CLASSIFYING HOST-GUEST TOPOLOGY WITH ION MOBILITY-MASS SPECTROMETRY AND MACHINE LEARNING

Quentin Duez[(a),*], Charlotte Lefebvre[(a)], Julien De Winter[(a)], Jérôme Cornil[(b)], Pascal Gerbaux[(a)]

[(a)] Organic Synthesis and Mass Spectrometry Laboratory, University of Mons, Place du Parc 23, 7000 Mons, Belgium.

[(b)] Laboratory for Chemistry of Novel Materials, University of Mons, Place du Parc 23, 7000 Mons, Belgium.

*Email: quentin.duez@umons.ac.be

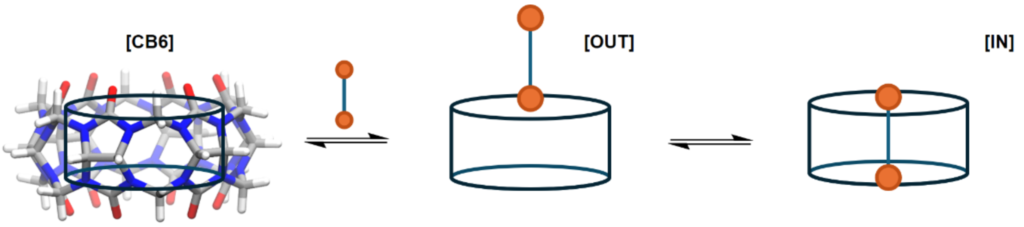
ABSTRACT: Elucidating the topology of host-guest complexes is essential for the rational design of supramolecular assemblies. Building on the recent success of data-driven approaches, we evaluate the combination of ion mobility–mass spectrometry (IMS–MS), density functional theory (DFT) featurization, and machine learning to predict and classify the binding modes of 1:1 complexes formed between cucurbit[6]uril (CB6) and diamine guests. Training a regression model with DFT-derived molecular descriptors and experimentally determined collisional cross sections (CCS) enables to predict the CCS of host-guest complexes with a diverse set of diamine guests. The predicted values naturally separate in two distinct groups corresponding respectively to inclusion and exclusion complexes, thereby enabling topology classification. This approach demonstrates that DFT-featurization and IMS–MS data capture well host–guest topology and provide a framework for the data-driven design of supramolecular assemblies.

Enzymes are generally considered to be the hallmark of reactivity under confinement, catalysing challenging reactions with remarkable efficiency and selectivity through interactions between their active pocket and the substrates.[1, 2] However, the intrinsic specificity of enzymes limits their broader applicability.[3] The demand for catalysts that retain the activity of enzymes through confinement effects, while offering a greater versatility, has drawn a significant interest in the development of artificial nanoenvironments featuring emergent behavior.[3, 4] These so-called 'molecular flasks'[3] alter chemical reactivity through confinement, either by encapsulating reactants within the cavity of containers such as cryptands,[5] cucurbiturils,[6] porphyrin cages,[7] coordination cages,[8] hydrogen-bonded capsules,[9] or by constraining them on a surface.[10] Conducting reactions within confined spaces can modulate chemical reactivity by increasing reaction rates,[3] stabilizing reactive intermediates[3] or improving reaction selectivity.[8, 11-15]

In this context, elucidating whether a guest binds within the cavity of a molecular flask presents a significant analytical challenge. For this purpose, ion mobility-mass spectrometry (IMS-MS) stands out as a particularly powerful tool.[16-18] Briefly, ion mobility spectrometry separates gaseous ions based on their mobility in a buffer gas under the influence of an electric field, which effectively enables to probe their size and shape as reflected by their collisional cross section (CCS). Combined with mass spectrometry, which gives access to the determination of complex stoichiometries based on detected *m/z* ratios, IMS-MS is a tool of choice for probing host-guest topology, typically the formation of isomeric inclusion (IN) *vs* exclusion (OUT) complexes.[19-24] Furthermore, IMS-MS combines high sensitivity with rapid analysis times, making it well-suited for detecting transient species within complex mixtures.[25] Traditionally, elucidating complex topology relies (*i*) on *a priori* knowledge of whether a given guest fits within the host cavity or; (*ii*) on comparing experimental CCS values with theoretical estimates obtained from atomistic simulations.[26-29]

In recent years, machine learning (ML) has been increasingly applied in organic chemistry to predict whether a reaction will occur[30] and to estimate reaction efficiency under varying conditions, such as changes in the substrate concentration, temperature and/or in the presence of additives.[31, 32] ML models incorporating chemically-informed descriptors - derived from density functional theory (DFT) calculations - have proven particularly valuable[33, 34] as they enable to identify key features governing reactivity trends, thus providing mechanistic insights and facilitating rational design.[35] Automated workflows leveraging the high throughput of MS and IMS-MS approaches have also been developed to screen the formation of self-assembled metal-organic complexes and coordination cages.[36, 37] Building upon the success of data-driven approaches, we hypothesized that ML could be integrated with IMS-MS measurements to classify guests based on their propensity to form inclusion or exclusion complexes with a specific host. To assess this, we sought to determine whether CCS measurements provide an adequate basis for training models able to perform such a classification.

Here, we examine the binding of protonated diamine guests to cucurbit[6]uril (CB6) as model complexes. CB6 hosts, characterized by a pumpkin-shaped structure, possess a hydrophobic cavity and carbonyl-lined portals that enable the formation of stable inclusion complexes with protonated (di)amine guests.[38, 39] Based on the proposed binding mechanism, the guest initially interacts with one of the carbonyl portals via its ammonium moiety before entering the CB6 cavity (**Scheme 1**).[40] In the case of protonated diamine guests, both terminal ammonium groups can engage with the portals, leading to the formation of interconverting 1:1 inclusion and exclusion complexes and of higher-order assemblies.[41, 42] Focusing on 1:1 complexes, we harness IMS measurements to train a ML model aiming at classifying guests by predicting the CCS of the corresponding host-guest complex. Relying on the Auto-QChem package developed by the Doyle Group, we chose to leverage chemically-informed descriptors derived from DFT calculations to capture relevant structural and electronic properties for the guest molecules.[31, 33, 35]



Scheme 1. *Representation of the binding of protonated diamine guests to a curcurbit[6]uril host (CB6). The protonated diamine is represented by a dumbbell whose orange disks correspond to ammonium groups. In the representation of CB6, grey atoms correspond to carbon atoms, blue to nitrogen atoms, red to oxygen atoms and white to hydrogen atoms.*

We first sought to construct a chemical space encompassing a diverse range of diamine guests to guide a selection of representative compounds for subsequent experiments. From a SciFinder search using the keyword 'diamine', 103 candidate molecules were arbitrarily preselected (**Figures S1-4**). DFT calculations were performed on multiple conformers for each compound, as generated by the Auto-QChem package (**Figure 1A**).[33, 35] The calculations yielded a broad set of molecular descriptors, including molar volume, hardness, dipole moment, HOMO/LUMO energies, atomic charges, and predicted nuclear magnetic resonance (NMR) shift for each atom. Dimensionality reduction was then applied to this large set of computed molecular descriptors, using either principal component analysis (PCA) or uniform manifold approximation and projection (UMAP), to visualize the associated chemical space in two dimensions (**Figure 1B** for UMAP, **Figure S7** for PCA).

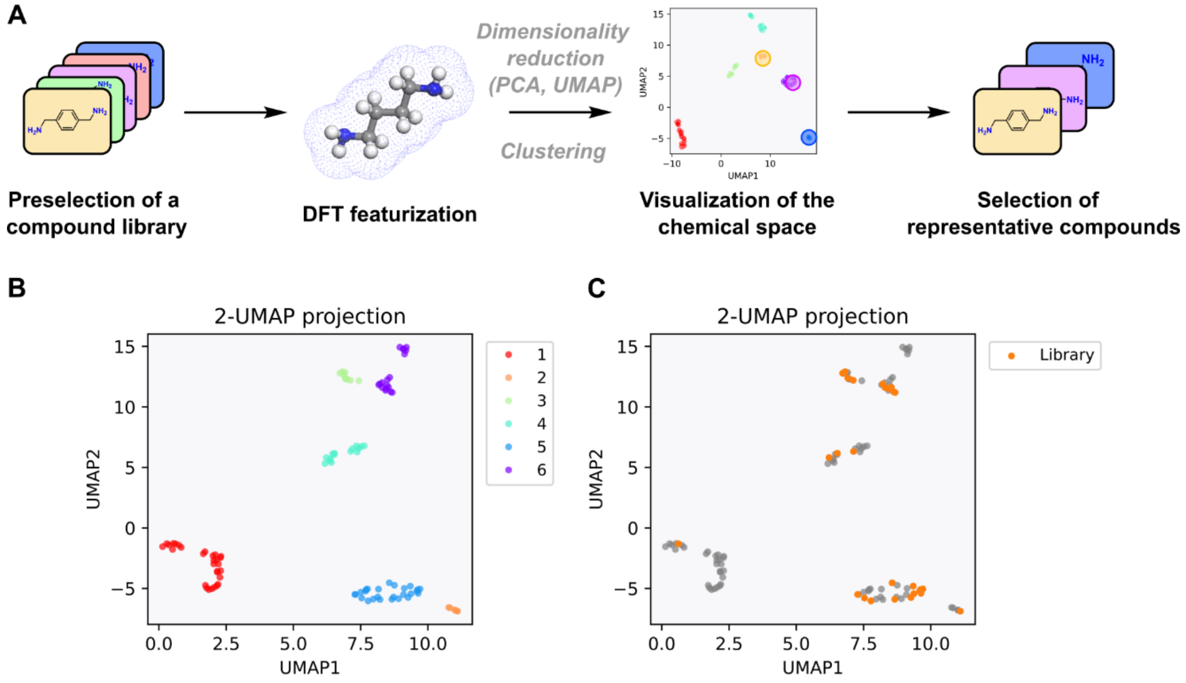


Figure 1. A. *Workflow for the construction of a chemical space from 103 candidate diamines and selection of representative compounds. Adapted from the Auto-QChem workflow in references.*[33, 35] ***B.*** *Visualization and clustering of the chemical space using uniform manifold approximation and projection (UMAP).* ***C.*** *Overlay between the chemical space and the library of selected representative diamines.*

Based on this low-dimensionality space, we then grouped the candidate diamines following agglomerative hierarchical clustering to identify representative compounds from each natural cluster. The optimal number of clusters was established by computing the Silhouette index across various levels of dimensionality reduction (**Figures S5-6**). The Silhouette index, ranging from -1 to 1, quantifies how well each data point fits within its own cluster compared to others.[43] Large positive values indicate that data points are best assigned with other members from their own group and do not match with others. On the other hand, negative values indicate that samples have been wrongly grouped together. The analysis of Silhouette indices reveals that dividing the chemical space into 6 clusters is optimal, with UMAP yielding consistently larger scores compared to PCA. The resulting clusters are visualized in two-dimensional space in **Figure 1B** for UMAP and in **Figure S7** for PCA. The molecules composing each group are shown in **Figures S8-13**. Gratifyingly, all groups exhibit distinct features to a chemist's eye, indicating that the molecular descriptors are adequate for differentiating the candidate molecules based on their structure. For instance, clusters 1 and 4 contain aromatic amines while 2 and 5 are constituted of aliphatic amines. Furthermore, clusters 1, 2 and 4 are made of primary amines while cluster 6 is primarily constituted of secondary amines.

We selected a library of 27 diamines across all 6 clusters (**Figure S14**), ensuring a comprehensive representation of the entire chemical space (see chemical space coverage in **Figure 1C** and **Figure S7**). IMS-MS experiments were then conducted on complexes between CB6 and the selected guests. To control the influence of mixing time on the equilibrium between inclusion and exclusion complexes, the host-guest adducts are generated under continuous flow conditions and analysed by ESI-MS after an arbitrarily selected mixing time. As described in the Experimental section, the complexes are continuously monitored with 1:2 host:guest stoichiometry after ~1.8 min of mixing in 80:20 H_2O:MeOH.

Representative ESI-MS spectra (**Figure 2A**) illustrate the complexation of CB6 with either 1,4-diaminobutane (C4) or N-(1-naphthyl)ethylenediamine (NED). The resulting speciation highlights distinct behaviours: for C4, only 1:1 and 1:2 guest:host complexes are detected, whereas for NED, both 1:1 and 2:1 assemblies are observed. This suggests that, despite its larger size, NED does not occupy the CB6 cavity, allowing for an additional NED molecule to bind. Even though the MS detection of larger assemblies provides information on whether the CB6 cavity is fully occupied, our discussion will focus exclusively on 1:1 host–guest complexes. As shown in **Figure 2B**, ion mobilograms of the doubly charged [CB6 + NED + 2H]$^{2+}$ ions reveal significantly longer arrival times compared to [CB6 + C4 + 2H]$^{2+}$ and empty hosts [CB6 + 2H]$^{2+}$ ions. The complexes are therefore characterized by different sizes and shapes, with the [CB6 + NED + 2H]$^{2+}$ ions characterized by a larger size than the other two ions. This observation is confirmed by the determination of CCS ($^{TW}CCS_{N2\rightarrow He}$) values from the experimental arrival time distributions (Blue, green and orange data points in **Figure 2C**).[44] The IMS data strongly suggest that NED is bound to one of the carbonyl portals outside of the CB6 cavity while C4 readily forms an inclusion complex, as the overall size of the resulting complex is nearly identical to the host alone.

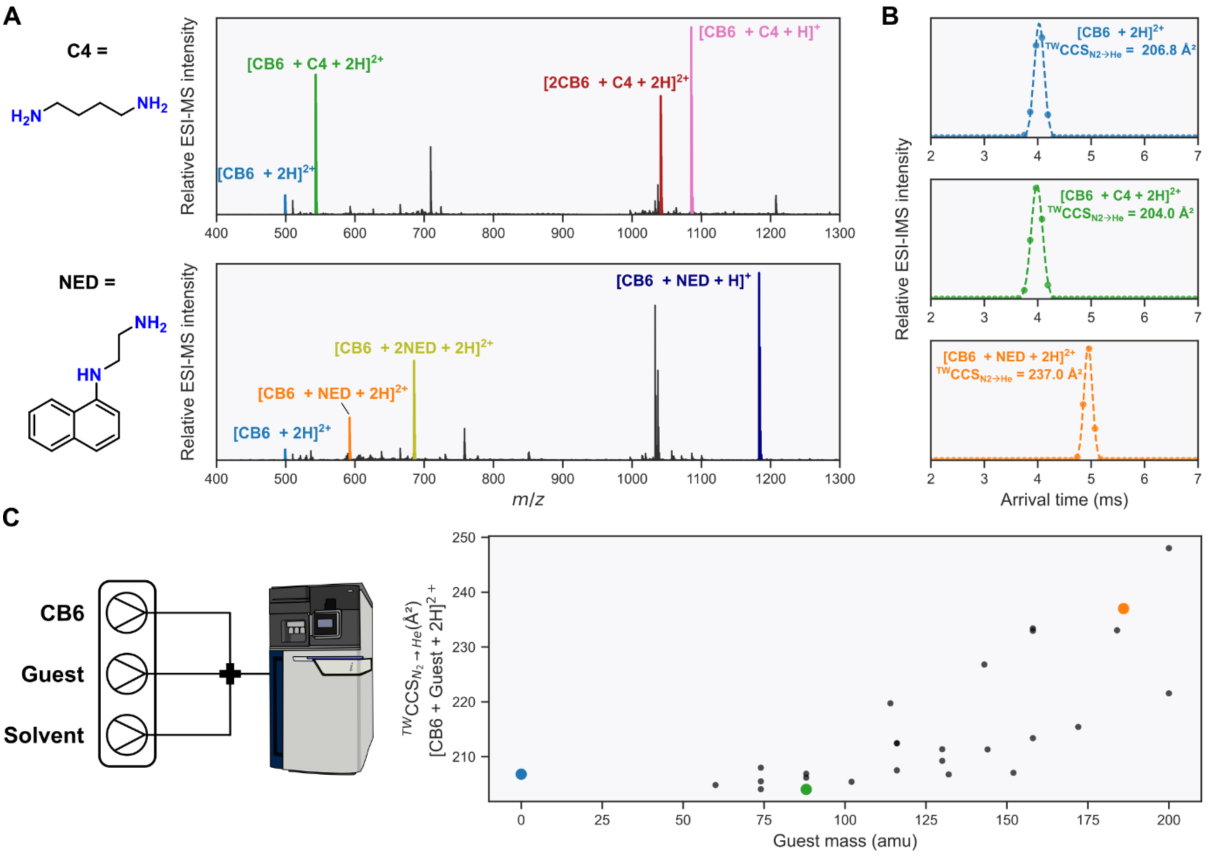


Figure 2. A. *ESI-MS spectra of 2:1 Guest:CB6 in 80:20 H_2O:MeOH measured after ~1.8 min of reaction, with the 1,4-diaminobutane (C4) or N-(1-naphthyl)ethylenediamine (NED) guests. The ions corresponding to complexes between CB6 and the guests are highlighted with colour.* ***B.*** *Ion mobilograms recorded for doubly charged ions [CB6 + 2H]$^{2+}$ and [CB6 + Guest + 2H]$^{2+}$ with the 1,4-diaminobutane or N-(1-naphthyl)ethylenediamine guests. Dots correspond to experimental data and dashed lines correspond to Gaussian fits.* ***C.*** *Left: Experimental setup interfacing a tubular flow reactor and an*

*ESI-MS instrument. Three syringe pumps independently deliver the CB6 host, the guest and solvent into a tubular reactor directly connected to the ESI source of the instrument. The reaction medium is monitored after a constant reaction time. Right: Collision cross sections ($^{TW}CCS_{N2\rightarrow He}$) measured for doubly charged ions corresponding to complexes between CB6 and various guests as a function of the guest mass. The three ions shown in panel **B** are highlighted by coloured dots and the entire dataset can be found in **Table S1**.*

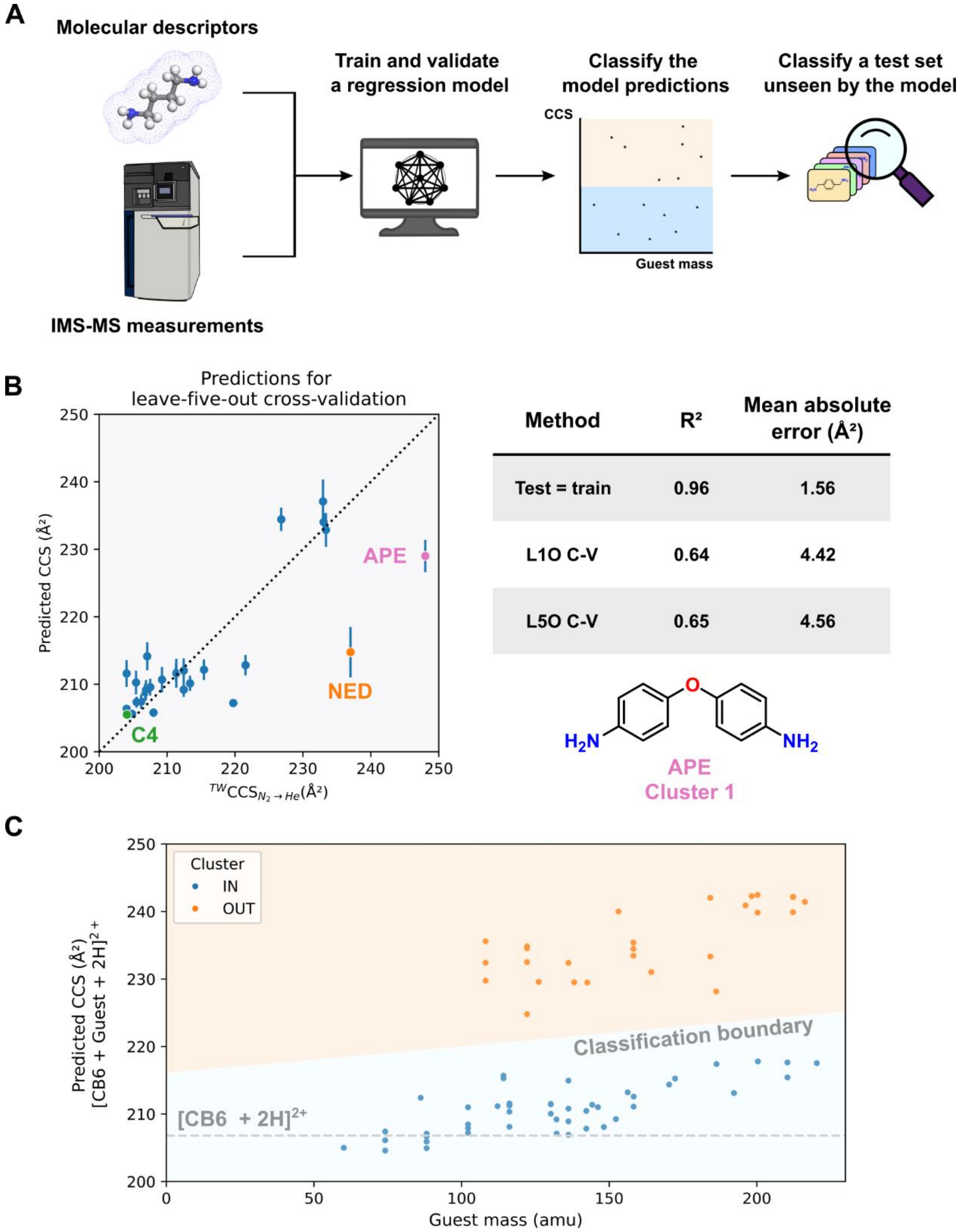
Comparing the $^{TW}CCS_{N2\rightarrow He}$ of $[CB6 + Guest + 2H]^{2+}$ ions formed by CB6 and the guest library as a function of guest mass reveals two distinct families of complexes (**Figure 2C**, full dataset in **Table S1**). Ions exhibiting similar $^{TW}CCS_{N2\rightarrow He}$ regardless of guest mass can be attributed to inclusion complexes, as the nature of the guest exerts minimal influence on the overall ion size. The ions characterized by larger $^{TW}CCS_{N2\rightarrow He}$ values (> 220 Å²) could be assigned to exclusion complexes, in which the guests remain bound to one of the carbonyl portals rather than being fully encapsulated. Based on this representation, host-guest topology may already be classified. However, this classification is ambiguous. For instance, the complex with 1,12-diaminododecane (m_{guest} = 200 Da ; $^{TW}CCS_{N2\rightarrow He}([CB6 + C12 + 2H]^{2+})$ = 221.6 Å²) exhibits a significantly higher $^{TW}CCS_{N2\rightarrow He}$ than the complex with 1,6-diaminohexane (m_{guest} = 116 Da ; $^{TW}CCS_{N2\rightarrow He}([CB6 + C6 + 2H]^{2+})$ = 207.5 Å²). This size difference could arise from either (*i*) 1,12-diaminododecane remaining outside the CB6 cavity; or (*ii*) partial encapsulation, where the guest is too large to fit completely in the cavity and extends beyond the host structure. Topology classification based on these experimental data can thus be hazardous.

To assess whether the set of molecular descriptors determined above could effectively classify complex topology based on IMS-MS data, we combined it with $^{TW}CCS_{N2\rightarrow He}$ values to train a regression model (**Figure 3A**). Prior to modelling, the DFT descriptors were pre-processed to remove low variance and correlated features, leaving 30 remaining descriptors.[33] As the size of the descriptor set was similar to that of the training set (27 guests), its dimensionality was reduced by PCA to mitigate the risk of overfitting. Reducing the descriptor set to 5 dimensions was chosen as it accounts for 90% of the total variance, as determined by PCA (**Figure S15**).

Using the 27 experimental $^{TW}CCS_{N2\rightarrow He}$ values as training set (**Figure 3A**), the model performance was validated through three train-test approaches: (*i*) training and testing on the full dataset (**Figure S16**), which constitutes an initial assessment of the model's performance but likely overestimates its accuracy since the model is evaluated on the same data it learned from; (*ii*) leave-one-out (L1O) cross-validation (**Figure S17**), where the model is trained on 26 values and tested to predict the remaining one, repeating this process across the 27 training values; and (*iii*) leave-five-out (L5O) cross-validation (**Figure 3B**), where five random values are excluded from the training set. The model is then trained on 22 values and tested to predict the remaining five, repeating this process 100 times with random train-test splits. As summarized in **Figure 3B**, the mean absolute error of the predictions is consistently lower than 5 Å², even for cross-validation, indicating that the model generalizes well to guests unseen during the training phase. Interestingly, the cross-validation predictions suggest that the CCS of complexes with N-(1-naphthyl)ethylenediamine (NED, $^{TW}CCS_{N2\rightarrow He}$ = 237.0 Å², cluster 3) and 4,4'-diaminodiphenyl ether (APE, $^{TW}CCS_{N2\rightarrow He}$ = 248.0 Å², cluster 1) are consistently underestimated when they are not included in the training set. On the contrary, the predicted CCS of complexes with 1,4-diaminobutane (C4, $^{TW}CCS_{N2\rightarrow He}$ = 204.0 Å², cluster 5) are always accurate.

We then applied the model for predicting the CCS values for $[CB6 + Guest + 2H]^{2+}$ ions across the chemical space (**Figure 1**). Since the model was trained on guests with molecular masses below 200 Da, we limited its application to guests with masses under 225 Da to avoid unreliable extrapolations for larger guests. The model yielded CCS predictions for 1:1 complexes with 75 guest molecules, which appeared to naturally separate in two distinct groups (**Figure 3C**, full dataset in **Table S2**). Agglomerative hierarchical clustering and Silhouette index analysis confirmed that grouping the predicted CCS into two clusters was optimal (**Figure S18**). We then used a classifier to provide a classification boundary between the predicted data points, which is shown in **Figure 3C**. Based on the guest mass and predicted CCS, the points below the boundary can be classified as inclusion complexes while data points above are identified as exclusion complexes. A comparison between the modelled classification boundary and the experimental CCS values is provided in **Figure S19**.

Host-guest complexes exhibiting a CCS similar to that of the free host (dashed horizontal line in **Figure 3C**) are indicative of inclusion topologies, where the guest fits entirely within the CB6 cavity and has minimal impact on the overall size of the complex. However, certain guests may only partially fit inside the host cavity and extend outward, resulting in a CCS increase as discussed above for $[CB6 + C12 + 2H]^{2+}$. The data shown in **Figure 3C** demonstrate that the regression model captures this effect, as the predicted CCS of the inclusion complexes with higher molecular weight guests are ~ 10 Å² larger than the host alone, reflecting the influence of partial protrusion.



Method	R²	Mean absolute error (Å²)
Test = train	0.96	1.56
L1O C-V	0.64	4.42
L5O C-V	0.65	4.56

Figure 3. A. *Workflow for training and testing a model for predicting the CCS of [CB6 + Guest + 2H]$^{2+}$ions based on IMS-MS measurements and molecular descriptors computed by DFT.* ***B.*** *Training and validation of the model. Left: Prediction of the model test set by leave-five-out cross-validation (L5O C-V). The L5O cross-validation predictions for complexes with 1,4-diaminobutane (C4), N-(1-naphthyl)ethylenediamine (NED) and 4,4'-diaminodiphenyl ether (APE) are highlighted with colour. Right: Model performance shown as the coefficient of determination scores (R²) and mean average errors on predicted CCS by testing the model on (i) the entire training set, (ii) leave-one-out cross-validation and (iii) leave-five-out cross-validation.* ***C.*** *Predicted CCS for 75 diamines originating from the chemical space shown in* ***Figure 1****. The data points naturally separate in two groups, corresponding to inclusion and exclusion complexes. The classification boundary is highlighted as the border between the two coloured areas, and the $^{TW}CCS_{N2\rightarrow He}$ of [CB6 + 2H]$^{2+}$ is shown as a dashed horizontal line.*

We finally evaluated model for classifying the topology of host-guest complexes involving guests that were not included in the training set. For this purpose, we excluded four guests from the training data, each forming [CB6 + Guest + 2H]$^{2+}$ species exhibiting two ion populations (**Figure 4**), suggesting the coexistence of inclusion and exclusion complexes with one population being more abundant.[27] For convenience, the IMS-MS mobilograms recorded for these complexes were converted into $^{TW}CCS_{N2\rightarrow He}$ distributions. The measurements reveal that the ion populations at higher $^{TW}CCS_{N2\rightarrow He}$ are most abundant for the guests o-phenylenediamine, 2-aminobenzylamine and N-phenylethylenediamine (**Figure 4A, B** and **D**), suggesting that the exclusion complexes are mainly produced. The opposite situation is observed for p-xylylenediamine, indicating that inclusion complexes are most abundant (**Figure 4C**).

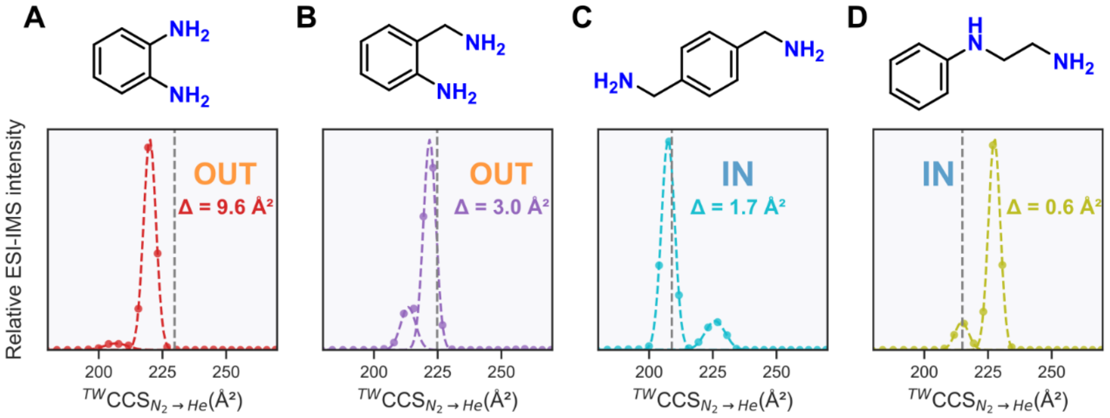


Figure 4. *Testing the model on **(A)** o-phenylenediamine, **(B)** 2-aminobenzylamine, **(C)** p-xylylenediamine and **(D)** N-phenylethylenediamine.* $^{TW}CCS_{N2\rightarrow He}$ *distributions are shown for [CB6 + Guest + 2H]$^{2+}$ involving these four guests, dots correspond to experimental data and dashed curves to Gaussian fits. CCS predicted by the model are represented by dashed lines. The model classification as inclusion (IN) or exclusion (OUT) topologies is also highlighted, along with the error between the experimental and predicted CCS values.*

When used as a test set, the model correctly classifies the topology of 1:1 complexes involving o-phenylenediamine, 2-aminobenzylamine and p-xylylenediamine based on the guest mass and predicted CCS of the host:guest complex. Moreover, CCS predictions highlighted by vertical dashed lines in **Figure 4** demonstrate the model accuracy for quantitative predictions. However, the model misclassified the complex with N-phenylethylenediamine (NED, cluster 3) as an inclusion complex (**Figure 4D**). This misclassification likely originates from the limited representation of compounds structurally similar to NED in the training set, despite four related compounds being included (cluster 3 - **Figure S10**). Furthermore, cross-validation indicated that the predicted CCS of complexes involving NED was significantly underestimated (**Figure 3B**).

To evaluate the misclassification of NED, predicted CCS values were coloured according to the cluster assignment of the corresponding guest, as shown in **Figure 1B**. This analysis, shown in **Figure S20**, reveals that compounds belonging to five of the six clusters are exclusively classified as inclusion or exclusion complexes. In contrast, compounds from cluster 3 exhibit both topologies, indicating that this cluster is not well captured by the model. This could either arise from an insufficient number of clusters, thereby grouping together structurally dissimilar compounds, or from a limited representation of cluster 3 in the training set. In the latter scenario, expanding the training set to include more structurally diverse compounds could probably improve the model accuracy for this subset of compounds. Nonetheless, the model ability in correctly classifying most of the test guests demonstrates that both DFT-derived molecular descriptors and CCS as informative physicochemical observables are adequate for characterizing and predicting the topology of host–guest complexes.

Classifying the topology of host-guest complexes often presents an analytical challenge, as it involves distinguishing subtle structural differences such as guests binding at different locations or guests partially protruding from the host cavity. Here, we developed a regression model that successfully predicts and classifies the topology of host-guest complexes formed by CB6 hosts and diamine guests by integrating DFT featurization and IMS-MS experiments. By selecting a training set composed of diamines representative of the chemical space and measuring the $^{TW}CCS_{N2\rightarrow He}$ of the corresponding 1:1 complexes, the model not only classifies known complexes but also generalizes to unseen guests.

As the host-guest complexes are analysed in continuous flow conditions, the integration of automated sample introductions with liquid chromatography injectors could significantly improve the experimental throughput, thereby expanding the chemical space coverage.[36, 45] Moreover, although this study only considered the binding of bifunctional guests to a rigid host, the proposed approach could be generalized to other complex systems, such a protein-ligand complexes or coordination cages,[46, 47] and ultimately paves the way for chemical models able to predict the topology of supramolecular assemblies based on IMS-MS measurements.

EXPERIMENTAL SECTION

Ion mobility-mass spectrometry. Mass spectrometry experiments were conducted with a Waters Synapt G2-S*i* (Wilmslow, UK) equipped with an ESI source operating in positive mode. Typical instrument parameters were a source voltage of 2.5 kV, sampling cone of 80 V, source offset of 100 V, source temperature of 100°C, desolvation temperature of 200°C. Ion mobility experiments were carried out with N_2 as buffer gas, with a Trap gas flow of 1 $mL.min^{-1}$, He gas flow of 200 $mL.min^{-1}$, IMS gas flow of 80 $mL.min^{-1}$, IMS wave velocity of 650 $m.s^{-1}$ and IMS wave height of 40 V. To mitigate space charge effects during IMS separation, the total signal was attenuated (pDRE) to 10 %. Arrival time distributions were fitted with Gaussian functions using Origin 9.0 and the apex values were converted to CCS in He ($^{TW}CCS_{N2\text{-->}He}$) following a calibration procedure described previously.[44]

Host-guest complexes were generated in continuous flow conditions using LABM8 syringe pumps (Nijmegen, The Netherlands) equipped with BD Plastipak plastic syringes. The syringes were connected to a tubular reactor made out of perfluoroalkoxy alkane (PFA - ~35 µL). All solutions were prepared in H_2O:MeOH 80:20. A 2.1 mM stock solution of CB6 was prepared in HCOOH:H_2O 50:50 and the stock diamine solutions were prepared in H_2O:MeOH 80:20. Because some diamines were poorly soluble, 1 % HCOOH was added to fully solubilize them. As shown in **Figure 2C**, the setup was constituted by three syringe pumps equipped with 3mL BD Plastipak syringes containing individually (*i*) 200 µM CB6, (*ii*) 200 µM diamine and (*iii*) H_2O:MeOH 80:20. The input flows of CB6, diamine and solvent were set to respectively 4.85, 9.7 and 4.85 $µL.min^{-1}$. The reactor outlet was directly connected to the ESI source of the MS instrument.

Computational details. DFT calculations were performed on neutral diamine molecules at the B3LYP/6-31** level[48] with the Gaussian 16 package.[49] The SCRF model was used to account for implicit water solvation.[50] DFT inputs were generated from SMILES strings corresponding to the 103 diamine compounds selected from a Sci-Finder search, using the gaussian_input_generator tool available in the Auto-QChem package.[33] To account for conformational heterogeneity, up to 5 conformers were computed for each compound. Based on their relative energies, Boltzmann weighting was applied to the molecular descriptors. The complete sets of descriptors are available in the 'global_data.xlsx' and 'atom_data.xlsx' files (see Data Availability).

Modelling details. All analysis, clustering and modelling tools relied on readily available Python packages including SciPy[51] and Scikit-learn.[52] Additional details can be found in Supporting Information and the complete workflow is reported in the related Jupyter Notebooks (see Data Availability).

AUTHOR CONTRIBUTIONS

QD conceived the research and methodology. QD and CL performed the experiments, DFT calculations and processed the data. QD wrote analysis software to construct and train the ML model based on the experimental data and wrote the initial draft of the manuscript. All authors discussed results and contributed to writing the manuscript. QD, JC and PG acquired funding.

DATA AVAILABILITY

Experimental CCS values, DFT data and Jupyter notebooks used for generating DFT inputs, parsing and processing the molecular descriptors, and constructing the ML model are available on GitHub (https://github.com/S2MOs/Topology_CCS_Predictor) and Zenodo (https://doi.org/10.5281/zenodo.15211371).

SUPPORTING INFORMATION

- Additional details for clustering and regression modelling.

- Molecular structures of all compounds selected for constructing the chemical space shown in **Figure 1** (**Figures S1 – 4**), Silhouette index as a function of the number of clusters for dimensionality reduction using UMAP and PCA (**Figures S5 – 6**), visualization and clustering of the chemical space using PCA (**Figure S7**), molecular structures of compounds in each cluster (**Figures S8 – 13**), library of compounds selected for IMS-MS experiments assigned to their respective cluster (**Figure S14**), explained variance of the descriptor set obtained from PCA analysis (**Figure S15**), training and validation of the model using the entire training set (**Figure S16**) and by leave-one-out cross-validation (**Figure S17**), Silhouette index for clustering the CCS predictions (**Figure S18**), comparison between predicted CCS and $^{TW}CCS_{N2\rightarrow He}$ (**Figure S19**), comparison between predicted CCS and cluster numbers determined by UMAP in **Figure 1** (**Figure S20**).

- Tables summarizing $^{TW}CCS_{N2\rightarrow He}$ (**Table S1**) and predicted CCS (**Table S2**).

ACKNOWLEDGEMENTS

This work was supported by Fonds National de la Recherche Scientifique (FRS-FNRS) postdoctoral fellowship to QD. JC. is a FNRS research director. Computational resources are provided by the Consortium des Équipements de Calcul Intensif (CÉCI), funded by the F.R.S.-FNRS under Grant No. U.G.018.18 and by the Walloon Region (LUCIA Tier-1 supercomputer; Grant Number 1910247). The S^2MOs and CMN labs are grateful to the FRS-FNRS for continuing support.

REFERENCES

1. A. L. Barabasi and Z. N. Oltvai, Network biology: understanding the cell's functional organization, *Nat. Rev. Genet.*, 2004, **5**, 101-113.
2. M. Cloutier and P. Wellstead, The control systems structures of energy metabolism, *J. R. Soc. Interface*, 2010, **7**, 651-665.
3. A. B. Grommet, M. Feller and R. Klajn, Chemical reactivity under nanoconfinement, *Nat. Nanotechnol.*, 2020, **15**, 256-271.
4. W. Liu and J. F. Stoddart, Emergent behavior in nanoconfined molecular containers, *Chem*, 2021, **7**, 919-947.
5. G. Qiu, P. Nava, A. Martinez and C. Colomban, A tris(benzyltriazolemethyl)amine-based cage as a CuAAC ligand tolerant to exogeneous bulky nucleophiles, *Chem. Comm.*, 2021, **57**, 2281-2284.
6. T. C. Lee, E. Kalenius, A. I. Lazar, K. I. Assaf, N. Kuhnert, C. H. Grun, J. Janis, O. A. Scherman and W. M. Nau, Chemistry inside molecular containers in the gas phase, *Nat. Chem.*, 2013, **5**, 376-382.
7. J. A. A. W. Elemans and R. J. M. Nolte, Porphyrin cage compounds based on glycoluril - from enzyme mimics to functional molecular machines, *Chem. Comm.*, 2019, **55**, 9590-9605.
8. Y. Nishioka, T. Yamaguchi, M. Yoshizawa and M. Fujita, Unusual [2+4] and [2+2] cycloadditions of arenes in the confined cavity of self-assembled cages, *J. Am. Chem. Soc.*, 2007, **129**, 7000-7001.
9. Q. Zhang, L. Catti and K. Tiefenbacher, Catalysis inside the Hexameric Resorcinarene Capsule, *Acc. Chem. Res.*, 2018, **51**, 2107-2114.
10. Q. Fu and X. Bao, Surface chemistry and catalysis confined under two-dimensional materials, *Chem. Soc. Rev.*, 2017, **46**, 1842-1874.
11. Q. Zhang and K. Tiefenbacher, Hexameric resorcinarene capsule is a Bronsted acid: investigation and application to synthesis and catalysis, *J. Am. Chem. Soc.*, 2013, **135**, 16213-16219.
12. P. Thordarson, E. J. A. Bijsterveld, A. E. Rowan and R. J. M. Nolte, Epoxidation of polybutadiene by a topologically linked catalyst, *Nature*, 2003, **424**, 915-918.
13. Q. Zhang and K. Tiefenbacher, Terpene cyclization catalysed inside a self-assembled cavity, *Nat. Chem.*, 2015, **7**, 197-202.
14. T. M. Bräuer, Q. Zhang and K. Tiefenbacher, Iminium Catalysis inside a Self-Assembled Supramolecular Capsule: Modulation of Enantiomeric Excess, *Angew. Chem. Int. Ed.*, 2016, **55**, 7698-7701.
15. T. M. Brauer, Q. Zhang and K. Tiefenbacher, Iminium Catalysis inside a Self-Assembled Supramolecular Capsule: Scope and Mechanistic Studies, *J. Am. Chem. Soc.*, 2017, **139**, 17500-17507.
16. L. Polewski, A. Springer, K. Pagel and C. A. Schalley, Gas-Phase Structural Analysis of Supramolecular Assemblies, *Acc Chem Res*, 2021, **54**, 2445-2456.
17. N. Geue, R. E. P. Winpenny and P. E. Barran, Ion Mobility Mass Spectrometry for Large Synthetic Molecules: Expanding the Analytical Toolbox, *J. Am. Chem. Soc.*, 2024, **146**, 8800-8819.
18. N. Geue, Modern Electrospray Ionization Mass Spectrometry Techniques for the Characterization of Supramolecules and Coordination Compounds, *Anal. Chem.*, 2024, **96**, 7332-7341.
19. E. Kalenius, M. Groessl and K. Rissanen, Ion mobility–mass spectrometry of supramolecular complexes and assemblies, *Nature Reviews Chemistry*, 2018, **3**, 4-14.
20. V. Gabelica and E. Marklund, Fundamentals of Ion Mobility Spectrometry, *Curr. Opin. Chem. Biol.*, 2018, **42**, 51-59.
21. V. Gabelica, A. A. Shvartsburg, C. Afonso, P. Barran, J. L. P. Benesch, C. Bleiholder, M. T. Bowers, A. Bilbao, M. F. Bush, J. L. Campbell, I. D. G. Campuzano, T. Causon, B. H. Clowers, C. S. Creaser, E. De Pauw, J. Far, F. Fernandez-Lima, J. C. Fjeldsted, K. Giles, M. Groessl, C. J. Hogan, Jr., S. Hann, H. I. Kim, R. T. Kurulugama, J. C. May, J. A. McLean, K. Pagel, K. Richardson, M. E. Ridgeway, F. Rosu, F. Sobott, K. Thalassinos, S. J. Valentine and T. Wyttenbach, Recommendations for reporting ion mobility Mass Spectrometry measurements, *Mass Spectrom. Rev.*, 2019, **38**, 291-320.
22. A. Kiesila, J. O. Moilanen, A. Kruve, C. A. Schalley, P. Barran and E. Kalenius, Anion-driven encapsulation of cationic guests inside pyridine[4]arene dimers, *Beilstein J. Org. Chem.*, 2019, **15**, 2486-2492.
23. O. H. Lloyd Williams, C. S. Cox, M. Y. Zhang, M. Lessio, O. Rusli, W. A. Donald, L. Jekimovs, D. L. Marshall, M. C. Pfrunder, B. L. J. Poad, T. Brotin and N. J. Rijs, Cation induced changes to the structure of cryptophane cages, *Dalton Trans.*, 2024, **53**, 18473-18483.
24. A. J. Arslanian and D. V. Dearden, in *Cucurbiturils and Related Macrocycles*, ed. K. Kim, Royal Society of Chemistry, 2019, DOI: 10.1039/9781788015967-00208, pp. 208-237.
25. J. Mehara and J. Roithová, Identifying reactive intermediates by mass spectrometry, *Chem. Sci.*, 2020, **11**, 11960-11972.
26. V. Lemaur, G. Carroy, F. Poussigue, F. Chirot, J. De Winter, L. Isaacs, P. Dugourd, J. Cornil and P. Gerbaux, Homotropic Allosterism: In-Depth Structural Analysis of the Gas-Phase Noncovalent Complexes Associating a Double-Cavity Cucurbit[n]uril-Type Host and Size-Selected Protonated Amino Compounds, *ChemPlusChem*, 2013, **78**, 959-969.
27. G. Carroy, C. Daxhelet, V. Lemaur, J. De Winter, E. De Pauw, J. Cornil and P. Gerbaux, Influence of Equilibration Time in Solution on the Inclusion/Exclusion Topology Ratio of Host-Guest Complexes Probed by Ion Mobility and Collision-Induced Dissociation, *Chem. Eur J.*, 2016, **22**, 4528-4534.
28. A. Kruve, K. Caprice, R. Lavendomme, J. M. Wollschlager, S. Schoder, H. V. Schroder, J. R. Nitschke, F. B. L. Cougnon and C. A. Schalley, Ion-Mobility Mass Spectrometry for the Rapid Determination of the Topology of Interlocked and Knotted Molecules, *Angew. Chem. Int. Ed.*, 2019, **58**, 11324-11328.
29. Q. Duez, S. Hoyas, T. Josse, J. Cornil, P. Gerbaux and J. De Winter, Gas-phase structure of polymer ions: Tying together theoretical approaches and ion mobility spectrometry, *Mass Spectrom. Rev.*, 2023, **42**, 1129-1151.
30. C. W. Coley, W. Jin, L. Rogers, T. F. Jamison, T. S. Jaakkola, W. H. Green, R. Barzilay and K. F. Jensen, A graph-convolutional neural network model for the prediction of chemical reactivity, *Chem. Sci.*, 2019, **10**, 370-377.
31. D. T. Ahneman, J. G. Estrada, S. Lin, S. D. Dreher and A. G. Doyle, Predicting reaction performance in C–N cross-coupling using machine learning, *Science*, 2018, **360**, 186-190.
32. N. H. Angello, V. Rathore, W. Beker, A. Wołos, E. R. Jira, R. Roszak, T. C. Wu, C. M. Schroeder, A. Aspuru-Guzik, B. A. Grzybowski and M. D. Burke, Closed-loop optimization of general reaction conditions for heteroaryl Suzuki-Miyaura coupling, *Science*, 2022, **378**, 399-405.

33. A. M. Żurański, J. Y. Wang, B. J. Shields and A. G. Doyle, Auto-QChem: an automated workflow for the generation and storage of DFT calculations for organic molecules, *React. Chem. Eng.*, 2022, **7**, 1276-1284.
34. P. Raghavan, B. C. Haas, M. E. Ruos, J. Schleinitz, A. G. Doyle, S. E. Reisman, M. S. Sigman and C. W. Coley, Dataset Design for Building Models of Chemical Reactivity, *ACS Cent. Sci.*, 2023, **9**, 2196-2204.
35. S. K. Kariofillis, S. Jiang, A. M. Zuranski, S. S. Gandhi, J. I. Martinez Alvarado and A. G. Doyle, Using Data Science To Guide Aryl Bromide Substrate Scope Analysis in a Ni/Photoredox-Catalyzed Cross-Coupling with Acetals as Alcohol-Derived Radical Sources, *J. Am. Chem. Soc.*, 2022, **144**, 1045-1055.
36. O. H. L. Williams, O. Rusli, L. Ezzedinloo, T. M. Dodgen, J. K. Clegg and N. J. Rijs, Automated Structural Activity Screening of beta-Diketonate Assemblies with High-Throughput Ion Mobility-Mass Spectrometry, *Angew. Chem. Int. Ed.*, 2024, **63**, e202313892.
37. A. R. Basford, A. H. Bernardino, P. C. P. Teeuwen, B. D. Egleston, J. Humphreys, K. E. Jelfs, J. R. Nitschke, I. A. Riddell and R. L. Greenaway, Development of an Automated Workflow for Screening the Assembly and Host-Guest Behavior of Metal-Organic Cages Towards Accelerated Discovery, *Angew. Chem. Int. Ed.*, 2025, **64**, e202424270.
38. W. L. Mock and N. Y. Shih, Structure and Selectivity in Host-Guest Complexes of Cucurbituril, *J. Org. Chem.*, 1986, **51**, 4440-4446.
39. H. J. Buschmann, L. Mutihac, K. Jansen and E. Schollmeyer, Cucurbit[6]uril as Ligand for the Complexation of Diamines, Diazacrown Ethers and Cryptands in Aqueous Formic Acid, *J. Incl. Phenom. Macrocycl. Chem.*, 2005, **53**, 281-284.
40. C. Márquez, R. R. Hudgins and W. M. Nau, Mechanism of Host-Guest Complexation by Cucurbituril, *J. Am. Chem. Soc.*, 2004, **126**, 5806-5816.
41. E. Masson, X. Ling, R. Joseph, L. Kyeremeh-Mensah and X. Lu, Cucurbituril chemistry: a tale of supramolecular success, *RSC Adv.*, 2012, **2**, 1213-1247.
42. X. Yang, R. Wang, A. Kermagoret and D. Bardelang, Oligomeric Cucurbituril Complexes: from Peculiar Assemblies to Emerging Applications, *Angew. Chem.*, 2020, **59**, 21280-21292.
43. P. J. Rousseeuw, Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, *Journal of Computational and Applied Mathematics*, 1987, **20**, 53-65.
44. Q. Duez, F. Chirot, R. Liénard, T. Josse, C. M. Choi, O. Coulembier, P. Dugourd, J. Cornil, P. Gerbaux and J. De Winter, Polymers for Traveling Wave Ion Mobility Spectrometry Calibration, *J. Am. Soc. Mass. Spectrom.*, 2017, **28**, 2483-2491.
45. D. Perera, J. W. Tucker, S. Brahmbhatt, C. J. Helal, A. Chong, W. Farrell, P. Richardson and N. W. Sach, A platform for automated nanomole-scale reaction screening and micromole-scale synthesis in flow, *Science*, 2018, **359**, 429-434.
46. D. Samanta, J. Gemen, Z. Chu, Y. Diskin-Posner, L. J. W. Shimon and R. Klajn, Reversible photoswitching of encapsulated azobenzenes in water, *Proc. Natl. Acad. Sci. U.S.A.*, 2018, **115**, 9379-9384.
47. J. Gemen, J. R. Church, T. P. Ruoko, N. Durandin, M. J. Białek, M. Weißenfels, M. Feller, M. Kazes, M. Odaybat, V. A. Borin, R. Kalepu, Y. Diskin-Posner, D. Oron, M. J. Fuchter, A. Priimagi, I. Schapiro and R. Klajn, Disequilibrating azobenzenes by visible-light sensitization under confinement, *Science*, 2023, **381**, 1357-1363.
48. A. D. Becke, A new mixing of Hartree–Fock and local density-functional theories, *The Journal of Chemical Physics*, 1993, **98**, 1372-1377.
49. G. W. T. M. J. Frisch, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman, D. J. Fox, Gaussian 16, *Gaussian 16*, 2016.
50. J. Tomasi, B. Mennucci and R. Cammi, Quantum mechanical continuum solvation models, *Chem. Rev.*, 2005, **105**, 2999-3093.
51. P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, I. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt and C. SciPy, SciPy 1.0: fundamental algorithms for scientific computing in Python, *Nat. Meth.*, 2020, **17**, 261-272.
52. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and É. Duchesnay, Scikit-learn: Machine Learning in Python, *J. Mach. Learn. Res.*, 2011, **12**, 2825–2830.

TABLE OF CONTENTS GRAPHIC