cross-validation procedure as estimates for the final model. The model, in addition to the training scripts are available on Github.[43]

High-Throughput Computational Discovery of Synthesisable POCs

To test the applicability of the MPScore, we investigated its performance within a high-throughput screen for the discovery of novel shape-persistent POCs. We repeated the screen with the SAScore and SCScore separately to compare the results. The scores were used

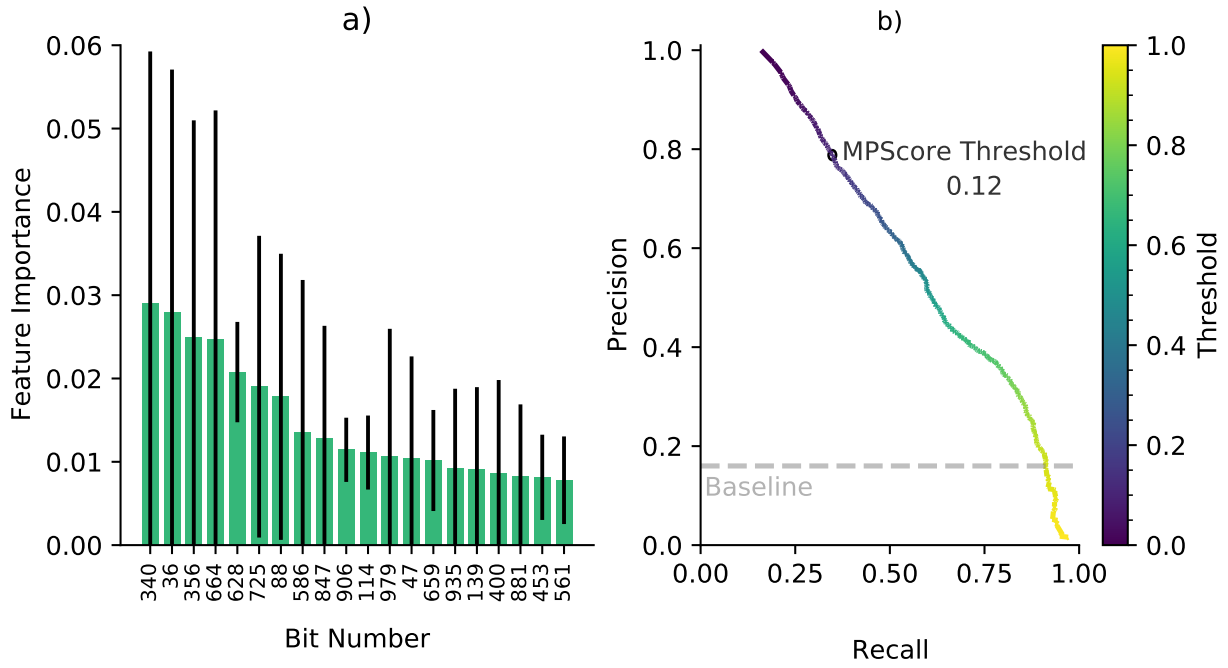

Figure 5: Mean importances for the 20 most important fingerprint bits (a) and the precision-recall curve (b) for our MPScore. Feature importances (a) were calculated across all 100 decision trees in the random forest, and the 95% confidence intervals across the 100 decision trees in the ensemble are shown by the black bars. Precision and recall scores (b) are calculated using the aggregated true positives, false positives and false negatives for each fold during cross-validation. The line colour indicates the probability threshold of classifying a molecule as difficult-to-synthesise, in addition to the baseline classifier shown in grey, defined as a classifier which assigns molecules to the easy- or difficult-to-synthesise class randomly. The final threshold used for the MPScore model, 0.12, is also shown.

to eliminate difficult-to-synthesise POC precursors from a database of diamines and tri-aldehydes. Using the remaining easy-to-synthesise precursors, we constructed and geometry optimised the POCs that could be formed from these precursors, analysing the pore structure of those that were predicted to be shape-persistent.

Precursor Database and Synthesisable Precursors

We explored the chemical space of **Tri^{4}Di^6** imine-based cages formed from a [4+6] condensation of a tritopic aldehyde with a ditopic amine.[33] The precursors library was based on the

eMolecules[42] and Reaxys[41,59] databases, thus containing literature-reported compounds of varying levels of synthetic difficulty. For the MPScore, we interpreted the probability value obtained from the RF model as a continuous score. Molecules with a higher probability of belonging to the easy-to-synthesise class were defined as being less synthetically difficult to make. We assigned an overall synthetic difficulty value to each pair of precursors that could be used to form a POC (diamine and trialdehyde). The aggregated score was defined as the sum of each synthetic difficulty score of each precursor in the pair, scaled between 0 and 1 to allow for easier comparison between scores. We calculated the aggregated scores for each pair using the SAScore, SCScore and MPScore, which can be seen in Figure 6. Overall, Figure 6 shows a similar distribution of precursor synthetic difficulties for the SAScore and MPScore, whereas the SCScore consistently assigns a greater synthetic difficulty.

The relatively broad distributions across all three scores imply that if considering precursors for experimental material synthesis, even restricting the initial database to the literature-reported Reaxys[59] database is insufficient. The synthesis of these molecules could require multiple difficult and low yielding reaction steps, toxic or expensive reagents, or challenging purification procedures; none of these points are considered if we simplify the existence of a molecule in the Reaxys[59] database to meaning it is 'synthesisable'. Even if those molecules *are* accessible via many-step syntheses, they are, most likely, unsuitable for materials discovery, unless one was extremely certain that a specific precursor was the only molecule that could infer a desired functionality that was of high value. Certainly, in the case of POC materials discovery programmes, we are not anticipating such scenarios, and instead high quantities of readily accessible precursors are required; meaning that cheap reagents and robust reactions are essential to create them.

For our screening workflow, we chose cut-off thresholds for each of the synthetic difficulty scoring methods that would remove 99% of all precursor combinations, leaving the 1% of precursors with the lowest synthetic difficulty values remaining. By using this approach, we hoped to identify any novel and previously undiscovered POCs with promising properties

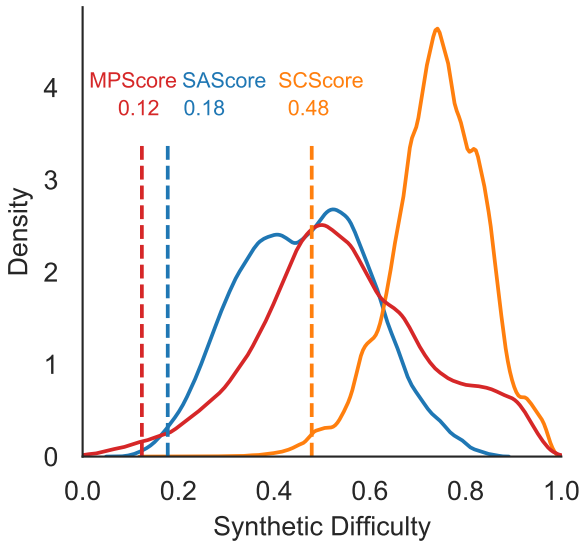

Figure 6: Distributions of the synthetic difficulty values for each POC precursor combination. Synthetic difficulty values are calculated using the SAScore (blue), SCScore (orange) and our MPScore (red). Vertical lines represent the first percentile for each score, the respective cut-off thresholds are given by the value above each vertical line, coloured according to score. A precursor pair is defined as easy-to-synthesise with a synthetic difficulty score lower than the threshold value for each score, indicated by the vertical dashed lines. A lower score for a precursor combination indicates both precursors in the combination are easier-to-synthesise, whereas a higher score indicates they are harder-to-synthesise. Precursors whose combined synthetic difficulty scores were below these threshold values (to the left of horizontal lines) were carried forward for POC construction and optimisation.

that could be made from only the easiest-to-synthesise precursors. As precursor synthesis is much more cost- and time-intensive than our POC screening workflow, 1% was chosen as a cut-off to maximise the chance that the precursors could be synthesised or are commercially available. The final cut-off thresholds were: 0.12 for the MPScore, 0.48 for the SCScore, and 0.18 for the SAScore. Precursor combinations with synthetic difficulty scores greater than these values were defined as too difficult-to-synthesise and disregarded. Combinations with scores less-than or equal-to these values were carried forwards for POC construction and further computational analysis.

Model Validation by Expert Chemists

Next, we evaluated the developed MPScore against the SAScore and SCScore methods as precursor filters, comparing molecules selected as 'easy-to-synthesise' and 'difficult-to-synthesise' by the different methods against data coming from an experienced materials chemist. Author RLG was blindly presented with 30 diamines and 30 trialdehydes with the lowest and highest synthetic difficulty values (calculated using the SAScore, the SCScore and the MPScore) from the POC precursor database and presented with the same criterion as previously: *can you make 1 g of this compound in under 5 steps?* The molecules and their calculated synthetic difficulties, can be found in Figure S2. We present the true positives, true negatives, false negatives and false positives for each model in Table 5.

Table 5 shows our MPScore was able to achieve the lowest number of false positives for the aldehyde task (no false positives) and performed identically to the SAScore for the amine ranking task (1 false positive). Generally, models were better at identifying difficult-to-synthesise molecules in agreement with RLG, as shown by the higher numbers of true negatives than true positives. While desirable, this is not a requirement of a synthetic difficulty model in our POC screening workflow, as scores are used to identify easiest-to-synthesise precursors, as opposed to identifying the hardest-to-synthesise. For our requirement of selecting precursors in agreement with expert chemists, this shows the MPScore is the most effective, given by the lowest cumulative number of false positives (1 false positive for both amines and aldehydes). However, we do note the sample size of 10 amines and 10 aldehydes for each model is too small to draw any concrete conclusions from.

Structural Analysis of Computationally Screened POCs

Following the analysis of the MPScore, we assessed the results of the computational screening workflow, when using the three different scoring methods to remove non-synthesisable precursors. After only using the 1% of precursors scored as most likely to be synthesisable from each of the methods, we automated cage construction and conformer searching using

Table 5: Performance metrics for the blind validation task used to assess the performance of the SAScore, SCScore and MPScore on predicting easy-to-synthesise POC precursors. True positive outcomes are defined as molecules with the lowest calculated synthetic difficulty scores that were also classified as easy-to-synthesise by RLG. For our MPScore model, we aim to minimise false positives, which are molecules classified as difficult-to-synthesise by RLG, but which a synthetic difficulty model assigned a low synthetic difficulty score, implying it is easy-to-synthesise. Values in bold correspond to the score that produced the fewest false positives.

	Trialdehydes			
	True Positives	**True Negatives**	**False Positives**	**False Negatives**
SAScore	4	8	6	2
SCScore	5	8	5	2
MPScore	10	9	**0**	1
	Diamines			
	True Positives	**True Negatives**	**False Positives**	**False Negatives**
SAScore	9	10	**1**	0
SCScore	2	9	8	1
MPScore	9	9	**1**	1

`stk` and `MacroModel`, and then analysed the windows and central cavity of the lowest energy conformation using `pyWindow`.[37,53,54] We then analysed the property distributions of the cages, with a focus on those that are shape-persistent (as defined by Equation 5). In other words, we sought molecules where the lowest energy conformations had a cavity large enough to host, at the very least, a hypothetical spherical probe with a diameter of 1 Å. Finally, we compared the cavity distributions of shape-persistent POCs from precursors filtered for synthetic difficulty with a control sample of randomly selected precursor combinations from the database, also 1% of all combinations.

In total, 28,176 precursor combinations were selected: 7,040 from the MPScore, 7,044 from the SAScore, 7,047 from SCScore and 7,045 from a control sample of randomly selected precursor combinations. A control sample was used to investigate whether the synthetic difficulty scores had any influence on the properties of the cages the precursor combinations formed. Following the duplicate precursor removal mentioned in the previous section, this

resulted in a total of 27,853 required optimisations: 6,917 from the MPScore, 6,876 from the SAScore, 7,015 from the SCScore and 7,045 from the control sample. In total, there were 1,448 overlapping precursor combinations between precursors combinations selected by all three synthetic difficulty scores and the control sample, the full breakdown of which can be seen in Table S5. The SAScore and MPScore had the most overlap in the number of selected precursors at 887. The SCScore had fewer overlaps, with 210 precursors in common with the SAScore and 165 for the MPScore. Optimisations were only performed once for each unique precursor combination, resulting in a total of 26,405 optimisations on unique precursors combinations. Two optimisations failed from precursor combinations selected by the SAScore, however, these were not explored further as we focus on a fully autonomous workflow with no human-input to screen failed optimisations. Table 6 shows the number of POCs optimised for each synthetic difficulty score, in addition to the percentages that remained shape-persistent.

Table 6: Number and percentage of shape-persistent cages formed from precursors selected by each synthetic difficulty model in the first three columns, in a control sample of randomly selected precursors, and in our previous work in the final column.[30] The absolute number of cages and percentages are shown.

	MPScore	**SAScore**	**SCScore**	**Control**	**Previous work[30]**
Cages	6,917	6,876	7,015	7,045	6,018
Shape Persistent	1,646 (25%)	770 (11%)	278 (4%)	347 (5%)	2,314 (38%)

In our previous work, out of 6,018 **Tri^{4}Di^6** POCs constructed from trialdehyde and diamine precursors designed by RLG to exhibit rigid precursor cores, 2318 (38%) were shape-persistent following geometry optimisation (Table 6).[30] As expected, a far lower proportion than 38% remained shape-persistent in our workflow now that we were also including a very strict consideration of synthetic difficulty of the precursors without manual assessment of precursors for their propensity to form a shape-persistent cage. As shown in Table 6, organic cages that remained shape-persistent from each synthetic difficulty score are: 25% from the MPScore, 11% from the SAScore, 4% from the SCScore, and 5% from the control sample.

Interestingly, the percentage of shape-persistent cages from precursor combinations selected by the MPScore was 19 percentage points greater than the control sample of randomly selected cages. Both percentages of the SAScore and SCScore were much closer to that of the control sample, with percentage point differences of 6 and 1 respectively. This large difference in the number of shape-persistent cages suggests that easy-to-synthesise precursors selected by the MPScore are more likely to form a shape-persistent POC, compared with precursors selected by other scoring methods. While our only aim for the MPScore was to filter for easy-to-synthesise POC precursors, its secondary effect of biasing towards precursors more likely to form a shape-persistent POC is nonetheless useful for POC chemists in the precursor selection process.

Figure 7 shows the distribution of cavity diameters for shape-persistent POCs screened using the three different synthetic difficulty models. Interestingly, precursors in the control sample formed POCs with the highest cavity diameter, with a mean of 8.5 Å, followed by the precursors with the lowest synthetic difficulty calculated by the MPScore (8.3 Å), the SAScore (5.9 Å) and finally the SCScore (4.1 Å). In our previous work, we found that out of 116 cages reported in the literature, cages with cavity sizes of 0-6 Å are most prevalent, with a general absence of cages with diameters larger than 16 Å.[28] Large cavity sizes are properties POCs rarely exhibit due to the propensity of larger cages to collapse upon desolvation or to catenate with other cages, making POCs with large cavity sizes an interesting target for a computational screening workflow. Precursors that form POCs with larger cavities typically contain greater numbers of atoms and bonds, resulting in more degrees of freedom, more competing synthesis pathways during their formation and greater flexibility, and thus more facile collapse mechanisms.

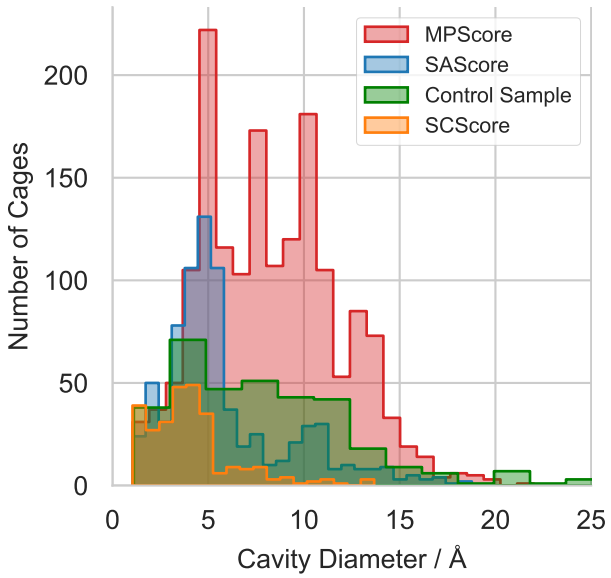

Figure 7: Distribution of cavity diameters for POCs predicted to be shape-persistent from the screening. Only POCs with a cavity diameter greater than 1 Å are included. The precursors of the POCs were those with the lowest 1% of synthetic difficulty scores calculated using the SAScore (blue), SCScore (orange), the MPScore (red) and a control sample of randomly selected precursors (green).

From this analysis, we conclude it is possible to discover shape-persistent POCs from precursors predicted to be easy-to-synthesise, without designing those precursors with structural features that are more likely to result in a shape-persistent POC. We show the MPScore is able to identify the largest proportion of shape-persistent POCs (1,646 cages), compared with other synthetic difficulty scores (770 from the SAScore and 278 from the SCScore). Using this approach, we discovered a total of 2,694 shape-persistent cages that could be formed from easy-to-synthesise precursors; a significant number in contrast to the hundreds of POCs that have already been discovered. Additionally, while the vast majority of shape-persistent POCs have a pore diameter of 5 Å, we show that this workflow is also able to identify a smaller number of cages with larger pore diameters. This is a promising discovery, as it shows precursors that could form larger POCs can be readily accessed.

Analysis of the Identified Promising Large Cavity Diameter POCs

From the computational screening workflow, 29 unique shape-persistent cages had a cavity size of 16 Å or greater, 23 selected using the MPScore, 9 with the SAScore and none with the SCScore. Three identical precursor combinations were present in the selection of precursors with the lowest synthetic difficulty scored by the MPScore and the SAScore. Figure 8 shows the six largest synthetically accessible cages, alongside their corresponding precursor pair and Table 7 summarises the properties of these cages. The remaining 23 of these POCs can be found in Figure S3 and their respective properties in Table S6. All six cages (labelled **1** through **6**, according to their size, with **1** being the largest cavity) in Figure 8 are predicted to have a cavity size greater than 16 Å, and five were from the screening using our new MPScore method. If using solely the SAScore for pre-screening, only cages **5** and **6** would have been included. Meanwhile, using only the SCScore, all of these cages would have been missed. We inspected the six precursor pairings to see whether they were indeed likely to be synthetically accessible. To assess whether a precursor was commercially available, we used the stock supplier feature on the ZINC database, one of the largest commercially available compound catalogues.[60,61] Diamines in cages **1, 3, 4, 5** and **6** are commercially available, with more than ten suppliers listed in the ZINC database. The enantiopure compounds **3** and **5** are relatively expensive to purchase. Diamine **2** meanwhile can be synthesised in five steps from a commercially available compound.[62]

The two trialdehyde precursors present in all six cages are all available via a single step from commercially available reagents. The 1,3,5-tris(phenylenevinylene)benzene precursor used in cages **2**-**6** was also present in 19 of the 23 remaining shape-persistent cages with cavity diameter greater than 16 Å (present six times in precursor combinations from the SAScore and 15 times from the MPScore), indicating this could be a promising precursor to choose when designing cages with a large cavity size. The relative ease with which these trialdehydes can be accessed and their propensity to form cages with large cavity sizes makes them a desirable precursor for targeting cages with large cavity sizes. Indeed, the 2,694

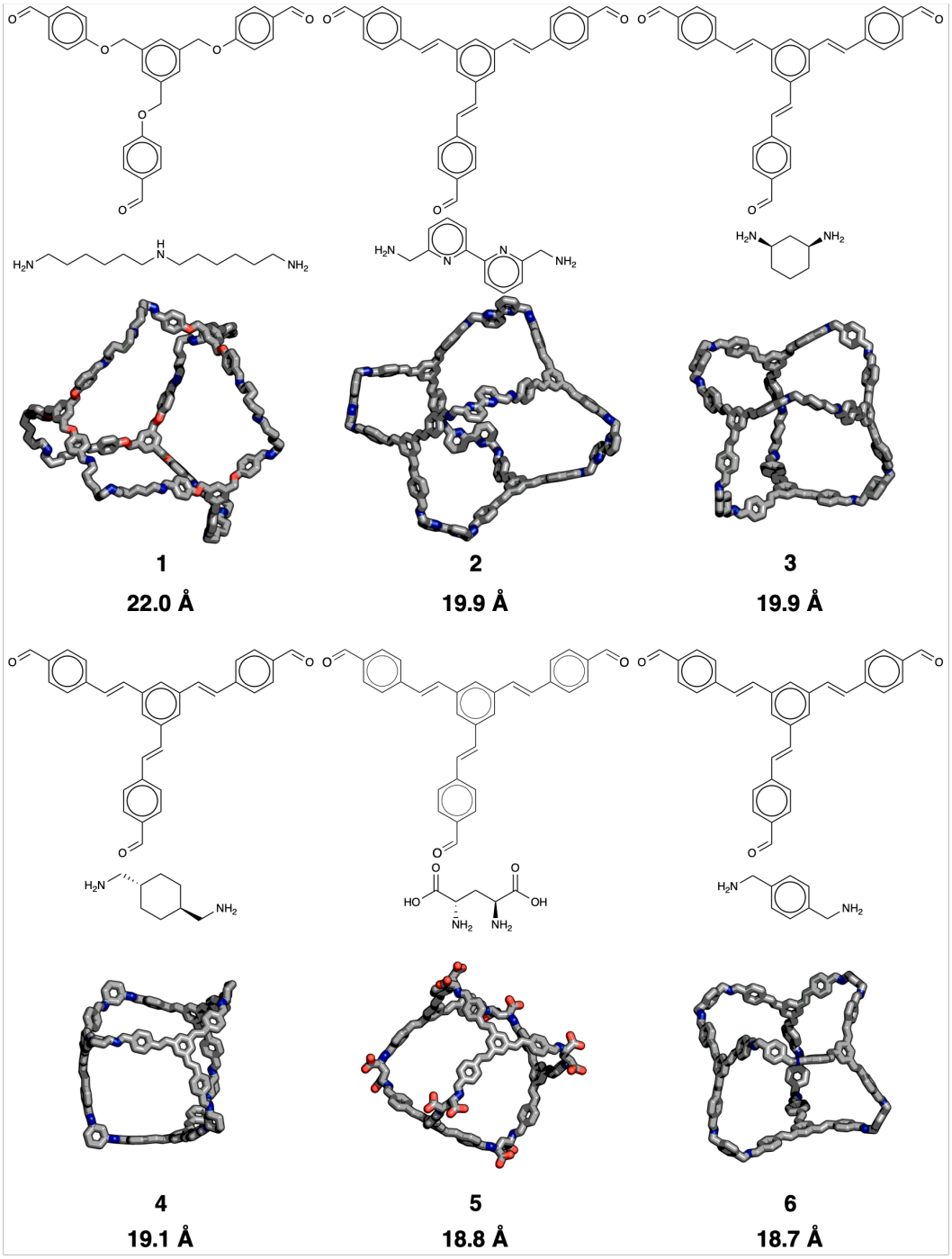

Figure 8: The six largest shape-persistent POCs identified from the screening workflow alongside their respective precursor pairings. The cages are labelled **1** through **6** in descending order of cavity diameter. Calculated cavity diameters are labelled below each POC. The corresponding features of the POCs are in Table 7. All but cage **5** were identified with our MPScore method for measuring synthetic difficulty, and the other was identified with the SAScore. Carbon atoms are shown in grey, nitrogen in blue and oxygen in red. Hydrogen atoms are omitted for clarity.

Table 7: Calculated cavity diameter and precursor synthetic difficulty scores (from each of the three scores) for the six largest shape-persistent POCs identified from any of the methods. A low score indicates a precursor is easy-to-synthesise, whereas a score nearer one indicates the synthesis is predicted to be challenging or even impossible. The synthetic difficulty scores in **bold** show that each precursor pair belongs in the lowest 1% of values for each scoring method.

		Precursor Synthetic Difficulty		
Number	**Cavity Diameter / Å**	SAScore	SCScore	MPScore
1	22.0	0.31	0.72	**0.09**
2	19.9	0.33	0.72	**0.11**
3	19.9	0.25	0.73	**0.07**
4	19.1	0.24	0.67	**0.09**
5	18.8	**0.18**	0.74	0.17
6	18.7	**0.16**	0.73	**0.10**

cages predicted to be shape-persistent are a great many more than those that have been experimentally realised to-date. Despite the poor recall of the MPScore potentially resulting in a large number of false negative precursor combinations, the score is still able to bias towards easy-to-synthesise precursors that are able to form shape-persistent POCs. Our computational workflow identified shape-persistent POCs made from easily synthesisable precursors. Promisingly, this also indicates that precursors that form cages with a cavity diameter of over 16 Å can be accessed using easy-to-synthesise precursors.

In this workflow, we only considered the ease with which cage precursors can be synthesised, however, there are a number of other challenges one must overcome when designing a POC. While we only screened for $\mathbf{Tri^4Di^6}$ topology cages, the thermodynamic driving force may steer the DCC reaction to form a diverse range of other topologies, or even form a polymer or oligomer. We also did not account for the precursor solubility, or the formation of insoluble products, which is also a significant challenge in POC synthesis.[34,35] Accounting for all of these factors computationally is extremely challenging (and expensive), which is why chemical knowledge from humans still remains irreplaceable in the field of POC development.

Conclusions

Our primary goal in this work was to develop a computational screening workflow that eases the transition between computational material prediction and experimental realisation, using porous organic cages (POCs) as an exemplar material. To achieve this, we needed an automated way to consider the ease with which the organic molecules that are the precursors to POCs can be synthesised. We compared existing methods of calculating synthetic difficulty computationally, showing how current methods are not able to replicate the decisions of materials chemists when selecting easy-to-synthesise precursors for material synthesis. As existing synthetic difficulty scores were unsuitable for our intended purpose of automated POC discovery, we created our own machine learning model to predict the ease of synthesis, the Materials Precursor Score (MPScore). We collected 12,553 pieces of labelled training data from three materials chemists in the field of POCs and more general organic synthesis, with the goal of developing a model capable of classifying POC precursors as easy-to-synthesise or difficult-to-synthesise. The data was collected based on the answer to the question: *can you make 1 g of this compound in under 5 steps?*.

We applied our MPScore to the task of identifying easy-to-synthesise precursors for a computational screening workflow aimed at identifying organic cages with permanent, shape-persistent cavities. We found that our MPScore performed slightly better than existing methods to score the synthetic accessibility of organic molecules, and was more likely to bias towards precursors that form shape-persistent POCs. We showed that even when limiting precursors to only those that are considered easiest-to-synthesise, we could still discover POCs with unconventional properties, including those with a cavity size of greater than 16 Å. In total, we predicted 2,694 shape-persistent cages, 29 of which have a cavity size greater than 16 Å, a property scarcely found in the literature. For the six largest cages discovered using our autonomous workflow, we confirmed that the precursors used to create these top-scoring cages are either commercially available or there are straightforward synthesis routes reported in the literature.

We provide the database of shape-persistent cages in the hope of future experimental validation of some of the computational predictions, which is accessible at doi.org/10.14469/hpc/8045. Our MPScore and training data are also open-source, and can be applied to the ranking of other potential POC precursors or expanded with additional training data to generalise to other material precursors. Code used to train and validate our MPScore is available at zenodo.org/badge/latestdoi/210332220. We hope the MPScore will also be of value for considering the ease of synthesis of organic precursors in the wider field of molecular materials, although the model has not yet been tested beyond POC precursors. In the future, we hope the MPScore can be used to help guide both experimental and computational design of new functional materials. Such validation is the subject of our ongoing collaboration with experimental chemists. We believe this workflow is a small step towards a more autonomous discovery of new porous molecular materials.

Acknowledgement

R. L. G. and K. E. J. thank the Royal Society for a Royal Society University Research Fellowships. K. E. J. and F. T. S. thank the Leverhulme Trust for a Leverhulme Trust Research Project Grant. S. B. thanks the Leverhulme Research Centre for Functional Materials Design for a Ph.D. studentship. We acknowledge funding from the European Research Council under FP7 (CoMMaD, ERC Grant No. 758370).

Supporting Information Available

The supporting information includes training set analysis; analysis of the labelled chemist data; data for the MPScore validation; and further information from the cage screening workflow. The labelled training dataset used to train the MPScore and the database of optimised cages are provided as separate files. The open-access code and MPScore training files can be found on Zenodo at zenodo.org/badge/latestdoi/210332220.[43] The dataset

of optimised cages can be found at `doi.org/10.14469/hpc/8045`.

References

(1) Pyzer-Knapp, E. O.; Suh, C.; Gómez-Bombarelli, R.; Aguilera-Iparraguirre, J.; Aspuru-Guzik, A. What Is High-Throughput Virtual Screening? A Perspective from Organic Materials Discovery. *Annu. Rev. Mater. Res.* **2015**, *45*, 195–216.

(2) 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.

(3) Oganov, A. R.; Saleh, G.; Kvashnin, A. G. *Computational Materials Discovery*; Royal Society of Chemistry, 2018.

(4) Szczypiński, F. T.; Bennett, S.; Jelfs, K. E. Can we predict materials that can be synthesised? *Chem. Sci.* **2021**, *12*, 830–840.

(5) Bennett, S.; Tarzia, A.; Zwijnenburg, M. A.; Jelfs, K. E. In *Machine Learning in Chemistry: The Impact of Artificial Intelligence*; Cartwright, H. M., Ed.; Royal Society of Chemistry, 2020; Vol. 17; p 280.

(6) Struble, T. J.; Alvarez, J. C.; Brown, S. P.; Chytil, M.; Cisar, J.; DesJarlais, R. L.; Engkvist, O.; Frank, S. A.; Greve, D. R.; Griffin, D. J.; Hou, X.; Johannes, J. W.; Kreatsoulas, C.; Lahue, B.; Mathea, M.; Mogk, G.; Nicolaou, C. A.; Palmer, A. D.; Price, D. J.; Robinson, R. I.; Salentin, S.; Xing, L.; Jaakkola, T.; Green, W. H.; Barzilay, R.; Coley, C. W.; Jensen, K. F. Current and Future Roles of Artificial Intelligence in Medicinal Chemistry Synthesis. *J. Med. Chem.* **2020**, *63*, 8667–8682.

(7) Barone, R.; Chanon, M. A new and simple approach to chemical complexity. Application to the synthesis of natural products. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 269–272.

(8) Ertl, P.; Schuffenhauer, A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *J. Cheminform.* **2009**, *1*, 8.

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

(10) Takaoka, Y.; Endo, Y.; Yamanobe, S.; Kakinuma, H.; Okubo, T.; Shimazaki, Y.; Ota, T.; Sumiya, S.; Yoshikawa, K. Development of a method for evaluating drug-likeness and ease of synthesis using a data set in which compounds are assigned scores based on chemists' intuition. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1269–1275.

(11) Coley, C. W.; Rogers, L.; Green, W. H.; Jensen, K. F. SCScore: Synthetic Complexity Learned from a Reaction Corpus. *J. Chem. Inf. Model.* **2018**, *58*, 252–261.

(12) Voršilák, M.; Kolář, M.; Čmelo, I.; Svozil, D. SYBA: Bayesian estimation of synthetic accessibility of organic compounds. *J. Cheminform.* **2020**, *12*, 35.

(13) Gao, W.; Coley, C. W. The Synthesizability of Molecules Proposed by Generative Models. *J. Chem. Inf. Model.* **2020**, *60*, 5714–5723.

(14) Genheden, S.; Thakkar, A.; Chadimová, V.; Reymond, J.-L.; Engkvist, O.; Bjerrum, E. AiZynthFinder: a fast, robust and flexible open-source software for retrosynthetic planning. *J. Cheminform.* **2020**, *12*, 70.

(15) Lajiness, M. S.; Maggiora, G. M.; Shanmugasundaram, V. Assessment of the consistency of medicinal chemists in reviewing sets of compounds. *J. Med. Chem.* **2004**, *47*, 4891–4896.

(16) Slater, A. G.; Cooper, A. I. Porous materials. Function-led design of new porous materials. *Science* **2015**, *348*, aaa8075.

(17) Morris, R. E.; Wheatley, P. S. Gas storage in nanoporous materials. *Angew. Chem. Int. Ed Engl.* **2008**, *47*, 4966–4981.

(18) Li, J.-R.; Kuppler, R. J.; Zhou, H.-C. Selective gas adsorption and separation in metal-organic frameworks. *Chem. Soc. Rev.* **2009**, *38*, 1477–1504.

(19) Zhang, J.; Chen, J.; Peng, S.; Peng, S.; Zhang, Z.; Tong, Y.; Miller, P. W.; Yan, X.-P. Emerging porous materials in confined spaces: from chromatographic applications to flow chemistry. *Chem. Soc. Rev.* **2019**, *48*, 2566–2595.

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

(21) Ma, L.; Abney, C.; Lin, W. Enantioselective catalysis with homochiral metal-organic frameworks. *Chem. Soc. Rev.* **2009**, *38*, 1248–1256.

(22) Brutschy, M.; Schneider, M. W.; Mastalerz, M.; Waldvogel, S. R. Porous organic cage compounds as highly potent affinity materials for sensing by quartz crystal microbalances. *Adv. Mater.* **2012**, *24*, 6049–6052.

(23) Wales, D. J.; Grand, J.; Ting, V. P.; Burke, R. D.; Edler, K. J.; Bowen, C. R.; Mintova, S.; Burrows, A. D. Gas sensing using porous materials for automotive applications. *Chem. Soc. Rev.* **2015**, *44*, 4290–4321.

(24) Giri, N.; Del Pópolo, M. G.; Melaugh, G.; Greenaway, R. L.; Rätzke, K.; Koschine, T.; Pison, L.; Gomes, M. F. C.; Cooper, A. I.; James, S. L. Liquids with permanent porosity. *Nature* **2015**, *527*, 216–220.

(25) Melaugh, G.; Giri, N.; Davidson, C. E.; James, S. L.; Del Pópolo, M. G. Designing and understanding permanent microporosity in liquids. *Phys. Chem. Chem. Phys.* **2014**, *16*, 9422–9431.

(26) 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.

(27) Moghadam, P. Z.; Li, A.; Wiggin, S. B.; Tao, A.; Maloney, A. G. P.; Wood, P. A.; Ward, S. C.; Fairen-Jimenez, D. Development of a Cambridge Structural Database Subset: A Collection of Metal–Organic Frameworks for Past, Present, and Future. *Chem. Mater.* **2017**, *29*, 2618–2625.

(28) Berardo, E.; Turcani, L.; Miklitz, M.; Jelfs, K. E. An evolutionary algorithm for the discovery of porous organic cages. *Chem. Sci.* **2018**, *9*, 8513–8527.

(29) Miklitz, M.; Turcani, L.; Greenaway, R. L.; Jelfs, K. E. Computational discovery of molecular C60 encapsulants with an evolutionary algorithm. *Communications Chemistry* **2020**, *3*, 10.

(30) Turcani, L.; Greenaway, R. L.; Jelfs, K. E. Machine Learning for Organic Cage Property Prediction. *Chem. Mater.* **2019**, *31*, 714–727.

(31) Slater, A. G.; Little, M. A.; Briggs, M. E.; Jelfs, K. E.; Cooper, A. I. A solution-processable dissymmetric porous organic cage. *Mol. Syst. Des. Eng.* **2018**, *3*, 223–227.

(32) Kulchat, S.; Chaur, M. N.; Lehn, J.-M. Kinetic Selectivity and Thermodynamic Features of Competitive Imine Formation in Dynamic Covalent Chemistry. *Chemistry* **2017**, *23*, 11108–11118.

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

(34) Briggs, M. E.; Cooper, A. I. A Perspective on the Synthesis, Purification, and Characterization of Porous Organic Cages. *Chem. Mater.* **2017**, *29*, 149–157.

(35) Lauer, J. C.; Zhang, W.-S.; Rominger, F.; Schröder, R. R.; Mastalerz, M. Shape-Persistent [4+4] Imine Cages with a Truncated Tetrahedral Geometry. *Chemistry* **2018**, *24*, 1816–1820.

(36) Abet, V.; Szczypiński, F. T.; Little, M. A.; Santolini, V.; Jones, C. D.; Evans, R.; Wilson, C.; Wu, X.; Thorne, M. F.; Bennison, M. J.; Cui, P.; Cooper, A. I.; Jelfs, K. E.; Slater, A. G. Inducing Social Self-Sorting in Organic Cages To Tune The Shape of The Internal Cavity. *Angew. Chem. Int. Ed Engl.* **2020**, *59*, 16755–16763.

(37) Turcani, L.; Berardo, E.; Jelfs, K. E. stk: A python toolkit for supramolecular assembly. *J. Comput. Chem.* **2018**, *39*, 1931–1942.

(38) Berardo, E.; Greenaway, R. L.; Turcani, L.; Alston, B. M.; Bennison, M. J.; Miklitz, M.; Clowes, R.; Briggs, M. E.; Cooper, A. I.; Jelfs, K. E. Computationally-inspired discovery of an unsymmetrical porous organic cage. *Nanoscale* **2018**, *10*, 22381–22388.

(39) Greenaway, R. L.; Santolini, V.; Pulido, A.; Little, M. A.; Alston, B. M.; Briggs, M. E.; Day, G. M.; Cooper, A. I.; Jelfs, K. E. From Concept to Crystals via Prediction: Multi-Component Organic Cage Pots by Social Self-Sorting. *Angew. Chem. Int. Ed Engl.* **2019**, *58*, 16275–16281.

(40) Greenaway, R. L.; Jelfs, K. E. Integrating Computational and Experimental Workflows for Accelerated Organic Materials Discovery. *Adv. Mater.* **2021**, e2004831.

(41) Reaxys database. `http://reaxys.com`, Accessed: 2019-2-1.

(42) eMolecules. `https://www.emolecules.com/`, Accessed: 2019-2-1.

(43) Bennett, S. Materials Precursor Score. 2021; https://zenodo.org/badge/latestdoi/210332220.

(44) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. Random forest: a classification and regression tool for compound classification and QSAR modeling. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1947–1958.

(45) Sheridan, R. P. Using random forest to model the domain applicability of another random forest model. *J. Chem. Inf. Model.* **2013**, *53*, 2837–2850.

(46) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

(47) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754.

(48) Landrum, G. RDKit. https://www.rdkit.org/, Accessed: 2019-2-1.

(49) Riniker, S.; Landrum, G. A. Open-source platform to benchmark fingerprints for ligand-based virtual screening. *J. Cheminform.* **2013**, *5*, 26.

(50) O'Boyle, N. M.; Sayle, R. A. Comparing structural fingerprints using a literature-based similarity benchmark. *J. Cheminform.* **2016**, *8*, 36.

(51) Ertl, P. An algorithm to identify functional groups in organic molecules. *J. Cheminform.* **2017**, *9*, 36.

(52) Harder, E.; Damm, W.; Maple, J.; Wu, C.; Reboul, M.; Xiang, J. Y.; Wang, L.; Lupyan, D.; Dahlgren, M. K.; Knight, J. L.; Kaus, J. W.; Cerutti, D. S.; Krilov, G.; Jorgensen, W. L.; Abel, R.; Friesner, R. A. OPLS3: A Force Field Providing Broad

Coverage of Drug-like Small Molecules and Proteins. *J. Chem. Theory Comput.* **2016**, *12*, 281–296.

(53) Schrödinger Release 2018-4: MacroModel, Schrödinger, LLC, New York, NY, 2020.

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

(55) Bonnet, P. Is chemical synthetic accessibility computationally predictable for drug and lead-like molecules? A comparative assessment between medicinal and computational chemists. *Eur. J. Med. Chem.* **2012**, *54*, 679–689.

(56) Raccuglia, P.; Elbert, K. C.; Adler, P. D. F.; Falk, C.; Wenny, M. B.; Mollo, A.; Zeller, M.; Friedler, S. A.; Schrier, J.; Norquist, A. J. Machine-learning-assisted materials discovery using failed experiments. *Nature* **2016**, *533*, 73–76.

(57) Saito, T.; Rehmsmeier, M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One* **2015**, *10*, e0118432.

(58) Nicolaou, C. A.; Watson, I. A.; Hu, H.; Wang, J. The Proximal Lilly Collection: Mapping, Exploring and Exploiting Feasible Chemical Space. *J. Chem. Inf. Model.* **2016**, *56*, 1253–1266.

(59) Lawson, A. J.; Swienty-Busch, J.; Géoui, T.; Evans, D. In *The Future of the History of Chemical Information*; American Chemical Society,, McEwen, L. R., Buntrock, R. E., Eds.; ACS Symposium Series; American Chemical Society: Washington, DC, 2014; Vol. 1164; pp 127–148.

(60) Sterling, T.; Irwin, J. J. ZINC 15–Ligand Discovery for Everyone. *J. Chem. Inf. Model.* **2015**, *55*, 2324–2337.

(61) ZINC. `zinc.docking.org`, Accessed: 2020-10-20.

(62) Wang, Z.; Reibenspies, J.; Motekaitis, R. J.; Martell, A. E. Unusual stabilities of 6,6-bis(aminomethyl)-2,2-bipyridyl chelates of transition-metal ions and crystal structures of the ligand and its copper(II) and nickel(II) complexes. *J. Chem. Soc. Dalton Trans.* **1995**, 1511–1518.