Imperial College London
Department of Chemistry

Computational modelling and topological design of porous organic cages

Valentina Santolini

Supervised by Dr Kim E. Jelfs

Submitted in part fulfilment of the requirements for the degree of
Doctor of Philosophy in Chemistry of Imperial College London
and the Diploma of Imperial College London

Declaration of originality

I hereby certify that the material of this thesis, which I now submit for the award of Doctor of Philosophy of Imperial College London, is entirely the result of my own work unless otherwise cited or acknowledged.

Valentina Santolini, March 2018

Copyright declaration

The copyright of this thesis rests with the author and is made available under a Creative Commons Attribution Non-Commercial No Derivatives licence. Researchers are free to copy, distribute or transmit the thesis on the condition that they attribute it, that they do not use it for commercial purposes and that they do not alter, transform or build upon it. For any reuse or redistribution, researchers must make clear to others the licence terms of this work.

Abstract

Porous organic cages are a class of materials constructed by intermolecular packing of discrete organic molecules, that contain void space in an internal cavity, and are typically solution processable. Many experimental and computational studies have been conducted on porous cages, however design and prediction of new molecular structures still represents a great theoretical challenge.

During my PhD, I developed a computational strategy to generate new porous organic molecules and test their experimental feasibility. An automated pipeline assembles building blocks with different numbers of reactive ends into cages with the underlying topology of geometric polyhedra. The topologies are defined and classified according to a new nomenclature, which is based on the topicity of the precursors contained in each cage. Once multiple topologies are generated from a pair of precursors, the geometry and energy of each molecule are investigated to understand which structure is the most experimentally likely to form. The structures that do not retain a cavity are "inflated" with a constrained molecular dynamics technique that simulate the scaffolding effect of the solvent, which is otherwise not taken into account.

The prediction strategy has been successfully tested on a number of experimentally available structures, and on a wide range of novel organic cages in the context of a computational-experimental collaboration with the group of Prof. A. Cooper (University of Liverpool). The experimental part was carried out by the Cooper's group, who used a high-throughput screening platform to carry out the synthesis of 78 new cages, starting from a selection of different functional precursors mixed in a combinatorial way. I carried out the computational part of the project, which was dedicated to the prediction of the same cages, and results were compared to rationalise the criteria for optimal design of porous organic cages. The pipeline can be now integrated with crystal structure prediction methods to suggest promising

novel porous organic materials for the synthesis.

List of publications resulting from this thesis

1. R. L. Greenaway, **V. Santolini**, M. J. Bennison, B. M. Alston, C. J. Pugh, M. A. Little, M. Miklitz, E. G. B. Eden, R. Clowes, A. Shakil, H. J. Cuthbertson, H. Armstrong, M. E. Briggs, K. E. Jelfs and A. I. Cooper, "High-Throughput Workflow for the Discovery and Characterisation of Organic Cages", Submitted

2. C.J. Pugh, **V. Santolini**, R. L. Greenaway, M. A. Little, M. E. Briggs, K. E. Jelfs and A. I. Cooper, "Solvent-mediated re-equilibration to form a large, covalent organic cage", *Crystal Growth and Design*, 2018, 18, 2759-2764

3. T. Hasell, M. A. Little, S. Y. Chong, M. Schmidtmann, M. E. Briggs, **V. Santolini**, K. E. Jelfs and A. I. Cooper, "Chirality as a tool for function in porous organic cages", *Nanoscale*, 2017, 9, 6783-6790

4. **V. Santolini**, M. Miklitz, E. Berardo, K.E. Jelfs, "Topological landscapes of porous organic cages", *Nanoscale*, 2017, 9, 5280-5298

5. P.S. Reiss, M.A. Little, **V. Santolini**, S Y. Chong, T. Hasell, K. E. Jelfs, M. E. Briggs and A. I. Cooper, "Periphery-functionalized porous organic cages", *Chemistry - a European Journal*, 2016, 22, 16547-16553

6. **V. Santolini**, G.A. Tribello, K.E. Jelfs, "Predicting solvent effects on the structure of porous organic molecules", *Chemical Communications*, 2015, 51, 15542-15545

Recognitions

- **2017** Royal Society of Chemistry Travel Grant (Materials Division)
- **2017** Recognition at Presidents Awards (Imperial College London)
- **2016** Talk Prize - Supramolecular Symposium (Southampton)
- **2016** Poster Prize - British Zeolites Association Meeting (Bath)
- **2014** Talk Prize - CCP5 Summer School (Manchester)

Acknowledgments

This thesis travelled with me in the past few months across Italy, continental Europe, and the UK. It was mainly written in three different countries, nine different cities, and eight different apartments (it's for the best thesis nowadays are digital, or half would have gone lost with all the moving). Each chapter, each paragraph, reminds me of a place and of the people I was sharing that space with, I'll try to thank them all.

First of all I would like to thank my supervisor, Kim, for all that she taught me and for what we learned together. I'm very happy about the scientific relationship we developed, and our eternal discussions about geometry, polyhedra and solids will stay with me. Thank you for your patience, support and guidance, in particular during the past few months of writing up. A very big thanks goes to the group of Prof Andy Cooper in Liverpool, and especially to Dr. Becky Greenaway, for the past few years of collaboration. I learned so much from all of you. This PhD wouldn't have been as much fun without the Jelfs' group, thank you Marcin, James and Ed for the company and the interesting discussions. Enrico, if I had to thank you for all the things you did for me I'd have to cover more than my longest chapter. I couldn't have wished for a better desk buddy and friend, I can't imagine a job in which you're not in a desk next to mine. Thank you to everyone in the office, and in particular to Claudia and Anna Sofia for making it such a pleasant place to spend one's days (although, yes, windows and natural light would be an improvement). For making my years at Imperial memorable I would like to thank Morgane, who never ceased to show support, love and encouragement, Will and João, who always had a wise perspective on things. Bruno, I miss our chats in the Queen's lawn, you are my fantastic seagull friend, and I hope we'll end up living again in the same city. And that there'll be a lawn between us. Eleonora, I wouldn't have written this

thesis if you had not dragged me to Swiss Cottage Library every single day. Thank you for giving me strength and motivation. And salted cashews! I look forward to the next coffee overlooking an olympic swimming pool.

Outside of Imperial, London has allowed me to pursue the vast majority of my creative dreams (and there were many), and to become a more cultured and open person. My years there were shared with a bunch of people I wish I could put in my pocket and always carry with me; now that I moved, my life is a bit emptier without them around. Christin, Marta, Lucie, you are a TerrificTrio, the happiest part of my brain is filled with memories of our dinners, our laughs, our hiking (and less-hiking more-lazy) trips and the "early" morning breakfasts at the farm. Danke, obrigada, and merci, I wouldn't have reached the end of this if it was not for your warmth, empathy and intelligence. Rossana, finding you again in London and living a bit Bologna style has been the best cure for homesickness. Thank you for all the time you made me pasta with zucchine.. and mainly for that time I was stressed and tired, and you gave me a bowl of mascarpone and nutella saying "Dip in". Maria, you are a continuous source of inspiration, grazie for being so full of contagious energy and happiness and for supporting me through the years. One day we will have our neighbouring houses and live happily ever after. Silvia, thank you for taking care of me so much during the writing up, and for making me caffe' d'orzo con un biscottino every time I got stuck and spent hours staring at a blank page. Thanks to Mouse & The Lotz for making the softest pancakes in London. Thesis and pancakes is so much better than plain thesis.

Moving into a new city where you don't know anyone can be hard, but I've been so lucky so as to meet fantastic people right at the start. My warmest thank you (gracias) goes to Dr. Aurora Cruz-Cabeza, who welcomed me in Manchester and made a place for me in her group. Having somewhere to go everyday to write has been fundamental to finish. Thanks to all the members of room D37, and in particular to Riccardo and Giulio, who made my days so nice I've almost forgotten (for hours at a time!) one is supposed to be stressed towards the end. Blair, Parvinder and Jaiya, I'm so glad we met you guys, and thank you for the Sunday roast on my last day of writing up. Each spoon of gravy was a bit of stress flowing away.

Last summer I had to move back to Bologna for a while and, although in a moment of great sadness, many people there gave me long moments of peace and delight. Thank you Valeria and Andrea for taking me to Sala Borsa and for helping me to write the first few sentences of this work, it unlocked the mechanism, and sharing

library time with you brought me back in time. Thanks to Marco for being always present with empathy and optimism, every time we meet again it's as if no time has passed and we're still 18. Thank you Eleonora for giving me determination, advice and cake. I can't think of a better combination, really. Thank you for your patience and for helping me through difficult times. Gabriela, you are the sweetest person, you brightened up my days with messages of encouragement.

Thank you Alessandra e Antonello, my AMGA and APGA, for spoiling me for a week in the magic Signor Congedo's house. The view from the gate, the grilled calamari, a starry night with guitars are cosy lights in my memory for when outside it's raining and the mood is a bit gloomy. Thank you Val, Malcolm and Monique for your enthusiastic support through the years, for meals, teas and literary discussions. A special acknowledgement goes to Malcolm, who went through the painful process of proof reading my Itanglian. Grazie!

Federica, Blanche, Lia, and Giulia, you four are my little pillars (as I think I said already many years ago now) and my family. Thank you for keeping me safe and warm. Thank you to my grandparents, Elsa and Renato, for holding me tight in their thoughts, and to my favourite aunt, Fab, for seeing the bright side of things and for making me see it as well. My parents built a perfect world for the three of us, where love, open-mindedness and happiness were part of every day of my life with them. I'm not good at describing things with words, but I know you know how important that has been for me, and how privileged I feel to have had such a happy time growing up. My dad is not here today to see that this work is finally finished. But he is in every word, in every page. And he will keep being, in every thought and every action, a kind, strong, happy voice playing in my ears. My mum is here, and she's been wonderful and positive and supportive. The two of us will carry on, travelling, enjoying life and eating avocado and poached eggs.

My last (not for importance) acknowledgement is for Clyde, who has a bright imagination, and draws fantastic imaginary worlds around me. He transforms every day into a story. If we survived writing up at the same time in a 40 squared meters flat (and then in a 4 squared meters room) it's because with our minds we were always somewhere more exotic. Your love led me here in one piece, you made my life more colourful and full of wonder.

For Claudio, my wonderful dad.

Contents

Abstract **5**

List of publications **7**

Recognitions **8**

Acknowledgements **9**

Contents **12**

List of Tables **17**

List of Figures **19**

List of Abbreviations **33**

Colour Code for Molecules **35**

1. Introduction **36**

1.1. Scope of the work 36

1.2. Porous materials and their properties 38

1.3. Porous organic cages 40

1.3.1. Historical background 42

1.3.2. Synthesis of porous organic cages 44

1.4. Porous organic cages topologies and challenges involved in their prediction 48

1.4.1. Nomenclature 56

1.4.2. Factors influencing the design of porous organic cages 57

1.5. Properties and characterisation of the solid state . . . 61
1.5.1. Experimental characterisation . . . 61
1.5.2. Properties of the solid state . . . 62
1.5.3. Computational investigation of the solid state . . . 68
1.5.4. Computational investigation on single molecules . . . 73
1.6. Structure of the thesis . . . 75

2. Methods and theoretical background 76
2.1. Potential energy surfaces . . . 76
2.1.1. Geometry optimisation . . . 78
2.2. Exploring the potential energy surface . . . 80
2.2.1. Forcefield Methods . . . 80
2.2.2. Conformer search with the low-mode search method. . . . 83
2.2.3. Molecular Dynamics . . . 84
2.3. Electronic structure theory . . . 86
2.3.1. Wave function methods . . . 87
2.3.2. Density Functional Theory . . . 89
2.3.3. Basis sets . . . 94
2.4. General pipeline adopted in this thesis . . . 97
2.4.1. Comparing the energy . . . 98
2.4.2. Void size calculation . . . 99

3. Predicting solvent effects 100
3.1. Introduction . . . 100
3.2. Outline of the work . . . 103
3.3. Methods . . . 104
3.3.1. Conformer search calculations to locate collapsed structures. . 104
3.3.2. Initial attempts to explore open molecular conformations . . . 105
3.3.3. Constrained MD simulations to locate metastable open molecular conformations. . . . 106
3.3.4. Constrained MD calculations with a cylindrical probe. . . . 108
3.3.5. Determination of spherical void size. . . . 109
3.4. Results . . . 109
3.4.1. **CC8.** . . . 109
3.4.2. Mastalerz endo-functionalised cage. . . . 112
3.4.3. [6+9] and [8+12] versions of the **CC3** imine cage. . . . 114
3.4.4. Macrocycles and cavitands. . . . 118

3.5. Conclusions . . . 122

4. Topological families and nomenclature of porous organic cages 124
4.1. Introduction . . . 125
4.2. A new nomenclature for porous organic cage topologies. . . . 125
4.2.1. Enumeration of organic cage topologies . . . 125
4.3. Topological families . . . 128
4.3.1. Tritopic + Ditopic topology family . . . 130
4.3.2. Tetratopic + Ditopic topology family . . . 134
4.3.3. Tritopic + Tritopic topology family . . . 137
4.3.4. Tetratopic + Tritopic topology family . . . 139
4.4. Conclusions . . . 141

5. Topological prediction of porous organic cages. 142
5.1. Introduction . . . 142
5.2. The prediction pipeline. . . . 143
5.3. Methods involved in the prediction pipeline. . . . 146
5.3.1. Methods benchmarking . . . 146
5.3.2. Conformer search simulations . . . 153
5.3.3. DFT methods . . . 153
5.3.4. Comparison of the energies . . . 154
5.3.5. Computational challenges for topology prediction. . . . 154
5.4. Results . . . 155
5.4.1. The Tritopic + Ditopic family . . . 156
5.4.2. The Tritopic + Tritopic family . . . 162
5.4.3. The Tetratopic + Ditopic family . . . 164
5.4.4. The Tetratopic + Tritopic family . . . 168
5.4.5. Hypothetical larger molecules . . . 170
5.5. Conclusions . . . 171

6. High-Throughput screening of porous organic cages 172
6.1. Introduction . . . 172
6.1.1. Overview of the project . . . 173
6.2. Methods . . . 182
6.3. Results . . . 184
6.3.1. Topology prediction for cages B1, B11, B23 . . . 184
6.3.2. HT screening - discussion of the results. . . . 190

6.3.3. Cages with $\mathbf{Tri}^2\mathbf{Di}^3$ topology . . . 193
6.3.4. Cages with $\mathbf{Tri}^4\mathbf{Di}^6$ topology . . . 197
6.3.5. Cages with $\mathbf{Tri}^4\mathbf{Tri}^4$ topology . . . 207
6.3.6. Criteria for the synthesis of promising candidates . . . 211
6.3.7. A serendipitous discovery, the pretzelane topology . . . 212
6.4. Conclusions . . . 213

7. Conclusions and future work **216**

Appendices **223**

Appendix A. Constrained MD input files **223**
A.1. Input files for Constrained MD calculations . . . 223
A.1.1. Spherical probe . . . 223
A.1.2. Cylindrical probe . . . 224

Bibliography **225**

List of Tables

3.1. Comparison of the relative energy of the lowest energy collapsed and open structures for each system, calculated using OPLS and given in kJ mol^{-1}. 123

4.1. Table of possible organic cage topologies and their key features. Ring (representing windows in the cages) sizes are defined as a^b, where a is the number of individual BBs that form part of the ring and b is the number of rings of size a 129
4.2. Experimentally reported organic cage topologies, synthesised via two-component DCC reactions. **Di** stands for ditopic precursors, **Tri** for tritopic, **Tet** for tetratopic. 130

5.1. Information on the lowest energy open conformations of the **CC3**-based Tritopic + Ditopic topologies. 157
5.2. Information on the lowest energy open conformations of the **CC5**- and **CC8**-based Tritopic + Ditopic topologies. 160
5.3. Information on the lowest energy open conformations of the **CC11**-based Tritopic + Tritopic topologies. 163
5.4. Information on the lowest energy open conformations of the Warmuth Tetratopic + Ditopic topologies. Values in brackets refer to formation energies per bond formed (kJ mol^{-1}), which were only calculated for cages $\mathbf{Tet}^2\mathbf{Di}^4$ A, B, and C. 165
5.5. Information about the window sizes of $\mathbf{Tet}_2^4\mathbf{Di}^8$, $\mathbf{Tet}^6\mathbf{Di}^{12}$ and $\mathbf{Tet}^8\mathbf{Di}^{16}$. 167
5.6. Information on the lowest energy open conformations of the Warmuth Tetratopic + Tritopic topologies. 169

6.1. Formation energies per bond formed, strain values and void diameters of cages with **Tri**^{2}**Di**^{3} topology (A/B/C 1-10). Orange cells refer to simulated structures that were very distorted at the end of the computational procedure; cyan cells highlight structures with a strain percentage greater than 2. 195
6.2. Formation energies per bond formed, strain values and void diameters of cages with **Tri**^{4}**Di**^{6} topology (A/B/C 11-19). 198
6.3. Formation energies, strain values and void diameters for the **Tri**^{2}**Di**^{3} and **Tri**^{4}**Di**^{6} topologies of cages (A/C)12, (B/C)14, (B/C)20 and (B/C)21. Cells in cyan highlight cages with a strain percentage value >2. 203
6.4. Formation energies per bond formed, strain values and void diameters of cages with **Tri**^{4}**Tri**^{4} topology (A/B/C 22-26). 208
6.5. Formation energies, strain values and void diameters for B13-**Tri**^{4}**Di**^{6} and B13-pretzelane 213

List of Figures

1.1. An example of a porous molecular organic material representing the crystal packing of Covalent Cage 3 (**CC3**).Three cages are differently coloured in blue, teal, and yellow to highlight the non-covalent nature of the packing. The same cages contain grey spheres to highlight the presence of a cavity. 36

1.2. Workflow integrating computational and experimental methods for the successful discovery of new porous organic materials. Steps 1-2-3: computational investigation of hypothetical single cages assembled from a pair of organic precursors; Step 4: promising candidates from step 3 are assembled into the solid state with crystal structure prediction methods; Step 5: promising candidates (whose properties can be studied with computational simulations) from step 4 are suggested for the synthesis. The colours of the cages and of the precursors do not reflect the atomic composition. . . 38

1.3. Examples of porous network materials, from left to right, a zeolite with Linde Type A (LTA) topology, MOF-5and COF-5.Zinc is in orange, silicon in beige. The images of the three systems are not to scale. 39

1.4. An example of extrinsic (grey) and intrinsic (cyan) porosity connected by the window-to-window arrangement of a selection of three **CC3** cages. 41

1.5. Geometries of the Dianin (Cambridge Structural Database (CSD) reference: BEGSOW), cryptand (CSD reference: ACEZUG), hemicarcerand (CSD reference: RAGXAY) compounds and of one imine cage from Warmuth and co-workers (no crystal structure available). . 43

1.6. Example of an imine condensation reaction between aldehydes and imines, in this instance for the synthesis of cages **CC1**, **CC2**, **CC3**, **CC4** and **CC13** 46
1.7. Examples of imine and boronic cages. Box A includes (from left to right and from top to bottom) cages **CC1**, **CC2**, **CC3**, **CC5**, and **CC13** synthesised by Cooper and co-workers. Box B includes (from top to bottom) three porous imine cages with the shape of a squared anti-prism, a distorted tetrahedron and an octahedron synthesised by Warmuth and co-workers. Box C contains two triptycene-based capsules (top row) and two triptycene-based tetrahderal (bottom row) molecules formed by Mastalerz and co-workers. Finally, box D contains two examples of boronic ester cubes synthesised by Mastalerz and co-workers (left) and Beuerle and co-workers (right). Crystal structures were identified for all cages, except for those synthesised by Warmuth and co-workers. 47
1.8. Platonic, Archimedean and Catalan polyhedra that can be related to organic cage topologies. 1) Tetrahedron, 2) Trigonal Prism, 3) Cube, 4) Octahedron, 5) Dodecahedron, 6) Square Antiprism, 7) Rhombic Dodecahedron, 8) Cuboctahedron, 9) Rhombicuboctahedron. 49
1.9. Selection of polyhedra that can describe the [6+8] organic cage synthesised by Warmuth and co-workers.1) a Rhombicuboctahedron, 2) an Octahedron, 3) a Cube 4) a Rhombic Dodecahedron. 50
1.10. Example of difference between geometrical shape and topology. The mug and the doughnut have the same topology (characterised by the presence of a hole), but different geometrical shapes, which allow one to describe them as "mug" and "doughnut". Image adapted from K. Crane, U. Pinkall and P. Schröder, *ACM Trans. Graph.*, 2013, 32-42. 51
1.11. An example of how two porous organic cages that both have an underlying topology of a tetrahedron adopt different geometric shapes, in one case maintaining a tetrahedron shape(bottom row), and in the other adopting an octahedral shape (top row).The underlying topologies are shown in purple and the geometric shapes in orange. 51
1.12. Example of emergent behaviour: mixing the central trialdehyde with cyclopentadiamine (top row) leads to the formation of the porous tetrahedral **CC5** imine cage and with cyclohexenediamine to the non-porous larger **CC8** imine cage (bottom row). 53

1.13. Example of emergent behaviour: mixing a trialdehyde with a diamine containing an even number of carbons (cyan) leads to the formation of tetrahedral cages (top), but mixing it with a diamine containing an odd number of carbons (yellow) leads to the formation of capsule-shaped cages (bottom). 54
1.14. The three different cages found by Warmuth and co-workers. The presence of chloroform led to the formation of an octahedral cage, of THF of a distorted [4+8] tetrahedral cage and of dichloromethane to a [8+16] square antiprismatic structure. 55
1.15. Examples of ditopic (**Di**), tritopic (**Tri**) and tetratopic (**Tet**) BBs reported for the synthesis of organic cages. Starred ditopic precursors have been used for the synthesis of boronate cages, where two hydroxy groups bond to one boron atom. Thus, I still labelled these precursors as ditopic rather than tetratopic. 58
1.16. Reversible on-off porosity switching in CC1 crystals can be triggered by exposure to specific solvents followed by guest removal. Exposure to ethyl acetate vapor and desolvation leads to the non-porous polymorph **CC1**-α; DCM exposure and evacuation affords **CC1**-β, which is non-porous to N_2, but adsorbs H_2. **CC1**-α cages are in yellow, whereas **CC1**-β are in blue. 63
1.17. A) Two cage molecules, **CC3** and **CC4**, can adopt isoreticular polymorphs by recrystallization in the presence of specific solvents. B) α-polymorphs of cages **CC2** and **CC13** and the β-forms synthesized by recrystallization in the presence of 1,4-dioxane, which makes them isoreticular to **CC3**-α. Figure adapted from Little, M. A.; Chong, S. Y.; Schmidtmann, M.; Hasell, T.; Cooper, A. I. *Chem. Comm.*, 2014, 50, 9465-9468, published by The Royal Society of Chemistry. 65
1.18. Schematic representation of the formation of racemic and quasiracemic imine cage cocrystals. Each cage is neighboured by four cages of the opposite chirality. A chemical representation of the **CC3**-R and **CC1**-S co-crystal is shown on the left hand side, with **CC3**-R in yellow and **CC1**-S in blue. 66

1.19. Box: chemical representation of cages **TCC**(1/2/3), including a 2-dimensional representation of the different aldehydes contained in each molecule. The aldehydes' carbons are in grey, amines' carbons in red, nitrogens in blue. Right: Crystal structure of cages **TCC2**-*R* and **CC3**-*S* consisting of layers of **CC3** molecules packed and pillared by the **TCC2** cages. Red and cyan sticks simplify, for clarity, the cyclohexyl groups, red sticks belonging to molecules of *R* chirality, cyan to *S* chirality. The intrinsic pore network is represented in yellow. The image has been adapted from A. G. Slater, M. A. Little, A. Pulido, S. Y. Chong, D. Holden, L. Chen, C. Morgan, X. Wu, G. Cheng, R. Clowes, M. E. Briggs, T. Hasell, K. E. Jelfs, G. M. Day and A. I. Cooper, *Nature Chemistry*, 2017, 9, 17-25. 67

2.1. One-dimensional representation of a potential energy surface. 77
2.2. General work-flow followed for the computational investigation of organic cages. Cyan boxes are for steps based on force field methods, orange on electronic structure theory. The watch-icons refer to the relative time required on average to complete each step. 98

3.1. Comparison of the collapsed (left hand side) and open (right hand side) geometries of **CC8**. Throughout the chapter, grey spheres are to highlight the presence of a cavity. The open geometry has a void radius of 4.6 Å. 101
3.2. Reaction scheme of the molecules inflated with constrained MD. . . . 104
3.3. Relative energies OPLS of the sampled structures (a constraint radius of 5 Å is applied in this case) before (red) and after (blue) geometry minimisation for the [8+12] version of **CC3**. 107
3.4. Relative energies of the lowest energy open conformers of **CC8**, calculated at the OPLS_2005 level and obtained with different probe sizes. 110
3.5. Overlays of **CC8**: A) the computed open structure (blue) and the solvate crystal structure (red); B) comparison of the computed open structure (blue) and the OPLS-minimised solvate crystal structure (dark red). 111
3.6. Comparison of the collapsed (left hand side) and open (right hand side) geometries of Mastalerz endo-functionalised cage. The cavity presents a void radius of 5.2 Å. 112

3.7. Relative energies of the lowest energy open conformers of the Mastalerz endo-functionalised cage, calculated at the OPLS_2005 level, and obtained with different probe sizes. 113
3.8. Overlays of the Mastalerz endo-functionalised cage: A) the computed open structure (blue) and the solvate crystal structure (red); B) comparison of the computed open structure (blue) and the OPLS-minimised solvate crystal structure (dark red). 114
3.9. Comparison of the collapsed (left hand side) and open (right hand side) geometries of **CC3**. A) [6+9] B) [8+12]. **CC3**-[6+9] has a void radius of 2.6 Å, whereas **CC3**-[8+12] of 4.1 Å. 115
3.10. The OPLS geometry optimised energies and final void sizes of CC3 [8+12] structures from constrained MD simulations at 300 K with a range of different constraint radii (as labelled at top). Each sub-plot shows a separate simulation run with a different constraint radii. Note that for the void radius plotted, the van der Waals radius of the nearest atom is subtracted, thus these values are typically smaller than the constraint sizes applied. . . 116
3.11. Relative energies of the lowest energy open conformers of **CC3**-[8+12] (in different colours), calculated at the OPLS_2005 level and obtained with different probe sizes. 116
3.12. Relative energies of the lowest energy open conformers of **CC3**-[6+9] (in different shades of red), calculated at the OPLS_2005 level and obtained with different probe sizes. The viewing angle is rotated for clarity for the structures inflated with 9 and 10 Å probes. 117
3.13. The OPLS geometry optimised energies and final void sizes of CC3 [6+9] structures from constrained MD simulations at 300 K with a range of different constraint radii (as labelled at top). Each sub-plot shows a separate simulation run with a different constraint radii. Note that for the void radius plotted, the van der Waals radius of the nearest atom is subtracted, thus these values are typically smaller than the constraint sizes applied. 118
3.14. A) Comparison of the collapsed (left hand side) and open (right hand side) geometries of CB[6]. The open geometry has a void radius of 2.6 Å. B) Left: overlay of the model (blue) and solvate crystal structure (red); Right: model (blue) and OPLS-minimised solvate crystal structure (dark red). 119

3.15. Comparison of the collapsed (left hand side) and open (right hand side) geometries of CB10. The open geometry has two cavities, each with a void radius of 2.5 Å; B) Left: overlay of the model (blue) and solvate crystal structure (red); Right: model (blue) and OPLS-minimised solvate crystal structure (dark red). 120
3.16. Comparison of the collapsed (left hand side) and open (right hand side) geometries of Cryptophane A (Xe-loaded).B) Left: overlay of the model (blue) and solvate crystal structure (red); Right: model (blue) and OPLS-minimised solvate crystal structure (dark red). . . . 121

4.1. Platonic, Archimedean and Catalan polyhedra that can be related to organic cage topologies. 1) Tetrahedron, 2) Trigonal Prism, 3) Cube, 4) Octahedron, 5) Dodecahedron, 6) Square Antiprism, 7) Rhombic Dodecahedron, 8) Cuboctahedron, 9) Rhombicuboctahedron. 126
4.2. Example of nomenclature for a cage of **TriDi** topology. The final $\mathbf{Tri}^4\mathbf{Di}^6$ cage is rotated by 90° respect to the tetrahedron for clarity. 127
4.3. Two nomenclature examples for the $\mathbf{Tri}_2^4\mathbf{Di}^6$ and the $\mathbf{Tet}_3^3\mathbf{Di}^6$ topologies. In the first case a double bond between two tritopic precursors occurs twice, therefore p=2; in the second case it occurs three times, therefore p=3. 128
4.4. The Tritopic + Ditopic topology family. Tritopic vertices are in blue, ditopic linkers in purple. 131
4.5. Top row: The range of geometrical shapes potentially accessible with a $\mathbf{Tri}^4\mathbf{Di}^6$ topology, from tetrahedron (left) to octahedron (right). The evolution of the shape is expressed through successive truncations of the vertices of the tetrahedron. The octahedron has been rotated into the page relative to the other shapes to give a clearer perspective. Bottom row: stages of truncation reproduced in hypothetical porous organic cages. The initial cage with a tetrahedral shape is shown in grey, precursors that lead to increasing truncation highlighted in teal and the final (completely truncated) cage with an octahedral shape is in teal. The cage molecules are not shown to scale. 132
4.6. The range of shapes possible for $\mathbf{Tri}^8\mathbf{Di}^{12}$ topology cages, from a vertex-folded shape(top), to a cube(bottom). 134
4.7. The Tetratopic + Ditopic topology family. Tetratopic vertices are in yellow, ditopic linkers in purple. 135

4.8. The range of shapes possible for a $\mathbf{Tet}^{6}\mathbf{Di}^{12}$ topology cage,from an octahedron (middle) to a cube with tetratopic linkers on the faces (right). The molecule reported on the left is an example of octahedral shaped cage. 136
4.9. Example of a $\mathbf{Tet}^{8}\mathbf{Di}^{16}$ topology cage reported by Warmuth and co-workers (left) with the shape of a square antiprism (right). 137
4.10. The Tritopic + Tritopic topology family. One of the tritopic precursors is in blue, the other in teal. 138
4.11. The range of shapes possible for $\mathbf{Tri}^{4}\mathbf{Tri}^{4}$ topology cages, from a tetrapod(top), to a cube(bottom). 139
4.12. The $\mathbf{Tet}^{6}\mathbf{Tri}^{8}$ topology for the Tetratopic + Tritopic family. Tetratopic precursors are in yellow, tritopic in blue. 140
4.13. The range of shapes possible for a $\mathbf{Tet}^{6}\mathbf{Tri}^{8}$ topology cage. Top: the $\mathbf{Tet}^{6}\mathbf{Tri}^{8}$ cagecan adopt the shape of a rhombic dodecahedron (middle) when considering both tetratopic and tritopic precursors as vertices of a solid. It has the shape of an octahedron (right) when only the tetratopic vertices are linked together. Bottom: the $\mathbf{Tet}^{6}\mathbf{Tri}^{8}$ cagecan be seen as a rhombic dodecahedron (middle, as before). It adopts the shape of a cube (right) when only the tritopic precursors are linked together. 140

5.1. Assembly of an octahedral molecule with $\mathbf{Tet}^{6}\mathbf{Di}^{12}$ topology, starting from the tetraformyl cavitand and ethylene-1,2-diamine precursors. **A** is the underlying topology, **B** is the final assembled molecule. . . . 144
5.2. Pipeline of the prediction approach.**A**: two precursors are chosen, in this example a ditopic (purple) and a tritopic (blue); **B**, precursors are assembled into molecular cages, according to a selection of possible topologies; **C**, Conformer search calculations are performed to locate the lowest energy conformer; if open ($\mathbf{C}_a$) the analysis proceeds to step **D**, if collapsed ($\mathbf{C}_b$) an open higher energy conformer is located with constrained MD; **D**: all the open conformers are analysed at the PBE and M06-2X level of theory; **E**: the relative internal energies per formula unit of all the cages, or their formation energies per bond, are compared to locate the lowest in energy, and therefore the topology which is most likely to form experimentally. **F**: if a crystal structure is available, the simulation and the crystal structure are overlaid 145

5.3. Conformer search analysis for cage **CC1**. Top: bar plot calculating the number of times each conformer was found in a simulation of 50000 steps (crimson) and in one of 100000 steps (orange); bottom: plot showing the number of unique conformations found per number of steps (teal, left Y axis) and the relative energies of the final collection of conformers (red, right Y axis). In the middle of the plot, the geometries of all the 26 conformers are shown as an overlay on the right and the lowest energy conformer is shown (at a different angle) on the left. 147
5.4. Left: bar plot showing the formation energies per bond of the imine condensation between [1:1] triformyl-benzene and ethanediamine, calculated at different levels of theory. Right: relative formation energy per [2+3] formula unit calculated for the $\mathbf{Tri}^{2}\mathbf{Di}^{3}$ and $\mathbf{Tri}^{4}\mathbf{Di}^{6}$ topologies of cage **CC1** (shown in the plot, carbon is in teal, nitrogen in blue). All energies are calculated with respect to the $\mathbf{Tri}^{2}\mathbf{Di}^{3}$ topology. 149
5.5. Fragment of cage **CC1** investigated at different levels of theory. Left: overlay of the PBE+D3 (in cyan) and MP2 (in purple) optimised geometries; Right: overlay of the M06-2X (in dark blue) and MP2 (in purple) optimised geometries. 150
5.6. Plot: Relative energies of the lowest energy conformers identified within a 30 kJ mol^{-1} with the OPLS-AA force field (in cyan) and the relative energies of the same conformers re-optimised at the PBE/TZVP+D3 level of theory (orange). The pink bars show the absolute energy difference between values obtained with the two methods. Bottom: overlays of the OPLS-AA (cyan) and PBE/TZVP +D3 (orange) geometries for conformers 1 (left), and 12 (right). 152
5.7. Molecular structures of the lowest energy open conformations of the Tritopic + Ditopic topologies for the **CC3** imine cage system, formed from the reaction of 1,3,5-triformylbenzene (**1**) and (*R*,*R*)-1,2-diaminocyclohexane (**2**). Tritopic BBs are in blue, ditopic precursors in purple, the imine bonds formed in the reaction in grey. 156
5.8. The relative energies of the lowest energy conformations of the **CC3**-based Tritopic + Ditopic topologies, as calculated by M06-2X. 158
5.9. Overlay of computed (blue) and crystal structure (red) of $\mathbf{Tri}^{4}\mathbf{Di}^{6}$. . 158

5.10. Molecular structures of the lowest energy open conformations of the Ditopic + Tritopic topologies for the **CC5** and **CC8** imine cage system, formed from the reaction of 1,3,5-triformylbenzene (**3**) and (1*S*,2*S*)-cyclopentanediamine (**4**) for **CC5** and (*R*,*R*)-1,2-diaminocyclohexene (**5**) for **CC8**. Tritopic BBs are in blue, ditopic precursors in purple, the imine bonds formed in the reaction are in grey. 159
5.11. The relative energies per [2+3] formula unit of the lowest energy conformations of the **CC5** and **CC8**-based Ditopic + Tritopic topologies, as calculated by M06-2X. 160
5.12. **CC5 - Tri^{4}Di^6, A)** Overlay of computed (blue) and crystal structure (red), **B)** Overlay of computed (blue) and minimised crystal structures (dark red). 162
5.13. Molecular structures of the lowest energy conformations of the Tritopic + Tritopic topologies for the **CC11** imine cage system, formed from the reaction of tris(2-aminoethyl)amine (**6**) and tris(4-formylphenyl)aldehyde (**7**). Tritopic BBs of **6** are in blue, tritopic BBs of **7** are in teal, the imine bonds formed in the reaction are in grey. 162
5.14. Left hand side: the relative energies of the lowest energy conformations of the **CC11**-based Tritopic + Tritopic topologies, as calculated by M06-2X; Right hand side: Overlay of computed (blue) and crystal structure (red) of **Tri^{4}Tri^4**. 163
5.15. Molecular structures of the lowest energy open conformations of the Tetratopic + Ditopic topologies, formed from the reaction of tetraformyl cavitand (**8**) and ethylene-1,2-diamine (**9**). Also shown with the **Tet^{2}Di^4** topology, are the lowest energy conformation with *m*-phenylenediamine (**10**) and *p*-xylylenediamine (**11**) as the ditopic BBs. Tetratopic precursors are in yellow, ditopic precursors in purple, the imine bonds formed in the reaction are in grey. 164
5.16. Two different conformers of **Tet^{8}Di^{16}**, the less inflated and lower in energy on the left, and the most inflated and higher in energy on the right. 165
5.17. The relative energies of the lowest energy conformations of the Warmuth Tetratopic + Ditopic topologies, as calculated by M06-2X. The error bar on point Tet^{8}Di^{16} shows the range of energies covered by conformations with differing degrees of inflation (see Figure 5.16, and Table 5.5), with more open conformers displaying higher energies. . . 166

5.18. Molecular structures of the lowest energy conformations of the Tetratopic + Tritopic topologies formed from the reaction of tetraformyl cavitand (**8**) and 1,3,5-tris(4-aminephenyl) benzene (**12**). Also shown with the $\mathbf{Tet}^{6}\mathbf{Tri}^{8}$ topology are the lowest energy conformations with 1,3,5-triaminobenzene (**13**) and 4,4',4-triaminotriphenylamine (**14**) as the tritopic BBs. Tetratopic precursors are in yellow, tritopic precursors in blue, the imine bonds formed in the reaction are in grey. . 168
5.19. Hypothetical structures of $\mathbf{Tri}^{20}\mathbf{Di}^{30}$, $\mathbf{Tet}^{16}\mathbf{Di}^{32}$ and $\mathbf{Tet}^{24}\mathbf{Di}^{48}$, all of which lack shape persistence. 170

6.1. Precursors involved in the HT screening and the presumed resultant topologies. The numbering is unrelated to that used in previous chapters. 176
6.2. Some of the criteria involved in the choice of the precursors; **a** is size of the precursors, **b** is length of the precursors, **c** precursors with different functionalities and **d** precursors that only differ by an atom. Carbon is in different shades of blue depending on which topology the precursors were targeted for, oxygen in red, nitrogen in blue, sulphur in yellow, boron in pink, bromine in dark red and chlorine in green. . 178
6.3. PBE/TZVP+D3 optimised geometries of cages generated with triamine A. $\mathbf{Tri}^{2}\mathbf{Di}^{3}$ cages (1-10) have carbon atoms in light yellow, $\mathbf{Tri}^{4}\mathbf{Di}^{6}$ cages (11-21) in dark yellow and $\mathbf{Tri}^{4}\mathbf{Tri}^{4}$ cages (22-26) in orange. Here, and throughout this chapter, nitrogens are in dark blue, oxygens in red, bromine in dark red, boron in pink, sulphur in yellow and chlorine in bright green. Cages A13, A14 and A17 collapsed after initial investigations, the starting open geometries being reported here. A7 is not shown as it was not included in the simulations. . . . 179
6.4. PBE/TZVP+D3 optimised geometries of cages generated with triamine B. $\mathbf{Tri}^{2}\mathbf{Di}^{3}$ cages (1-10) have carbon atoms in pink, $\mathbf{Tri}^{4}\mathbf{Di}^{6}$ cages (11-21) in red and $\mathbf{Tri}^{4}\mathbf{Tri}^{4}$ cages (22-26) in dark red. B7 is not shown as it was not included in the simulations. 180
6.5. PBE/TZVP+D3 optimised geometries of cages generated with triamine C. $\mathbf{Tri}^{2}\mathbf{Di}^{3}$ cages (1-10) have carbon atoms in cyan, $\mathbf{Tri}^{4}\mathbf{Di}^{6}$ cages (11-21) in teal and $\mathbf{Tri}^{4}\mathbf{Tri}^{4}$ cages (22-26) in steel blue. C7 is not shown as it was not included in the simulations. 181

6.6. Example of a dihedral angle (atoms 1 to 4) containing an imine bond. All atoms in the example cage (B11) are in grey; only atoms forming part of the dihedral are coloured, carbons in purple and nitrogen in blue. 184
6.7. Possible topologies arising from the combination of triamine B and aldehydes 1, 11 and 23. The topologies in the boxes are those targeted in the initial plan. 185
6.8. B1: relative energies of the different topologies calculated at the PBE/TZVP+D3 level. All energies are calculated with respect to the $\mathbf{Tri}^2\mathbf{Di}^3$ topology. 186
6.9. B1-$\mathbf{Tri}^2\mathbf{Di}^3$: A) overlay of the simulation (blue) and the crystal structure (red); B) overlay of the simulation (blue) and the OPLS3-minimised crystal structure (dark red). 187
6.10. B11: relative energies of the different topologies calculated at the PBE/TZVP+D3 level. All energies are calculated with respect to the $\mathbf{Tri}^4\mathbf{Di}^6$ topology. Circles represent open conformers, whereas triangles collapsed conformers. 187
6.11. B11-$\mathbf{Tri}^4\mathbf{Di}^6$: A) overlay of the simulation (blue); the crystal structure (red) and B) overlay of the simulation (blue) and the OPLS3-minimised crystal structure (dark red). 188
6.12. B23: relative energies of the different topologies calculated at the PBE/TZVP+D3 level. All energies are calculated with respect to the $\mathbf{Tri}^4\mathbf{Tri}^4$ topology. 189
6.13. B23-$\mathbf{Tri}^4\mathbf{Tri}^4$: A) overlay of the simulation (blue) and the crystal structure (red); B) overlay of the simulation (blue) and the OPLS3-minimised crystal structure (dark red). 189

6.14. Colour maps summarising the results obtained by the HT screening. 1) Description of the experimental outcome as per the legend below the colour map. 2) Colour map showing the formation energies per bond formed (calculated at the PBE/TZVP +D3 level) of each simulated cage assembled into the targeted topology; warmer colours represent more unstable cages, colder colours more stable cages, whereas grey squares are for cages that collapsed following high temperature MD and therefore their formation energy was not calculated. 3) Colour map showing the computationally calculated void diameters of the cages that successfully formed experimentally; shades of red represent smaller cavities, shades of blue larger cavities, grey squares cages that did not experimentally form, and crossed grey squares cages that formed but did not retain a cavity. In both plots 2) and 3), the squares corresponding to cages containing precursor 7 are barred, as simulations were not performed. 191
6.15. Organic cages that formed as a result of the HT screening, optimised at the PBE/TZVP +D3 level. The colour code is as per Figures 6.3, 6.4 and 6.5. Cages marked with a star are those that formed partially, with the presence of residual aldehyde starting material or other oligomeric side-products. . . 192
6.16. Colour map containing the formation energies per bond formed calculated at the PBE/TZVP +D3 level, of cages with $\mathbf{Tri}^2\mathbf{Di}^3$ topology. 193
6.17. C1: A) overlay of the simulation (blue) and the crystal structure (red); B) overlay of the simulation (blue) and the OPLS3-minimised crystal structure (dark red). 196
6.18. Overlays of cages B2 (left) and C2 (right). The simulations are in blue, the crystal structures in red. 196
6.19. Overlays of cages B9 (left) and C9 (right). The simulations are in blue, the crystal structures in red. 197
6.20. B15: A) overlay of the simulation (blue) and the crystal structure (red); B) overlay of the simulation (blue) and the OPLS3-minimised crystal structure (dark red). 199

6.21. Preference for **Tri**^{4}**Di**^{6} over **Tri**^{2}**Di**^{3} topology for cages (A/C)12, (B/C)14, (B/C)20 and (B/C)21 where a negative value indicates an energetic preference for **Tri**^{4}**Di**^{6} and a positive value an energetic preference for **Tri**^{2}**Di**^{3} topology (the preferred topologies according to the simulations are circled in grey for clarity). All were experimentally found to result in a **Tri**^{2}**Di**^{3}, rather than the planned **Tri**^{4}**Di**^{6} topology, with the exception of B14 and C21, in which both topologies were observed. 202
6.22. Example of two intramolecular hydrogen-bonds forming in cage C12-**Tri**^{2}**Di**^{3} between the hydroxyl group in aldehyde 12 and the nitrogen of the imine bond. All atoms are in grey, except for the atoms involved in the hydrogen-bond (carbon is in cyan, oxygen in red and nitrogen in blue). 202
6.23. π-stacking between the anthracene units of cage B14-**Tri**^{2}**Di**^{3}. All atoms are in grey, except for the aromatic rings involved in the π-stacking, which are in two different shades of purple. 204
6.24. C14-**Tri**^{2}**Di**^{3}: A) overlay of the simulation (blue) and the crystal structure (red); B) overlay of the simulation (blue) and the OPLS3-minimised crystal structure (dark red). 205
6.25. A) overlay of the simulation (blue) and the crystal structure (red) of C21-**Tri**^{2}**Di**^{3}; B) overlay of the simulation (blue) and the crystal structure (red) of C21-**Tri**^{4}**Di**^{6}. 206
6.26. Colour map containing the formation energies per bond formed calculated at PBE/TZVP+D3 level of theory for cages with **Tri**^{4}**Tri**^{4} topology. 209
6.27. Formation energy difference between cages A and C (22-26) calculated at the PBE/TZVP +D3 level. A positive difference in formation energy corresponds to C cages being more stable than A cages. 210
6.28. Overlays of cages C23, B24 and B26. The simulations are in blue, the crystal structures in red. 211
6.29. Left: topology scheme of B13-pretzelane shows two interlocked units (in yellow and grey) bridged in two positions (blue); Middle: B13-pretzelane coloured in grey, yellow and blue to make the connectivity more understandable; Right: geometry of B13 pretzelene, with carbons in teal, oxygens in red and nitrogens in blue. 212
6.30. Pie charts representing the percentages of success of the experiments for each topological family. 214

7.1. Molecules inflated with constrained MD in Chapter 3, in each box the molecule on the left is collapsed, whereas that on the right is open, with a grey sphere highlighting the cavity. Orange boxes surround cages reported in the literature, red boxes hypothetical cages. 216
7.2. The four topological families classified in Chapter 4, and a scheme showing the difference between topology and geometrical shape. . . . 217
7.3. The molecules investigated in Chapter 5. Orange boxes highlight the molecules reported in the literature. 218
7.4. Cages investigated in Chapter 6, together with the plots reporting results from experiments (left), and simulations (right). 219
7.5. Workflow integrating computational and experimental methods for the successful discovery of new porous organic materials. Steps 1-2-3: computational investigation of hypothetical single cages, assembled from a pair of organic precursors; Step 4: promising candidates from step 3 are assembled into the solid state with crystal structure prediction methods; Step 5: promising candidates (of which properties can be studied with computational simulations) from step 4 are suggested for the synthesis. 220
7.6. **CP1**, **CP3**, and **CP4**, and their parent imine cages, **CC1**, **CC3**, and **CC4**. The numbers below each cage refer to the formation energy per bond formed calculated at the PBE/TZVP +D3 level of theory. . 221
7.7. Clusters showing interactions between **CP** molecules in the solid state, as predicted by crystal structure prediction. **CP3** molecules are in different colours for clarity: left) magenta and purple, middle) the *R* enantiomer is in magenta and the *S* in teal, right) *R*-**CC3** is in orange and *S*-**CP3** in teal. 221

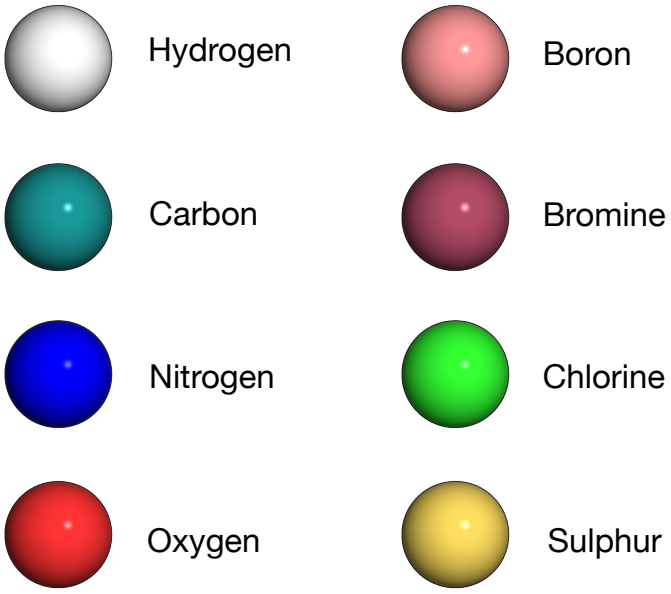
List of Abbreviations

BB	Building Block
BET	Brunauer-Emmett-Teller
CB	Cucurbituril
CC*n*	Covalent Cage n
COF	Covalent Organic Framework
COM	Centre of Mass
CP	Cage Pot
CSD	Cambridge Structural Database
CSP	Crystal Structure Prediction
D3	Grimme's DFT-D3 dispersion correction
DCC	Dynamic Covalent Chemistry
DFT	Density Functional Theory
Di	Ditopic
FF	Force Field
HT	High Throughput
MD	Molecular Dynamics
MOF	Metal Organic Framework

MOP	Metal Organic Polyhedra
PES	Potential Energy Surface
POC	Porous Organic Cage
PXRD	Powder X-ray Diffraction
RMSD	Root Mean Square Deviation
SA	Simulated Annealing
SCXRD	Single Crystal X-ray Diffraction
Tet	Tetratopic
Tri	Tritopic

Colour Code for Molecules

Unless otherwise specified in the figures' captions, the colours below are going to be adopted throughout the thesis. In this work multiple bonds and hydrogens will not be shown for clarity.



Chapter 1
Introduction

1.1. Scope of the work

Porous organic cages are wholly organic molecules characterised by the presence of a cavity[1] and by the ability to pack together in the three-dimensional (3-dimensional) space via intermolecular non-covalent interactions (see Figure 1.1). This latter characteristic separates them from network materials, such as zeolites,[2] metal-organic frameworks (MOFs),[3–5] and covalent organic frameworks (COFs).[6]

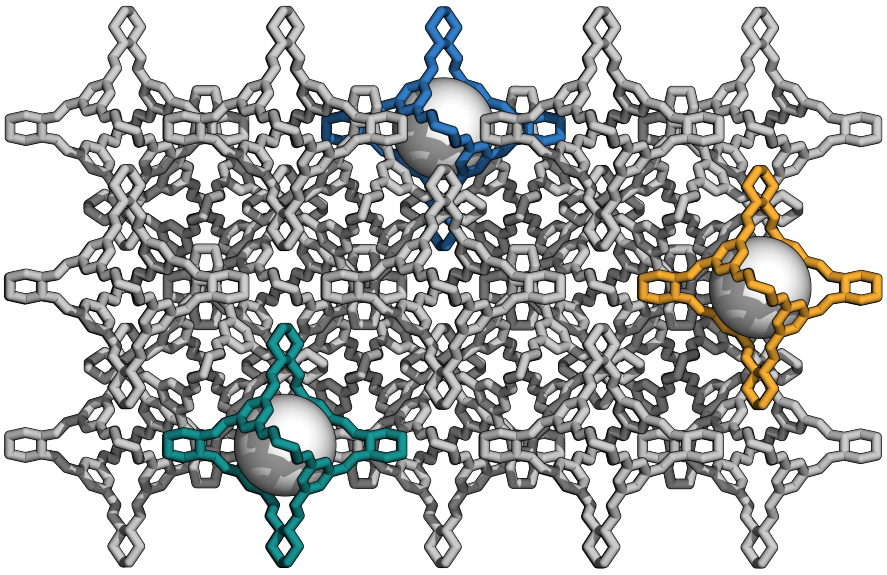


Figure 1.1.: An example of a porous molecular organic material representing the crystal packing of Covalent Cage 3 (**CC3**).[7]Three cages are differently coloured in blue, teal, and yellow to highlight the non-covalent nature of the packing. The same cages contain grey spheres to highlight the presence of a cavity.

My interest in porous organic cages was initially stimulated by the fact that the design of new porous materials is still an emerging field, and that scientists are striving to obtain new structures that can be used for a wide range of applications. Additionally, I found porous materials inherently beautiful because of their symmetry and spatial organisation. However, the time required for the synthesis and characterisation of porous organic cages still slows down the research, as long periods of time can be dedicated to the study of structures that decompose or are not suited for applications. As a computational chemist, I tried to address the following questions at the beginning of my work: i) can the computational prediction of the most likely reaction outcomes be improved, in order to direct the synthesis towards a set of feasible starting materials? ii) if so, can those methods be applied to rationalise the geometric criteria for the design and development of new porous organic materials?

This thesis, which collects the main projects I carried out during my PhD, aims not only to answer those questions, but also to build the basis of a novel way of describing and classifying porous organic cages, in the hope that it will be applied to future works. Ideally, the project would fit in a joint computational/ experimental pipeline, where, as shown in Figure 1.2, the techniques developed in this PhD are at the basis of a prediction workflow, which aims to suggest, from an initial collection of hypothetical molecules, feasible candidates for the synthesis. The pipeline would initially involve the assembly of organic precursors into hypothetical organic cages of multiple sizes and topologies (steps 1-2 in Figure 1.2), which would then be computationally investigated to test whether a) they maintain a cavity (step 3 in Figure 1.2) and b) are likely to be the outcome of an experimental reaction. Feasible candidates from step 3 can then be computationally arranged into crystalline or amorphous solids (step 4 in Figure 1.2) and be investigated for properties. Materials with promising properties will be suggested for synthesis and characterisation. My PhD was focused on developing computational procedures to cover steps 1-3 from the assembly of organic precursors to the prediction of shape persistence and topological outcome of individual molecules that can be either assembled in the solid state computationally, or of which the synthesis can be attempted straightaway. An outline of the thesis will be given in the last section of this chapter.

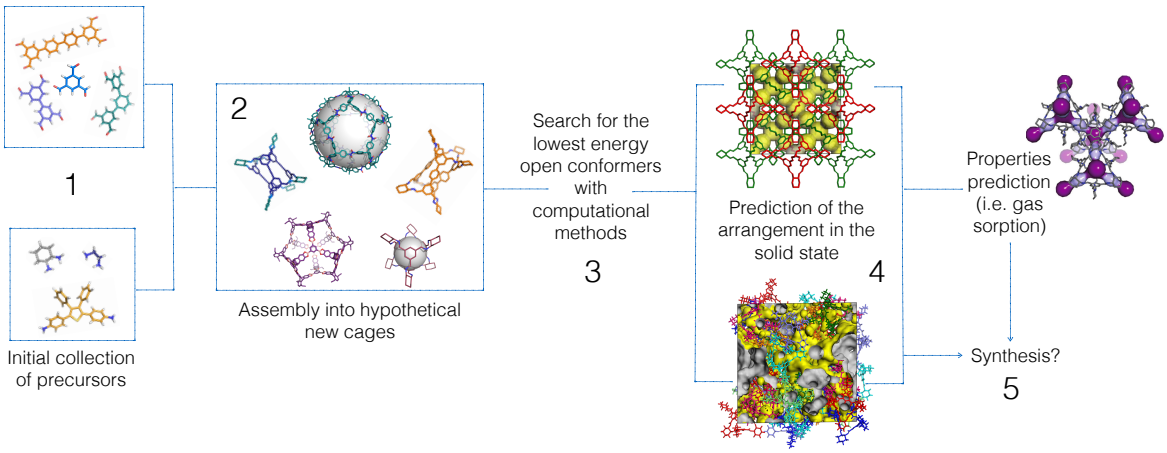


Figure 1.2.: Workflow integrating computational and experimental methods for the successful discovery of new porous organic materials. Steps 1-2-3: computational investigation of hypothetical single cages assembled from a pair of organic precursors; Step 4: promising candidates from step 3 are assembled into the solid state with crystal structure prediction methods; Step 5: promising candidates (whose properties can be studied with computational simulations) from step 4 are suggested for the synthesis. The colours of the cages and of the precursors do not reflect the atomic composition.

At the beginning of this introductory chapter, I will briefly discuss porous materials, of which porous organic cages are a subclass; then I will cover the latest developments in porous organic materials, their properties and the important factors that can affect both the synthesis and the prediction of experimental outcomes, including a discussion of porous organic cages' topologies. Even though my work was entirely focused on individual cage molecules, I will give a brief overview of the solid state properties of cage materials, and I will conclude by outlining the content of the chapters that follow. I hope this work will provide an interesting insight into the world of porous organic cages.

1.2. Porous materials and their properties

Porous materials are involved in many different aspects of our lives, playing crucial roles in the fields of catalysis, separation and ion exchange.[8,9] The most common examples are zeolites,[2] MOFs[5] and COFs,[6] which are illustrated in Figure 1.3. Zeolites (from the Greek for "boiling stone") are crystalline aluminosilicates containing pores and cavities of molecular dimensions[8] and, although many are present in nature, synthetic varieties are widely used as sorbents, catalysts and ion-exchange materials. Zeolites in their acid form probably were the most important heteroge-

neous acid catalysts used in industry. Despite their wide application, zeolites are entirely inorganic materials and their chemistry can only afford limited control over the final product. In order to obtain a greater degree of flexibility, organic and inorganic components were combined to form MOFs (term introduced by Omar Yaghi in 1995), crystalline networks formed by metal "nodes" connected via organic "linkers".[10] Adjusting the chemistry and the size of the two components helped customise the crystal pores and therefore, in principle, tailor the material for a given application. The first MOFs were reported in 1999 and were HKUST-1, containing copper-based clusters and benzene tricarboxylate linkers,[11] and then MOF-5 (shown in Figure 1.3), containing zinc-based clusters and benzene dicarboxylate linkers.[12] To date, tens of thousands of different MOFs have been generated, which have been used for applications ranging from catalysis to drug delivery, hydrogen storage, and reversible adsorption of CO_2.[5,13] Differently, COFs are made by organic building blocks covalently bonded to form 2- and 3-dimensional porous crystalline solids.[14] They can mainly be applied for gas storage (hydrogen and methane), optics and carbon capture.[14]

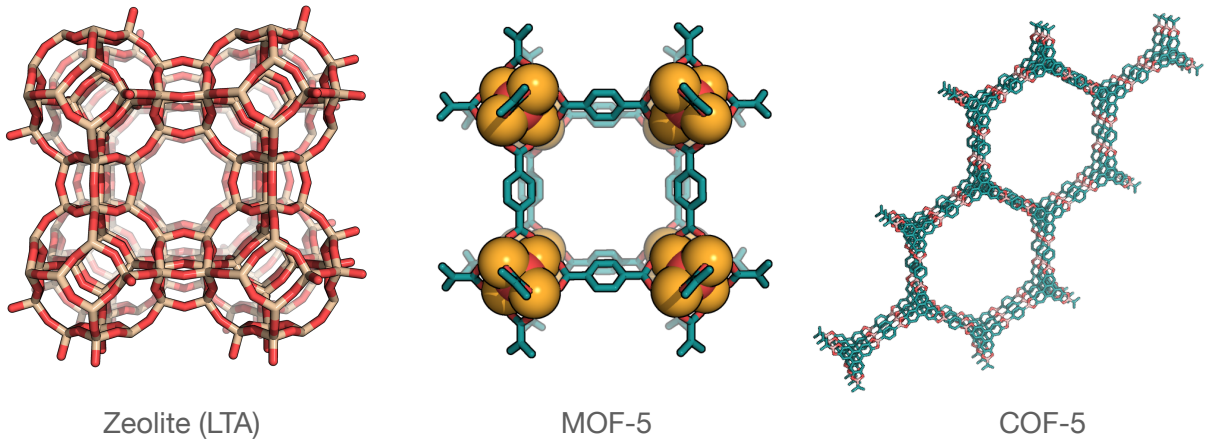


Figure 1.3.: Examples of porous network materials, from left to right, a zeolite with Linde Type A (LTA) topology, MOF-5[15] and COF-5.[16] Zinc is in orange, silicon in beige. The images of the three systems are not to scale.

Molecular porous materials differ from network materials in that they are composed of discrete entities bonded together in the solid state via weaker intermolecular interactions and are comparatively rare.[1,17–23] This is mainly because of two reasons: i) poor packing is relatively rare - most molecules tend to pack efficiently to remove as much void space as possible; ii) weakly bonded materials have become a serious

focus of research interest only quite recently and therefore have been the object of fewer investigations.

The network materials in Figure 1.3 have been optimised for more than 20 years to fulfil a broad range of applications; it is, therefore, not easy to generate novel molecular materials with improved features. Network materials are applied as catalysts and adsorbents;[21] they have been processed in usable forms and the raw materials have become in time less expensive, thus, more economically attractive. However, commercially available materials do not fulfil all the functions that modern technology requires, such as applications in the fields of electronics (in particular electronic materials), light harvesting energy transduction, proton conductors and molecular sensing. All these could be satisfied by the development of novel porous structures. The assembly of hosts for supramolecular chemistry has been of continuing interest in recent years and, in the case of network materials, enumeration of possible framework topologies[24–27] has allowed for the targeting of given topologies through the design of ligands and nodes that meet the necessary geometric requirements. This "reticular design" has assisted in the synthetic realisation of >91 different MOF topologies[28] from a potential of >2500.[27] A further work by Wilmer *et al.* involved the assembly of an even larger number (∼138k) of hypothetical MOFs from a library of 102 building blocks,[29] which led to the identification of over 300 MOFs with good methane-storage capacity and many structure-property relationships, but it did not cover the prediction of synthetic feasibility.

Although reticular design has many notable successes, such as the series of IR-MOFs with cubic-based topology of increasing ligand length,[30] too often the topological outcome of a synthesis remains a matter of serendipity, rather than successful *a priori* design. For finite molecular hosts, a wide range of different motifs have been realised, including rings, cycles, knots, polyhedra and catenanes.[31–33] Here, I will focus my attention on the design techniques that have been developed and can be applied to molecular porous materials.

1.3. Porous organic cages

The most recent definition of a cage, which I am going to refer to throughout this thesis, states that a porous organic cage is "a discrete covalently bonded polycyclic molecule that possesses an enclosed intrinsic pore, which is accessible via windows in the organic skeleton. The molecular pore and windows are of a size that is large enough to be occupied or traversed by guest molecules".[34] This definition could

be extended to some cryptophane molecules,[35] but it excludes most macrocyles (*e.g.* cucurbiturils and cyclodextrines), which have a cavity that is not completely enclosed by atoms, and the fullerene family, whose framework does not include accessible windows. Throughout this work, a cage is going to be defined as *shape-persistent*, when it is characterised by a defined cavity large enough to host other molecules and that does not collapse to a more dense or twisted structure upon solvent removal.[33] In terms of their porosity, porous organic cage materials (and porous molecular materials in general) can be characterised by two kinds of porosity, extrinsic and intrinsic. The first is given by inefficient packing of the molecules in 3-dimensional space, creating channels guest molecules can penetrate (in grey in Figure 1.4), whereas the latter is provided by the internal cavities of the cage molecules (in cyan in Figure 1.4).[36] These two kinds of porosity can work together in a cooperative way and can be connected, depending on the arrangement of the single entities within the crystal structure.

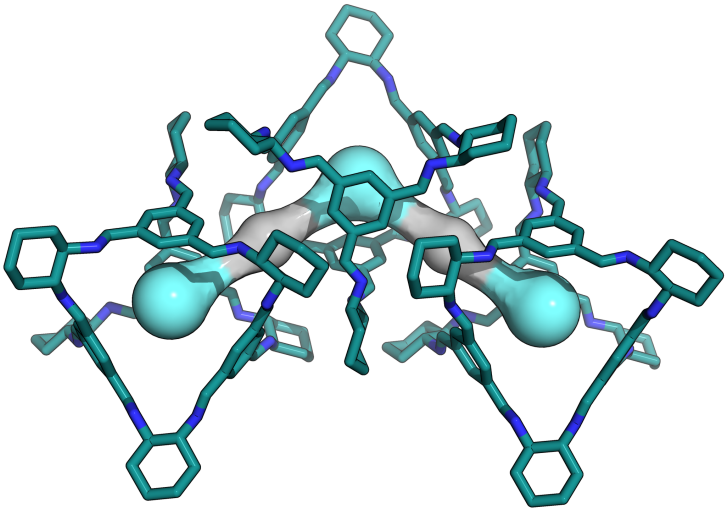


Figure 1.4.: An example of extrinsic (grey) and intrinsic (cyan) porosity connected by the window-to-window arrangement of a selection of three **CC3** cages.

Porous molecular materials have potential advantages compared to porous framework materials, such as MOFs, zeolites and activated carbons,[1] in that their inherently modular nature may allow for property-tuning through compositional variation of a material with different ratios of molecular modules.[37] Further, some have been shown to be water stable,[38] acid and base stable[39] and their solubility in common solvents has processing advantages; these materials have already been processed to form thin film membranes[40] and hierarchical architectures.[41]

In the next few sections, I will discuss in more detail the latest development in research on porous organic cages, firstly by giving a short historical background on the synthesis of the first cages, and then by discussing the ways porous organic cages can be synthesised and characterised.

1.3.1. Historical background

The discovery of porous molecules goes back to the 1960s, when synthetic routes for hydrocarbons shaped like Platonic polyhedra were devised and outlined.[42] Later on, in 1976, Barrer and Shanson reported the synthesis of a small organic molecule, Dianin's compound (shown in Figure 1.5), which crystallised from solution, was able to form inclusion complexes and was able to adsorb gases.[43] It was initially described as an "organic" zeolite, a solid with a porous crystalline lattice, but characterised by a higher degree of flexibility (compared to zeolites), which allowed guest molecules to penetrate the Dianin's cavities, even though no windows large enough seemed to be available. This was the first evidence of what Barbour later named "porosity without pores", namely the ability of a material with permanent voids, but no guest-sized channels connecting them, to be permeable to those guests.[23] It is now a recognised feature of many molecular porous materials.

The Dianin's compound, together with tris(o-phenylenedioxy)cyclotriphosphazene (TPP),[44] were the first examples of organic molecular materials with extrinsic porosity, meaning that accessible void channels form in the crystal because of the molecules' inefficient packing. Sozzani *et al.* demonstrated that TPP materials were characterised by permanent porosity;[45] however, a work from Hulliger and co-workers calculated their Brunauer-Emmett-Teller (BET) surface area to be quite small (< 200 m^2 g^{-1}).[46] Later on, McKeown and co-workers published some useful criteria to predict the assembly of such molecules in the solid state and showed that a number of possible candidates were available to fulfil those criteria.[47] One of these was 3,3,4,4-tetrakis(trimethylsilylethynyl)biphenyl (TTEB), which showed permanent porosity and was permeable to H_2.[47] However, the surface area of this material was still relatively low (278 m^2 g^{-1}), and it was not until hydrogen-bonded frameworks (HOFs) prepared from molecules such as triptycene trisbenzimidazolone (TTBI)[48] were assembled that materials with a surface area more than one order of magnitude higher were obtained by Mastalerz and co-workers (BET surface area = 2796 m^2 g^{-1}).[48] More recently, crystal structure prediction studies on triptycene-

based materials have been conducted by Pulido *et al.*, who investigated the stability of possible packing arrangements via energy-structure-function (ESF) maps.[49]

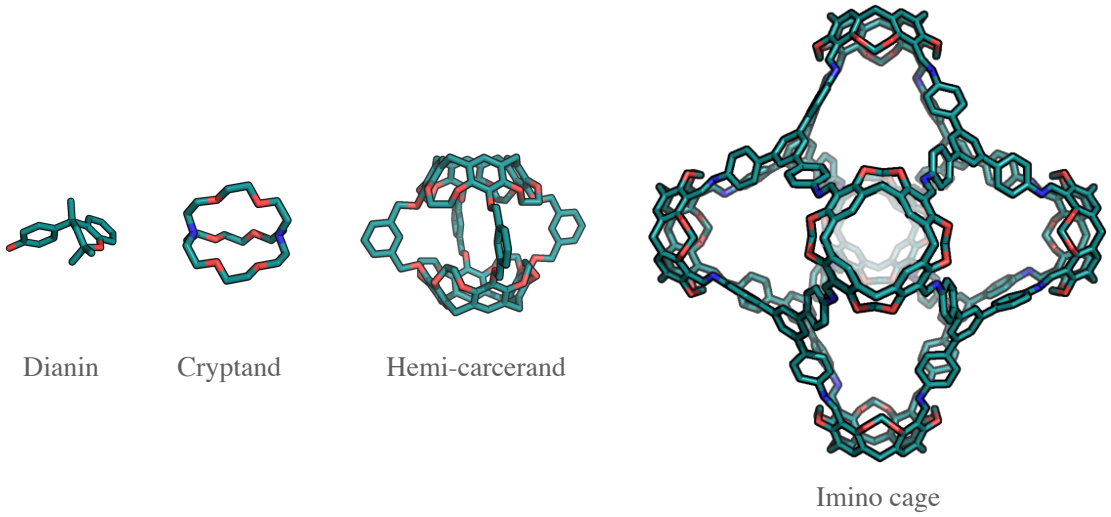


Figure 1.5.: Geometries of the Dianin (Cambridge Structural Database (CSD) reference: BEGSOW), cryptand (CSD reference: ACEZUG), hemi-carcerand (CSD reference: RAGXAY) compounds and of one imine cage from Warmuth and co-workers (no crystal structure available).[50]

A few years after the Dianin's compound was discovered, two new porous molecular compounds were synthesised, namely the cup-shaped calixarene molecules (1979)[51,52] and the ring-shaped cucurbituril molecules (1981). More recently, in 2006, the paddle-wheel shaped cage Noria was synthesised and discovered to be able to pack and display porosity both in the crystalline and amorphous state.[53] In 1984, the first shape persistent cage was synthesised to fulfil a specific application, a compound that acted as a siderophore and was shown to outperform EDTA for Fe^{3+} binding.[54] The same year, Cram and co-workers synthesised the first carcerand molecule[55] by covalently bonding two cavitand molecules, but it was obtained in very low yields, mainly due to the irreversible nature of the chosen reaction method. In 1987, Lehn successfully synthesised the first cryptands (a "3-dimensional extension" of a crown ether, shown in Figure 1.5), which were initially used for the selective recognition of cations.[56] It was again the group of Cram and co-workers that, a few years later, took advantage for the first time of the reversible nature of the imine condensation reaction, which allowed them to synthesise an imine covalently bonded porous hemi-carcerand in high yields (shown in Figure 1.5), starting from a mixture of amines and aldehydes.[57]

Imine condensation belongs to the family of dynamic covalent chemistry (DCC) reactions,[58] which has proved particularly fruitful for producing highly symmetrical molecular assemblies. The reversible nature of these reactions allows for error correction over time and, combined with sufficiently fast kinetics and long reaction times, should generate a thermodynamically controlled product distribution. This means, assuming the molecular cage assemblies are the most stable product, that DCC allows for initial kinetic oligomeric or polymeric products to be gradually converted over time into molecular cage structures. Thus far, the majority of the reported organic cage structures are imine cages (of which an example from Warmuth and co-workers is shown in Figure 1.5), although the current record-holder for solid state porosity is a boronate ester-based cage from Mastalerz and co-workers, with a BET surface area of >3500 m^2 g^{-1},[33]comparable to that of porous network materials such as COF-102 (BET surface area = 3472 m^2 g^{-1})[59] or MOF MIL-101 (BET surface area = 4100 m^2 g^{-1}).[60] There are also examples of other DCC reactions involving disulfide bridge formation, alkene metathesis and aldehyde condensation reactions.[61] There are rare examples of cage molecules synthesised without resorting to DCC, including the synthesis of an alkyne cage by Doonan and co-workers, where two precursors of the correct geometry were preassembled and then fused through carbon-carbon bond formation in a single step.[62] More details will be provided in the next section.

1.3.2. Synthesis of porous organic cages

Porous organic cages have so far been synthesised according to two different approaches, one based on reversible, and the other on irreversible bond formation. Molecules synthesised with irreversible reactions are generally characterised by higher chemical stability, but are rarely isolated in high yields because of the kinetic nature of the reactions involved.[33] By contrast, cages formed via reversible bond formation are generally thermodynamically controlled, which means that they should lead to the most thermodynamically stable product; they can be formed in very high yields in a single step from relatively simple precursors,[33] but are, however, less chemically stable.

Irreversible bond formation

Cages synthesised with irreversible bond formation are comparatively rare in the literature as, generally, they do not form in high yields, as reported by the early works

on cages by Vögtle and Raymond.[63,64] Nevertheless, some reactions have been tailored to produce organic cages in higher yields and some examples can be discussed. It was found in the literature[65,66] that reactions such as amide bond formation (in high dilution) and nucleophilic aromatic substitution[67] could be adopted for the synthesis of small capsule-like cages, but that larger templates would be required for the formation of larger cages. The formation of carbon-carbon bonds was adopted by both Moore and co-workers,[68] and Doonan and co-workers[62] for the synthesis of a macrobicycle compound and a small capsule-like cage. Still, these few cases represent exceptions and reversible chemistry is now preferred to obtain cages in higher yields and quantities.

Reversible bond formation

Most cages are currently synthesised according to reversible routes, the imine bond condensation being the most commonly adopted and of which a scheme is reported in Figure 1.6. Reversible DCC reactions are particularly convenient because they are based on a *self-healing* mechanism, which involves consecutive bond making and breaking steps until, ideally, the most thermodynamically stable product is obtained.[58] The main advantage of these reactions is the fact that complex structures can be obtained via a one-pot reaction in high yield and have therefore become widely used in the literature. Acid catalysts are often used in the synthesis to enhance the reversibility of the bond formation. Most organic cages currently in the literature have been synthesised via imine condensation between aldehydes and amines, the example in Figure 1.6 reporting the precursors used for a group of cages (Covalent Cages 1, 2, 3, 4 and 13 - **CC1**, **CC2**, **CC3**, **CC4** and **CC13**) synthesised by Cooper and co-workers.[7,69–73]Remarkably, the synthesis of **CC1** was optimised to reach a yield of almost 100%.[74] After the first imino-based carcerand was synthesised by Cram, the synthesis of porous organic cages evolved towards molecules with the underlying topology of a tetrahedron, the first examples synthesised by Skowronek and Gawronski in 2008[75] and Cooper in 2009.[7] The cage was renamed by Cooper and co-workers as Covalent Cage 3 (**CC3**) and packs into a window-to-window diamondoid network.[7] A few examples of organic imine cages with multiple topologies are shown in boxes A, B, and C of Figure 1.7.

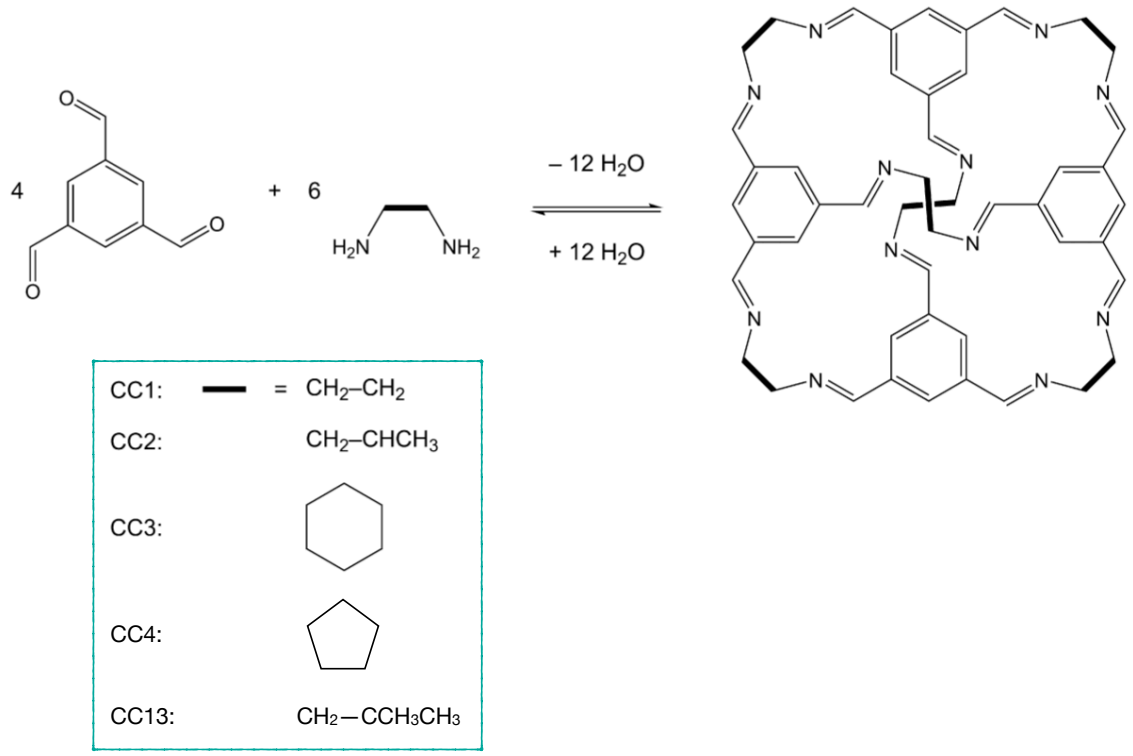


Figure 1.6.: Example of an imine condensation reaction between aldehydes and imines, in this instance for the synthesis of cages **CC1**, **CC2**, **CC3**, **CC4** and **CC13** .

In the years following the publication of the first organic cages, Cooper and co-workers focussed their attention on the synthesis and characterisation of a series of imine cages built by applying small changes in the chemistry and geometries of an initial set of starting precursors,[7,69–72] Mastalerz published a series of triptycene-based imine cages (in box C of Figure 1.7)[76–79] and Warmuth a series of cages obtained combining cavitands and both ditopic and tritopic linkers according to multiple topologies (in box B of Figure 1.7).[80,81] More details and a literature background on the shape and topologies of imine cages will be discussed in the next section and in Chapter 4, but have also been covered in multiple recent excellent reviews.[21,22,33,34,61,82,83]

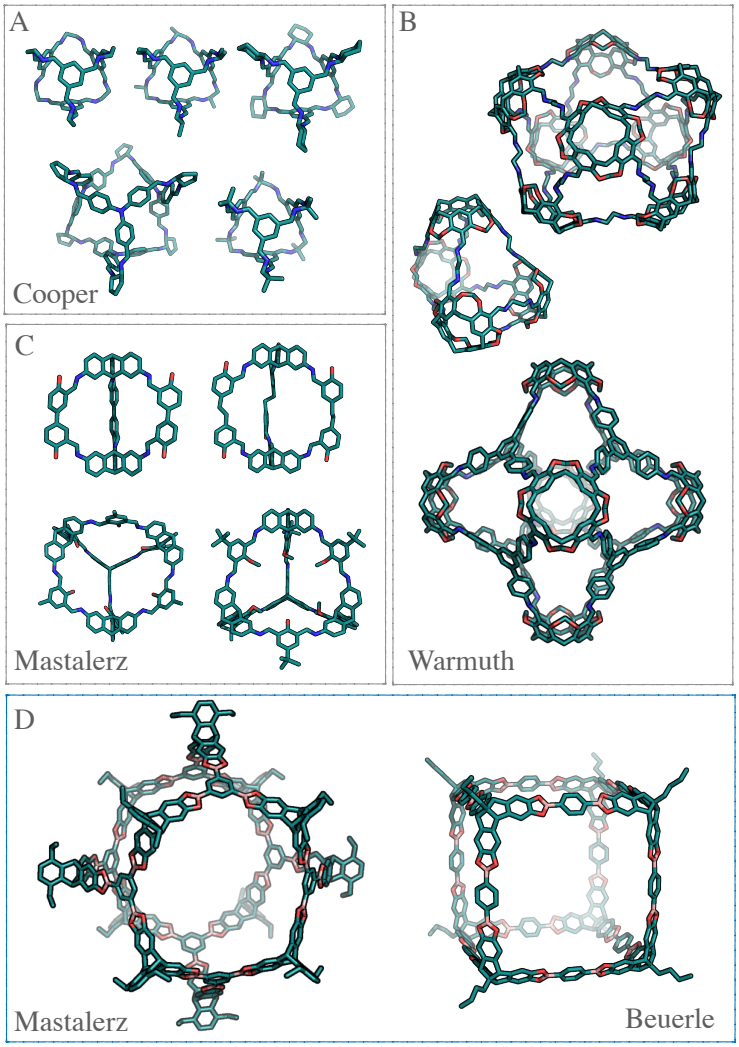


Figure 1.7.: Examples of imine and boronic cages. Box A includes (from left to right and from top to bottom) cages **CC1**, **CC2**, **CC3**, **CC5**, and **CC13** synthesised by Cooper and co-workers. Box B includes (from top to bottom) three porous imine cages with the shape of a squared anti-prism, a distorted tetrahedron and an octahedron synthesised by Warmuth and co-workers. Box C contains two triptycene-based capsules (top row) and two triptycene-based tetrahderal (bottom row) molecules formed by Mastalerz and co-workers. Finally, box D contains two examples of boronic ester cubes synthesised by Mastalerz and co-workers (left) and Beuerle and co-workers (right). Crystal structures were identified for all cages, except for those synthesised by Warmuth and co-workers.

Some other cages formed according to DCC reactions other than imination are based on boronic ester formation obtained by mixing alcohols and boronic acids. This reaction has been used in particular for the synthesis of porous organic cubes, which are, generally, relatively hard to obtain due to the orthogonal arrangements

required in the starting tritopic precursors necessary to match the corners of the solid. Mastalerz reported the synthesis of a large [12+8] boronic cube (shown on the left hand side of box D in Figure 1.7) obtained by the 48-fold condensation of 12 trypticene tretraols and 8 boronic acids, which was reported to have the remarkable BET surface area of 3758 m^2 g^{-1}, currently the highest reported for a molecular organic crystal.[84] Pore size distribution calculations[85] reported a pore size comprised between 2.03 and 2.30 nm (depending on the method adopted for the calculations), confirming the mesoporous classification of the cage (pores > 2 nm are considered mesoporous). More organic nanocubes were reported by Beuerle and co-workers by mixing 8 tribenzonquinacene catechol precursors with 12 phenyldiboronic acid linkers (shown on the right hand side of box D in Figure 1.7).[86–88]

Other DCC reactions used in the literature include the formation of capsules obtained with template-free alkyne methastesis,[89] disulphide formation and phenol/aldehyde condensation, although the last two are comparatively less common and have only been used in a few cases.[61,82] A combination of different kinds of DCC reactions have been described for the synthesis of cages combining imine and boronic ester condensation[90] or combining imine and olefin metathesis.[91]

1.4. Porous organic cages topologies and challenges involved in their prediction

In this section I will discuss the necessary background on the topology of porous organic cages, in order to introduce the prediction challenges that were at the basis of my research and the work that will be discussed in the next few chapters. Part of my PhD was dedicated to understanding what is the most likely topological outcome of a reaction, starting from a pair of precursors assembled into multiple stoichiometric ratios. As already mentioned, porous organic cages are a relatively new kind of porous material and, whilst both computational and experimental characterisations of such systems are progressing quickly, an unambiguous description of their topology had not yet been defined. Framework materials, for example, can be described by their underlying extended topology, as in the case of the MOF-5 topology used for reticular chemistry;[5] however, the same can not be said for porous organic cage materials as they are not covalently bonded in 3-dimensional space. The underlying topology of molecular materials, and of cages in particular, can include polyhedra such as Platonic, Archimedean and Catalan solids, some of which are shown in

Figure 1.8.

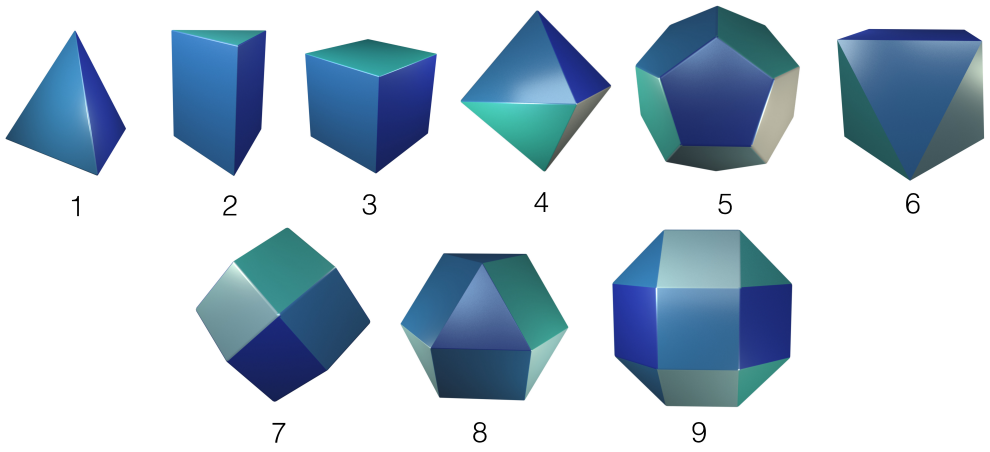


Figure 1.8.: Platonic, Archimedean and Catalan polyhedra that can be related to organic cage topologies. 1) Tetrahedron, 2) Trigonal Prism, 3) Cube, 4) Octahedron, 5) Dodecahedron, 6) Square Antiprism, 7) Rhombic Dodecahedron, 8) Cuboctahedron, 9) Rhombicuboctahedron.

The polyhedra work as an underlying scaffold upon which the cages are built. The final structure is obtained when the vertices, edges or faces of the chosen solid are replaced with molecular building blocks (BBs). Capsule-like topologies, such as cavitands or molecules connected by multiple linkers, used to be more common. However, in the past few years, the increased focus on the area and the improved understanding of favourable experimental conditions has made the generation of cages with more complex topologies possible; examples include tetrahedra, cubes (more common), square antiprisms and cuboctahedra (less common). In the next few paragraphs, I will describe some plausible polyhedral topologies for organic cages and explain how topology differs from geometrical shape, a term that is more commonly used to describe molecules, but can lack accuracy when objectively trying to describe a system. Let us consider the case of the organic cage molecule synthesised by Warmuth and co-workers.[92] The cage is made up of 6 tetratopic and 8 tritopic precursors and can be described as (see Figure 1.9):

1. A Rhombicuboctahedron, if one considers tetratopic precursors to lie on every other square face of the solid and tritopic precursors to lie on the triangular faces.

2. An Octahedron, if one links together the precursors with the highest coordination number, in this case the tetratopic cavitands.

3. A Cube, if one links together the precursors with the lowest coordination number, in this case the tritopic molecules.

4. A Rhombic Dodecahedron, if one places both tetratopic and tritopic precursors on the vertices of the solid.

However, which solid represents the underlying topology and which the geometrical shape of the molecules? The answer to this question will be the focus of Chapter 4.

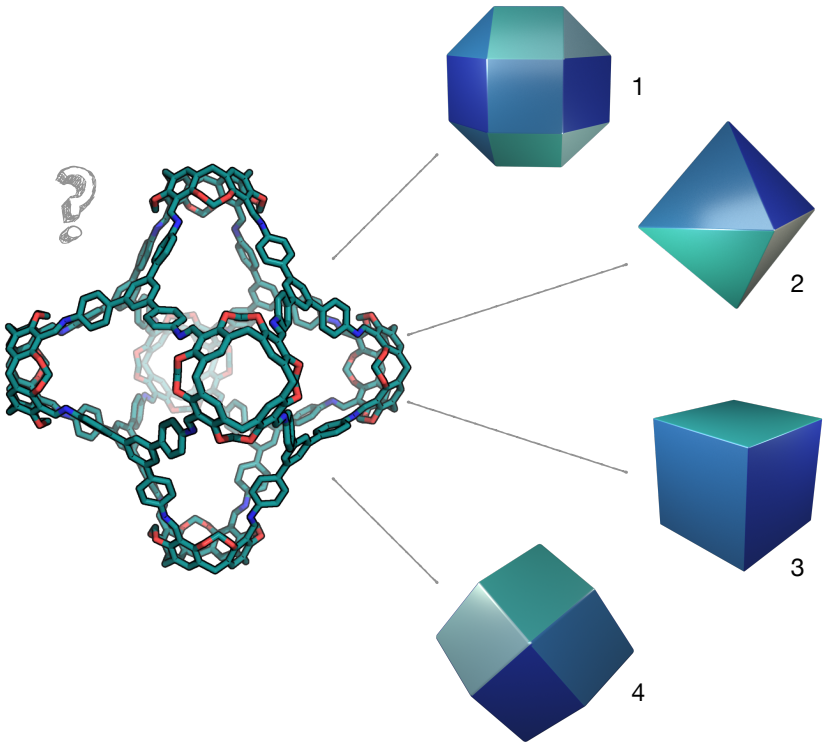


Figure 1.9.: Selection of polyhedra that can describe the [6+8] organic cage synthesised by Warmuth and co-workers.[92] 1) a Rhombicuboctahedron, 2) an Octahedron, 3) a Cube 4) a Rhombic Dodecahedron.

The topology is an intrinsic component of every cage and is independent of the geometrical shape the molecule adopts. This is because, mathematically, the topology is the region of space that is preserved upon deformation, as Figure 1.10 explains.[93] The mug and the doughnut have indeed the same underlying topology (characterised by the presence of a hole), but clearly different geometrical shapes. This means that a molecule (or more generally an object) is characterised by a single topology, but it may adopt different geometric shapes, mainly depending on the geometry of its precursors, their flexibility and the angles they will adopt in the assembled structure.

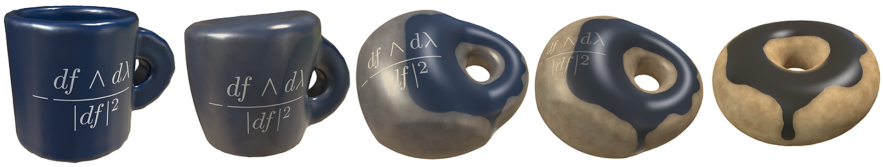


Figure 1.10.: Example of difference between geometrical shape and topology. The mug and the doughnut have the same topology (characterised by the presence of a hole), but different geometrical shapes, which allow one to describe them as "mug" and "doughnut". Image adapted from K. Crane, U. Pinkall and P. Schröder, *ACM Trans. Graph.*, 2013, 32-42.

Figure 1.11 gives a visual description of the difference between topology and geometrical shape applied to organic cages. It shows in particular how a molecule with the underlying topology of a tetrahedron can adopt the shape of a tetrahedron or that of an octahedron (or any intermediate), depending on the angle of the component BBs.

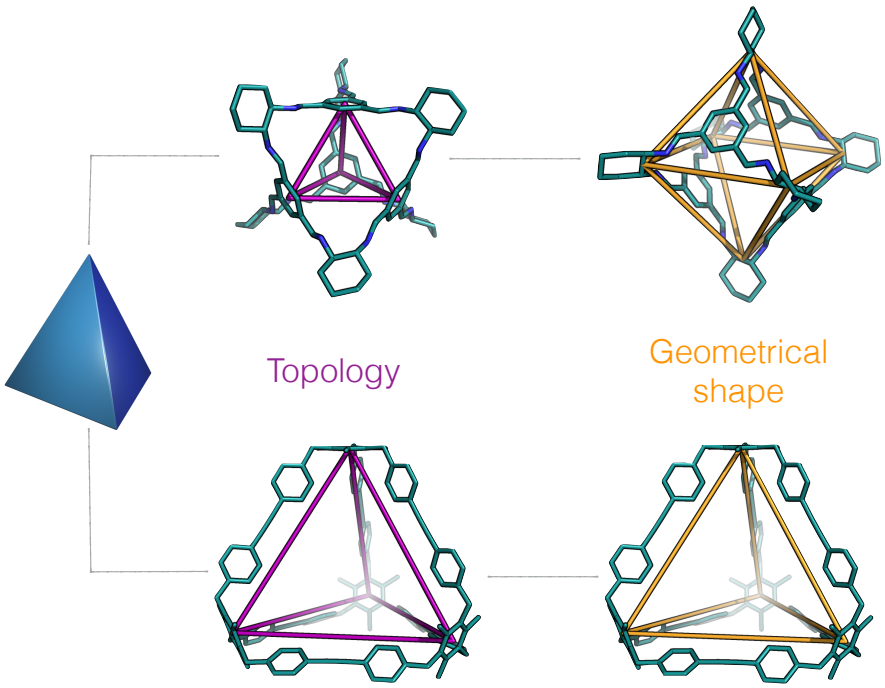


Figure 1.11.: An example of how two porous organic cages that both have an underlying topology of a tetrahedron adopt different geometric shapes, in one case maintaining a tetrahedron shape[94] (bottom row), and in the other adopting an octahedral shape (top row).[7] The underlying topologies are shown in purple and the geometric shapes in orange.

The topology plays a crucial role in the way molecules can be used for applications, as cavities of different shape and size can be more suited for one application than another. Some cavities will show better results for specific guest encapsulation, whereas another cage with different size, shape and number of windows may be more useful for separation, sensing or porous liquids. All these factors are obviously strictly related to the BBs that make up the cage, which not only have an effect on the guest the cage could encapsulate/separate, but also, depending on their geometry and flexibility, on the shape of the host itself. In the solid state, the geometry of the cage can further affect the way the molecules pack into the crystal or amorphous structure and then, more specifically, factors such as guest mobility or binding.

At the moment, it is possible to design a specific cage topology for a new cage by using fairly rigid BBs, even though the rigidity of the precursors is not a guaranteed element of success. In fact, unexpected outcomes may form or, if the hypothetical structure ends up very strained, no structure may form at all. Some examples in which chemical intuition led to a correct prediction of the final structure involve BBs that were designed to meet the specific geometrical requirements of the solid. This was the case of the already mentioned molecular cube synthesised by Beuerle and co-workers,[86,95] where 8 precursors with angles of $\sim 90^{\circ}$ replaced the vertices of the cube and 12 linear BBs linked them via the edges. Another example was the cuboctahedron synthesised in the Warmuth group, which was generated by tritopic and tetratopic precursors assembled on the faces of the solid via molecular panelling.[92] Still, the correct prediction of the topological or geometrical outcome can be very challenging. Frequently, the synthesis of a particular polyhedral cage can lead to the formation of a family of cages with related polyhedral topologies that can be obtained by modifying the initial precursors. A good example of this was given by the series of tetrahedral [4+6] imine cages by Cooper and co-workers discussed in the previous section.[7,36,72] However, this strategy was not always successful and some examples showed that small changes in the initial conditions led to unexpected outcomes. In one case, a change from a cyclopentane diamine to a cyclohexene diamine resulted in the formation of a large [8+12] cubic molecule rather than the tetrahedral molecule that was expected (see Figure 1.12).[96]

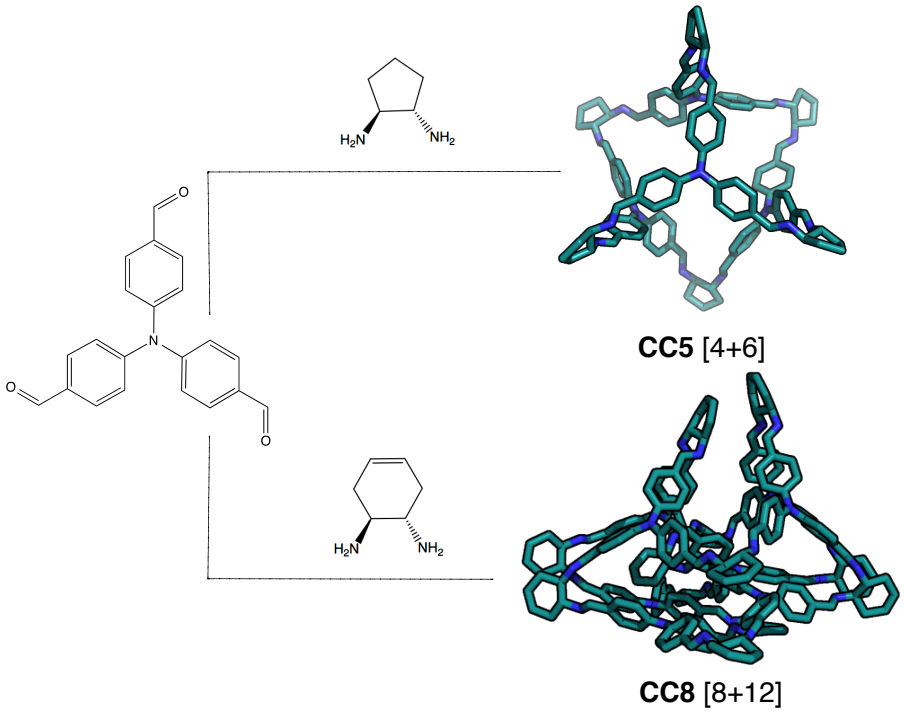


Figure 1.12.: Example of emergent behaviour: mixing the central trialdehyde with cyclopentadiamine (top row) leads to the formation of the porous tetrahedral **CC5** imine cage and with cyclohexenediamine to the non-porous larger **CC8** imine cage (bottom row).[96]

The expectation of a [4+6] outcome was based on the fact that a [4+6] tetrahedral molecule formed when the cyclopentane diamine was used, which differed from the cyclohexene diamine by a single carbon atom.[96] Furthermore, whilst the tetrahedral molecule formed a porous material with a BET surface area of >1300 m^2 g^{-1}, the larger one collapsed upon desolvation and therefore generated a non-porous amorphous material. This showed how such a small change in the starting precursors had such a considerable effect on the reaction outcome and consequently on the properties of the final molecular material. A similar behaviour was observed in the metal-organic polyhedra reported by Fujita and co-workers.[97] In their study they showed how Pd^{2+} (**M**) and a dipyridylthiophene ligand (**L**) gave $M_{24}L_{48}$, whereas dipyridylfuran gave the much smaller $M_{12}L_{24}$ polyhedron. This was described by Fujita as *emergent behaviour*, a phenomenon that clearly can play a central role in the topological design of organic cage materials. Another example of emergent behaviour upon small changes in the starting precursors is given by the "odd-even" effect discovered by Jelfs *et. al.* In their computational study they discovered that upon assembly of imine cages, when the diamine precursor contained an even

number of carbon atoms, a tetrahedral [4+6] cage was predicted and observed, whereas an odd carbon chain length (such as propane-, pentanediamine) resulted in the formation of a [2+3] capsule-like cage (as shown in Figure 1.13).[98]

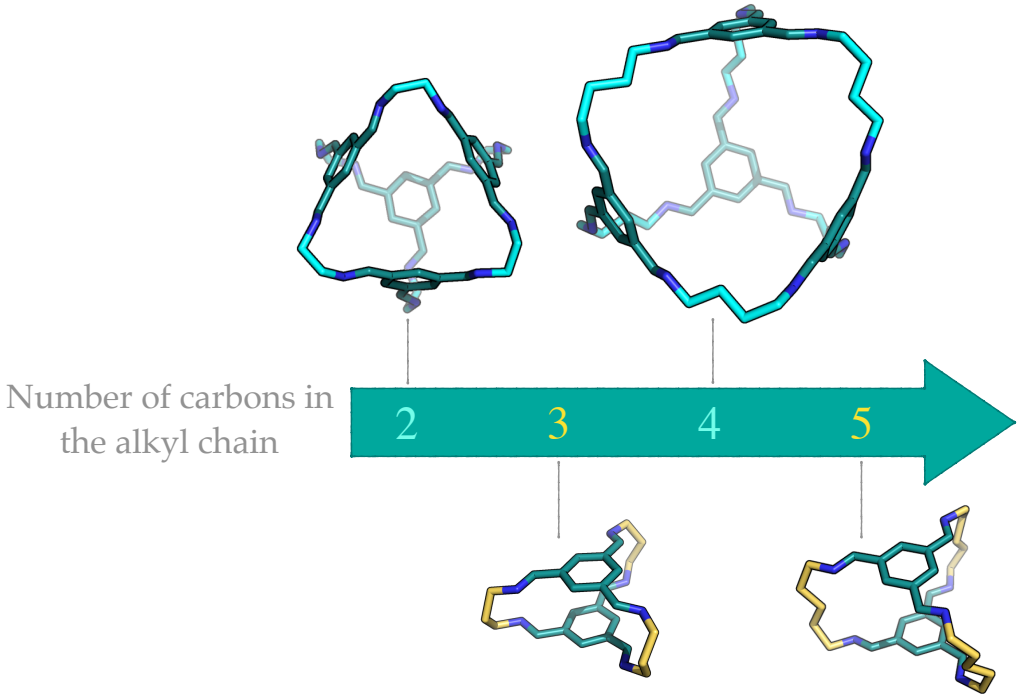


Figure 1.13.: Example of emergent behaviour: mixing a trialdehyde with a diamine containing an even number of carbons (cyan) leads to the formation of tetrahedral cages (top), but mixing it with a diamine containing an odd number of carbons (yellow) leads to the formation of capsule-shaped cages (bottom).

Additionally, one should consider that the successful synthesis of the final outcome is not only influenced by modifications on the starting BBs, but also by other experimental factors such as solvent or temperature. Some reports showed how, in particular, the choice of solvent could control the topological outcome. An example was given by the work from Liu and Warmuth, where a tetraformylcavitand was mixed with an ethylene diamine in an imine condensation reaction in the presence of tetrahydrofuran, chloroform or dicholoromethane. The presence of the first solvent led to the formation of a distorted [4+8] tetrahedral cage, of the second to an octahedral [6+12] cage and of the third to a [8+16] square antiprismatic structure (see Figure 1.14).[99]

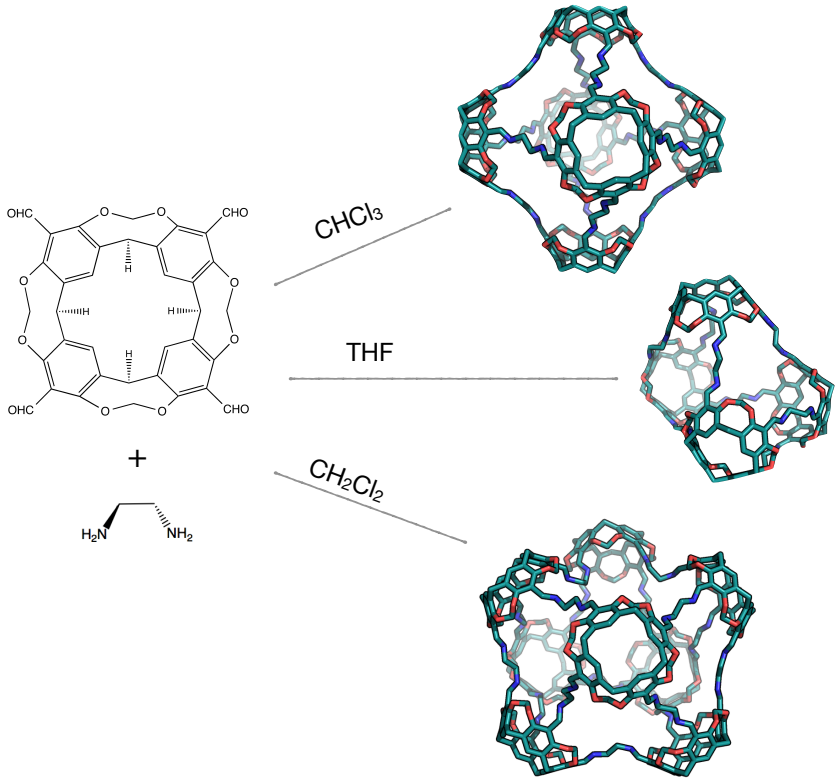


Figure 1.14.: The three different cages found by Warmuth and co-workers. The presence of chloroform led to the formation of an octahedral cage, of THF of a distorted [4+8] tetrahedral cage and of dichloromethane to a [8+16] square antiprismatic structure.[99]

As already stated, the use of DCC reactions should give thermodynamic control over the product distribution, meaning that the molecules with the lowest free energy should be the main reaction product. However, this strongly depends on the reaction mixture being able to reach this product, avoiding kinetic intermediates along the way. Depending on the reaction's characteristics, kinetic products could include a variety of oligomeric products, polymeric materials, alternative molecules with different topologies or catenated molecules. Porous organic cages rarely catenate, however some cases were reported in which the interlocked form was the thermodynamic outcome of a reaction. Some examples came from Hasell *et al.*[70] and Zhang *et al.*[100] They suggested that stabilising intermolecular interactions, such as π-π stacking and alkyl-π interactions, could cause the "self-templation" of the catenated form rather than the formation of the monomeric one.

1.4.1. Nomenclature

Even though porous organic cages are a field of growing interest, a uniform classification system does not yet exist. As stated by Brunner in 1981, the "synthesis of new structures requires not only chemical skill but also some knowledge of the principal topological possibilities".[101] The number of synthetically feasible topologies for porous organic cages is surely lower than the number of possible 3-dimensional nets available for network materials, however there are still some unrealised topologies and shapes that could offer new material properties or stimulating synthetic challenges. Similar topology descriptions are available in the literature for metal organic polyhedra (MOPs) and come from the groups of Yaghi[102] and Stang.[32] In their work, they listed all the MOP topologies synthetically realised, including tetrahedra, cubes, cuboctahedra and even a larger dodecahedron.[103] This was followed by a paper by Hay and co-workers, where they discussed design strategies targeted at MOPs and metal-organic polygons, and stressed that describing molecules by their shape might lead to multiple classifications for the same underlying topology.[104] This could be particularly confusing for face-directed polyhedra, as the description of the final shape could change depending on how metals and ligands were related to vertices, edges or faces of a polyhedron. To resolve these inconsistencies, Hay defined a nomenclature based on the number of vertices, edges or faces included in the structure and its point group.

In my PhD, I worked on a new nomenclature system for porous organic molecules that takes inspiration from the approach of Hay and co-workers. Whilst for MOPs the metal could be systematically assigned to the highest coordination site of a polyhedron, for wholly organic molecules the classification was more challenging, as it was not always obvious which BB should have been assigned to a vertex, edge or face. In Chapter 4, the topologies synthetically reported to date will to be examined and divided into families, depending on the coordination number of the precursors involved in their making. This is firstly, to generate a consistent nomenclature for organic cages, then to rationalise the criteria that could help the design of the desired topologies and shapes. The work discussed will not only provide a way of describing existing systems, but will also shed light on a broader range of topological outcomes that might be encountered when targeting a particular topology.

1.4.2. Factors influencing the design of porous organic cages

In the next few sections I will list some relevant factors that may influence the synthesis and *a priori* design of cage molecules, which I became aware of whilst carrying out my research.

Stoichiometry of the building blocks

When aiming at a particular topology, the starting building blocks (BBs) have to be mixed in the correct stoichiometry to assemble into the correct structure. For example, a cube with 8 tritopic vertices and 12 ditopic edges can be formed if the reactants are mixed in a 2:3 ratio. An excess of one BB could change the topological landscape and lead to the formation of alternative, different topological products, where not all the functionalities have formed covalent bonds. An example was reported in the literature by Warmuth and co-workers.[92] They attempted the formation of a [6+8] rhombicuboctahedral cage by mixing 6 tetratopic BBs and 8 tritopic BBs in a 3:4 ratio; when the ratio was exact, the rhombicuboctahedral cage formed, however an excess of the tritopic BB led to the formation of a [2+4] capsule, in which one of the 3 reactive end groups on the tritopic precursors was left unreacted.

Number of reactive end groups

Two ditopic BBs can only form single cycles, whereas BBs with a larger number of reactive ends, such as tritopic and tetratopic BBs, can form the more complex topologies related to the solids reported in Figure 1.8. Some examples of BBs commonly used for the synthesis of porous organic cages are shown in Figure 1.15. Generally, BBs with at least a C_n axis, with n equal to the number of reactive end groups, are employed, however their availability decreases as n increases. This means that many symmetric ditopic and tritopic ligands are commercially available, but far fewer tetratopic with a C_4 symmetry, with the exception of cavitands and porphyrins. Even fewer symmetric pentatopic or hexatopic BBs are available, so in this work only precursors with 2, 3 and 4 reactive ends are included. This is the main distinction between porous organic cages and MOPs, where the presence of metal nodes allows one to use precursors with higher coordination numbers.

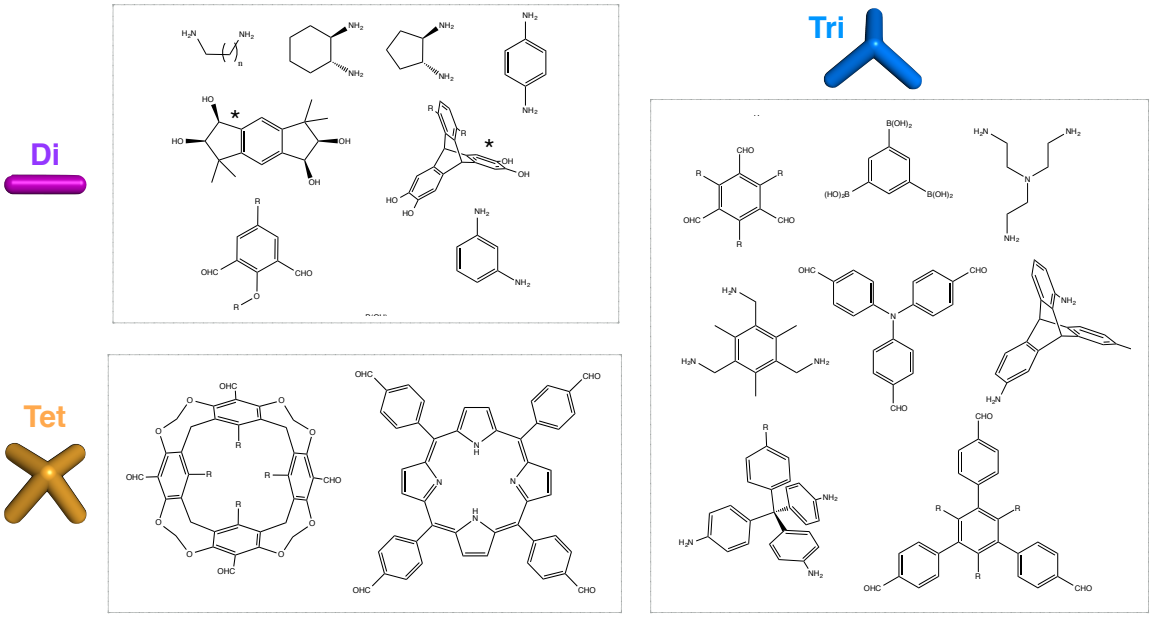


Figure 1.15.: Examples of ditopic (**Di**), tritopic (**Tri**) and tetratopic (**Tet**) BBs reported for the synthesis of organic cages. Starred ditopic precursors have been used for the synthesis of boronate cages, where two hydroxy groups bond to one boron atom. Thus, I still labelled these precursors as ditopic rather than tetratopic.

Building blocks geometry and rigidity

Other factors to consider for topological design are the geometry of the starting precursors and the direction of their reactive end groups. To maximise the chance of producing a cage unit (rather than a periodic network), the angle between the reactive end groups and the centre of at least one BB has to be less than 180°. Let us consider the design of a molecule shaped like a cube and with the underlying topology of a cube. Ideally, one would use tritopic ligands with (end group)-(centre of mass)-(end group) angles of 90° for the corners and ditopic ligands of (end group)-(centre of mass)-(end group) angles of 180° for the edges. This strategy was successfully adopted by both the Warmuth[95] and Beuerle[86] groups to design two molecular cubes. However, for MOPs, examples in the literature[102] showed that even if the angles did not exactly match the desired topology, the structure could still form, with deviations up to 10° likely to be tolerated. Larger deviations could be compensated for by adaptations between matching geometries of the BBs, whereas significant deviations led to distorted shapes that could no longer resemble the appearance of the initial topology. Rules for topology and geometrical shape design of MOPs were proposed by the groups of Yaghi[102] and Stang,[32] who based their strategy

on complementary end group vector selection. However, Young and Hay recently discovered, whilst analysing the crystal structures of MOPs, that the vector approach worked for only a few cases, but not for the majority.[104] They attributed unsuccessful results to either conformational flexibilities that led to deviations from the original bonding vector or concerted rotations about the symmetry axes of the molecules.[104]

Of course, defining design rules could be even more complicated for porous organic cages compared to MOPs, as they are characterised by greater flexibility and weaker directional bonding. The BBs may change their geometry when reacting; for example, an amine functionality in a BB may be different from the geometry it adopts when covalently bonded in a cage, and can be again different if a reduction functionalises the imine after the synthesis occurs. A parameter that was useful for the design of both porous organic cages and crystalline network materials, was the rigidity of the starting BBs, where more rigid BBs better fit the targeted underlying topology.[30] However, even if the rigidity had been successfully applied to the design of organic cages, it was not a fundamental requirement and some materials relying on dynamic motion of the final assembly (such as host-guest response behaviour) could be made of more flexible components. For this reason, more flexible BBs are not excluded from this work, which made defining rules and the design itself more challenging. When the rules were applied to test cases, computational modelling was required to predict the conformations of the possible reaction outcomes. On the other hand, one should bear in mind that extreme flexibility is not desirable, as it is linked to low product yields[33] and makes loss of porosity more probable. Together, geometries, end-group orientations and flexibility strongly affect the topology of the final outcome and the molecule's shape, making the use of computational tools even more relevant. This, for example, has worked for the design of cubes by Warmuth and co-workers and Beuerle and co-workers.

Solvent effect

Solvent can play a crucial role in the synthesis of a specific cage, as the non-covalent interactions between solvent "guest" molecules and "host" cages have a potential effect on both the thermodynamics and kinetics of the product formation.[105,106] Factors such as the limited solubility of the reactant molecules or the formation of kinetic intermediates can lead to low yields or precipitation during the reaction process.[107] Regarding the formation of final products, the assembly of molecular capsules could be obtained in high yields, according to Rebek and co-workers, when the internal cavity had a sufficient size to host at least one or more solvent molecules.

The optimal binding was then obtained when 55% ± 9 of the pore volume was occupied.[108].

Later, Warmuth and co-workers suggested that molecules of solvent enclosed in a cavity could establish strong interactions with the BBs by interacting with the molecule's internal surface over long periods (in the range of 10^{-3} to 10^{3} s).[105] This allowed the solvent trapped in the cavity and in the windows of the cage to influence the geometry of the BBs and drive the assembly towards a specific topology. According to this idea, the major contribution from solvent would consist of an improper solvation of the flexible BBs.[99] Describing the role of solvent in cage formation is not trivial and it can be said to be dependent on three main factors:

- The entropic cost associated with the molecular reorganisation, which involves collecting n BBs from the solvent phase.
- The energy cost of the BB - solvent interactions (*e.g.* hydrophobic effect).
- The amount of strain arising for each BB when bonded into the assembled structure.

Due to the complex and variable character of the solvent, it is incredibly challenging to computationally predict these interactions and to quantify them. As mentioned above, Warmuth and co-workers estimated the differences in solvent effects that led to the formation of different imine cages in different yields (see Figure 1.14) to be ∼17 kJ mol^{-1} per imine bond (which is quite significant). Additonally, they showed how macrocyclic assemblies preferentially formed over oligomeric products, which experienced an energetic penalty for each functional group that remained unreacted.[99,109–111]

Synthetic methods

Another factor that should be considered is the way organic cages are synthesised. The vast majority of porous organic molecules are generated with batch chemistry at high dilution combined with slow addition of the reactants, as this generally leads to good formation yields. However, more recently, it was found that synthesising cages via continuous flow methodology showed a good degree of success too.[112,113] The method involved shorter reaction times and a lower quantity of solvent, and led, nonetheless, to high purities and yields. Additionally, methods such as mechanochemistry[114] provided innovative synthetic routes to specific topologies

by acting on the kinetics of the reaction and therefore on the possible reaction pathways. Other experimental variables influencing the outcome of the reaction may include temperature, concentration, rate of mixing, order of addition of precursors, use of catalysts, pH and template molecules.

1.5. Properties and characterisation of the solid state

1.5.1. Experimental characterisation

A variety of techniques are applied, alone or combined, to investigate the formation of porous organic cages, the most used being listed below.

- Mass Spectroscopy (MS), Gel Permeation Chromatography (GPC) and Nuclear Magnetic Resonance (NMR) spectroscopy are normally applied to investigate the chemical composition and the purity of the reaction outcome.

- Diffusion NMR and Diffusion-Ordered Spectroscopy (DOSY), are exploited to define the size and shape of molecules formed in solution - quite useful if the cage collapses, decomposes or generates an amorphous material upon solvent removal. To gain information about the shape and size of molecules the average size and maximum diameters of the cages can be calculated from simulations and compared to the experimentally determined solvodynamic diameters, thus allowing one to determine the topology of the cages. If there is agreement between the measured size and the computed average size, the combination of diffusion NMR and simulations can provide a rapid and direct measure of the success of predictions.

- Crystallography is in general applied to study the molecular connectivity inside the crystal structure, and to investigate the 3-dimensional shape of the molecule in more detail. In particular:

    1. Single crystal diffraction (SCXRD) can provide an accurate description of the molecules in the solid state, assuming a crystal of suitable size is available for the analysis. It can also give information about the position of ordered guests inside the crystal channels/cavities and about the particular composition of the crystals. This is possible when the product can be isolated and crystallised, which can be particularly problematic

if the cages do not form cleanly in solution and a mixture of products is present. SCXRD does not give information about the purity of the structure. If disordered guests are there, they can either be modelled *in silico*, or processed with software to estimate the amount of solvent present inside the structure (sometimes this can be exploited in combination with other characterisation techniques, *e.g.* thermogravimetric analysis).

2. Powder diffraction (PXRD) provides structural information about the degree of crystallinity of the material or about the presence of additional crystal phases (both impurities or cages). It is, however, insensitive to amorphous contaminants.

3. *In situ* diffraction experiments can be performed to observe changes such as adsorption or desorption of guests and can therefore obtain information about the host-guest interaction or about the molecular/packing flexibility.

- Microscopy techniques, such as Scanning Electron Microscopy (SEM) and Transmission Electron Microscopy (TEM), are frequently adopted to analyse the morphology of the sample, to identify the presence of amorphous phases, or to gain information about the surface structure of the material.

- Solid State NMR can provide structural information about amorphous solid materials.

In general, the choice of technique to be used depends on the kind of experiment and on the system under investigation. A combination of different methods can be applied, for example, to gain information about the material before and after a phenomenon occurs (*e.g.* gas sorption).

1.5.2. Properties of the solid state

In this section I will describe some of the properties of both crystalline and amorphous porous organic materials, in order to provide an overview of their advantages and characteristics.

On-off Porosity

With the term "on-off porosity", Jones *et al.* described the possible switch between two or more polymorphs of the same materials, which are characterised by different

gas-adsorption selectivity.[115] Imine cage **CC1** was initially isolated into two different forms after solvent removal,[7] **CC1**-α (non-porous to both N_2 and H_2 - displaying window-to-arene packing, on the left hand side of Figure 1.16) and **CC1**-γ (porous to both N_2 and H_2). Firstly, it was discovered that it was possible to switch from the α-polymorph to the γ-polymorph in the solid state by exposing the first to ortho-xylene vapour, followed by solvent removal.[115] It was then proved that **CC1**-α could be converted into a third polymorph, **CC1**-β, which was porous to H_2, but not to N_2, by exposing the first to dichloromethane vapour followed by activation. The process was reversible by exposing the β-polymorph to ethyl acetate, followed by solvent removal and crystallinity was retained throughout. **CC1**-γ could also be converted into the β-polymorph by exposure to dichloromethane vapour, although this second process was not reversible. MD simulations were applied to rationalise the selectivity of **CC1**-β for hydrogen: the trajectories displayed connectivity between the intrinsic pores (see Figure 1.16 - right hand side) when the probe radius was equal to the kinetic radius of H_2 (1.42 Å), but not when equal to the kinetic radius of N_2 (1.82 Å). Being able to selectively switch porosity upon solvent exposure can be a particularly useful tool for applications requiring selective guest-adsorption or permeation barrier materials.

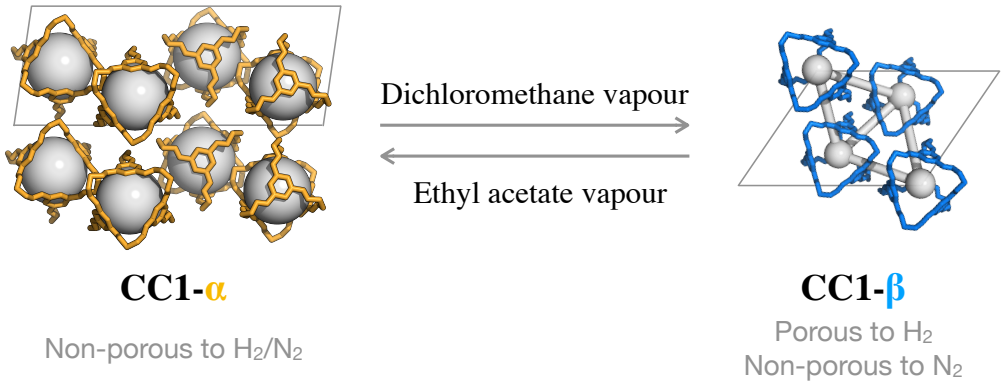


Figure 1.16.: Reversible on-off porosity switching in CC1 crystals can be triggered by exposure to specific solvents followed by guest removal. Exposure to ethyl acetate vapor and desolvation leads to the non-porous polymorph **CC1**-α; DCM exposure and evacuation affords **CC1**-β, which is non-porous to N_2, but adsorbs H_2. **CC1**-α cages are in yellow, whereas **CC1**-β are in blue.

Isoreticular Cages

The ability to generate isoreticular MOF materials, which share a common underling network topology, is considered one of the biggest successes of crystal engineering.[116,117] The procedure is possible because networks are based on metallic "nodes" connected by ditopic organic "linkers", which represent units that can be substituted (by tuning chemistry and length of the precursor) without disrupting the material connectivity, assuming the angle between the reactive end groups is preserved. Given the different nature of porous organic cage materials, which are not bonded in 3-dimensional space, the formation of isoreticular crystals remained a challenging goal for a long time. Cooper and co-workers generated the first isoreticular porous organic cages by studying the recrystallisation of cage **CC3** and **CC4** in different solvents,[118] switching their arrangements from window-to-window to window-to-arene and *vice versa*. They performed the study by conducting a high-throughput screening in the presence of 30 solvents and discovered that **CC3**-α and **CC4**-β both crystallise into window-to-window arrangements when the first is exposed to most solvents and the second to paraxylene. By contrast, **CC3**-β and **CC4**-α become isoreticular and pack in a window-to-arene arrangement, when the first is exposed to diethyl ether and the second to most solvents, as shown in Figure 1.17-A. Similarly, the presence of dioxane allows the conversion of cages **CC2**-α and **CC13**-α into their β-polymorph, which packs in a window-to-window arrangement, making them also isoreticular to **CC3**-α, as shown in Figure 1.17-B. In all cases, the solvent promotes the formation of diamondoid networks when it binds in the extrinsic space by stabilising the window-to window-interaction, whereas it promotes a window-to-arene arrangement when it frustrates the alignment of neighbouring cages.

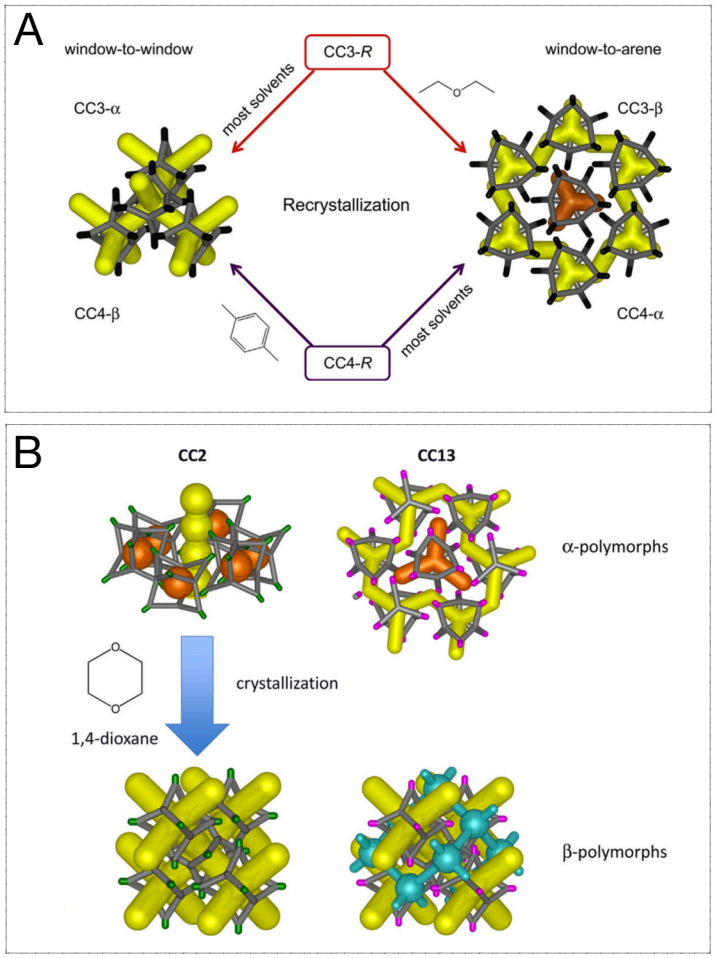


Figure 1.17.: A) Two cage molecules, **CC3** and **CC4**, can adopt isoreticular polymorphs by recrystallization in the presence of specific solvents. B) α-polymorphs of cages **CC2** and **CC13** and the β-forms synthesized by recrystallization in the presence of 1,4-dioxane, which makes them isoreticular to **CC3**-α. Figure adapted from Little, M. A.; Chong, S. Y.; Schmidtmann, M.; Hasell, T.; Cooper, A. I. *Chem. Comm.*, 2014, 50, 9465-9468, published by The Royal Society of Chemistry.

"Mix & Match" strategies

The "mix & match" property of porous organic materials allows one to potentially "customise" the porosity and the properties of the final material by combining different types of cages into the same crystal structure.[72,119] The assembly of the multi-component crystal is driven by a combination of multiple weak intermolecular interactions; however, if dominant supramolecular interactions, such as chiral recognition (as observed in the previous section) or solvent driven effects, can be identified and applied, then complex assemblies of cages can be formed. The first "mix-&-match"

example of porous organic cages was for the the co-presence of differently functionalised cages of opposite chirality packed into a diamondoid window-to-window fashion. Four examples of multicomponent organic crystal structures are shown in Figure 1.18.

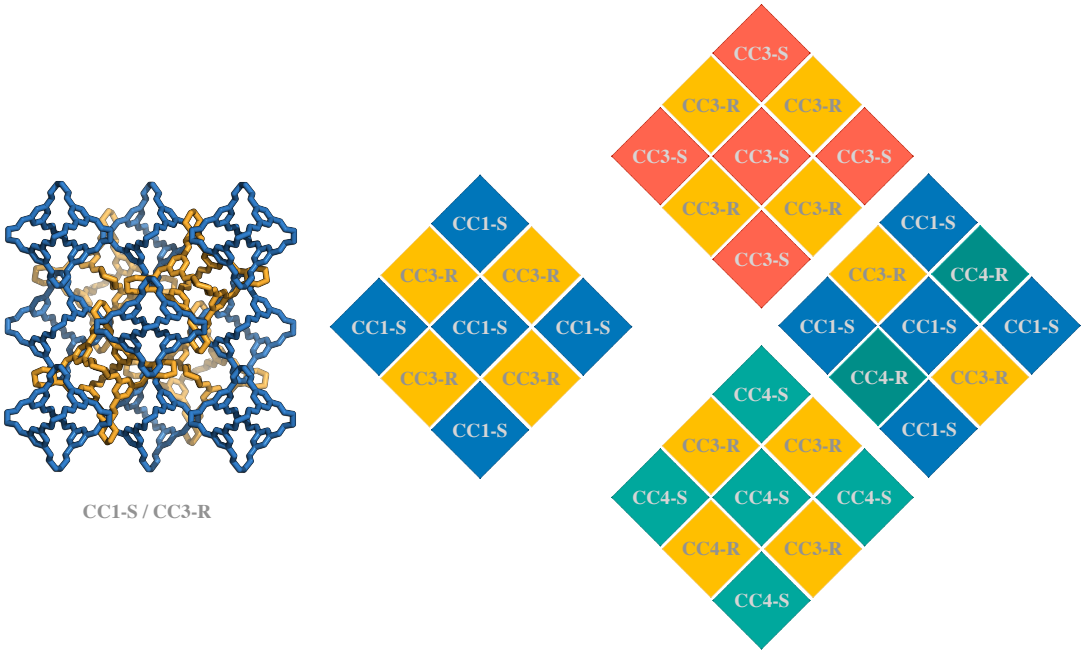


Figure 1.18.: Schematic representation of the formation of racemic and quasiracemic imine cage cocrystals. Each cage is neighboured by four cages of the opposite chirality. A chemical representation of the **CC3**-*R* and **CC1**-*S* co-crystal is shown on the left hand side, with **CC3**-*R* in yellow and **CC1**-*S* in blue.

The first attempt made regarded the crystallisation of a racemic crystal containing both the *R* and the *S* enantiomers of cage **CC3** (yellow and orange in the figure). It led to the discovery (via computational simulations) that not only was the racemic crystal more stable than the two homochiral parent crystals, but also that varying the experimental conditions allowed the crystal size and morphology to be controlled, therefore allowing its porosity to be tuned. Secondly, it was discovered that **CC3**-*R* and **CC1**-*S* could co-crystallise (yellow and blue scheme in Figure 1.18), the first being able to control the chirality of the second, which in the homochiral phase tends to interconvert into the other enantiomeric form.[71] **CC3**-*R* was also capable of co-crystallising with cage **CC4**-*S* (yellow and teal scheme in Figure 1.18) and to form a ternary co-crystal in which each window matched a window of opposite chirality belonging to either **CC1**-*S* or **CC4**-*R* (yellow, blue, and teal scheme in

Figure 1.18).

More recently, co-crystals assembling into a network different from the diamondoid could be obtained with the newly discovered tubular covalent cages (**TCC**) 1/2/3 which, if combined with cage **CC3**, can form a reticular structure in which **CC3** cages act as nodes and **TCC** cages as linear ditopic linkers.[120] Heterochiral **TCC1** and **TCC2** cages (which are also chiral) can pack with **CC3** into a solid that contains layers of one enantiomer of **CC3** pillared by **TCC** cages of opposite chirality, as shown in Figure 1.19.[120]

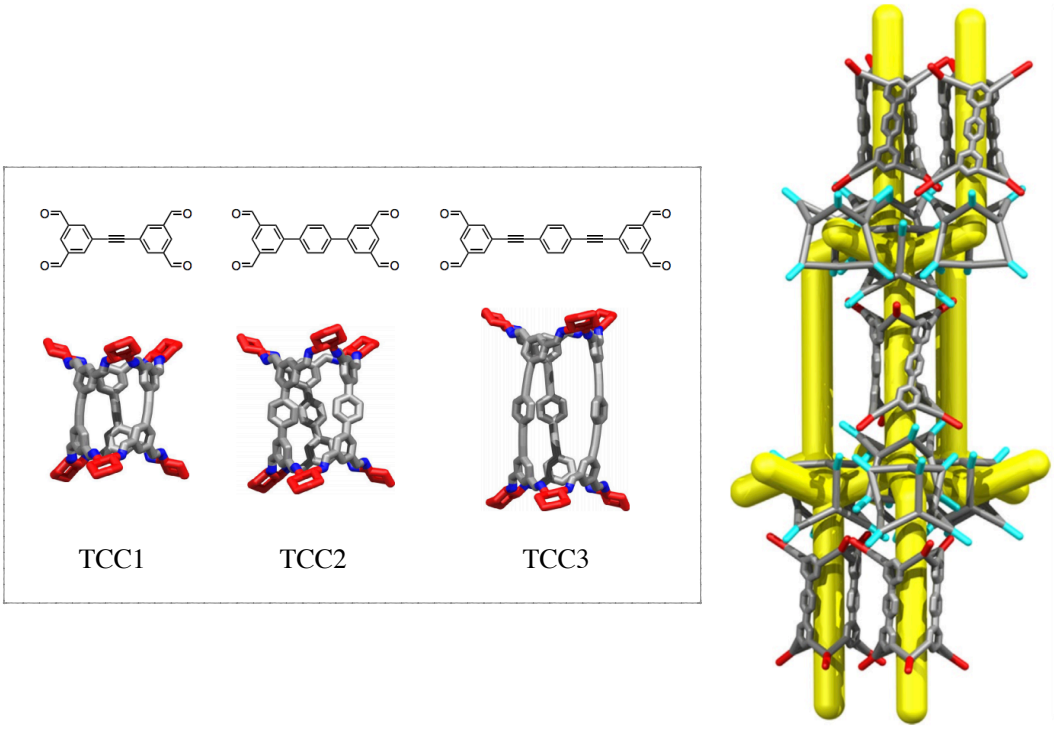


Figure 1.19.: Box: chemical representation of cages **TCC**(1/2/3), including a 2-dimensional representation of the different aldehydes contained in each molecule. The aldehydes' carbons are in grey, amines' carbons in red, nitrogens in blue. Right: Crystal structure of cages **TCC2**-*R* and **CC3**-*S* consisting of layers of **CC3** molecules packed and pillared by the **TCC2** cages. Red and cyan sticks simplify, for clarity, the cyclohexyl groups, red sticks belonging to molecules of *R* chirality, cyan to *S* chirality. The intrinsic pore network is represented in yellow. The image has been adapted from A. G. Slater, M. A. Little, A. Pulido, S. Y. Chong, D. Holden, L. Chen, C. Morgan, X. Wu, G. Cheng, R. Clowes, M. E. Briggs, T. Hasell, K. E. Jelfs, G. M. Day and A. I. Cooper, *Nature Chemistry*, 2017, 9, 17-25.

Porous Liquids

Porous liquids, a relatively recent promising field, allow some properties of microporous adsorbents to be combined with the transport properties of fluids.[121–123] Three definitions of porous liquid have been given to date: the first, in 2007, described it as a neat liquid comprising rigid porous molecular hosts, the second as a solution of molecular hosts and the third as a dispersion of porous nanoparticles. Porous organic cages were not ideal to make porous liquids of the first type, as they would generally decompose before melting.[124]They were, however, suitable for type 2 and different studies[122,125] have demonstrated their application to improve methane solubility, with scrambled cages performing particularly well as hosts, improving not only methane solubility, but also of N_2, CO_2, and Xe.[126]

1.5.3. Computational investigation of the solid state

Sometimes the experimental techniques available are not sufficient for a full description of a system and certain characteristics need to be analysed and/or understood with computational simulations. In other cases, simulations can be adopted to predict the possible outcome of a reaction or behaviours of the solid state; they are therefore adopted to direct the synthesis towards specific targets, with the additional goal of reducing the time spent for the unsuccessful synthesis or characterisation of materials. Details about the computational description of porous organic cages will be discussed in detail in the Methods section and in the Results chapter of this work. However, an overview of the principal techniques used to analyse single molecules and to predict the structures and properties of the solid state are briefly covered in the next few sections.

Crystal Structure Prediction (CSP)

The intuitive *a priori* prediction of the packing arrangements molecules can adopt in the solid state is extremely challenging. CSP techniques aim at the reliable prediction of how molecules will crystallise in the solid state, with only the chemical structure of a single molecule and the crystallisation conditions known.[127–130] This would allow chemists to analyse and investigate the structures and properties of the solid state before the synthesis is attempted and could direct the experiments towards the most promising candidates.

Typically, the first step to perform CSP involves using Density Functional Theory (or other methods) to compute the starting geometry of the single molecules and its

multipole charges, followed by annealing Monte Carlo (MC) force field simulations or quasi-random sampling to generate some potential packing arrangements. The computational cost is acceptable if the search is initially performed on the most commonly observed space groups. After this step, the structures are ranked depending on their lattice energies at an atomistic level then, in most cases, re-ranked with higher-level methods in order to gain a more accurate energetic description.[129]
It is particularly useful for a number of applications pertinent to porous organic cages, some of which are listed below:

- Identifying the crystal arrangements of different polymorphs and calculating the difference in their lattice energies.[118,131]
- Understanding from an energetic point of view why some systems only have one dominant polymorphic form.
- Given that some porous organic cages are chiral, CSP can be applied to determine whether an enantiopure or racemate solid state form is preferred.
- Gaining insight about the way instrinsic and/or extrinsic pores are connected and whether they form guests-accessible channels.
- Predicting the effect of solvent on the way molecules pack together, comparing solvent-loaded and desolvated crystal structures.[132] The procedure is non-trivial, as the presence of solvent might interact with the crystal structure in a plethora of ways, *e.g.* by changing its polymorphic phase or by altering the intermolecular interactions. Additional simulations on individual molecules or dimers can be performed to calculate specific adsorption energies, to compare the stability of solvent present within a cavity or in the window-to-window space.
- Predicting the full crystal energy landscapes.

More recently, computational crystal structure prediction has been combined with property prediction to build energy-structure-function maps describing the possible structures and properties that are available to a specific molecule.[49] The likelihood of each structure being stable and experimentally feasible is linked to its predicted lattice energy. The maps were used to identify a highly porous solid, **T2**, a trypticene-based porous molecular crystal with benzomidazolone functionalities, which had the

lowest density reported for a molecular crystal so far. The turning point of this work was that both the structure of the crystal and its physical properties were predicted using the molecular structure as the only input.[49]

Simulations on amorphous materials

Simulations on amorphous phases are normally carried out because non-crystalline states are experimentally much harder to characterise and therefore simulations can provide insights into phenomena occurring at the microscopic level. The realistic prediction of the amorphous state is non-trivial, however different techniques have been tested in the literature.[23,79,119] The earliest examples involved simulations of amorphous graphene sheets, as interconnected carbon fragments or as carbon discs in order to obtain porous carbon systems.[133] A simple approach was developed by Cooper and co-workers to generate a model for the amorphous "scrambled" mixture of cage molecules, obtained by combining the amine precursors used for cages **CC1** and **CC3** in 3-component "scrambled" cages.[134] In that case, a cell was loaded with a distribution of the cage species using a MC simulation, followed by consecutive molecular dynamics simulations until the targeted experimental density was reached. The results shed light on microscopic information not available from simple experimental characterisation; for example, it was found that 80% of the void volume came from extrinsic porosity given by the inefficient packing of the cages, justifying the higher level of porosity of the amorphous mixture when compared to the porosity of the parent "unscrambled" crystalline cages, **CC1** and **CC3**. The problem at this stage was that simulations crucially relied on experimental target density, which is unknown if the aim is to design hypothetical molecules or accurately simulate microporous solids.

In the following years, Colina and co-workers generated a simulation procedure to generate models of hyper cross-linked polymers, which included the generation of the structure followed by a 21-step compression and slow decompression of the system that led to densities close to those of the experiments.[135] Properties such as density and pore volume emerged from the simulation and therefore did not have to come from the experiments; the approach performed well on polymers with intrinsic microporosity[136] and structurally similar amorphous porous organic molecules.[137] In the latter case the molecules were packed and arranged at either a low or specified final density and were either compressed and relaxed with molecular dynamics or annealing.[137]

A later work from Jiang *et al.* showed that molecular dynamics simulations could

be used to rationalise gas selectivity in amorphous solids and to gain insights on properties such as self-diffusion, gas diffusion trajectories and gas hopping mechanism.[138] Their procedure was based on four steps. Initially a cell was loaded at low density with single cage molecular structure taken from SCXRD structures in which the molecules were treated as rigid bodies; molecular dynamics combined with a microcanonical ensemble was used to stabilise the environment, the cage molecules being then fully flexible to maximise favourable intermolecular interactions; an NPT simulation followed to compress the cell, the last configuration of which was geometry optimised.[138] Molecular dynamics was then applied again to study the properties of the material in terms of gas sorption. Doonan and co-workers improved the strategy by using GPU-accelerated molecular dynamics and Voronoi network analysis to provide insight into structure-property relationships.[139] The study was applied to a selection of porous organic cages obtained with both reversible and irreversible bond formation.

Properties investigation with MD

Once a crystal structure is available, either from experiments or simulations, modelling can be performed to investigate the properties of the porous materials, both to understand the selectivity towards certain kind of guests and to get more insights about the way guest molecules move within the crystal structure. Molecular dynamics simulations can be used to better understand behaviours dependent upon host flexibility. In these circumstances, treating the host molecules as "rigid" can neglect factors that need to be considered when rationalising the observed porosity of the solid[140–143] and in order to obtain more information about structural changes on a short timescale. MD trajectories have been used to visualise the accessible pores within a structure and to rationalise the presence of transient access to formally disconnected voids[140,142] and also to understand the diffusion of adsorbents in guest-loaded structures, which allowed researchers to describe not only the uptake of gaseous guests,[115,125,138,141] but also how cages performed in dynamic molecular separations and ion conduction.[141,144,145] MD simulations generally last in the order of tens of nanoseconds, unless larger guests need to be studied in which case longer times should be used to observe rare diffusion events.[141,146] With larger guests, umbrella sampling[147] or metadynamics techniques[146] can also be applied to investigate the diffusion mechanisms.

Properties investigation with stochastic methods

Monte Carlo (MC) simulations combined with force field methods can be applied to study guest sorption by generating random moves to sample possible sorption sites. This allows thermodynamic information to be calculated at a range of different pressures. Geometry optimisation can be used to study host-guest interactions, whereas a full adsorption isotherm can be calculated with grand-canonical MC (GCMC) simulations, which could help understanding which parts of the structure are contributing to the porosity of the material. When compared with simulations on network materials, MC simulations on porous molecular materials have an additional challenging variable, which depends on the higher molecular mobility of the molecules within the solid state structure. MC simulations are usually carried out on the rigid host to avoid structural changes in response to guest sorption, which can significantly alter the properties of the material.[83] Calculations of host-guest binding energies can be useful to predict the selectivity of cages for specific guests[145] and to gain more insight into the effect of supramolecular interactions.[73] Furthermore, GCMC simulations have been used to predict the selectivity of various cage molecules towards different gaseous guests, not only in the solid state[141,148] but also as porous liquids.[123,125] MC docking calculations were used to rationalise the experimentally observed N_2 uptake in cage **CC2**. The dynamic motion in the crystal was presumed to allow gas sorption, even though the void seemed isolated according to the Connolly surface area representation.[7] Multiple GCMC simulations of gaseous guests on the calixarene molecule TBC4 and on the TTB system[149,150] were useful to discover that GCMC tends to overestimate the gas uptake in porous materials, probably because crystal packing effects are excluded in the simulation and treating the host as a rigid body does not take into account important motion factors.

GCMC and MD studies have been recently used to study the separation properties of **CC3**. Simulations (that were used for gas diffusion in the crystal structure and to calculate adsorption isotherms) showed that **CC3**-α preferentially adsorbed krypton, xenon and radon, due to a precise size match between the rare gas and the organic cage cavity.[141] Within the same work, simulations were used to show that **CC3** can also be applied for the enantioselective separation of chiral organic molecules.[141] Both methods were again combined to study the tuning of gas sorption behaviour in the co-crystal of cages **CC1**, **CC3** and **CC4**. GCMC was used to load the material, by simulating the gas adsorption, whereas MD was adopted to study the the diffusion mechanism that allowed gas to access formally occluded voids.[142]

The simulations led to the discovery that gas diffusion in the material was controlled by the cage vertices and that the cyclopentane diamines of **CC4** allowed N_2 to diffuse more easily than the cyclohexane diamine vertices of **CC3**.[142]

High-throughput screening

High-throughput screening of molecules can be used to investigate the structure-property relationship of materials. The computational prediction of functional properties has been more thoroughly performed in recent years via high-throughput screening methods.[151] In the past, more than 100,000 hypothetical MOFs have been screened for rare gas,[152] CO_2 selectivity[153] and methane storage capacity,[154] which allowed the rationalisation of design principles that could be followed to target specific properties.[152,153] McKeown *et. al.* screened the Cambridge Structural Database (CSD) in order to find promising candidates for porous solids. Candidates were obtained by targeting only structures with a density lower than 0.9 g cm^{-3} (porous molecular materials usually have low density) and rejecting molecules containing saturated hydrocarbons and inorganic fragments (which are normally non-porous).[47] In a similar way, Mastalerz and co-workers screened the CSD database looking for BBs to be used for extrinsically porous materials. This led to the discovery of TTBI (mentioned earlier in the chapter), which displayed exceptional porosity.[48] In 2016 Evans and co-workers screened the CSD for porosity, combining high-throughput screening and machine learning techniques, and identified 481 potential porous organic crystals from a start of 150,000.[17]

1.5.4. Computational investigation on single molecules

Single cages can be assembled and analysed with a combination of computational methods in order to understand their likelihood of collapse or of remaining shape persistent upon desolvation or, more recently, to describe their open geometry in solution (details will be discussed in Chapter 3). Some properties of organic molecular materials can be predicted from simulation of the individual cage building block, which reduces substantially the computational cost required to perform CSP modelling and the time eventually required for the synthesis of cages without any prior knowledge of their shape persistency. Different studies have been carried out to predict the final topology of a reaction outcome, even though, as discussed in Section 1.4, emergent behaviour is a major challenge in the prediction. Generally, conformer search methods are applied on the initial guess in order to sample the potential en-

ergy surface (PES) surrounding it and to locate the lowest accessible minimum.[83] If these calculations are performed in the gas phase, the lowest energy geometry corresponds to the experimental structure upon desolvation.[83] Calculations including solvent are comparatively non trivial and more rare, and they will be discussed more extensively in Chapter 3. Computational methods can, additionally, be applied to assemble precursors into different molecular sizes and shapes in order to compare the relative stability of a group of molecules generated from the same pair of precursor, but characterised by different underlying topologies. This can provide information about their likelihood of experimental formation or about their chemical/ physical properties.[83] Such analysis requires a combination of classical and quantum mechanics methods, as will be discussed in Chapters 2 and 5.[83]

Computational investigations on single cages have been used in the past to predict solid state properties such as separation[145] and sensing.[155] Additionally to the studies described in Section 1.4, investigations on the large [8+12] cage **CC7** showed that the cage collapsed via a vertex folding mechanism (where the vertices of the cage collapse towards the centre of the cavity, creating stabilising π-interactions).[96] However, the collapse could be prevented by redesigning the cage to be more rigid, thereby reducing the number of torsional degrees of freedom that led to the collapse.[96] Simulations have also been used to study the conformer interconversion of cage **CC1**, which can influence the properties of the material in the solid state.[115] The interconversion is based on the rotation of each of the twelve imine bonds in the cage; it was discovered to be characterised by very low energy barriers, additionally lowered if solvent was present. In the solid state, the ability to interconvert is reflected by the polymorph transition described in Section 1.5.2. The chiral interconversion from the *R* to the *S* enantiomer of cages **CC1**, **CC2** and **CC13** was further investigated in a more recent work.[156] Chiral interconversion was found to occur in solution for cages **CC1** and **CC13**, whereas the single vertex methyl groups lock the chirality in place for **CC2**.[156] Vanka *et. al.* used DFT methods to calculated the energy barriers involved in the dehydrogenation of ammonia borane catalysed by cage **CC1** and suggested modifying the cage with phosphorous to improve the catalytic efficiency.[157]

In a recent paper,[158] Sholl and co-workers suggested predicting the relative stability of organic cages not by comparing the relative energies of the final products, but by calculating at the DFT level the cost of each step towards the assembly of the final cage structure, starting from a fragment including only the two initial precursors. This procedure could help understand whether, even though a cage is

predicted to be shape persistent, the energetic cost of its assembly is too high and does not lead to its formation in solution.[158]

1.6. Structure of the thesis

In the next few chapters I will try to give a broad overview of the work I have carried out on porous organic cages in the past few years. The idea behind the project was to develop *in silico* methods, not only to predict the shape persistency of a reaction outcome, but also to predict which topology was preferred. This involved considering multiple factors from solvent effect to the most feasible topologies organic cages could adopt. Then, it involved implementing a computational prediction pipeline, which needed to be applied first on cages already available in the literature in order to validate the approach, then on a broader variety of systems in order to understand if it could be used to direct the synthesis towards selected promising candidates.

In this context, Chapter 2 will provide the necessary theoretical background to understand the computational methods applied throughout the work. Chapter 3 will be dedicated to the investigation of the scaffolding effect of the solvent on porous organic cages, which involved devising a computational strategy to "inflate" collapsed cages in order to study the geometry they adopt when in solution. In Chapter 4, I will define a nomenclature for the classification of porous organic cage molecules, enumerating the 20 most probable topologies they may adopt experimentally, 12 of which have been synthetically realised to date. In Chapter 5 I will discuss the implementation of the prediction pipeline, and how it was tested on multiple systems obtained from a series of 10 different building blocks from the literature. The successful prediction obtained on experimentally available systems, made me move forward to the next step, which involved predicting the topology and shape persistency of systems not yet synthetically available. The final step of the work will be described in Chapter 6, where I will discuss how the computational prediction methods were applied to a synthetic high-throughput screening of 78 new cages and how both topology and shape persistence prediction performed when compared with the experimental work. In the last chapter of this thesis, I will outline conclusions and future work.

Chapter 2
Methods and theoretical background

In this chapter I will describe the principal methods involved in the computational investigation of the organic cages studied in this thesis. In the first few sections, the definition of potential energy surface and of the main classical mechanics methods used to explore it will be covered. Then, I will discuss electronic structure theory methods, which were mainly used in this work to refine the results obtained with classical mechanics to a higher level of theory. Finally, I will describe the general computational pipeline I developed for the investigation of porous organic cages, which then was modified in each chapter depending on the focus of the project discussed.

2.1. Potential energy surfaces

Potential energy surfaces (PES) describe the energy of a molecule as a function of its coordinates and, therefore, they can provide information about the kinetic and the thermodynamic properties of a system. The motion of the atoms' nuclei on the PES can either be described classically (without explicitly accounting for electron interactions) or with quantum mechanics; both methods will be discussed in the next sections.

If there are N nuclei in a molecule, the dimensionality of the PES is $3N$, meaning that there are $3N$ nuclear coordinates that define the geometry. Of these coordinates, three describe the translation of the atom/molecule and three the rotation of the atom/molecule with respect to the three dimensional axes, leaving $3N$ - 6 modes to describe the vibrations of the system occurring through atomic displacement. Given all these variables, calculating the complete PES of systems containing

more than a few atoms is not feasible. The goal is, thus, to only explore the chemically relevant regions, which can be done by locating selections of stationary points on the PES corresponding to significant geometric arrangements of the molecule. A one dimensional representation of a hypothetical PES with such significant points highlighted is shown in Figure 2.1. The most relevant parts of the PES are usually nuclear arrangements characterised by low energies and, specifically, the minimum energy points that relate to metastable configurations of the system. Among the minima located on the PES, the global energy minimum is the point where the molecule is most stable, *i.e.* it has the lowest energy.

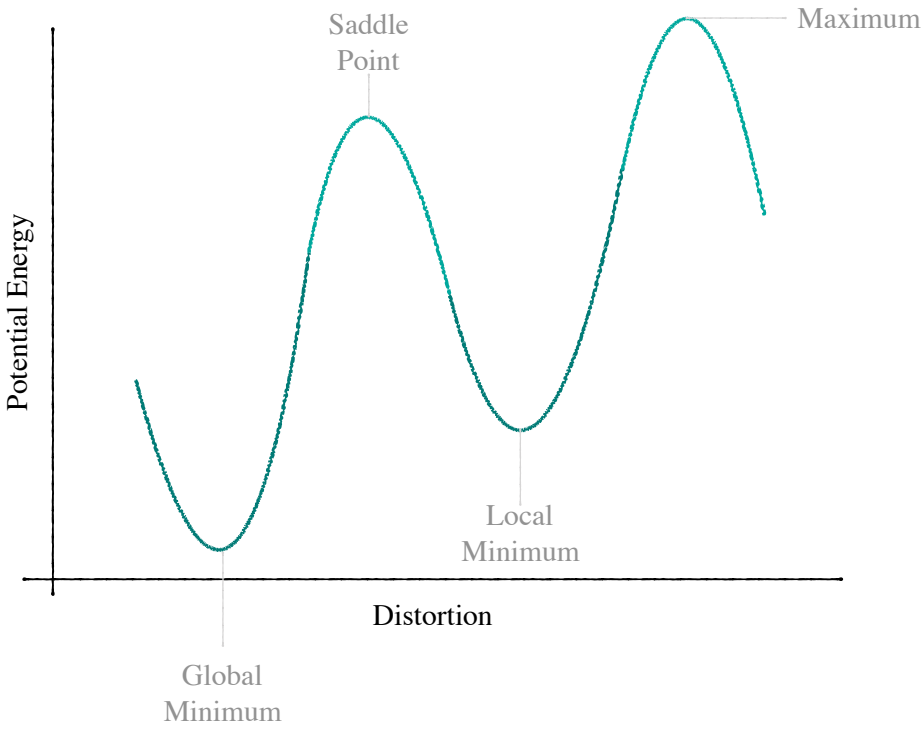


Figure 2.1.: One-dimensional representation of a potential energy surface.

A stationary point is reached when the forces acting on the atoms are zero and, if such stationary point is a minimum, any displacement would cause an increase in energy. This implies that at this point on the PES, the first derivative of the force is zero and the second derivative is positive (motion along any of the $3N$-6 coordinates increases the gradient):

$$\frac{\partial f}{\partial x_i} = 0 \qquad \frac{\partial^2 f}{\partial x_i^2} > 0 \tag{2.1}$$

Minima are connected on the PES via saddle points, *i.e.* stationary points in which the first derivative of the PES is zero and the second derivative is negative. Saddle points correspond to intermediate configurations molecules need to cross in order to reach more stable states. Transition states correspond to saddle points where only one of the possible $3N$-6 vibrational modes is negative, which, if followed, will lead to a decrease in energy. If all the vibrational modes are negative, the molecule stands on a global maximum on the PES (the highest point in energy), and any mode would lead to a lower energy configuration.

Computational methods have been developed to search the PES of molecules and different kinds of simulations can be chosen depending on the size and chemistry of the system investigated. Geometry optimisations are used to locate the closest minimum with respect to the initial guess, whereas saddle points search methods identify transition states. The PES can be more thoroughly explored with simulation techniques such as molecular dynamics (MD) and Monte Carlo.

2.1.1. Geometry optimisation

Geometry optimisation (or energy minimisation) aims at decreasing the energy of the system by applying a displacement on the coordinates of the atoms, with the goal of finding a minimum. Finding the global minimum in a PES is always a very challenging goal; however, if there is any chance to locate it, *a priori* knowledge of the equilibrium geometry is recommended and a good initial guess geometry should be defined to try and be in the minimum proximity.

PESs can not be explored using analytical methods, as calculations would be too computationally expensive (unless only very small molecules are involved) and generally, numerical methods are applied instead. Many algorithms are available for this purpose and they are used to compute derivatives of the energy of the atoms to understand how to displace the coordinates: the first derivative (the gradient) provides information about the steepness of the curve and thus about the direction to follow to find the minimum, whereas the second derivative (the Hessian) provides information about the curvature of the slope. Different methods calculate the function and its derivative with finite precision, which depends on the accuracy/cost balance of each particular implementation. Generally, the optimisation can be considered converged if the gradient reaches a suitable cut-off value, or if the change in the energy function between two subsequent steps is small enough.

There are three main methods that can be adopted to minimise the energy of a

system, the first two based on computation of the gradient only (this is less accurate and efficient, but cheaper), the last taking into account the second derivative as well:

1. Steepest Descent
   All steps are computed in the direction of the gradient. The steepest descent is a simple algorithm and it only requires storage of one gradient vector per step; however, it is not very accurate and it is normally just used to "relax" a poor initial guess before more advanced algorithms take over.

2. Conjugate Gradient
   The search for the minimum is not performed along the gradient direction, but along a line that is "conjungate" to the previous search directions. The conjugate gradient method has much better convergence characteristics than the steepest descendent method and it works at its best with quadratic, or quasi-quadratic surfaces. In this thesis, the Polak-Ribiere conjugate gradient (PRCG)[159] is used, combined with the OPLS forcefield[160] within the Macromodel (version 10.3.015, Schrödinger, LLC, New York, NY, 2015-3) package, to minimise low energy conformers before feeding them to more accurate electronic structure methods.

3. Newton-Raphson
   In this method, the Hessian (the matrix containing the mass weighted second derivatives) is diagonalised and stored at each step to obtain information about the curvature of the surface. The method is by nature more accurate and efficient than the previous two, however, computing the full Hessian matrix can be very demanding, particularly if a large number of atoms (of the order of 10^6) are involved. Quasi-Newton methods, which do not compute the full Hessian at every step and that update it over successive iterations, can be adopted instead.

In this thesis, every time a structure was optimised by using electronic structure methods within the CP2K software,[161] the Broyden-Fletcher-Goldfarb-Shanno (BFSG)[162] Quasi-Newton method was adopted. Within BFSG, the differences between previous and current atomic positions are calculated and used to build an approximation of the Hessian matrix.

2.2. Exploring the potential energy surface

In the previous section I described potential energy surfaces and how to locate minima in proximity of the initial guess. In this section, I will discuss the methods used throughout this work to explore the PES more thoroughly in order to locate not only the closest minimum, but also to obtain a broader view of the conformational space surrounding the starting point. In this thesis, force field methods based on classical mechanics were adopted to explore the PES to locate the lowest energy minimum of the molecules investigated, and this was done with either the low-mode search method or molecular dynamics, which are both described in the next following subsections. In the molecular dynamics section I will also discuss constrained molecular dynamics, which was used to locate higher energy open conformations of molecules when the lowest energy minima corresponded to a collapsed structure.

2.2.1. Forcefield Methods

Classical mechanics methods treat atoms in molecules as spheres connected by bonds and represented by springs, whose energy is defined by angles and torsions. They do not explicitly take into account the motion of the electrons and, like the quantum mechanical methods described later, implicitly adopt the Born-Oppenheimer approximation, which states that the motion of the nuclei is so much slower than that of the electrons (being that the mass of a proton is about 1800 times that of the electron) that only nuclei positions need to be considered, the electrons adapting instantly. Treating molecules as atoms connected by springs allows CM methods to be able to simulate large numbers of atoms (up to tens of millions) with a relatively small computational expense.

Force fields (FFs) use classical mechanics to describe the potential energy of a system as a sum of intermolecular and intramolecular terms, ignoring the electronic degrees of freedom. The contributions to the total potential energy can be in the form of self-energy, two-body, three-body or four-body terms, each describing the energy contribution of a displacement. Terms larger than four are generally not included in the series, as their contribution to the total energy is not significant.

FFs are parameterised so as to reproduce either experimental properties or the results of *ab initio* calculations, which has the downside of making different FFs specific to the chemistry of particular systems, when it would be ideal to have FFs that can be as transferable as possible. When using FF methods, the potential

energy (V) of a system is typically calculated as:

$$V = \sum E_{bond} + \sum E_{angle} + \sum E_{torsions} + \sum E_{non-bond} \tag{2.2}$$

where:

$$E_{bond} = \frac{1}{2}\, k_{ab}\,(l - l_0)^2 \tag{2.3}$$

$$E_{angle} = \frac{1}{2}\, k_{abc}\,(\theta - \theta_0)^2 \tag{2.4}$$

$$E_{torsions} = \frac{1}{2}\, k_{abcd}\,(l - l_0)\, cos\omega \tag{2.5}$$

The first three terms are intramolecular terms (respectively two-, three-, and four-body) describing the deviation of, respectively, bond lengths l (Equation 2.3), angles θ (Equation 2.4) and torsions ω (Equation 2.5), from the ideal equilibrium value parametrised for that FF; k is Hook's law's constant.

The last term of Equation 2.2 includes two terms: the van der Waals forces and the Coulombic electrostatic interactions. Van der Waals forces include:

- The description of the long-range dispersive forces coming from the instantaneous dipoles forming in particles, which then induce dipoles in the neighbouring particles, resulting in an overall attraction between all particles.
- The description of the short-range repulsive forces that forbid electrons from two different atoms to occupy the same region of space (in agreement with Pauli's exclusion principle).

Van der Waals interactions are commonly described by the Lennard-Jones potential and have the shape of a potential energy well:

$$E_{van\,der\,Waals} = 4\epsilon \left[\left(\frac{\sigma}{r_{ab}} \right)^{12} - \left(\frac{\sigma}{r_{ab}} \right)^{6} \right] \tag{2.6}$$

where ϵ determines the depth of the well and σ is the point where the van der Waals energy is zero (collision diameter). This description makes sure that i) when the distance between the particles is infinite, the interaction is zero; ii) when the particles approach, the energy is lowered to a minimum corresponding to the equilibrium distance; iii) as the atoms get closer, the energy rapidly increases because of repulsion when the distance between the two particles is too small.

The long-range Coulomb interactions are described by Coulomb's law:

$$E_{Coulomb} = \sum_{a=1}^{a=N} \sum_{b=a}^{b=N} \frac{q_a\, q_b}{4\pi\, \epsilon_0\, r_{ab}} \tag{2.7}$$

where q_a and q_b are the atoms' point charges, ϵ_0 is the constant for permeativity of free space and r the distance (inversely proportional to the electrostatic energy) between atoms a and b.

Two kind of force fields were used in this work, the Universal Forcefield (UFF)[163] and the Optimized Potentials for Liquid Simulations (OPLS) forcefield.[161] The first is very general and contains parameters for every atom, which were estimated using general rules about the atomic elements and their characteristics (such as hybridisation, connectivity and so on). Given its potential general use, it is not particularly accurate and, during my PhD, I only used it to "clean" the geometries of very distorted hand-drawn molecules, before applying higher level methods.

The OPLS forcefield was used previously for the description of porous organic molecules,[98,164,165] as its parametrisation was found to suitably describe the imine bond. It was initially developed to fit experimental properties of liquids, such as density and heat of vaporization, then for the description of organic large systems, such as proteins or nucleic acids. In the OPLS force field the Lennard-Jones potential is in the form:

$$E_{van\,der\,Waals} = \frac{A}{r^{12}} - \frac{C}{r^6} \tag{2.8}$$

Where A and C are parameters that can be calculated exactly for specific systems. The project described in Chapter 3 was based on calculations performed with OPLS-2005,[166] the 2005 version of the OPLS all-atoms FF,[167] in which all the atoms are included explicitly. In the 2005 version, the bond stretching and angle bending parameters for peptides were mostly adapted from the AMBER all-atoms FF,[168] the torsional parameters were derived from *ab initio* Hartree-Fock calculations for more than 50 organic molecules and ions and the non-bond terms were developed via stochastic Monte Carlo simulations for 34 pure organic liquids.

All the simulations for the remaining projects were carried out with the more recent OPLS3 FF,[160] which, compared with the previous version, offers a better description of intermolecular interactions and of some dihedral parameters, among which the one containing the imine bond, crucial for the description of organic imine cages. The amount of reference data and associated parameters is an order of magnitude greater than that of the previous version,[160] offering a more accurate

description of the chemical groups contained in organic molecules.

2.2.2. Conformer search with the low-mode search method.

Conformer search methods are usually employed to explore a large set of local minima, with the aim, in my case, to be able to locate the one that corresponds to the lowest possible energy for the molecules investigated. The lowest energy conformers I found in the gas-phase are considered here equivalent to the experimental desolvated geometry of an organic cage and their identification helped in the prediction of reaction outcomes.

The low mode search, initially devised by Kolossováry and Guida in 1996,[169] was only used for some of the projects involved in this work, as, although accurate, it was sometimes computationally expensive, particularly for systems with a large number of atoms (>200) and with many large macrocycles/rings. It was thus replaced, in more recent calculations, by molecular dynamics carried out at high temperatures, which could run on a shorter time-scale.

The low mode search is an eigenvector-following method and it proceeds according to the following steps:

1. The simulation starts at an arbitrary energy minimum conformer. The non-mass-weighted Hessian matrix is diagonalised to calculate eigenvalues and their associated eigenvectors (vibrational modes) and the latter are stored in an array, where the number of eigenvectors stored is set by the user, but that, generally, includes all the vibrations within 250 cm^{-1}.

2. The geometry is then distorted following the direction of a selected low-frequency mode, generating a new structure.

3. The energy of the new structure is evaluated and step 2 is repeated (without rediagonalising) until the energy difference per step is greater than a threshold set by the user.

4. The geometry is then reoptimised to a minimum, potentially finding a new local minimum. If a new minimum is found, step 1 is started again. If not, a new eigenvector is selected and the initial geometry is distorted accordingly (step 2).

5. Once all eigenvectors have been followed in both directions, a new direction is chosen by randomly selecting a combination of eigenvectors at step 2, where

the search starts again.

Diagonalisation only occurs once per minimum, making the method much cheaper than traditional eigenvector-following methods, in which the Hessian matrix is diagonalised at each step. Within the projects that used the low mode method, all conformations sampled within 25 kJ mol^{-1} from the lowest energy minimum were considered.

2.2.3. Molecular Dynamics

Molecular dynamics (MD) methods solve Newton's equation of motion for atoms on the PES, which is generally used to obtain information about the dynamic evolution of a system in time. The local area around the starting point can be explored if sufficiently high temperatures are adopted to increase the energy and overcome the barriers. Different local minima can be located by sampling configurations at suitable intervals during the time of the simulations (normally in the scale of pico- or nanoseconds, as small steps need to be used to integrate Newton's equation) and that can be subsequently optimised.

During the simulation, the nuclei are treated as classical bodies subjected to Newton's second law, $F{=}m_a\boldsymbol{a}$, where $\mathbf{a}$ is the acceleration. The law is solved as a differential equation to calculate the trajectory:

$$\frac{\partial^2 \mathbf{r}_a}{\partial t^2} = \frac{F}{m_a} \tag{2.9}$$

Where $\boldsymbol{r}_a$ is the position of the atom a with mass m_a and F is the force acting on it. The equation of motion is not normally integrated directly, but via finite difference methods to calculate the trajectory of the particles:

1. The total force on the atom is calculated and, subsequently, the acceleration.

2. Knowing the current position $\mathbf{r}$ and the velocities $\mathbf{v}$, the acceleration $\mathbf{a}$ can be calculated and used to compute the positions and velocities at a time step in the future.

3. If the force is considered constant and the time step is sufficiently small, an accurate description of the time-dependent evolution of the system is obtained.

The most commonly used algorithms for MD are the Leapfrog Verlet and the Velocity Verlet algorithms, both based on the original Verlet algorithm.[170] The latter

used the positions and accelerations at time t and the positions of the previous step, to calculate those of the next step (although the method lacked precision, as velocities were not explicitly calculated). The Leapfrog Verlet algorithm improves the Verlet by implicitly including the velocities; the Velocity Verlet by having positions, velocities and accelerations known at each point.

In the projects discussed in this thesis, MD is used to sample the PES with the OPLS3 FF in the Macromodel software, which adopts the Velocity Verlet algorithm. A temperature of 1000 K is applied to better sample the configurational space, the simulations last 100 ns, with a time step of 1 fs, each structure being sampled and geometry optimised every 10 ps. The simulations are carried out in the NVE ensamble, which means that the number of particles (N), the volume (V), and the energy (E) are kept constant, whereas temperature and pressure can vary.

Constrained Molecular Dynamics

The methods described so far, focus on the lowest energy reaction path (as they were used to locate the lowest energy minimum). However, higher energy areas of the free energy surface (on which the activation energy of a reaction depend) can also be explored with dynamical methods. The sampling of the higher energy reaction path with the methods described in the previous sections is not feasible, as they normally (unless calculations gets unfeasibly long and expensive) cover the space within 25-50 kJ mol^{-1} from the lowest energy minimum and rarely visit higher energy regions of the surface.

In order to reach these higher energy regions, the sampling (which can be still performed with MD) must be biased towards those regions by adding constraints to the simulations, which means that one forces the free energy of the system to change as a function of a particular reaction coordinate. This can be achieved by applying a penalty acting on the energy surface as a biasing potential (U):

$$V'(\mathbf{r}) = V(\mathbf{r}) + U(\mathbf{r}) \tag{2.10}$$

which, for example, can have the form of a harmonic function centred in $\mathbf{r}_0$:

$$U(\mathbf{r}) = k_U(\mathbf{r} - \mathbf{r}_0)^2 \tag{2.11}$$

where k_U is the width of the harmonic well. If the biasing potential is steep enough (*i.e.* if k_U is large), the region of the free energy surface far from $\mathbf{r}_0$ will be so high in energy that only the region surrounding $\mathbf{r}_0$ will be explored. If one performs a series

of simulations locating the biasing potentials in different positions, then the Potential Mean Force (PMF) is simulated, *i.e.* the free energy along the reaction path. The main drawback of the method is that the location of the biasing potential must be selected manually, requiring some *a priori* knowledge of the reaction coordinate.

Constrained MD simulations were performed for the project described in Chapter 3, with the aim of locating higher energy open conformers of densely or partially collapsed geometry. Biasing potentials were applied with the PLUMED plugin to MD simulations,[171] with the purpose of favouring the exploration of the higher energy regions of the conformational space where open conformers could be found. More details about the specific procedure that was developed will be given in Chapter 3.

2.3. Electronic structure theory

When performing calculations based on electronic structure theory, the classical description of the systems is lost, and quantum mechanics is applied to calculate the energy. Properties can in principle be derived by solving the Schrödinger equation:

$$\hat{H}\Psi = E\Psi \tag{2.12}$$

where ψ is the wave function of the three spatial and one spin coordinates for each particle and $\hat{H}$ is the Hamiltonian, whose eigenvalues are the energy levels of the system. The Hamiltonian takes the form:

$$\hat{H} = \hat{T}_e + \hat{T}_n + \hat{V}_{ne} + \hat{V}_{nn} + \hat{V}_{ee} \tag{2.13}$$

where the kinetic component $\hat{T}$ is divided into $\hat{T}_e$, the kinetic energy of the electrons, and $\hat{T}_n$, that of the nuclei. The potential energy $\hat{V}$ takes into account three kinds of interaction: i) $\hat{V}_{ne}$, nucleus-electron; ii) $\hat{V}_{nn}$, nucleus-nucleus, iii) $\hat{V}_{ee}$, electron-electron. The Schrödinger equation can be solved exactly only for very simple systems (like the hydrogen atom), due to its $3N$ dependency on the number of particles ("many-body problem"). If larger many-electrons systems are considered, approximations need to be taken into account. The Born-Oppenheimer approximation (mentioned already in Section 2.2.1), allows simplification of the problem, by considering the motion of nuclei and electrons uncoupled and thus neglecting the motion of the former. According to this approximation, the kinetic energy of the nuclei is considered zero and their potential energy is constant for a fixed geome-

try. Thus, the energy of the system parametrically depends on the coordinates of the nuclei. The time-independent Schrödinger equation for a system containing N electrons (with coordinates $\mathbf{r}$) and M nuclei (with coordinates $\mathbf{R}$), the Schrödinger equation can be expressed as:

$$\hat{h}_e\psi(\mathbf{R},\mathbf{r}) = \epsilon(\mathbf{R})\psi(\mathbf{R},\mathbf{r}) \tag{2.14}$$

where $\psi(\mathbf{R},\mathbf{r})$ is the wave function, $\epsilon(\mathbf{R})$ the electronic energy for the nuclear coordinates $\mathbf{R}$, and $\hat{h}_e$ is the electronic hamiltonian, which now adopts the simplified form:

$$\hat{h}_e = \hat{T}_e+\hat{V}_{ne}+\hat{V}_{ee} = -\frac{\hbar}{2m}\sum_{i=1}^{N}\nabla_i^2-e^2\sum_{A=1}^{M}\sum_{i=1}^{N}\frac{Z_A}{|\mathbf{R}_A-\mathbf{r}_i|}+\frac{e^2}{2}\sum_{i=1}^{N}\sum_{j>1}^{N}\frac{1}{|\mathbf{r}_i-\mathbf{r}_j|} \tag{2.15}$$

where $\mathbf{r}_i$ are the coordinates of the electron i, $\mathbf{R}_A$ are the coordinates of the atom A, $\mathbf{Z}_A$ is the mass of the atom A.

Even though the Born-Oppenheimer approximation simplifies the problem, the equation still can not be solved as it is, due to the influence of electron and exchange correlation on the wave function. Electron correlation is the tendency of the electrons to repel each other (in agreement with the Coulomb's law), so that when two electrons are close, the probability density is zero. The exchange correlation requires two electrons belonging to the same orbital to have opposite spin (in agreement with Pauli's exclusion principle) so that their wave functions are antisymmetric with respect to the exchange. Many methods have been developed to calculate an approximate solution to the electronic Schrödinger equation, and these can be divided in two main groups: the wave function based methods (Hartree-Fock and post Hartree-Fock), and those based on Density Functional Theory, which are going to be discussed in the next sections.

2.3.1. Wave function methods

Wave function based methods rely on the variational principle, which states that the energy of a trial wave function is always greater than the exact ground state solution of the wave function. This allows an approximate wave function to be improved by minimising its energy with respect to a set of parameters. The most simple method is the Hartree method, which states that the overall wave function is the product of

individual one-electron wave functions:

$$\Psi(\mathbf{r}) = \Pi_{i=1}^{N}\psi(\mathbf{r}_i) \tag{2.16}$$

This description is then improved by the Hartree-Fock approximation, which replaces the one-electron wave functions, $\psi(\boldsymbol{r}_i)$, with a Slater determinant. This results in a wave function, Ψ_{HF}, that is antisymmetric and hence in agreement with Pauli's principle. This anti-symmetrisation is responsible for the exchange effect seen in Hartree-Fock theory. The Slater determinant has the form:

$$\Psi_{HF} = \frac{1}{\sqrt{N!}}\begin{vmatrix} \chi_1(1) & \chi_2(1) & ... & \chi_N(1) \\ \chi_1(2) & \chi_2(2) & ... & \chi_N(2) \\ ... & ... & ... & ... \\ \chi_1(N) & \chi_2(N) & ... & \chi_N(N) \end{vmatrix} \tag{2.17}$$

where N is the total number of electrons and each $\chi_i(j)$ term is a one-electron spin-orbital (the product of a spatial orbital and an electron spin eigenfunction), where j represents electron j. The energy associated with a guess Slater determinant is then minimised with respect to its spin orbitals, which, in accordance with the variational principle, improves the estimate of the energy of the system. Inserting the expression for the Slater determinant into the electronic Schrödinger equation leads to the Hartree-Fock equation (HF) for each spin-orbital (shown for an example electron 1):

$$\hat{F}\chi_i(1) = \epsilon_i\chi_1(1) \tag{2.18}$$

where $\hat{F}$, the Fock operator, containing in turn the Coulomb operator $\hat{J}_j$ and the exchange operator $\hat{K}_j$, is:

$$\hat{F} = \hat{h}_1 + \sum_{j=1}^{N}(\hat{J}_j - \hat{K}_j) \tag{2.19}$$

In order to calculate $\hat{F}$, all the spin-orbitals of all the electrons need to be known. The Fock operator can then be used to calculate the HF energy:

$$E_{HF} = \sum_{i=1}^{N} E_{ii} + \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}(J_{ij} - K_{ij}) \tag{2.20}$$

The equation can be solved iteratively in a few steps:

1. An initial set of spin-orbital is generated from the initial guess.

2. The Fock operator is calculated.

3. The HF equation is solved to obtain a new set of spin-orbitals.

4. Steps 1-3 are repeated until the spin-orbitals in steps 1 and 3 are considered self-consistent, according to the self consistency field method.

This method accounts for 99% of the total energy of the system, but excludes the correlation energy (which is the difference between the HF energy and the energy one would obtain from the exact solution of the Schrödinger equation), which is quite relevant for the description of chemical processes.

Post-Hartree-Fock (post-HF) methods expand the HF wave function into a multi-determinant representation and use the latter to calculate the correlation energy. Among these, the configuration interaction (CI) methods use the multi-determinant to describe the energy of both occupied and unoccupied orbitals, although only some unoccupied states of chemical relevance can be included for calculations to remain computationally feasible for any but the smallest chemical system. CI methods are labelled according to the kinds of excitation described in the expansion, e.g. CISD includes all the single (S) and double (D) excitations, whereas in CISDT the wave function is truncated after triple (T) excitations. Other post-HF methods are based on perturbation theory, where a perturbation *H'* with weight Λ is added to the HF hamiltonian ($\hat{H}_0$):

$$\hat{H}_{perturbed} = \hat{H}_0 + \Lambda\hat{H}' \tag{2.21}$$

In the Møller-Plesset methods, for example, ($\hat{H}_0$) is a sum of Fock operators and correlation is added from the doubly excited determinants (as the first order would just be the HF energy). The more excitations that are considered, the more accurate is the method, although a good accuracy/computational cost balance always needs to be considered.

2.3.2. Density Functional Theory

Density Functional Theory (DFT) describes the properties of many-electron systems with functionals (functions of functions) of the electronic density, $\rho(\mathbf{r})$. DFT methods are the most versatile and have been applied in both condensed-matter and computational physics, and chemistry. They have been developed following the initial theorems proposed by Hohenberg and Kohn in 1964,[172] and then computationally implemented by Kohn and Sham in 1965.[173]

The first theorem by Hohenberg and Kohn, called the existence theorem, proves that the ground state density of a system determines the external potentials V_{ext} (which in a molecule corresponds to the attraction force to the nuclei) and that the energy of a system could be written as:

$$E[\rho(\mathbf{r})] = F[\rho(\mathbf{r})] + \int \hat{V}_{ext}(\mathbf{r})\rho(\mathbf{r})d\mathbf{r} \tag{2.22}$$

where $F[\rho(\mathbf{r})]$ is a functional containing the kinetic energy $\hat{T}[\rho(\mathbf{r})]$, the classical Coulomb contribution $J[\rho(\mathbf{r})]$, and the non-classical exchange-correlation functional $E_{XC}[\rho(\mathbf{r})]$. Among these, the Coulomb component represents a significant contribution to the electron-electron interaction and it is the only one that can be calculated exactly, whereas the other two terms must be described with approximations. The Coulomb interaction has the form:

$$J[\rho(\mathbf{r})] = \frac{1}{2}\int\int \frac{\rho(\mathbf{r}_1)\rho(\mathbf{r}_2)}{|\mathbf{r}_2 - \mathbf{r}_1|}d\mathbf{r}_1\mathbf{r}_2 \tag{2.23}$$

The formalism proposed by Kohn and Sham in 1965 relies on a fictitious system of N non-interacting electrons, which can be expressed as one-electron Kohn-Sham orbitals:

$$\rho(\mathbf{r}) = \sum_{i}^{N} |\phi(\mathbf{r})|^2 \tag{2.24}$$

where N is the total number of occupied orbitals for a closed shell system. Such a system of non-interacting electrons allows the kinetic energy to be expressed as a function of the density, adopting the form:

$$\hat{T}_{KS} = -\frac{\hbar}{m}\sum_{i=1}^{N} < \phi_i|\nabla^2|\phi_i > \tag{2.25}$$

I will now cover the remaining steps necessary to describe the final form of the Kohn-Sham functional. The one-electron Schrödinger equation can be solved to obtain the Kohn-Sham orbitals:

$$\Big(-\frac{\hbar}{2m}\nabla^2 + V_{KS}(\mathbf{r})\Big)\phi_i(\mathbf{r}) = \epsilon_i\phi_i(\mathbf{r}) \tag{2.26}$$

where $V_{KS}(\mathbf{r})$ is the external potential of the fictitious system of non-interacting electrons. In turn, $V_{KS}(\mathbf{r})$ can be derived by minimising the Kohn-Sham functional

with respect to the electronic density (which is possible according to the second Hohenberg-Kohn theorem which states that the energy of the ground state can be found variationally):

$$V_{KS} = V_{ext}(\mathbf{r}) + \int \frac{\rho(\mathbf{r}_2)}{|\mathbf{r}_1 - \mathbf{r}_2|} d\mathbf{r}_2 + V_{XC}(\mathbf{r}) \tag{2.27}$$

where the last term V_{XC} is the exchange-correlation potential obtained as a derivative of the exchange-correlation energy with respect to the electronic density.

The final form of the Kohn-Sham functional then becomes:

$$E_{KS}[\rho(\mathbf{r})] = \hat{T}_{KS} + \int \rho(\mathbf{r}) V_{ext}(\mathbf{r}) d\mathbf{r} + J[\rho(\mathbf{r})] + E_{XC}[\rho(\mathbf{r})] \tag{2.28}$$

The Kohn-Sham equation needs to be resolved iteratively, given the dependency of V_{KS} on the electronic density. The iterative cycle can be divided into a few steps:

1. An initial guess of the Kohn-Sham orbitals is used to calculate $\rho(\mathbf{r})$.
2. V_{KS} is calculated.
3. The Kohn-Sham equation is solved to calculate a new $\rho(\mathbf{r})$.
4. Steps 2-3 are repeated until V_{KS} and $\rho(\mathbf{r})$ are self-consistent.
5. The Kohn-Sham energy is calculated.

Approximations for the exchange and correlation functionals.

The exact ground state energy could be known if the exchange-correlation contributions to the energy could be calculated, however approximations are necessary as the formulation of this term is unknown. The three most used approximations are the Local Density Approximation (LDA) the Generalised Gradient Approximation (GGA) and the meta-GGA approximation, which are discussed below.

- LDA
  The value of E_{XC} depends uniquely on the local value of the density and not on its derivatives. E_{XC} can be divided into two contributions, one for the exchange and one for the correlation. The first can be calculated exactly from the homogeneous electron gas, whereas the second one has been derived from quantum Monte Carlo simulations. LDA performs better on metallic systems, where the electron density varies slowly, and it works reasonably well

at both the molecular and solid state level. Its main drawbacks are caused by the underestimation of binding energies and bond lengths. Generally, the exchange energies are only slightly underestimated, whereas the correlation energies are largely overestimated; good performance are obtained thanks to the fact that the first parameter is about an order of magnitude larger than the second.

- GGA
  According to the GGA approximation, E_{XC} depends on both the density and its derivative, providing, compared with LDA, a better description of both binding energies and geometries. Many functionals developed in the past twenty years are based on the GGA approximation, among which is the PBE functional, used for most calculations in this thesis. PBE was developed by Perdew, Burke and Ernzerhof in 1996,[174] and it is a parameter-free GGA, which means that the gradient correction to the LDA functional was calculated without introducing experimentally fitted parameters, which makes it fairly accurate for a wide range of systems.

- Meta-GGA
  In meta-GGA functionals, the functional depends on the kinetic energy density as well, which leads to a better performance with respect to normal GGAs. The M06-2X Minnesota functional[175] is based on the meta-GGA approximation and it was used to perform the calculations described in Chapter 5. M06-2X is an hybrid functional, where the correlation contribution is calculated with DFT, whereas the exchange contribution is given by a combination of DFT and HF (in this case 46% DFT and 54% HF) and it is known to perform particularly well for the calculation of non-covalent interactions of organic molecules (very large systems, such as proteins, excluded).[175]

Dispersion corrections

DFT methods cannot accurately describe the long-range London dispersion interactions occurring between atoms/ molecules not directly bonded to each other. They are, as already discussed in Section 2.2.1, temporary attractive forces that occur when the electrons in two adjacent atoms occupy positions that make the atoms form temporary dipoles; for this reason they can be also called induced dipole - induced dipole forces. They are known to be of both physical and chemical importance for many molecular systems, among which are organic cages.

Many kinds of dispersion corrections have been developed in the past 20 years and can be divided into four groups:

1. Van der Waals density functionals
   Van der Waals density functionals (vdW-DFs) compute the dispersion energy of a system in a non-empirical way, relying on their electron density only. According to this approach, E_{XC} is calculated as follows:

$$E_{XC}^{vdW} = E_x^{LDA/GGA} + E_c^{LDA/GGA} + E_c^{NL} \tag{2.29}$$

   where $E_x^{LDA/GGA}$ and $E_c^{LDA/GGA}$, the standard exchange and correlation components of E_{XC} of the LDA or GGA approximations, are used to describe the short-range interactions, whereas E_c^{NL} is a non-local term describing the dispersion energy. It is an add-on to the density functional energy and it is computed non-self-consistently in the form:

$$E_c^{NL} = \frac{1}{2} \int \int \rho(\mathbf{r}_1)\phi(\mathbf{r}_1, \mathbf{r}_2)\rho(\mathbf{r}_2)d\mathbf{r}_1 d\mathbf{r}_2 \tag{2.30}$$

   Correction methods differ in the choice of the non-local correlation kernel $\phi(\mathbf{r}_1, \mathbf{r}_2)$ they use. In vdW-DFs methods dispersion effects are included via the charge density.

2. Parametrised density functionals
   Part of this group are those DFT functionals that, although not being dispersion corrected, can well describe dispersion forces of equilibrium geometries of molecules that are not too large (e.g. they are not recommended for solids or biomolecules that require long-range more accurate descriptions). Among these, the M06-2X functional is one of the most accurate,[175] together with other meta-GGA highly parametrised functionals that include kinetic energy density descriptions.

3. Semiclassical corrections (DFT-D)
   DFT-D methods are based on an atom pairwise additive treatment of the dispersion energy, which is just added to the Kohn-Sham DFT energy in the form:

$$E_{disp}^{DFT-D} = \sum_{AB} \sum_{n=6,8,..} s_n \frac{C_n^{AB}}{R_{AB}^n} f_{damp}(R_{AB}) \tag{2.31}$$

where R_{AB} is the internuclear distance between atoms A and B and s_n is a scaling factor to adjust repulsion and C_n^{AB} is the Casimir-Polder dispersion coefficient. C_n^{AB} describes induced dipole-induced dipole interactions for n=6, dipole-quadrupole interactions for n=8 and higher multipole interactions for n=10. Damping functions $f_{damp}(R_{AB})$ are included to determine the range of the dispersion corrections. C_6 coefficients are known exactly and provide a very accurate description of most long-range weak interactions present in molecular systems.

In this work, the DFT-D3 corrections from Grimme and co-workers[176] is combined with the PBE functional in CP2K.[161] DFT-D3 includes atom pairwise-specific dispersion coefficients computed from first-principle methods and well describes dispersion for all the atoms of the first two rows of the periodic table. It works with any kind of DFT functional and it has been tested for both molecules and solids.[176]

4. One-electron corrections
   Compared to the other dispersion corrections, one-electron corrections have a more empirical character, as each atom type need a specific parametrisation. These corrections rely on potentials rather than non local terms, which produce changes on the electronic charge density.

For a more detailed description of the dispersion corrections, the reader is referred to the review by Grimme reported in reference 177.

2.3.3. Basis sets

Basis sets are mathematical functions developed to describe the wave functions, so that either the HF or DFT problems can be solved within an algorithm. The mathematical functions can be used to describe the radial and angular distribution of the electronic density. According to the basis set approximation, orbitals can be described as a linear combination of basis functions:

$$\Phi_i = \sum_{\mu=1}^{M} c_{i\mu}\phi_\mu(\mathbf{r}) \tag{2.32}$$

where $\phi_\mu(\mathbf{r})$ is the basis function, $c_{i\mu}$ is the weight coefficient of the orbital μ and M is the size of the basis set. The basis set, in theory, should be infinite, however this is not applicable in practice. When included in the algorithm, the basis set is

truncated so as to include enough basis functions to provide an accurate result. The first kind of basis sets to be developed were called Slater Type Orbitals (STOs), they described the angular distribution with a spherical harmonic function and the radial distribution with an exponential function. Even though they describe fairly well the decay of the electronic radial distribution (with a cusp close to the nuclei and slow decay moving away from the nuclei), they cannot be integrated analytically and are thus difficult to handle.

The Gaussian Type Orbitals (GTOs) described in the next subsection substitute the exponential function with a gaussian function, which on one side partially worsen the description of the electronic radial distribution decay (too flat close to the nuclei, too steep moving away from the nuclei), but on the other makes the basis sets easier to handle as they can be computed analytically.

Gaussian basis sets

GTOs are very commonly used and rely on an interpretation of the overall wave function as a molecular orbital (MO) described by a linear combination of atomic orbitals (AO). The general form of a GTO in atom-centred cartesian coordinates is:

$$\Phi(x, y, z; \zeta, l_x, l_y, l_z) = N x^{l_x} y^{l_y} z^{l_z} e^{-\zeta r^2} \tag{2.33}$$

where l_x, l_y, l_z are the angular momenta, the sum of which describes the orbital type (e.g. if the sum =0 orbitals are spherical, if =1 p-type and so on), whereas $e^{-\zeta r^2}$ is the gaussian function describing the decay of the electronic radial distribution. GTO functions are combined to make the description of the electronic radial decay better, and more similar to that of a STO. The number of gaussian functions included in the basis set determines how accurately the decay is reproduced, but one should also bear in mind that large numbers of functions will make the calculation expensive.

Minimal basis sets (the smallest possible) only include one function per atomic orbital, which means that the function does not have the flexibility to describe diverse molecular environments, as it cannot expand or contract to better reproduce the radial electronic distribution. More than one basis function is normally required to describe each orbital. Double zeta (DZ) basis sets contain two basis functions per orbital, which is described by the linear combination of one contract gaussian function (narrower), and one diffuse gaussian function (wider), allowing one, for example, to describe the anisotropy of p-type orbitals. Triple zeta (TZ) basis sets include three gaussians per orbital, making the description of the contract/diffuse character of the

orbital even better, although, of course, they are more computationally expensive than the two other kinds. Polarisation basis functions, which have higher angular momentum, can be added to a basis set to further improve the description of the anisotropic electron distribution. This is done by adding p-type orbitals to H/He, d-type orbitals to first-row atoms and so on.

In this work the Triple Zeta Valence Polarization TZVP MOLOPT basis set[177] is used in combination to the PBE DFT functional; V indicates that only the valence electron are described by triple zeta functions, whereas P means that polarisation functions are embedded in the basis set. The Pople basis set 6-311G* is used in combination with the M06-2X functional;[175] 6 is number of GTOs used to describe the core electrons, whereas the valence orbitals are split in three basis functions (making this too a triple zeta basis set) obtained, respectively, by the combination of 3, 1 and 1 GTOs. The star indicates that polarisation functions are added on the first-row elements.

Plane-wave basis sets

Plane-wave basis sets are non-atom-centred basis sets, which are generally adopted for the description of the solid state, as they implicitly consider the periodicity of a system. They are based on Bloch's theorem, which states that the wave function of an electron $\psi_{\mathbf{k}}(\mathbf{r})$ in a periodic potential must be periodic and have periodicity consistent with that of the lattice, and can be written as:

$$\psi_{\mathbf{i},\mathbf{k}}(\mathbf{r}) = e^{i\mathbf{k}\cdot\mathbf{r}} u_{\mathbf{k}}(\mathbf{r}) \tag{2.34}$$

where $\mathbf{k}$ is a wave vector of the reciprocal space, and $u_{\mathbf{k}}(\mathbf{r})$ is a function with the same periodicity of the unity cell, $u_{\mathbf{k}}\ (\mathbf{r}) = u_{\mathbf{k}}\ (\mathbf{r} + \mathbf{a}_i)$ ($\mathbf{a}_i$ being a lattice vector). The final form of the wave function is a linear combination of plane waves:

$$\Psi_{j,\mathbf{k}}(\mathbf{r}) = \sum_{\mathbf{G}} c_{j,\mathbf{k}} + \mathbf{G}\ e^{i(\mathbf{k}+\mathbf{G})\cdot\mathbf{r}} \tag{2.35}$$

where $\mathbf{G}$ are the vectors for the reciprocal lattice. Ideally, an infinite number of plane waves should be considered to solve the equation. However, within this approach, the accuracy increases with the number of plane waves involved, so that when two sampling points are close enough, the wave function can be considered the same. In this way, the summation can be truncated accordingly to a threshold value and only the plane waves with kinetic energy lower than the defined cutoff need to be

included:

$$E_{cut} \geq \frac{1}{2}|\mathbf{k} + \mathbf{G}|^2 \tag{2.36}$$

A drawback of the procedure is that high energy plane waves need to be considered for the description of the core electrons (which increases the number of **G** vectors and the cost of the calculations), therefore pseudopotentials (discussed in the next section) can be adopted to reduce the computational cost. Plane wave basis sets were adopted in this work within the CP2K package, although periodic interactions were turned off as my calculations only investigated single molecules.

Pseudopotentials

Generally only valence electrons are of chemical relevance, as they are involved in chemical bonding, whereas core electrons are chemically unimportant, but nonetheless require a large number of basis functions for an accurate description of their orbitals. Pseudopotentials (or effective core potentials, ECPs) are linear combinations of specially designed gaussian functions that model the core electrons. The latter are represented as a whole by an effective potential, which means that no basis functions for the core are needed, and that the computational cost of the calculation can be further reduced. Goedecker - Teter - Hutter (GTH)[178] pseudopotentials are used in this work in combination with the PBE/TZVP setting in the CP2K package.

2.4. General pipeline adopted in this thesis

A general procedure (shown in Figure 2.2) was devised to computationally investigate all the cages part of this PhD study and was then modified to analyse the properties of relevance of each individual project. Details of the computational settings adopted in each case will be discussed in the method sections of each result chapter, however the general work-flow procedure can be described as below:

1. Hypothetical geometries are assembled and minimised.

2. Molecules that retain a cavity after step 1 are investigated with conformer search methods to locate their lowest energy conformation.

3. Molecules that retain a cavity after step 2 are geometry optimised with DFT methods, to obtain more accurate results with respect to both energy and geometry.

4. Molecules that do not retain a cavity after step 2 are fed into a constrained MD procedure to locate higher energy open conformers. As this is the most time-consuming step of the procedure, it is normally applied only if the open conformers are interesting for the sake of the project. If higher energy open conformers can be successfully located, those are also geometry optimised with DFT methods.

5. The set of open molecules optimised with DFT methods is further analysed, and the energy of molecules assembled with different ratios using the same set of precursors are compared, to predict which is the most thermodynamically stable. The diameters of the cavities are calculated for promising candidates for the synthesis.

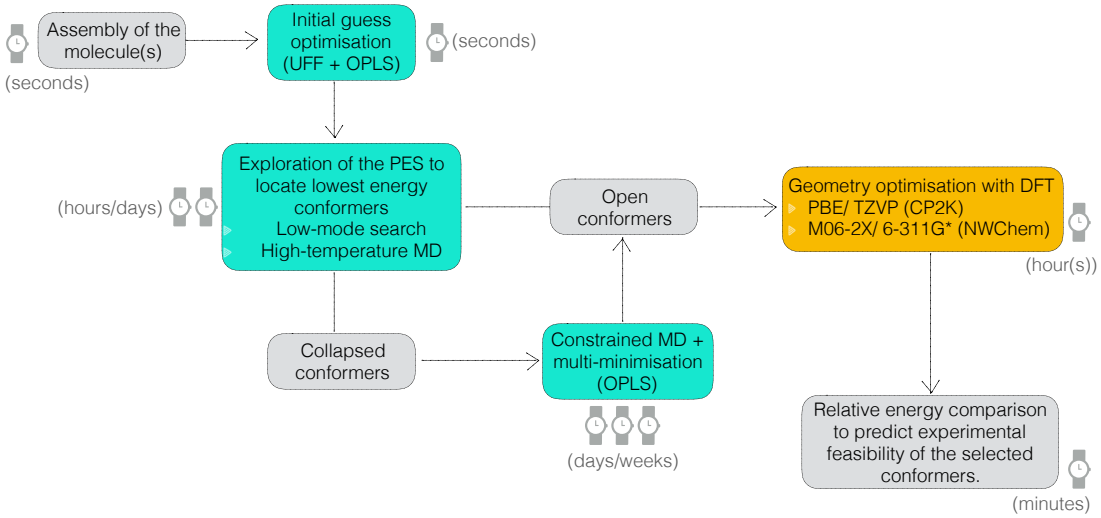


Figure 2.2.: General work-flow followed for the computational investigation of organic cages. Cyan boxes are for steps based on force field methods, orange on electronic structure theory. The watch-icons refer to the relative time required on average to complete each step.

2.4.1. Comparing the energy

The energies of the molecules obtained from step 4 of the previous section's pipeline can be compared in two different ways, as explained in the next two subsections.

Comparison of relative energies

If the molecules investigated are formed according to multiples of stoichiometric $[m+n]$ ratios, their relative energies can be compared. When calculating the relative

energies, the energy of one molecule is arbitrarily taken as a reference and all the others are calculated with respect to that. Each molecular size is then divided by its formula unit so that values are comparable. Thus, if we take as an example four cages generated with increasing molecular sizes, [2+3], [4+6], [6+9], [8+12], the energies of the latter three cages are divided respectively by 2, 3 and 4, so that their energies are comparable.

Comparison of relative formation energies

In some projects, cage molecules that do not contain multiples of a formula unit are compared. In those cases, their relative formation energies are instead compared. All the cages generated for this work are based on the DCC imine formation reaction, where an aldehyde and an amine react to form an imine and a molecule of water. To calculate the formation energy per bond formed, the following equation is used:

$$E_{formation} = \frac{(E_{cage} + xE_{water}) - (mE_{aldehyde} + nE_{amine})}{x} \tag{2.37}$$

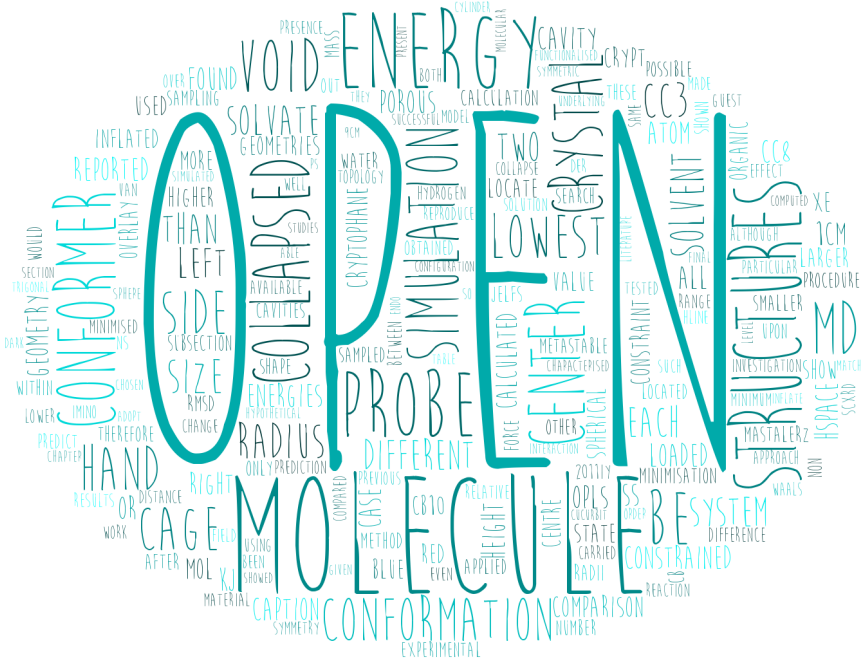
where E_{cage} is the energy of the cage formed, E_{water} is the energy of the water produced in the reaction, $E_{aldehyde}$ is the energy of the aldehyde and E_{amine} is the energy of the amine. The number of aldehyde reactants is m, n is the number of amine precursors and x is the number of imine bonds formed, equivalent to the number of water molecules produced in the reaction.

2.4.2. Void size calculation

When the void diameters are calculated they are considered as the largest sphere that can be built from the centre of mass of the cage, without any of the atoms penetrating it. The void radius of the sphere is thus considered the distance between the centre of mass and the closest atom to it, taking into account the van der Waals radius of that atom.

Chapter 3
Predicting solvent effects

The work in this chapter led to the publication of:
Santolini V., Tribello G.A. and Jelfs K.E., *Chem. Commun*, 2015, 51, 15542-15545.



3.1. Introduction

In this chapter I am going to cover the topic of predicting the effect of solvent on porous organic cages, with the goal of reproducing the geometry that collapsed desolvated systems may adopt when solvent plays a scaffolding effect on the structure. All the porous molecules that I am going to describe were synthesised via dynamic

covalent chemistry reactions experimentally carried out in solution. Solvent plays a crucial role both in the formation of the final product geometries and in the properties associated with those geometries. The influence of solvent on the reaction outcome is still computationally very hard to determine and predict and for these reasons, when running simulations, it is often either ignored or treated as a continuum. When trying to synthesise new porous organic cages and in particular cages with larger cavities, a common obstacle is the potential collapse of the structures. In most cases this occurs upon solvent removal, resulting in a loss of porosity.[96] The left hand side of Figure 3.1 shows an example of a collapsed molecule, **CC8**, which can be synthesised via an imino chemistry condensation of tris(4-formylphenyl)amine and (R,R)-1,2-cyclohexanediamine in a [8+12] ratio.[96] The "open" conformation of **CC8** (where the word *open* is used in the text to describe molecules characterised by the presence of a cavity) is shown on the right hand side of Figure 3.1 and it exhibits a void radius of 4.6 Å.

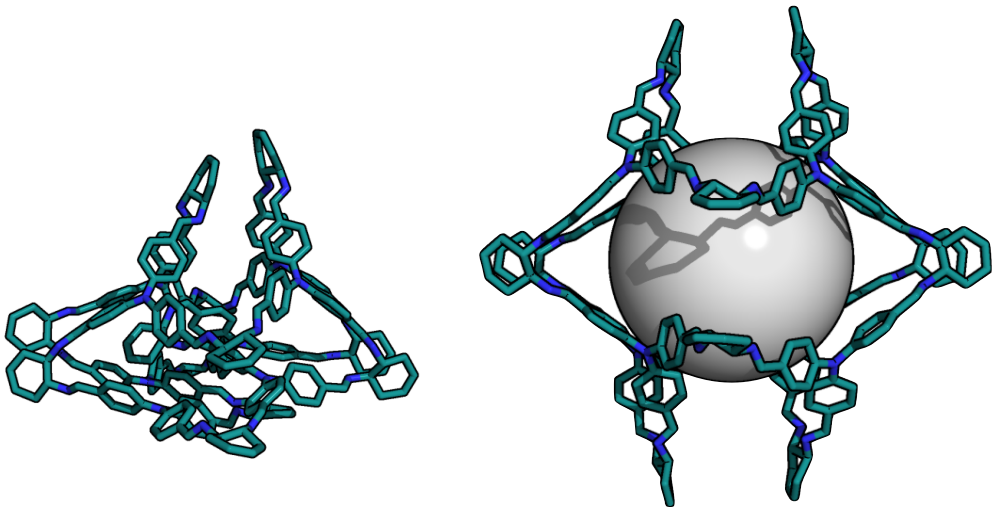


Figure 3.1.: Comparison of the collapsed (left hand side) and open (right hand side) geometries of **CC8**. Throughout the chapter, grey spheres are to highlight the presence of a cavity. The open geometry has a void radius of 4.6 Å.

Collapsed materials may be erroneously considered of little interest, however, even if they are not porous upon desolvation, they can still find different applications in solution, in the fields of porous liquid materials,[125] encapsulation, separation[145] or optical sensing.[179] For molecules that are porous in solution, the prediction of the open geometry could be vital to rationally design synthetic modifications that could prevent the loss of the cavity.[39] Simulations have been used so far to predict reaction

outcomes in the gas-phase,[83] the resulting molecules corresponding to experimental desolvated structures. Even if it is known that solvent can affect molecular mass[99] and can induce polymorphism phenomena,[118] so far solvent effects have not been successfully included or accurately described in computational studies of porous organic cages. However, simulations in the presence of solvent have been successful for the prediction of solvent stoichiometry and geometry of pharmaceutical compounds, where the investigated molecules were encapsulating a maximum of 1 or 2 solvent molecules each.[180] Describing the solvent explicitly dramatically increases the cost of the simulation and it would make calculations on porous organic cages incredibly expensive, bearing in mind that some of the systems described in this chapter can host up to 70 solvent molecules per host molecule.[96] To reproduce explicit solvent interactions, one would firstly have to greatly increase the number of atoms in the simulations by the necessity of including hundreds of solvent molecules.[96] Secondly, one would then need to sample the conformations of the molecules surrounding the cage to consider the dynamic nature of solvation. For these reasons, an accurate simulation of porous organic cages in the presence of explicit solvent is not currently tractable. Other methods have been applied to different porous materials, such as zeolites and MOFs. In the former case *de novo* design of templates to obtain certain topologies showed some degree of success,[181,182] although the same method was not working equally well for other systems such as MOFs or porous molecular materials. MOFs in particular show phase-changes from "open" to "closed" configurations in response to solvent loading and/or other experimental factors such as temperature or pressure. An example reported in the literature is MIL-53.[183] The external factors required for the transition to happen can be predicted through a thermodynamic framework if both the "open" and "closed" structure (before and after the phase-change) are known.[184] If force fields are available it is possible to predict structural collapse via atomistic simulations. Nonetheless, the simulation is only possible because advanced studies on MOFs' stability problems are largely available in the literature.[185] More challenging is to predict the structure of the guest-loaded "open" phase.[186] An alternative implicit solvation approach was tested, where a dielectric constant was applied to reproduce the dielectric screening a molecule in solution experiences; however it did not reproduce the scaffolding effect of the solvent and preliminary tests have found this to make no difference to the relative energies of cage topologies.

In this chapter I will show how with a new procedure involving constrained MD it is possible to predict the "open" geometries of many collapsed porous organic

molecules, by locating higher energy open metastable states that correspond to reported or hypothetical solvate structures. Higher energy open conformations typically lie more than 100 kJ mol^{-1} above the lower energy collapsed structures, and they cannot be adequately sampled or at all located by normal conformer search algorithms. Even by performing very extensive searches, one cannot be sure an open conformer will be within the selected range. The technique used here allows one to locate and select higher energy conformers that retain a cavity, within a time frame and computational resources more modest than the ones required by conformer search methods or traditional MD.

3.2. Outline of the work

Constrained MD was tested on porous organic molecules characterised by the presence of a cavity when in a solvate state. The open conformation is a metastable energy state, which lies higher in energy with respect to the lowest energy collapsed conformer. The description does not take into account crystal packing forces. The reaction schemes of the systems included in this study are shown in Figure 3.2. The objects of the investigation are two reported imine cages, two hypothetical imine cages, and three molecules belonging to two other classes of porous organic molecules, namely cucurbit[n]urils[187] and cryptophanes.[188] The two imino-cages reported in the literature and of which solvate crystal structures are available for comparison are **CC8**,[96] with the underlying topology of a cube and synthesised by the Cooper Group, and Mastalerz's endo-functionalised tetrahedral cage.[77] The two hypothetical imino cages consist of the [6+9] and [8+12] versions of the reported [4+6] **CC3** molecule,[7] forming respectively a tubular shaped molecule with the underlying topology of a trigonal prism and a molecule with the underlying topology of a cube. Regarding cucurbit[n]urils (glycoluril-based macrocycles), investigation was focussed on CB[6][189] and bis-nor-sec-CB10 (CB10-ns), the latter of which lacks two methylene bridges compared to the better known CB10, resulting in a molecule characterised by two interconnected cavities rather than one developed around the centre of mass.[190] As was the case for the imine cages, solvate crystal structures were available for comparison. Finally Cryptophane A (made up by two interconnected cyclotriguaiacylene - CGT - cups) was studied and compared with Xe and water-loaded SCXRD structures that were reported in the literature.[191] Cryptophane A is the only case for which a collapsed conformation has been experimentally reported. The lowest energy structures of all molecules are collapsed geometries (correspond-

ing to collapsed desolvated structures), although both Mastalerz's cage and CB[6] preserve a small cavity.

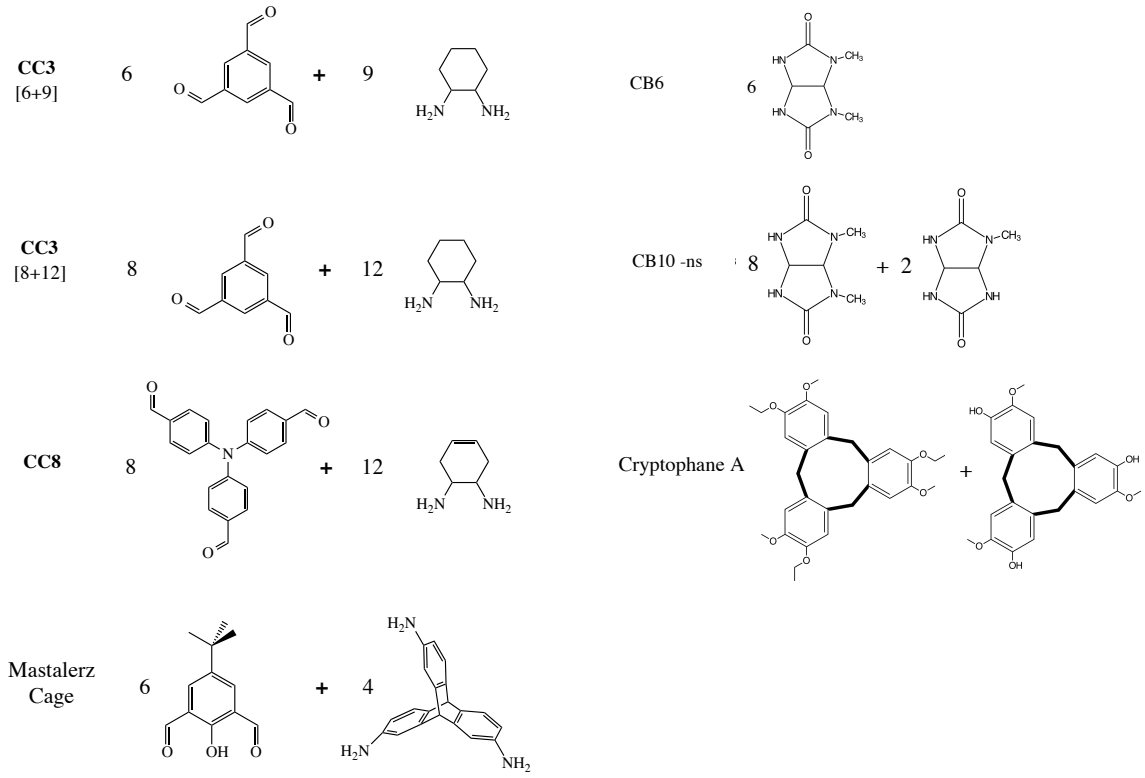


Figure 3.2.: Reaction scheme of the molecules inflated with constrained MD.

3.3. Methods

In the next few paragraphs I will outline the methods that were used to locate the lowest energy conformers of the cages, the arguments suggesting constrained MD was the best way to locate "open" conformations, and finally the steps involved in the procedure.

3.3.1. Conformer search calculations to locate collapsed structures.

Conformer search calculations were applied to a subset of systems to investigate the potential energy surfaces (PES) surrounding the initial guess and to locate the lowest energy conformers. As expected, all the low energy conformations within 50

kJ mol^{-1} of the global minimum were found to be collapsed, lacking an internal void. As discussed in Chapter 2, the low-mode conformer search[169] within the Macromodel package (version 9.6, Schrödinger, LLC, New York, NY, 2013) was chosen for imino-cages, as it had already been largely tested on such systems.[98] However, the size of the systems involved in this work (from 66 to 272 non-hydrogen atoms), and the number of cycles included in each molecule, would have made the calculations excessively slow, therefore an alternative faster approach involving MD calculations was tested. On each molecule a 1 ns MD simulation was carried out with the OPLS_2005 force field,[166] with sampling every 1 ps, at the temperature of 500 K. Each sampled structure was geometry optimised with the same force field within Macromodel, applying a conjugate gradient minimisation algorithm with a convergence criterion of a remaining root mean square (RMS) force of 0.05 kJ mol^{-1} Å^{-1}. A full conformer search for large molecules would require a much larger number of steps (∼50 ns, sampling every 0.7 ps). However these calculations were only performed to check the collapsed geometries of the lower energy states, and to estimate the energy difference with the higher energy open structures. More accurate simulations would reveal a higher energy gap as even lower energy non-shape persistent structures could be found.

3.3.2. Initial attempts to explore open molecular conformations

Given that it was not possible to locate higher energy open states with standard conformer search methods (that normally sample conformations within a range of 50 kJ mol^{-1}), some alternative approaches were tested. Firstly MD simulations were run on a single molecule with explicit solvent molecules (dichloromethane was chosen, as it is commonly used in the experimental reactions). The simulation was not successful, as it was not possible to locate any open conformations over a multi-nanosecond timescale. In a further attempt, an MD simulation was started from an open solvent-loaded conformation (which was known to be a local minimum on the PES),[96] but nonetheless the molecule collapsed. Subsequently, another approach based on symmetry criteria was tried, which was inspired by the observation that the geometries of the open states are generally highly symmetric. Steered MD[192] was used to find configurations in which the faces of the cages were ordered around the molecule's centre of mass.[193] Simulations produced molecules with ordered faces, but that unfortunately did not retain a void. In each case the molecule adopted a

collapsed configuration in which the void region was occupied by some of the atoms of the structure.

3.3.3. Constrained MD simulations to locate metastable open molecular conformations.

When none of the above methods proved fruitful, I moved towards a simpler approach that involved applying a rigid spherical probe around the centre of mass of the molecule, x_c, acting as an energetic penalty for any atom trying to collapse within the whole volume of the sphere. In order to apply the penalty, the minimum distance, s, was calculated between each atomic position, x_i, and the centre of mass, x_c, with the equation below:

$$s = \frac{\beta}{\log\left[\sum_{i=1}^{n} \exp\left(\frac{\beta}{\mathbf{x}_i - \mathbf{x}_c}\right)\right]} \tag{3.1}$$

where β ensures the function's derivatives with respect to the atomic positions are continuous. Then a harmonic restraint was defined:

$$V(s) = \begin{cases} \frac{1}{2}k(s - s_0)^2 & s \leq s_0 \\ 0 & s > s_0 \end{cases} \tag{3.2}$$

where k is a force constant included in the Hamiltonian for the system, in order for s (minimum distance) to take on particular values. This equated to the insertion of a spherical probe of radius s_0 into the cavity of the molecule and, as a consequence, during the MD simulation no atom entered the volume of the sphere. Combining this strategy with Simulated Annealing (SA) allowed the system to improve the sampling of configurational space. The lowest energy open conformers were successfully located by repeating the simulations with multiple $\mathbf{s}_0$ values, generally in the range 3 - 8 Å. This approach allowed the geometry the molecule could adopt in the presence of solvent to be predicted, without needing experimental input.

An example input file for constrained MD calculations involving a spherical probe is in Appendix A.1.1. The modelling can start from a collapsed conformation optimised to a minimum, although the starting point is not important if the SA procedure is followed. For these calculations the OPLS_2005 force field was used. All the MD simulations were carried out at 300 K with DL_POLY2.20 combined with a Velocity Verlet algorithm, a time step of 0.7 fs and a cutoff of 12 Å. The constraints

were applied with the PLUMED2 metadynamics plug-in. The force constant was scaled from 0 to 50000 kJ mol^{-1} nm^{-2} during the first 14 ps of a 10 ns MD run. Then the calculations were carried out with a range of different probe sizes (constraints) and sampled every 0.35 ps. The 30,000 sampled structures were all subsequently geometry optimised using MacroModel, and only those retaining a spherical void of the desired size (or larger) were kept. The minimisation step is fundamental, not only to confirm the presence of a local minimum, but also to see if a structure that is open during the sampling stays open after minimisation. The structures as sampled are thousands of kJ mol^{-1} higher than those obtained after minimisation and therefore could be dramatically changed by the process of relaxation. This is shown by the plot in Figure 3.3, which reports the energies of structures as sampled (red) and after geometry optimisation (blue). The average decrease in energy upon geometry optimisation is 3066 kJ mol^{-1}. The open metastable conformers located after minimisation are on average 100 kJ mol^{-1} higher than the lowest energy collapsed conformers.

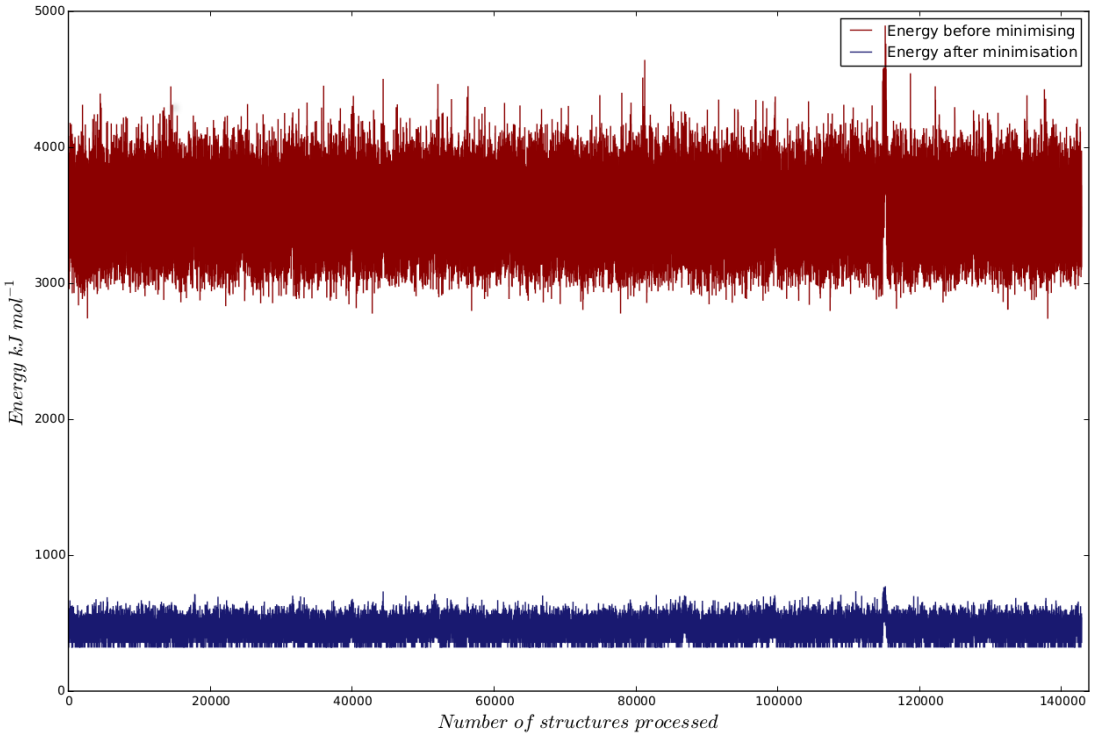


Figure 3.3.: Relative energies OPLS of the sampled structures (a constraint radius of 5 Å is applied in this case) before (red) and after (blue) geometry minimisation for the [8+12] version of **CC3**.

After removing duplicates (by selecting only one conformation out of a collection of degenerate structures - details regarding this procedure are explained later in the text) and inflated structure that collapsed upon minimisation, ∼30% of the molecules in the initial collection were found to correspond to metastable open states. Finally, SA was applied on the lowest energy open conformers found after the 300 K simulations. The SA consists in 500 ps simulations carried out at each temperature between 300 and 500 K, with increments of 50 K. This procedure was repeated a minimum of 10 times.

3.3.4. Constrained MD calculations with a cylindrical probe.

After testing systems that possess radial symmetry and that are therefore suited to be inflated with a spherical probe, a cylindrical probe was determined to be necessary for molecules with more elongated tubular shapes. The input file slightly varies from the previous one. Initially, two positions need to be chosen in order to give a direction c to the long axis of the cylinder. The beginning and end of the vector defining the orientation of the cylinder are in this case two centres of mass, $\mathbf{x}_a$ and $\mathbf{x}_b$. One of the centres of mass is calculated as the centroid of 3 atoms on the top of the molecule, the other one as the centroid of 3 atoms on the bottom of the molecule. The direction of the long axis of the cylinder is therefore calculated as:

$$\mathbf{c} = \frac{(\mathbf{x}_a - \mathbf{x}_b)}{|\mathbf{x}_a - \mathbf{x}_b|} \tag{3.3}$$

Where the bars in the denominator indicate the norm of the vector is considered. The distance, d, projected in the plane perpendicular to c between an atom at position y and the centre of the cylinder, can be calculated using the following cross product:

$$d = |(\mathbf{y}) - \mathbf{x}_a) \times \mathbf{c}| \tag{3.4}$$

The minimum distance projected in the plane normal to c, between the centre of the cylinder and all the atoms in the molecule can be calculated with an equation analogous to the one used for the spherical probe. The input file for this type of calculation is in Appendix A.1.2.

3.3.5. Determination of spherical void size.

At the end of the constrained MD procedure the void size of each conformation was calculated, in order to understand which were more likely to be open and symmetric and which collapsed upon minimisation. A PERL script was applied to all the conformers with the goal of determining the largest sphere that could fit the cavity. The script starts by calculating the centre of mass of each molecule, then it measures the shortest distance between the centre of mass and any atom in the molecule (with a procedure similar to the one that applies the constraint to the MD simulation). The radius of the largest sphere that can fit the cage's cavity is defined as the distance between the centre of mass and the nucleus of the closest atom, minus the Van der Waals radius of the atom (C = 1.70 Å, H = 1.20 Å, N = 1.55 Å, and O = 1.52 Å). Molecules were considered "open" if the radius was larger than a threshold value, that varied across the different sized cages. Specifically: 3.5 Å (**CC8**), 2.5 Å (Mastalerz cage), 2.0 Å (CC3 [6+9]) and 2.0 Å (CC3 [8+12]). Open structures for cucurbiturils and cryptophanes were manually selected due to their more asymmetrical shapes or because of specific features that were isolated. Cavities of this size are able to encapsulate small gaseous guests such as H_2, N_2, CO_2, Xe and some other common organic solvents. Finally, OPLS energies and void sizes were compared to remove duplicates. Conformers were considered duplicates when the energy difference between them was within ± 0.01 kJ mol^{-1} and the difference in their cavity radii ± 0.05 Å. The number of duplicates found for the lower energy structures was used to evaluate the accuracy of the sampling (the greatest the number, the better the sampling).

3.4. Results

Below are reported all the results obtained on the selection of molecules that were chosen for this study.

3.4.1. CC8.

The first molecule to be "inflated" was **CC8**, of which a solvated crystal structure is available for comparison.[96] The collapsed (left hand side) and open (right hand side) geometries of **CC8** are in Figure 3.1. **CC8** is a fairly large system, the largest that was investigated for the purpose of inflation, and it is made up of 272 non-hydrogen atoms. Experiments have previously shown that the material is non-porous

upon desolvation.[96] Simulations found the lowest energy conformer to collapse via a vertex-folding mechanism and no open structure was found within 50 kJ mol^{-1} from the lowest energy conformer. Constrained MD simulations were used to inflate the collapsed conformer and therefore to locate the lowest energy open conformer available. Different probe sizes were tested, with radii ranging from 3 to 9 Å. Many metastable higher energy conformers were located, however it was found that the lowest energy open conformer could be isolated with a probe radius of 6 Å, all the other probes found more unstable conformers, characterised by a lower degree of symmetry (see Figure 3.4).

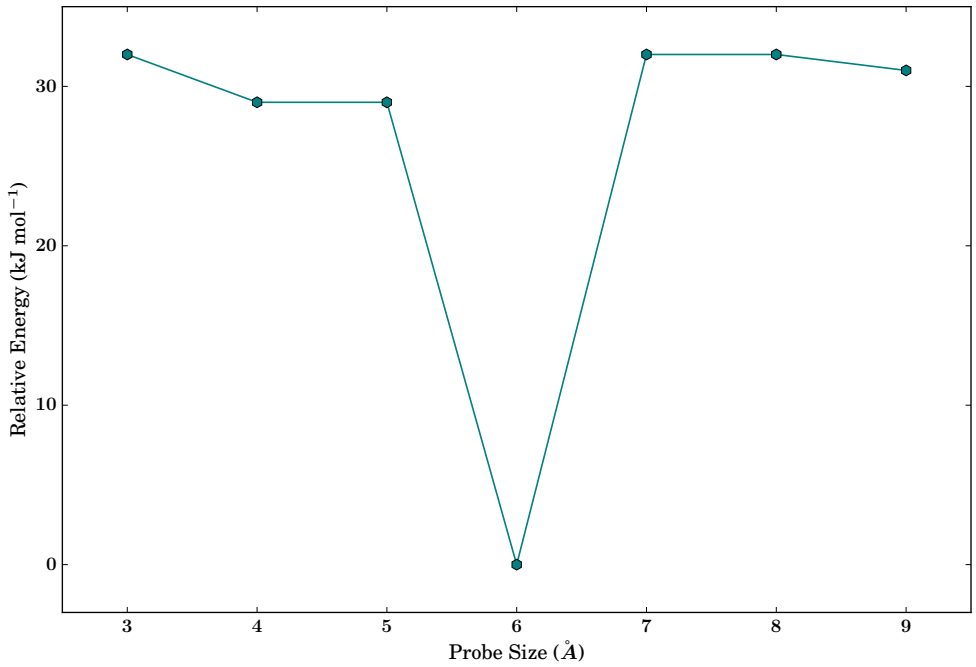


Figure 3.4.: Relative energies of the lowest energy open conformers of **CC8**, calculated at the OPLS_2005 level and obtained with different probe sizes.

The lowest energy open configuration (right hand side of Figure 3.1) exhibited a final void size of 4.6 Å, which is smaller than the probe radius used to inflate the molecule. This is because van der Waals radii are subtracted when calculating the void sizes, but not when calculating the initial constraint. Further investigation shows that the open metastable state lies more than 113 kJ mol^{-1} above the lowest energy collapsed state, it has the underlying topology of a cube and it exhibits pseudo-octahedral symmetry. The structure was compared with the experimental structure, which is also pseudo-octahedral, presumably because of the high level of solvation it allows for (about 70 solvent molecules per cage) and because of crystal

packing effect that make this conformation particularly favoured. In Figure 3.5-A the simulated open geometry and the solvated crystal structure are overlaid to see whether the open molecule can reproduce the shape the experimental molecule has when solvent is present. The calculated RMSD (Root Mean Squared Deviation - unless otherwise specified, the RMSD values are calculated taking into account both hydrogens and non-hydrogens atoms) value shows a difference of 1.67 Å between the two structures, mainly given by a larger contraction of the vertex pairs in the simulated structure. It is to be taken into account, though, that this simulation does not include crystal packing effects or the effects of solvent interactions and that therefore, a more accurate comparison should be done with an OPLS minimised solvate crystal structure.

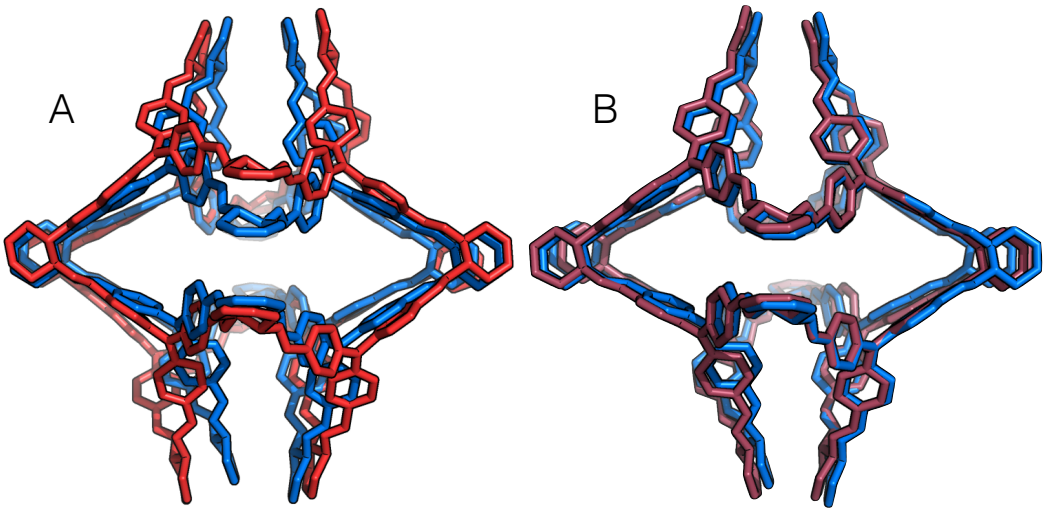


Figure 3.5.: Overlays of **CC8**: A) the computed open structure (blue) and the solvate crystal structure (red); B) comparison of the computed open structure (blue) and the OPLS-minimised solvate crystal structure (dark red).

In Figure 3.5-B the modelled structure is overlaid with the minimised solvate crystal structure. Interestingly, upon minimisation, vertex contraction occurs in the experimental structure too, thus reducing the RMSD value to 0.63 Å. Nevertheless, the prediction of conformation, shape and void size is in good agreement with experiment.

3.4.2. Mastalerz endo-functionalised cage.

The imino cage synthesised by Mastalerz[77] (Figure 3.6, right hand side), is a symmetric tetrahedral cage of 170 non-hydrogen atoms of which a solvate crystal structure is available. MD simulations showed that, upon desolvation, the tetrahedral symmetry of the molecule was lost. This can be caused by the collapse of one methyl group towards the centre of the cavity, which led to the folding of one of the tetrahedron vertices and therefore to a partial collapse of the void (Figure 3.6, left hand side).

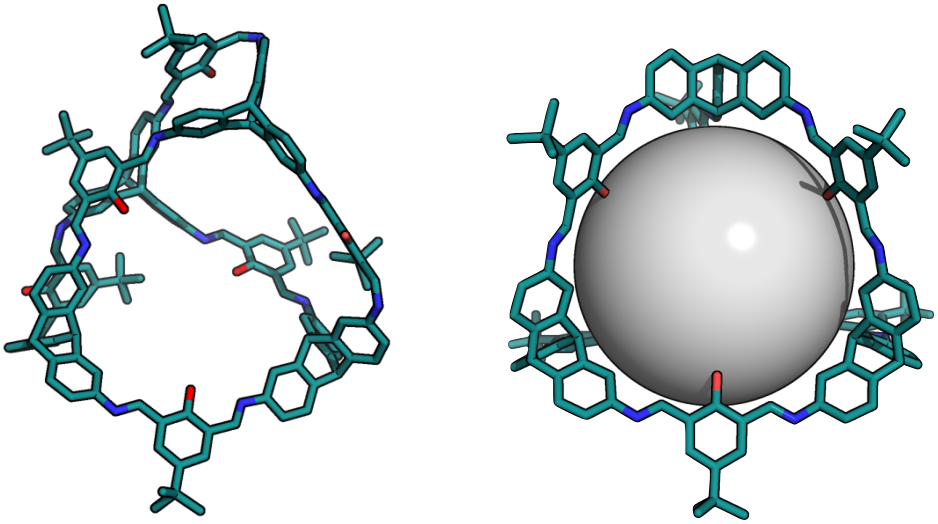


Figure 3.6.: Comparison of the collapsed (left hand side) and open (right hand side) geometries of Mastalerz endo-functionalised cage. The cavity presents a void radius of 5.2 Å.

A smaller void of 2.1 Å is preserved, although much smaller than the 5 Å one calculated for the solvate crystal structure. This would explain why the reported BET surface area of 1377 m^2 g^{-1} [78] is much smaller than the computed accessible surface area of 4000 m^2 g^{-1} for a N_2 gas in the solvated crystal structure (considering radii of 1.55 - 1.82 Å, spanning the van der Waals and kinetic radius of N_2).[194] As in the previous case, the molecule was investigated with constrained MD simulations, using probe sizes ranging from 1 to 9 Å. A plot showing the energies of the lowest energy open conformers found with each probe is in Figure 3.7.

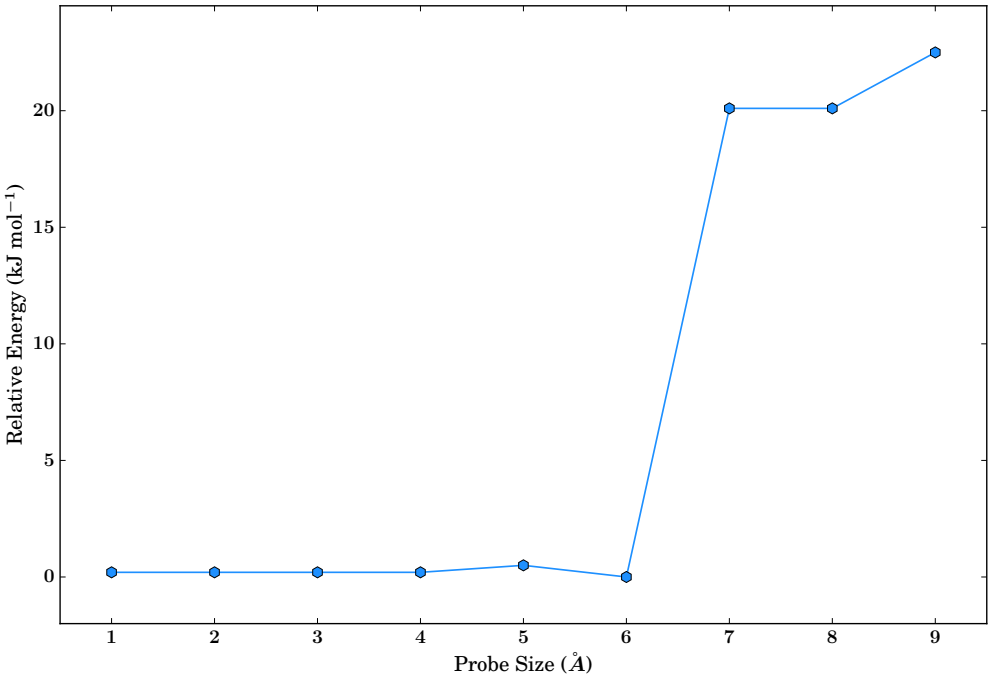


Figure 3.7.: Relative energies of the lowest energy open conformers of the Mastalerz endo-functionalised cage, calculated at the OPLS_2005 level, and obtained with different probe sizes.

It was possible to locate the lowest energy open metastable state with any probe equal or smaller than 6 Å. When using a larger probe, higher energy structures were found ($>$ 20 kJ mol^{-1} higher than the lowest energy open conformer). The inflated molecule lies 158 kJ mol^{-1} above the partially collapsed lowest energy conformer and has a void radius of 5.2 Å.

Again, the model reproduces well the solvate crystal structure, presenting an RMSD value of 2.20 Å. Minimising the crystal structure does not lead to a significant difference in results. The minor torsional differences that can be seen in Figure 3.8 can be caused by either minor force field inaccuracies or crystal packing effects, that can not be accounted for in my simulations.

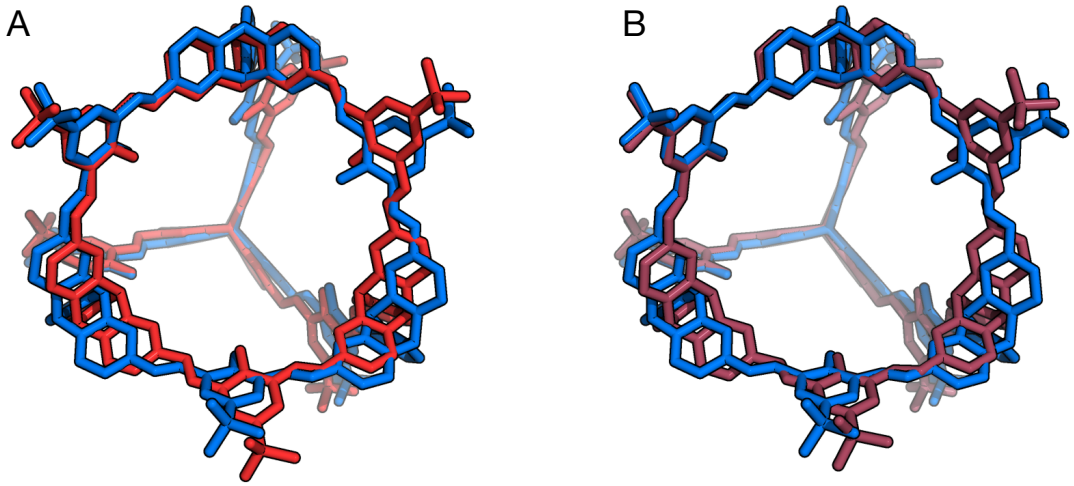


Figure 3.8.: Overlays of the Mastalerz endo-functionalised cage: A) the computed open structure (blue) and the solvate crystal structure (red); B) comparison of the computed open structure (blue) and the OPLS-minimised solvate crystal structure (dark red).

3.4.3. [6+9] and [8+12] versions of the CC3 imine cage.

After observing successful results with the previous studies, two hypothetical imino cages were studied, all reported in Figure 3.9. These are built according to the same chemistry of the experimentally reported [4+6] **CC3** cage, but precursors are mixed into a [6+9] (Figure 3.9-A) and [8+12] (Figure 3.9-B) ratios, giving, respectively, molecules with the underlying topologies of a trigonal prism and a cube. These two hypothetical structures are of particular interest for the topology prediction of systems built with a combination of tritopic and ditopic precursors, as I will discuss in Chapter 5. Different topologies coming from the same couple of precursors are going to be compared in order to confirm that the tetrahedral topology of **CC3** is the most likely to form. In this work, only systems with a cavity can be compared, as collapsed structures are not interesting for the purpose of the study.

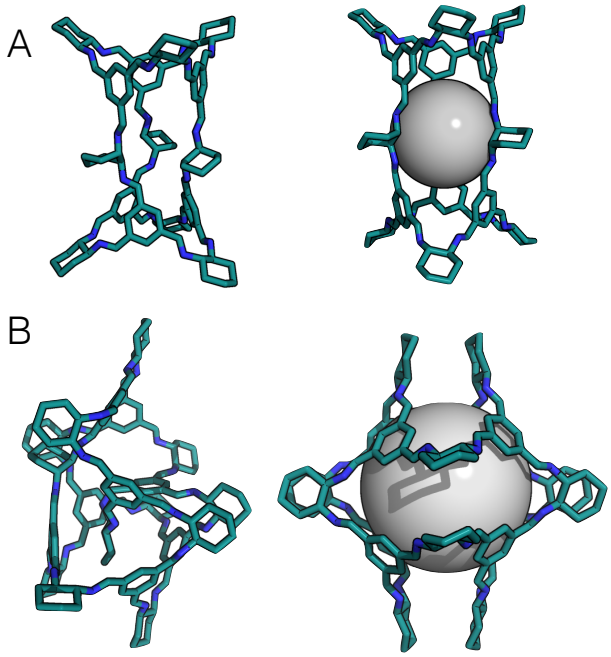


Figure 3.9.: Comparison of the collapsed (left hand side) and open (right hand side) geometries of **CC3**. A) [6+9] B) [8+12]. **CC3**-[6+9] has a void radius of 2.6 Å, whereas **CC3**-[8+12] of 4.1 Å.

CC3-[8+12] imine cage

For the **CC3**-[8+12] case, all structures were very well sampled by constrained MD. The lowest number of conformers was located by the smallest (3 Å) and the biggest (7 Å) constraints adopted, finding respectively ∼5800 and ∼14300 open conformations each. Spherical probes with intermediate radii (4 to 6 Å) can locate from ∼22500 to ∼26500 open structures each, exception made for the constraint with a radius of 5.5 Å that was able to locate ∼65000 open structures, therefore providing the best sample of the PES. The open conformations with the lowest energy for each probes are reported in Figure 3.11. As in the previous cases, the use of different probe sizes led to different structures. As it can be seen from Figure 3.11, smaller probes (3-4 Å) failed to inflate the structures properly and gave very high energy partially collapsed structures, whereas larger probes (6-7 Å) tended to over-inflate the structures leading to strained higher energy conformers. Symmetric open conformations with pseudo-octahedral symmetry were successfully found with medium-sized probes (4.5-5.5 Å).

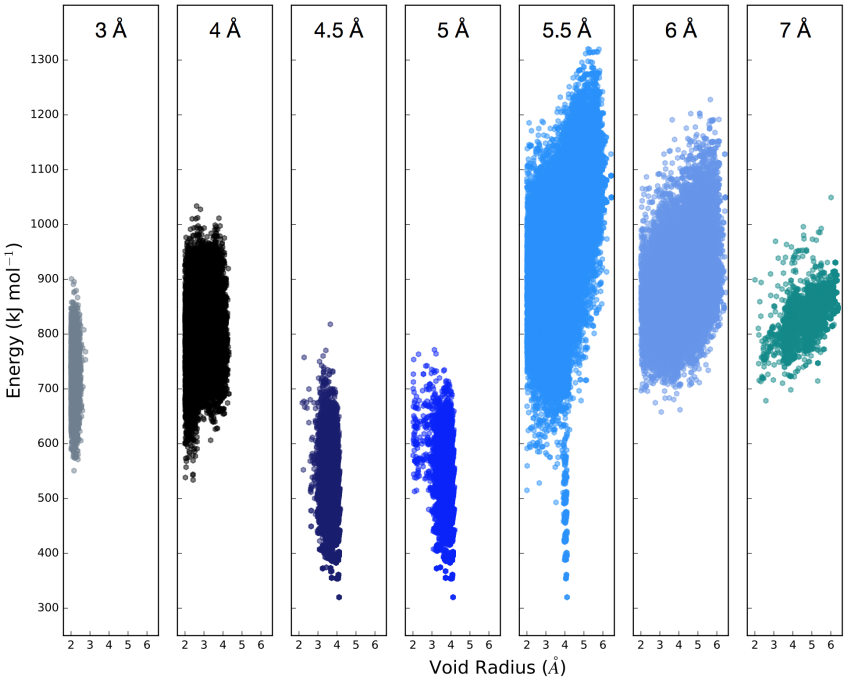


Figure 3.10.: The OPLS geometry optimised energies and final void sizes of CC3 [8+12] structures from constrained MD simulations at 300 K with a range of different constraint radii (as labelled at top). Each sub-plot shows a separate simulation run with a different constraint radii. Note that for the void radius plotted, the van der Waals radius of the nearest atom is subtracted, thus these values are typically smaller than the constraint sizes applied.

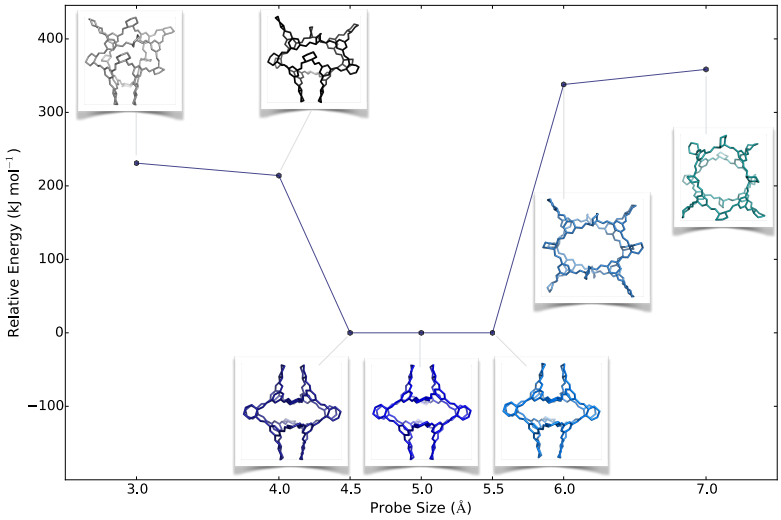


Figure 3.11.: Relative energies of the lowest energy open conformers of **CC3**-[8+12] (in different colours), calculated at the OPLS_2005 level and obtained with different probe sizes.

CC3-[6+9] imine cage

Differently, **CC3**-[6+9] could potentially adopt a trigonal prism shape, so it was inflated in the first place with a cylindrical probe, however it was found it was not crucial to find prismatic structures, as trigonal prism shaped molecules were located regardless using spherical probes. Results obtained with the spherical probes are in Figure 3.12.

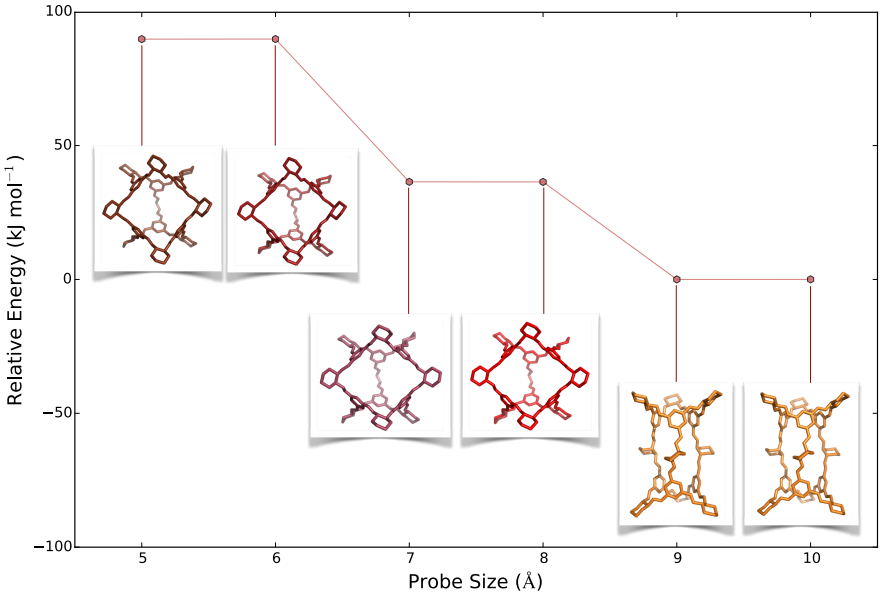


Figure 3.12.: Relative energies of the lowest energy open conformers of **CC3**-[6+9] (in different shades of red), calculated at the OPLS_2005 level and obtained with different probe sizes. The viewing angle is rotated for clarity for the structures inflated with 9 and 10 Å probes.

Consistently with the **CC3**-[8+12] case, it was observed that probes with a radius larger than 9 Å led to an over-inflation of the structures. However, it is shown that the over inflated structure minimised to symmetric open conformations, even if the sampling in this instance was quite poor, as it can be seen in Figure 3.13. Smaller spherical probes (5 and 6 Å) are, in fact, able to locate only ∼100 structures each; intermediate size (7 and 8 Å) ∼490. The number almost doubles when the two biggest probes are used (9 and 10 Å), sampling respectively 1037 and 1804 conformers each, sign that larger constraints sample the PES of this kind of structures better. The poor sampling could, in general, be attributed to the strained angle three of the diamines have to adopt when fitting the long edges of the trigonal prism, making these geometries more strained and less likely to form. In this case too, the lowest

energy open structures located by each probe are reported, with their corresponding energies, in Figure 3.12.

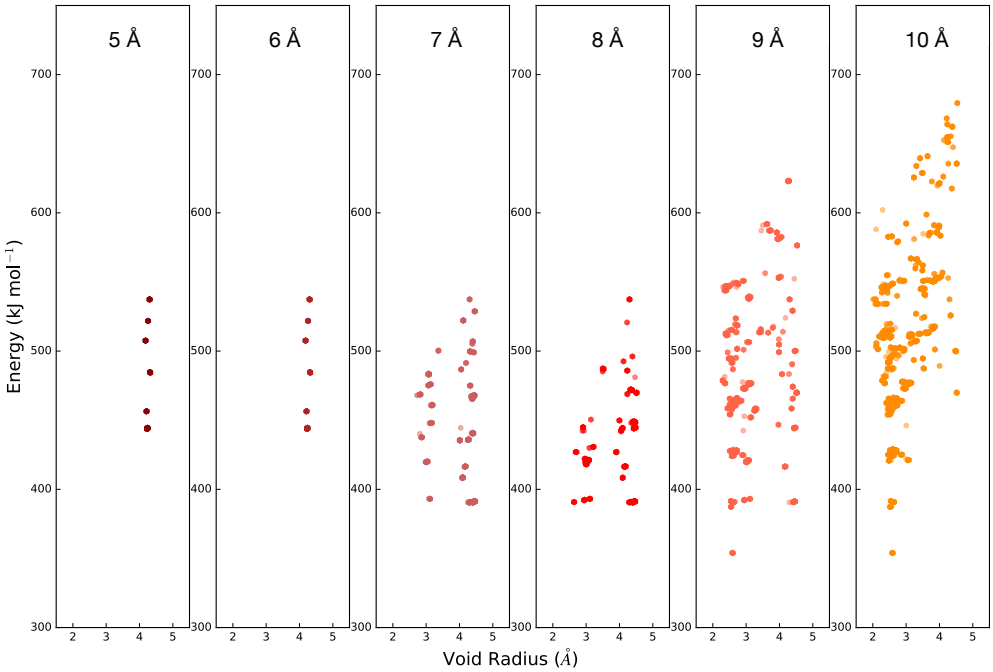


Figure 3.13.: The OPLS geometry optimised energies and final void sizes of CC3 [6+9] structures from constrained MD simulations at 300 K with a range of different constraint radii (as labelled at top). Each sub-plot shows a separate simulation run with a different constraint radii. Note that for the void radius plotted, the van der Waals radius of the nearest atom is subtracted, thus these values are typically smaller than the constraint sizes applied.

All the structures of this kind lack spherical cavities but are characterised by a tubular-shaped void. Lower probe sizes over-inflate the structures, although not enough to provoke further relaxation and lead to strained high-energy conformations.

3.4.4. Macrocycles and cavitands.

The constrained MD approach was initially conceived of only for cage systems, but given the successful results and the adaptability of the procedure to molecules with different chemistry and shapes, the inflation method was tested on macrocyles and cavitands too. The investigation began with cucurbit[6]uril (CB[6])[189] and cucurbit[10-ns]uril (CB10-ns),[190] then moved on to Cryptophane A.[191]

Cucurbit[n]urils.

Conformer search investigations on the smallest macrocycle studied, CB[6] (Figure 3.14-A), showed that, in the lowest energy conformation, every other glycouril is inverted (Figure 3.14-A, left hand side).

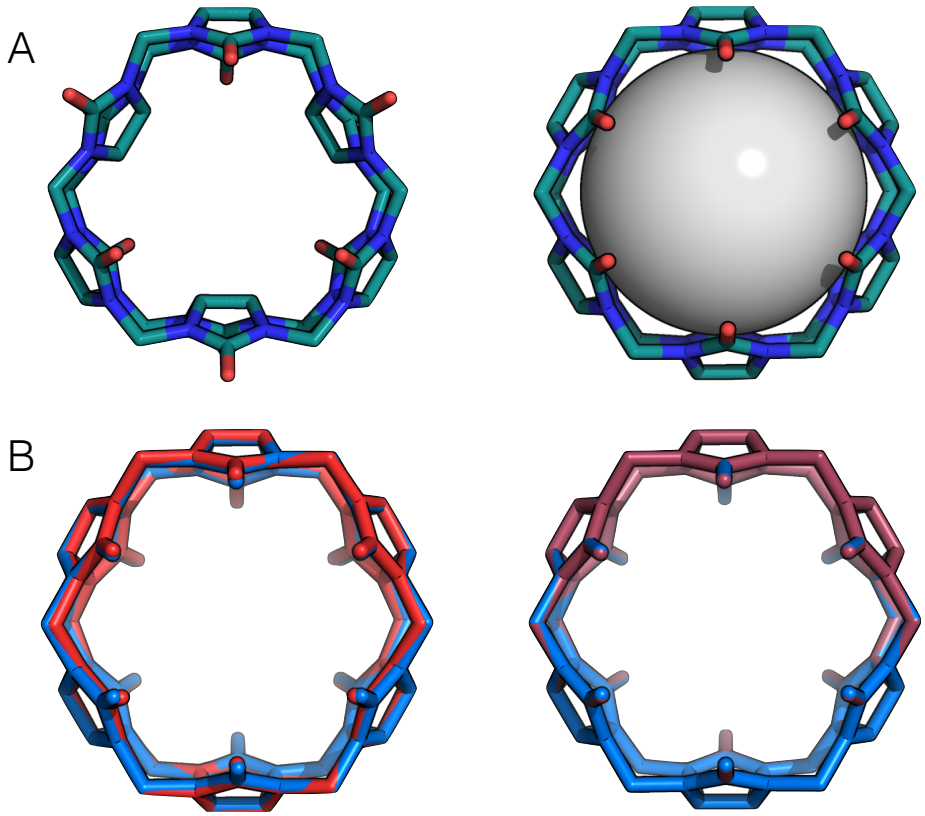


Figure 3.14.: A) Comparison of the collapsed (left hand side) and open (right hand side) geometries of CB[6]. The open geometry has a void radius of 2.6 Å. B) Left: overlay of the model (blue) and solvate crystal structure (red); Right: model (blue) and OPLS-minimised solvate crystal structure (dark red).

This confirms previous experimental studies that showed potential inversion of glycouril units.[195] Constrained MD simulations were used in this case to open the geometry by flipping glycouril units, so that in the final structure they are all oriented in the same way. Any probe above 2.5 Å gave open conformations with no glycouril inversion. A comparison was carried out with the available solvate crystal structure, which showed an excellent match (Figure 3.14-B, left hand side), with an RMSD value of 0.10 Å. Minimising the crystal structure removes even the smallest differences between the two molecules (Figure 3.14-B, right hand side), reducing the

RMSD value to 0.00 Å (which corresponds to a perfectly reproduced structure).

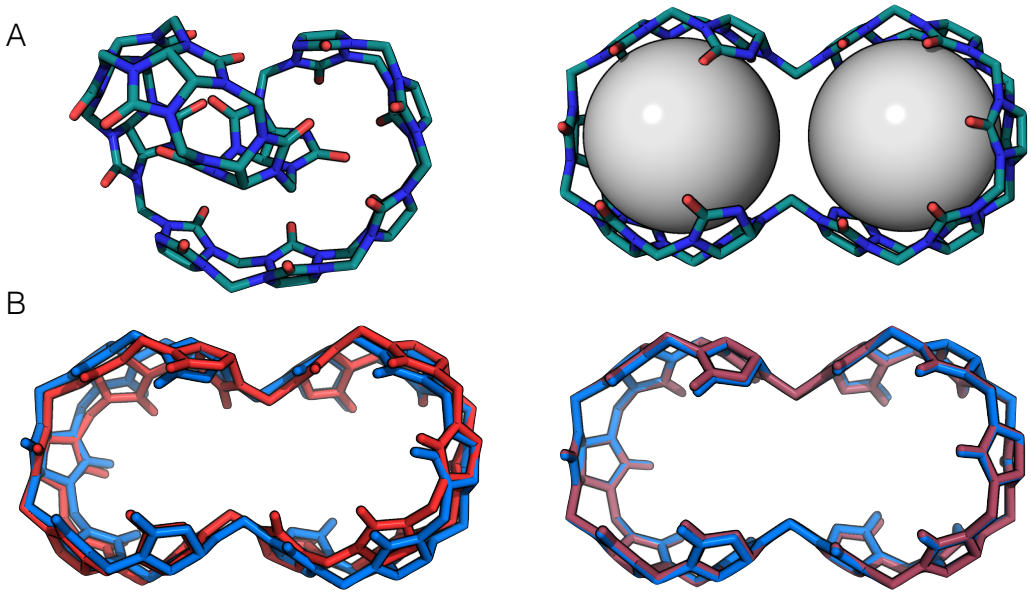


Figure 3.15.: Comparison of the collapsed (left hand side) and open (right hand side) geometries of CB10. The open geometry has two cavities, each with a void radius of 2.5 Å; B) Left: overlay of the model (blue) and solvate crystal structure (red); Right: model (blue) and OPLS-minimised solvate crystal structure (dark red).

When moving to bigger cucurbiturils such as CB10-ns, the lowest energy collapsed conformation was found to lack an internal cavity, as shown in Figure 3.15-A, on the left hand side. The solvate crystal structure shows molecules with two distinct cavities (not corresponding to the host molecule's centre of mass), rather than a ring shape. When applying constrained MD simulations, it was possible to exclusively find open conformations with the reported dual symmetrical interconnected cavities when using probes larger than 16 Å (Figure 3.15-A, on the right hand side). I then proceeded to overlay the simulated structure to the SCXRD structure, as shown in Figure 3.15-B (left hand side), finding a good overlay with an RMSD of 0.66 Å, which, as in other cases, was decreased to 0.13 Å upon minimisation of the SCXRD structure (same Figure - right hand side), showing again an excellent match with experiments.

Cryptophane A.

The literature for cavitand Cryptophane A,[191] reported a collapsed structure as well as a number of experimental crystal structures loaded with different guests, such as

xenon and water.[191] The collapsed conformation is shown on the left hand side of Figure 3.16-A. Constrained MD calculations on the collapsed structure correctly predict two of the possible open conformations of the molecule, one of which is reported on the right hand side of Figure 3.16-A.

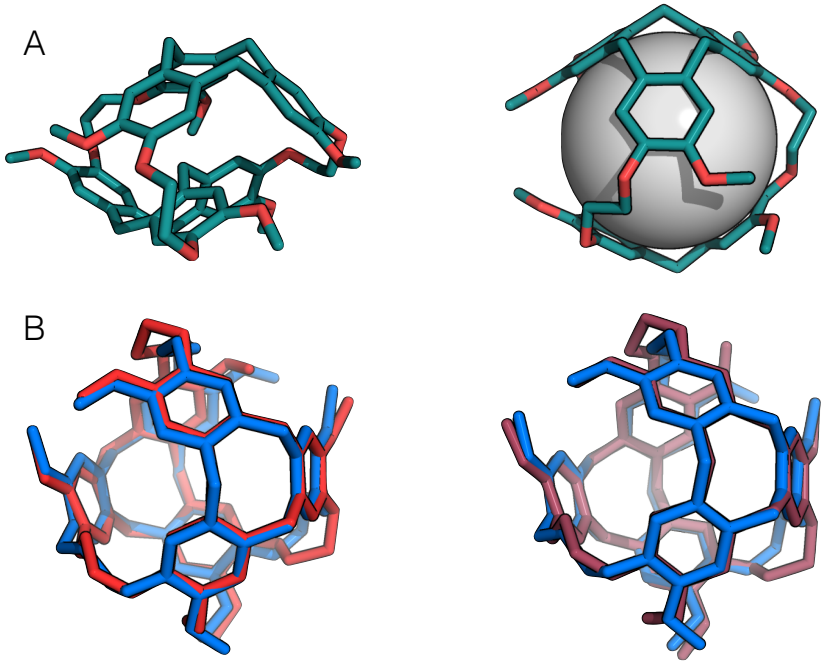


Figure 3.16.: Comparison of the collapsed (left hand side) and open (right hand side) geometries of Cryptophane A (Xe-loaded).[191] B) Left: overlay of the model (blue) and solvate crystal structure (red); Right: model (blue) and OPLS-minimised solvate crystal structure (dark red).

The cavity of cryptophane can change depending on the lenght and/or composition of the two alkane chains linking the two CGT cups. Also the internal volume changes depending on the intermolecular interactions occurring with the solvent and therefore with the kind of guest they load.[196,197] Previous studies show that Cryptophane A preferentially binds Xe over a wider selection of solvent (although the reason for this is not yet understood).[197–199] Particularly favourable interactions happen when water is present. In fact Cryptophane A adopts a more compact geometry when in water, which should increase favourable van der Waals contacts with Xe.[191]

When carrying out the overlays with the SCXRD structures, it was found that different probe sizes were able to match both the Xe-loaded (Figure 3.16-B, left hand side) or the more "squeezed" water-loaded crystal structures. In particular a smaller

5 Å constraint could match the water-loaded crystal structure, a larger 7 Å probe the Xe-loaded crystal structure. This was possible because of the wide range of conformations that were sampled, with spherical cavities ranging from 2.0 to 2.5 Å in radius. The structures obtained from constrained MD simulations were overlaid with the SCXRD structures loaded with water and Xe, and a match was found for both cases. The RMSD values are of 0.62 Å for the overlay with the Xe-loaded structure, and 0.68 Å for the overlay with the water-loaded structure. Given the great flexibility of the methyl groups, the RMSD value is not improved by minimising the SCXRD structure, and it just changes to 0.66 Å for the overlay with the Xe-loaded structure (Figure 3.16-B, right hand side) and 0.70 Å for the overlay with the water-loaded structure. Also a comparison of the void radii of each was performed, confirming that simulated and experimental geometries are extremely similar. The calculated void radius for the geometry inflated with a 5 Å probe is 2.0 Å, just 0.04 Å smaller than the void of the SCXRD of the water-loaded cryptophane; the void radius of the structure inflated with a larger 7 Å probe gives a void radius of 2.5 Å, 0.13 Å larger than the void of the SCXRD of the Xe-loaded cryptophane. The work from Taratula *et al.* shows how the volume of Cryptophane A can increase by 20 % when loading its cavity with solvents of increasing size, ranging from methanol to chloroform.[191] My investigation was focused on comparing two different structures, but I do not exclude that constrained MD methods could reproduce the cavities of a wider range of solvent-loaded cryptophane capsules.

3.5. Conclusions

The table below summarises this work's results, listing for each molecule the number of non-hydrogen atoms, the OPLS_2005 energy of the collapsed structure, the OPLS_2005 energy and the void sizes of the open structures. All the collapsed structures selected for the study were successfully "inflated", and metastable open conformers were located. The energies of the open conformers were always >100 kJ mol^{-1} higher than the ones of their lower energies collapsed conformations, confirming that normal conformer search algorithm would not have easily located the open states.

Table 3.1.: Comparison of the relative energy of the lowest energy collapsed and open structures for each system, calculated using OPLS and given in kJ mol^{-1}.

System	non-hydrogen atoms	Collapsed	Inflated	Inflated void radii (Å)
CC8 [8+12]	272	395	508	4.6
Mastalerz [6+4]	170	1499	1657	5.2
CC3 [6+9]	128	187	353	2.6
CC3 [8+12]	168	220	320	4.1
CB[6]	72	2025	2129	2.8
ns-CB10	118	2977	3165	two cavities of 2.5
Cryptophane A (Xe-loaded)	66	399	854	2.5

This part of the work was dedicated to the development of a new approach which allows one to predict the open metastable conformations of porous organic molecules in solution or in the solvate solid state. The procedure starts by determining the lowest energy conformers of the chosen systems and proceeds to its inflation with the aid of constrained MD simulations. Probes of different sizes and shapes can be used to suit different molecules' geometrical characteristics. Thanks to its simplicity and flexibility, the method has the potential to be used for the prediction of different classes of porous systems, in particular for guest responsive materials, such as MOFs. The investigation shows that in the field of porous organic materials, the solvent mainly acts as a "scaffold", and its effect on the geometry of the molecule is reproduced here without taking into account any "host-guest" chemical interaction. In conclusion, I believe this technique constitutes a significant advance for the *in silico* design of porous organic cages, as successful prediction of the solvated structures was achieved for a broad variety of systems.

Chapter 4
Topological families and nomenclature of porous organic cages

The work in this chapter led to the publication of:
Santolini V., Miklitz M., Berardo E. and Jelfs K.E., *Nanoscale*, 2017, 9, 5280-5298.



4.1. Introduction

In this chapter I will discuss the process that led to a new classification system for porous organic cages topologies. It involved the development of a nomenclature that aimed at the description of the connectivity and geometry of porous organic cages, whilst understanding the criteria to improve design strategies. The 20 most probable topologies were defined, 12 of which had already been experimentally synthesised. The nomenclature hereby described is going to be applied to all the cages discussed in the following chapters. The topic of topology is going to be further discussed in Chapter 5 where, by comparing the internal energies of cages of different topology, I am going to show how I predicted, with the aid of computational simulations, the most plausible experimental reaction outcome for a series of 10 different building block combinations.

4.2. A new nomenclature for porous organic cage topologies.

As already discussed in Section 1.4 in the Introduction chapter, at the beginning of my PhD, a uniform classification system of the topologies of porous organic cages did not exist. The aim of the work discussed here is both to provide a way of describing systems existing in the literature and to rationalise the number of feasible topological outcome one may encounter when attempting the design and synthesis of porous organic molecules. In the next section the topologies synthetically reported to date are going to be examined and divided into families, depending on the coordination number of the precursors involved in their making. This is to, first of all, generate a consistent nomenclature for organic cages, then to rationalise the criteria that can help the design of the desired topologies and shapes.

4.2.1. Enumeration of organic cage topologies

Below are listed the main criteria for the enumeration of plausible organic cage topologies:

1. The topology must be a polycyclic cage molecule, with the potential for a 3-dimensional internal cavity. Macrocycles such as cyclodextrins, crown ethers, cucurbiturils and 2-dimensional topologies (that would correspond to polygons such as triangles, squares, pentagons and so on) are excluded.

2. The topology must be obtained from a two-component reaction of BBs through DCC.

3. The two BB components can be ditopic, tritopic and tetratopic only.

4. BBs should have a C_n symmetry axis with n equal to the number of reactive end groups.

5. For topologies with larger numbers of components, only highly symmetric structures are included. These are structures relatable to Platonic, Archimedean or Catalan solids (shown again in Figure 4.1) that have been already reported for MOPs, as they are more feasible targets for organic cage molecules.

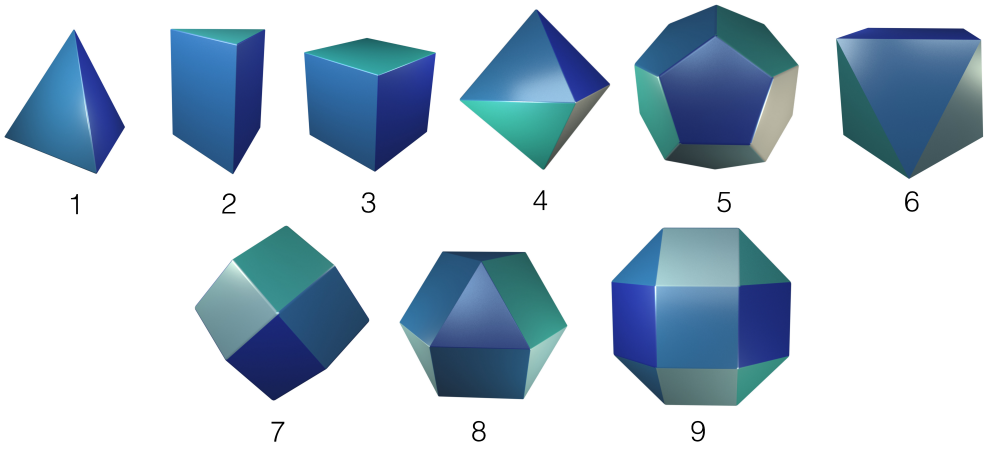


Figure 4.1.: Platonic, Archimedean and Catalan polyhedra that can be related to organic cage topologies. 1) Tetrahedron, 2) Trigonal Prism, 3) Cube, 4) Octahedron, 5) Dodecahedron, 6) Square Antiprism, 7) Rhombic Dodecahedron, 8) Cuboctahedron, 9) Rhombicuboctahedron.

These rules did not lead to an exhaustive set of topologies, but included the 20 most plausible underlying topologies for porous organic cages. Almost two thirds of these had been synthetically reported. The mathematicians from the group of Reinhard developed an algorithm to generate maps of stable planar cages and produced >400,000 unique maps.[200] However, they did not restrict themselves to two-components systems and in the majority of the cases the maps were too complex to be applied to organic cages, which made the selection of 20 possible topologies more practical. Nonetheless, the classification system described in this section can be used in the future to define the underlying connectivity of multi-component systems as well, of which multiple examples already exist in the literature.[87,88,91,134] Some

also involved the use of orthogonal DCC reactions, including the formation of both boronate and imine bonds.[114,201]

A new nomenclature for organic cages was introduced, and each structure was labelled according to it:

$$\mathbf{X_p^m Y^n}$$

where **X** and **Y** are the two different component BBs that constitute the cage. **X** and **Y** are labelled **Di** if they are ditopic, **Tri** if they are tritopic and **Tet** if they are tetratopic. The first BB, **X**, has the highest number of reactive end groups and it will lie at the vertices of the polyhedra it relates to. The second BB, **Y**, can have a number of reactive end groups less than or equal to **X**. Where the number of reactive end groups for **X** and **Y** is equal, which is denoted **X** and which is denoted **Y** is arbitrary. The superscripts, **m** and **n**, denote the number of each BB incorporated in the topology for **X** and **Y** respectively. A schematic representation of the nomenclature explained so far is reported in Figure 4.2.

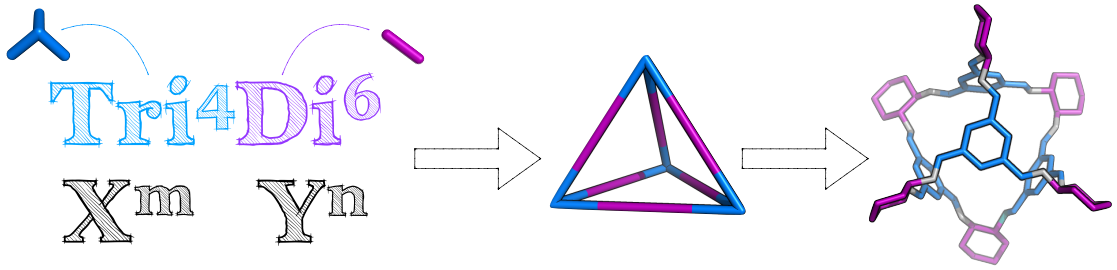


Figure 4.2.: Example of nomenclature for a cage of **TriDi** topology. The final **Tri**^{4}**Di**^{6} cage is rotated by 90° respect to the tetrahedron for clarity.

The majority of the time, **X**-type BBs are connected to other **X**-type BBs through only one **Y**-type BB; in this case no subscript **p** will be given. However, if two **X**-type BBs are directly connected through links with two distinct **Y**-type BBs, then they have $\mathbf{p} = 2$. The subscript, **p**, thus gives the number of double connections between BB pairs within a topology (see Figure 4.3).

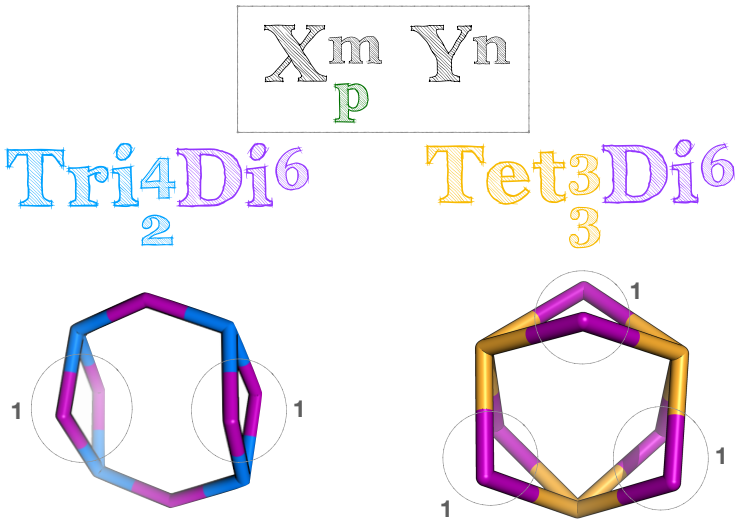


Figure 4.3.: Two nomenclature examples for the $\mathbf{Tri}_2^4\mathbf{Di}^6$ and the $\mathbf{Tet}_3^3\mathbf{Di}^6$ topologies. In the first case a double bond between two tritopic precursors occurs twice, therefore p=2; in the second case it occurs three times, therefore p=3.

These multiple links are also reported for periodic nets[25] and result in topologies with multiple ring sizes. For some of the smaller topologies with only two **X**-type BBs, there is triple or quadruple linking of those BBs; in this case no subscript is given as there is no alternative connectivity for that topology. Finally, it is advisable to use the prefix **c**− in the case of catenated cages, equivalent to its usage as a prefix with interpenetrated periodic nets.[27]

4.3. Topological families

In this section I am combining BBs with different connectivity into four different topological families. The topologies and their features are individually discussed, and are summarised in Table 4.1. Topologies that have already been synthetically reported are in Table 4.2. *

*The literature search for the table was done by Marcin Miklitz.

Table 4.1.: Table of possible organic cage topologies and their key features. Ring (representing windows in the cages) sizes are defined as a^b, where a is the number of individual BBs that form part of the ring and b is the number of rings of size a

Topology	Edge-directed form	Face-directed form	Solid type	Multiple connections	Point group in high symmetry form	Ring sizes
$\mathbf{Tri}^2\mathbf{Di}^3$				1 triple	D_{3h}	4^3
$\mathbf{Tri}^4\mathbf{Di}^6$	Tetrahedron		Platonic		T_d	6^4
$\mathbf{Tri}_2^4\mathbf{Di}^6$				2 double	D_{2h}	4^2, 8^2
$\mathbf{Tri}^6\mathbf{Di}^9$	Triangular prism				D_{3h}	6^2, 8^3
$\mathbf{Tri}^8\mathbf{Di}^{12}$	Cube		Platonic		O_h	8^6
$\mathbf{Tri}^{20}\mathbf{Di}^{30}$	Dodecahedron		Platonic		I_h	10^{12}
$\mathbf{Tet}^2\mathbf{Di}^4$				1 quadruple	D_{4h}	4^4
$\mathbf{Tet}_3^3\mathbf{Di}^6$				3 double	D_{3h}	4^3, 6^2
$\mathbf{Tet}_2^4\mathbf{Di}^8$				2 double	C_{2v}	4^2, 6^4
$\mathbf{Tet}_4^4\mathbf{Di}^8$				4 double	D_{4h}	4^4, 8^2
$\mathbf{Tet}_5^5\mathbf{Di}^{10}$				5 double	D_{5h}	4^5, 10^2
$\mathbf{Tet}^6\mathbf{Di}^{12}$	Octahedron		Platonic		O_h	6^8
$\mathbf{Tet}^8\mathbf{Di}^{16}$	Square antiprism				D_{4d}	6^8, 8^2
$\mathbf{Tet}^{16}\mathbf{Di}^{32}$	Cuboctahedron		Archimedean		O_h	6^8, 8^6
$\mathbf{Tet}^{24}\mathbf{Di}^{48}$	Rhombicuboctahedron		Archimedean		O_h	6^8, 8^{18}
$\mathbf{Tri}^1\mathbf{Tri}^1$					C_{3v}	2^3
$\mathbf{Tri}_2^2\mathbf{Tri}^2$				2 double	D_{2h}	2^2, 4^2
$\mathbf{Tri}_2^3\mathbf{Tri}^3$				2 double	C_1	2^2, 4^4
$\mathbf{Tri}^4\mathbf{Tri}^4$		Tetrahedron	Platonic		T_d	4^6
$\mathbf{Tet}^6\mathbf{Tri}^8$		Rhombic-dodecahedron	Catalan		O_h	4^{12}

Table 4.2.: Experimentally reported organic cage topologies, synthesised via two-component DCC reactions. **Di** stands for ditopic precursors, **Tri** for tritopic, **Tet** for tetratopic.

Topology	Reported?	CSD reference codes for crystal structures of archetypal examples.
$\mathbf{Tri}^2\mathbf{Di}^3$	yes	ZUYPUG,[202] AJOHUD,[203] SATJAA[204]
$\mathbf{Tri}^4\mathbf{Di}^6$	yes	PUDXES,[7] TOVWUY,[205] EKUKUR[77]
$\mathbf{Tri}_2^4\mathbf{Di}^6$	no†	-
$\mathbf{Tri}^6\mathbf{Di}^9$	no	-
$\mathbf{Tri}^8\mathbf{Di}^{12}$	yes	KATJAS,[96] REYMAL,[206] ZIRCIO[84]
$\mathbf{Tri}^{20}\mathbf{Di}^{30}$	no	-
$\mathbf{Tet}^2\mathbf{Di}^4$	yes	LUXVAB[207]
$\mathbf{Tet}_3^3\mathbf{Di}^6$	yes	VILCEZ[208]‡
$\mathbf{Tet}_2^4\mathbf{Di}^8$	yes	no reported crystal structures[99]
$\mathbf{Tet}_4^4\mathbf{Di}^8$	yes	AVAFIN[80]
$\mathbf{Tet}_5^5\mathbf{Di}^{10}$	no	-
$\mathbf{Tet}^6\mathbf{Di}^{12}$	yes	no reported crystal structures[99]
$\mathbf{Tet}^8\mathbf{Di}^{16}$	yes	no reported crystal structures[99]
$\mathbf{Tet}^{16}\mathbf{Di}^{32}$	no	
$\mathbf{Tet}^{24}\mathbf{Di}^{48}$	no	
$\mathbf{Tri}^1\mathbf{Tri}^1$	yes	no reported crystal structures[209]
$\mathbf{Tri}_2^2\mathbf{Tri}^2$	no	-
$\mathbf{Tri}_2^3\mathbf{Tri}^3$	no	-
$\mathbf{Tri}^4\mathbf{Tri}^4$	yes	VOFROZ[210]
$\mathbf{Tet}^6\mathbf{Tri}^8$	yes	QUFYIB, QUFYOH[211]

†This topology has been reported for a single component alkene metathesis reaction.[212]

‡This topology has been reported with a bifunctional ditopic ligand and a cyclisation reaction.

4.3.1. Tritopic + Ditopic topology family

Tritopic and ditopic precursors can be mixed in a 2:3 ratio to give organic cages. The smallest cage of this family is a capsule obtained by reacting two tritopic BBs with three ditopic BBs in a [2+3] reaction. This led to the $\mathbf{Tri}^2\mathbf{Di}^3$ topology shown in Figure 4.4, where all the **TriDi** topologies are reported. This topology was reported multiple times in the literature and some examples, including reference codes for

structures in the CSD, are given in Table 4.2. These capsule-like cages have been used for many different applications, from gas separation[213] to sensing.[179] If the constituent BBs have a fairly rigid geometry, the cages had shape-persistent cavities, which were accessible via three possible windows. All the windows of **Tri^{2}Di^3** cages have ring size 4 (where 4 is the number of BBs that constitute the ring).

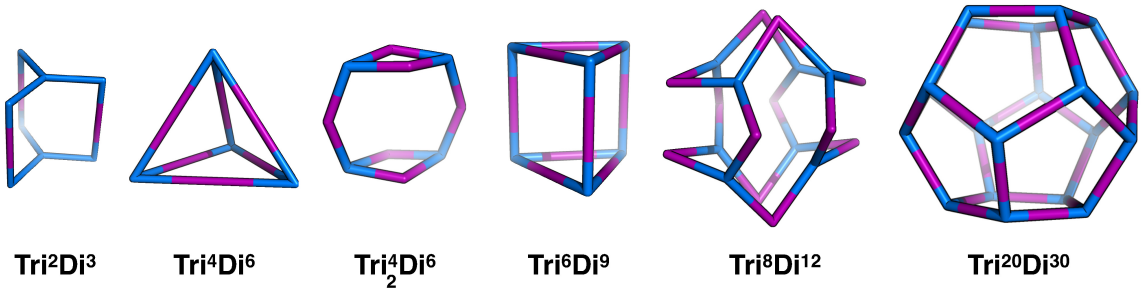


Figure 4.4.: The Tritopic + Ditopic topology family. Tritopic vertices are in blue, ditopic linkers in purple.

When mixing 4 tritopic BBs with 6 ditopic BBs, cages with a **Tri^{4}Di^6** topology are generated. They form a tetrahedron via an edge-directed assembly, with the tritopic BBs on the vertices and the ditopic BBs on the edges. If the cage is shape-persistent, the cavity can be accessed by four potential windows, each of which has a ring size of 6. This topology was also often reported in the literature, the cages coming from Cooper and co-workers[7] being used for molecular separation[141,145,148] and as porous liquids,[125] the ones from Mastalerz and co-workers for sensing.[214] This topology was also reported in a triply-interlocked catenated form of cages **CC1** (Covalent Cage 1), **CC2** and **CC4**, which was labeled **c-Tri^{4}Di^6**.[70] A molecule of **Tri^{4}Di^6** topology, if shape-persistent, can adopt final shapes ranging anywhere between a tetrahedron at one extreme to an octahedron at the opposite extreme. Some examples from the literature of **Tri^{4}Di^6** topology are also shown in Figure 1.11. The tetrahedral shape formed when the tritopic BBs adopted a capping vertex-like conformation, with (end group)-(centre of mass)-(end group) angles of ∼60°, and when the ditopic BB were linear. By contrast, the octahedral shape formed when the tritopic BBs were planar, with (end group)-(centre of mass)-(end group) angles of ∼120°, and the ditopic BB bent with an angle of ∼60°. Nevertheless, these geometric requirements are not essential to obtain a specific geometrical shape, as compensation in each of the BBs can still give the desired results. The transformation of a tetrahedral shape can happen through degrees of truncation of the tetrahedron's vertices. The trans-

formation is shown in Figure 4.5, which illustrates the truncation alongside a series of hypothetical organic cages with $\mathbf{Tri}^4\mathbf{Di}^6$ topology. The changing geometries of the BBs show one possible way in which this range of shapes can be accessed. Starting from an initial molecule with both the shape and the topology of a tetrahedron (on the left hand side of Figure 4.5), one can cause "flattening" of the corners of the molecule by swapping the starting tritopic BB with a planar tritopic precursor. Increasing the size of the planar tritopic BBs can lead to geometric shapes with higher degrees of truncation. Finally, increasing the "bending" of the linear ditopic precursor can lead to an octahedral geometrical shape. The key point is that all the molecules shown in the figure have different geometrical shapes, but only one $\mathbf{Tri}^4\mathbf{Di}^6$ topology.

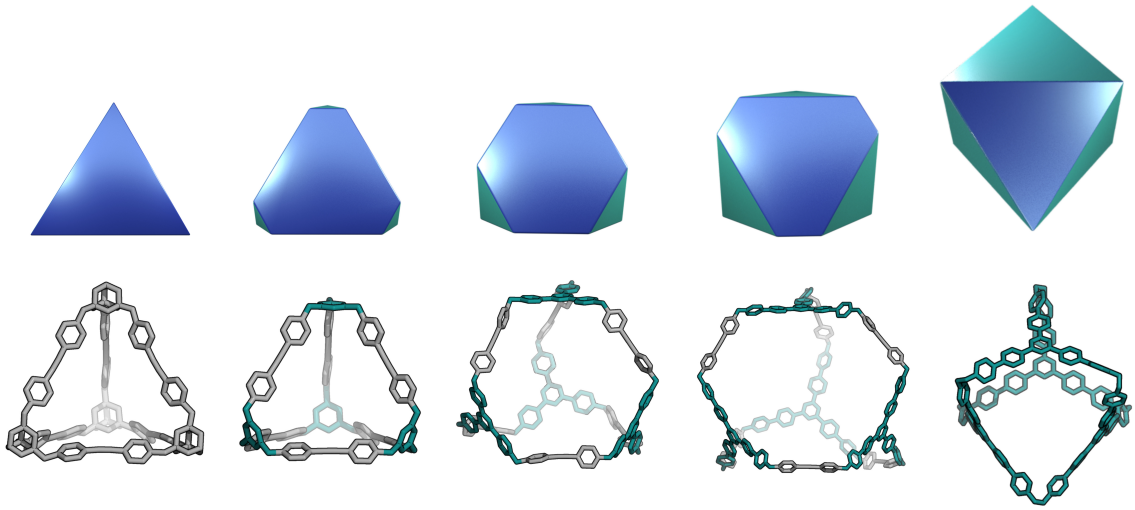


Figure 4.5.: Top row: The range of geometrical shapes potentially accessible with a $\mathbf{Tri}^4\mathbf{Di}^6$ topology, from tetrahedron (left) to octahedron (right). The evolution of the shape is expressed through successive truncations of the vertices of the tetrahedron. The octahedron has been rotated into the page relative to the other shapes to give a clearer perspective. Bottom row: stages of truncation reproduced in hypothetical porous organic cages. The initial cage with a tetrahedral shape is shown in grey, precursors that lead to increasing truncation highlighted in teal and the final (completely truncated) cage with an octahedral shape is in teal. The cage molecules are not shown to scale.

A [4+6] reaction can also form the $\mathbf{Tri}_2^4\mathbf{Di}^6$ topology, which contains two doubly connected tritopic BBs as well as two singularly connected BBs. The double connections form four windows of two distinct window sizes, two rings of size 4 and two rings of size 8. This topology has not been reported in the literature for DCC

reactions, but a molecule with this topology was synthesised in 2014 by Wang *et al.* via a one-component alkene metathesis reaction.[212] Therefore, it should be possible to achieve this topology through two-component DCC reactions.

The **Tri**^{6}**Di**^{9} topology can be obtained with a [6+9] reaction where BBs are substituted on the vertices and edges of a triangular prism, via an edge-directed approach. The final topology has five potentially accessible windows, two with ring size 6 and three with ring size 8. To my knowledge, this topology has not been synthetically reported yet. It requires tritopic BBs that have two (end group)-(centre of mass)-(end group) angles of 90° and one of 60°, to form both triangular and square connectivity rings, or BBs that can compensate for the differing angles required. It could be easier to access this topology with lower symmetry BBs. A hypothetical molecule with this topology was computationally investigated, and was described in Chapter 3.

An [8+12] reaction can lead to a **Tri**^{8}**Di**^{12} topology, which is related to a cube via an edge-directed assembly and has 6 potential access windows, all with ring size 8. Multiple examples of this structure were available in the literature.[86,100,215] Although most of them seemed to lack an internal cavity, Zhang and co-workers reported a shape-persistent catenated version, **c-Tri**^{8}**Di**^{12}.[100] Different shapes can be associated with the **Tri**^{8}**Di**^{12} topology, ranging from a cube (bottom right hand side of Figure 4.6) to the vertex-folded topology shown on the top right hand side of Figure 4.6. To obtain the cubic shape, the tritopic BBs adopted a conformation resembling a capping vertex, with (end group)-(centre of mass)-(end group) angles of ∼90°, whereas the ditopic BB were linear. By contrast, the vertex-folded shape formed when the tritopic BB were more planar with (end group)-(centre of mass)-(end group) angles of ∼120°, and the ditopic BB bent.

Finally, a larger topology is suggested for the **TriDi** family, a **Tri**^{20}**Di**^{30} molecule from a [20+30] reaction. The topology relates to a dodecahedron via an edge-directed assembly and has twelve potential access windows, all of ring size 10. No organic cage with this topology has been reported yet, although an example of a dodecahedral MOP was synthesised by Stang and co-workers.[103] This topology remains a challenging synthetic target, with potential experimental issues involving the solubility of the intermediates and the need to find a driving force that leads to a dodecahedron, rather than more stable, smaller, topologies or a polymeric product. Furthermore, one needs to take into account that in the vast majority of cases, even if the dodecahedron forms and the precursors are not particularly flexible, the large size of the structure and the number of possible rotations may lead to collapse and

subsequent loss of porosity. The most obvious way to design this topology is to have an (end group)-(centre of mass)-(end group) angle of ∼108° in the tritopic BB to match that of a dodecahedron's vertices, and to combine this with a linear ditopic BB. The required angle of ∼108° is close to that of a tetrahedral carbon, thus trifunctionalised methane molecules, such as used by Olenyuk *et al.* for synthesis of the dodecahedral MOP,[103] can be used for organic cages too.

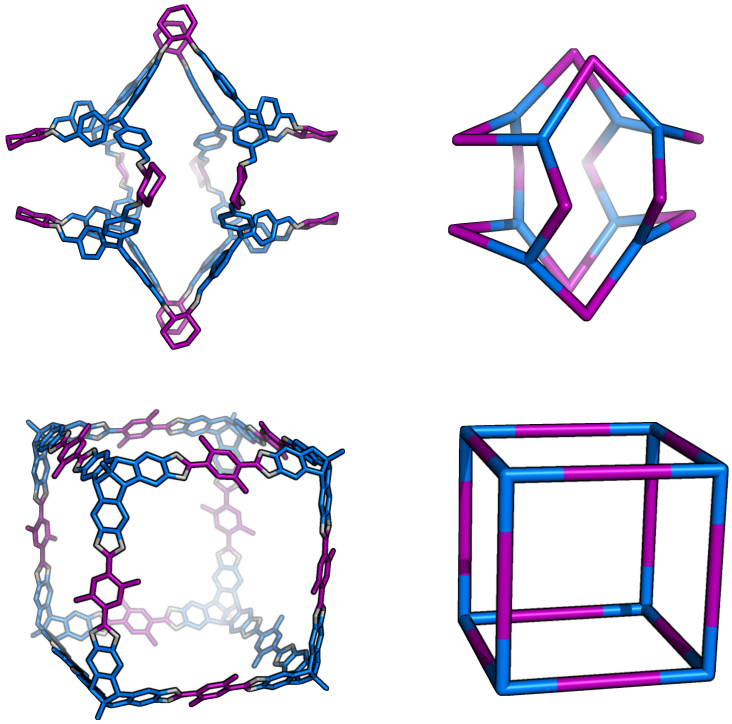


Figure 4.6.: The range of shapes possible for **Tri**^{8}**Di**^{12} topology cages, from a vertex-folded shape[215] (top), to a cube[86] (bottom).

4.3.2. Tetratopic + Ditopic topology family

A different kind of organic cage can be obtained by mixing tetratopic and ditopic BBs in a 2:4 ratio, leading to the formation of **TetDi**. As in the case of **TriDi** molecules, the smallest topology is a **Tet**^{2}**Di**^{4} capsule, formed from a [2+4] reaction of two tetratopic BBs linked by four ditopic BBs. This topology is shown in Figure 4.7, where all the topologies of this family are reported. The capsule has four windows of ring size 4. It has been reported in the literature multiple times (see Table 4.2 for references) and if the BBs involved are fairly rigid, it is often shape-persistent.

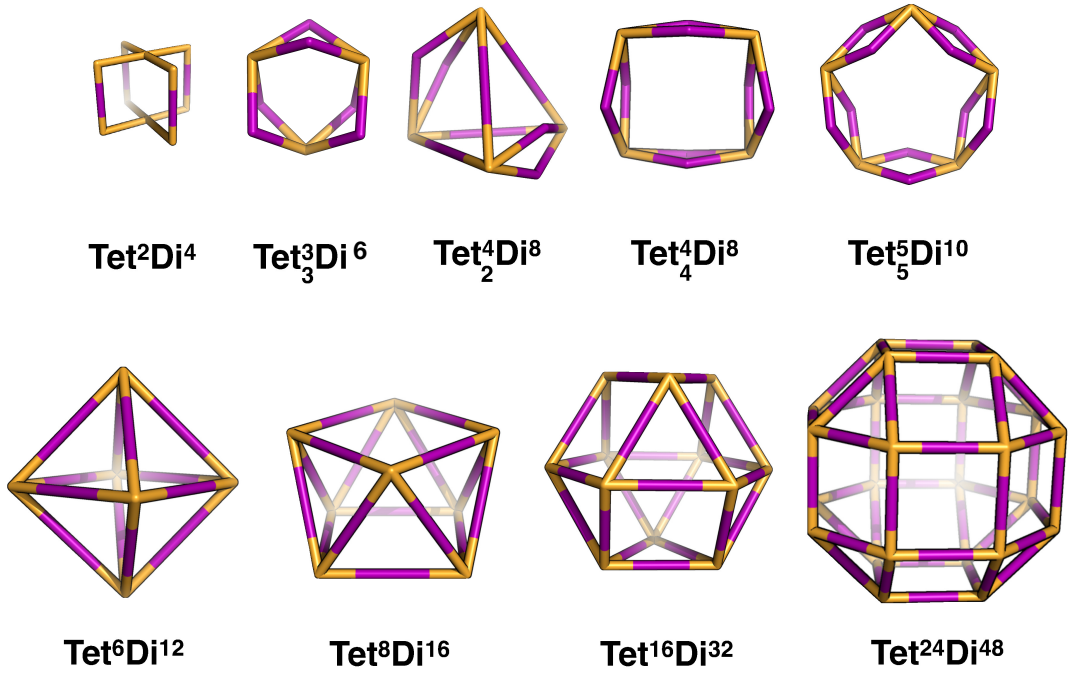


Figure 4.7.: The Tetratopic + Ditopic topology family. Tetratopic vertices are in yellow, ditopic linkers in purple.

A [3+6] reaction forms cages with a $\mathbf{Tet}_3^3\mathbf{Di}^6$ topology. This kind of cage includes three sets of doubly connected tetratopic BBs, which lead to the presence of five windows of two distinct window sizes, three rings of size 4 and two rings of size 6. A cage with this underlying topology was synthesised using a bifunctional ditopic ligand with a cyclisation reaction.[208]

The $\mathbf{Tet}_2^4\mathbf{Di}^8$ topology, which resembles a distorted tetrahedron, can be obtained with a [4+8] reaction and contains two double connections, such that each tetratopic BB is part of one double-link and two single-links. The double connections form two distinct window sizes, two rings of size 4 and four rings of size 6. Warmuth and co-workers reported an example of this cage, which was synthesised by mixing tetratopic cavitands with ethane diamines in a solvent of tetrahydrofuran. Although the structure was characterised with mass spectroscopy, no crystal structure was reported.[99]

The $\mathbf{Tet}_4^4\mathbf{Di}^8$ topology is also formed from a [4+8] reaction, but in this case there are four double connections, such that each tetratopic BB is doubly connected to its two neighbours. This topology also has two distinct window sizes, four windows of ring size 4 and two of ring size 8. This structure too was reported by Warmuth and co-workers and again it involved mixing tetratopic cavitands wih ethane diamines;

however the mix give a folded structure lacking an internal void.[80]. A [5+10] reaction forms a $\mathbf{Tet}_5^5\mathbf{Di}^{10}$ topology, which has all building blocks doubly connected to each of their neighbours. It creates two distinct windows, 5 with ring size 4 and two with ring size 10. This topology is not yet available in the literature, and it is highly likely to collapse (as the $\mathbf{Tet}_4^4\mathbf{Di}^8$ might do). This is the reason for which larger topologies with this kind of multiple connectivities were excluded.

The $\mathbf{Tet}^6\mathbf{Di}^{12}$ topology is the first in the **TetDi** family to relate to a polyhedron. The topology comes from a [6+12] reaction and it leads to an octahedron via an edge-directed assembly that places tetratopic BBs on the vertices and ditopic BBs on the edges. The structure has eight windows of ring size 6. Although no crystal structure is yet available, a molecule with this underlying topology was reported again by Warmuth and co-workers, from the same BBs as the $\mathbf{Tet}_2^4\mathbf{Di}^8$ topology, but using chloroform as a solvent, rather than tetrahydrofuran.[99] Figure 4.8 shows the possible shapes that molecules with this topology could adopt, ranging from an octahedron to a cube. The octahedral shape can be obtained by using tetratopic BBs in a capping vertex conformation (with (end group)-(centre of mass)-(end group) angles of ∼60°), and linear ditopic BBs. For the cubic shape to form, the tetratopic BBs need to be more planar, with (end group)-(centre of mass)-(end group) angles of ∼90° and the ditopic BB to bend to angles of ∼90°.

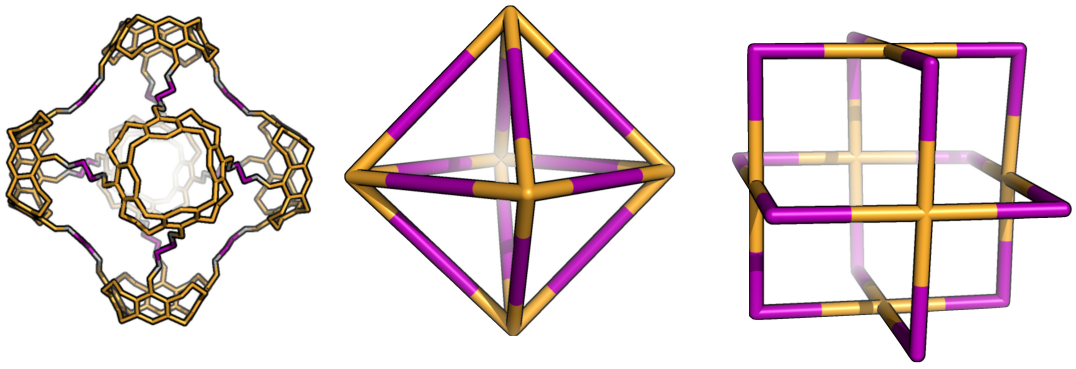


Figure 4.8.: The range of shapes possible for a $\mathbf{Tet}^6\mathbf{Di}^{12}$ topology cage,[99] from an octahedron (middle) to a cube with tetratopic linkers on the faces (right). The molecule reported on the left is an example of octahedral shaped cage.

The $\mathbf{Tet}^8\mathbf{Di}^{16}$ topology is formed from a [8+16] reaction and can be related to a square antiprism through an edge-directed assembly of tetratopic BBs on vertices and ditopic BBs on edges. This topology has eight windows of ring size 6 and two larger windows of size 8. No crystal structures are available yet for this topology,

however it was reported by Warmuth and co-workers for the same building blocks as the $\mathbf{Tet}_4^4\mathbf{Di}^8$ and $\mathbf{Tet}^6\mathbf{Di}^{12}$ topologies, but using a dichloromethane as the solvent.[99] Molecular mechanics simulations suggested that the molecule was able to preserve a square antiprism shape like the one shown in Figure 4.9.[99]

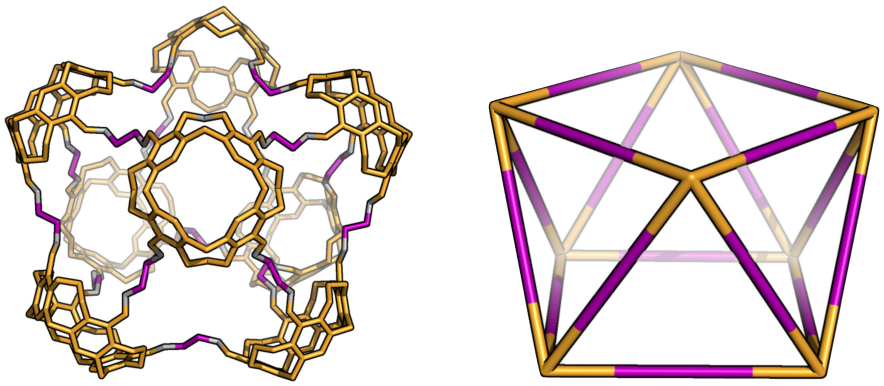


Figure 4.9.: Example of a $\mathbf{Tet}^8\mathbf{Di}^{16}$ topology cage reported by Warmuth and co-workers (left) with the shape of a square antiprism (right).[99]

A [16+32] reaction leads to the formation of the $\mathbf{Tet}^{16}\mathbf{Di}^{32}$ topology, which relates to a cuboctahedron via an edge-directed assembly of tetratopic BBs on the vertices and ditopic BBs on the edges. There are two possible window sizes, eight of size 6 and six of size 8. This topology has not yet been synthetically reported. Finally, the largest topology for this family is the $\mathbf{Tet}^{24}\mathbf{Di}^{48}$ topology, which is obtained from a [24+48] reaction and that can be related via an edge-directed assembly to the Archimedean solid rhombicuboctahedron. It is characterised by two window sizes, eight of size 6 and eighteen of size 8. This topology also has not yet been synthesised, however Fujita and co-workers reported the synthesis of a MOP with this topology using square planar Pd^{2+} as the tetratopic vertex and a dipyridylfuran ligand with an angle 127° as the edge,[97] but, regarding porous organic molecules, this structure remains an alluring and challenging synthetic target.

4.3.3. Tritopic + Tritopic topology family

This family of topologies involves the use of BBs with the same tritopic coordination, which can be combined in a 1:1 ratio to form organic cages. The smallest topology consists of an asymmetric capsule with $\mathbf{Tri}^1\mathbf{Tri}^1$ topology, coming from a [1+1] reaction, where the two tritopic BBs are connected directly to one another. This

forms three windows of ring size 2. This topology is shown in Figure 4.10, where all the topologies of the **TriTri** family are reported. Organic cages with this topology may be shape persistent, provided the starting precursors have a concave shape. A $\mathbf{Tri}^1\mathbf{Tri}^1$ was reported by Kataoka *et al.* via a condensation reaction of a bowl-shaped triboronic acid and a bowl-shaped tri-diol molecule. The resultant molecule was able to encapsulate small guests.[209]

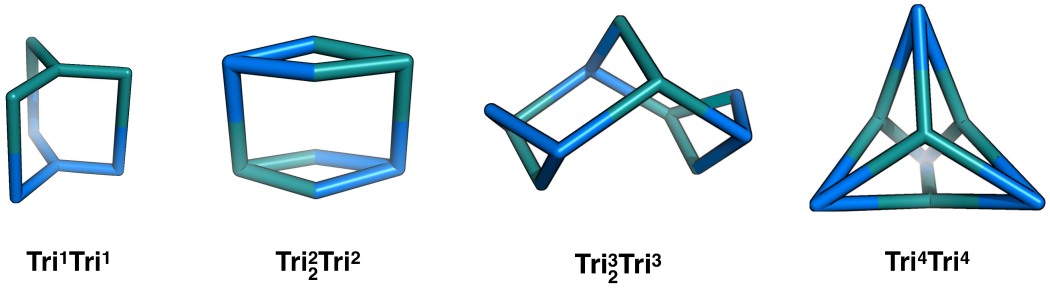


Figure 4.10.: The Tritopic + Tritopic topology family. One of the tritopic precursors is in blue, the other in teal.

Mixing tritopic precursors in a [2+2] reaction leads to the $\mathbf{Tri}_2^2\mathbf{Tri}^2$ topology, where each BB has one set of double connections and one single connection to the opposite building block. All of the BBs have the same connectivity environment and the final molecule presents two windows of ring size 2 and two of ring size 4. This molecule has not yet been reported in the literature. The $\mathbf{Tri}_2^3\mathbf{Tri}^3$ topology is formed from a [3+3] reaction, with two pairs of BBs having double connections and the others having single connections, resulting in two windows of size 2 and four of size 4. Once again, there are no synthetic reports of this molecule.

The $\mathbf{Tri}^4\mathbf{Tri}^4$ topology is the first of this family relating to a Platonic solid. It is formed from a [4+4] reaction and has the underlying topology of a tetrahedron, which is obtained via a face-directed assembly, where one kind of tritopic BB is placed on the vertices and the other on the faces of the solid. When applied to the synthesis of MOPs, this kind of face-directed assembly is named "molecular panelling".[216] Within the topology all the BBs are individually linked, leading to the formation of six windows of ring size 4. The $\mathbf{Tri}^4\mathbf{Tri}^4$ topology was used for the synthesis of porous organic cages by both Warmuth and co-workers[76] and Cooper and co-workers.[210] Molecules with this topology could adopt different shapes ranging from cubes[76] to tetrapods,[210] as Figure 4.11 illustrates. Both tritopic BBs needed to have similar conformational geometry to form a cube, with (end group)-(centre of mass)-(end group) angles of ∼90°. In this way they assembled into the 8 corners of a

cube. On the other hand, tetrapods were more commonly formed when one tritopic BB was planar (with angles of ∼120°) and the other had more narrow angles, so to be able to fit each "foot" of the tetrapod.

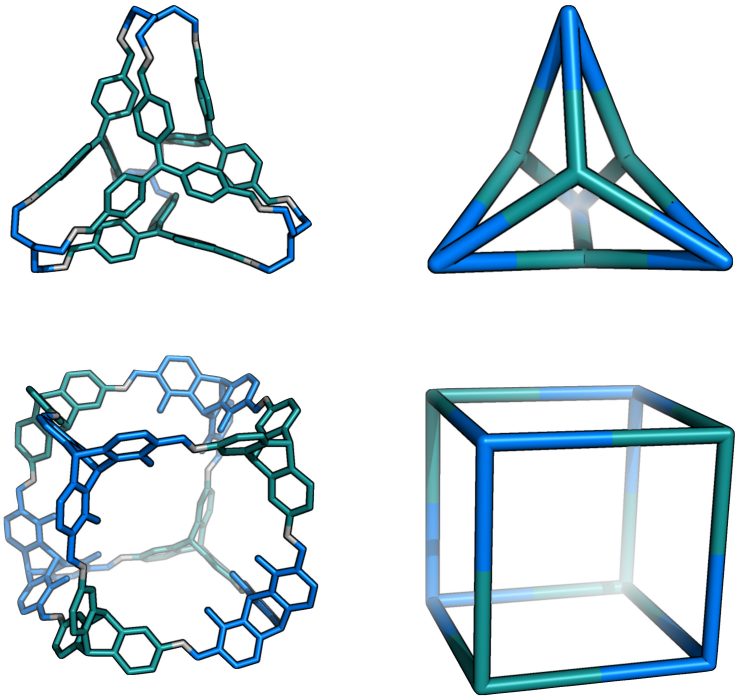


Figure 4.11.: The range of shapes possible for **Tri**^{4}**Tri**^{4} topology cages, from a tetrapod[210] (top), to a cube[76] (bottom).

4.3.4. Tetratopic + Tritopic topology family

The last family of topologies sees the combination of tetratopic and tritopic BBs in a 3:4 ratio to form molecular cages. It was possible to identify only one synthetically plausible topology for this combination, **Tet**^{6}**Tri**^{8} (see Figure 4.12). It is formed from a [6+8] reaction and it relates to a rhombic dodecahedron by a vertex-directed assembly, where both types of BB lie on the vertices of the solid. This topology has twelve windows of ring size 4. Alternative topologies generated with a smaller ratio of the BBs do not form organic cages, but rather open bowl topologies or systems that require interweaved connections, therefore they were excluded from this work.

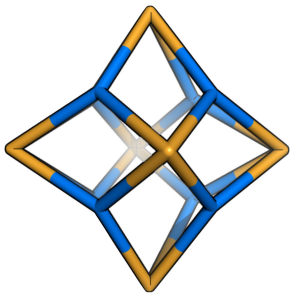


Figure 4.12.: The **Tet**^{6}**Tri**^{8} topology for the Tetratopic + Tritopic family. Tetratopic precursors are in yellow, tritopic in blue.

Cages with a **Tet**^{6}**Tri**^{8} topology were obtained by Warmuth and co-workers by mixing a tetratopic cavitand and a triphenylamine[92] (although no crystal structure was reported), and by Hong *et al.* by mixing porphyrin building blocks and triamines.[211] For the latter case, crystal structures are available, and the final structure displayed selectivity for small gases. This topology can adopt a wide range of shapes, including rhombic dodecahedra,[92] cubes,[211] and octahedra (see Figure 4.13).

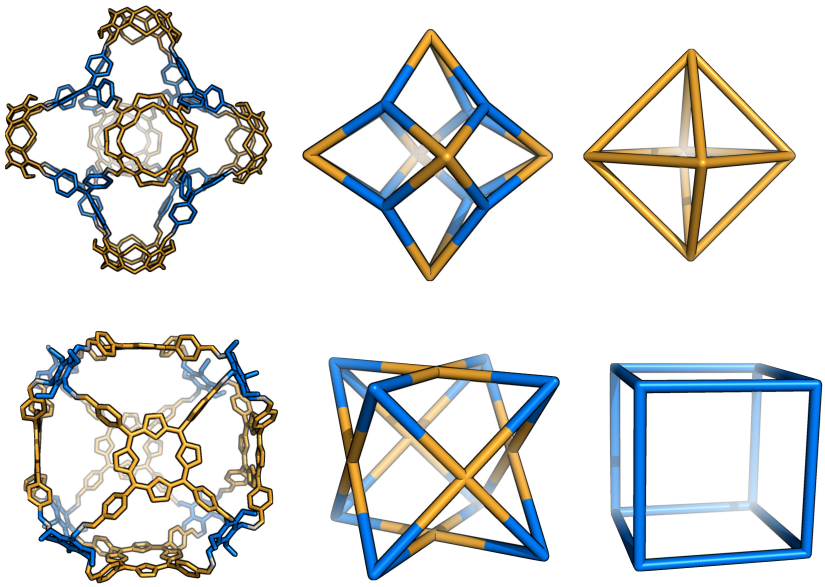


Figure 4.13.: The range of shapes possible for a **Tet**^{6}**Tri**^{8} topology cage. Top: the **Tet**^{6}**Tri**^{8} cage[92] can adopt the shape of a rhombic dodecahedron (middle) when considering both tetratopic and tritopic precursors as vertices of a solid. It has the shape of an octahedron (right) when only the tetratopic vertices are linked together. Bottom: the **Tet**^{6}**Tri**^{8} cage[211] can be seen as a rhombic dodecahedron (middle, as before). It adopts the shape of a cube (right) when only the tritopic precursors are linked together.

The rhombic dodecahedron shape is more likely to occur when the tritopic and tetratropic BBs have (end group)-(centre of mass)-(end group) angles of ∼60° and ∼90°, respectively. On the other side, for a cubic shape to form, the tetratopic BB should preferably be planar, with angles of ∼90° and the tritopic BB shaped like a cube corner, with angles of ∼60°.

4.4. Conclusions

This chapter defined a new nomenclature for 20 topologies that can be accessed via two-component DCC syntheses of organic cage molecules. The topologies were classified into four families, depending on the number of reactive groups possessed by the starting BBs. Of the topologies analysed, 12 have been reported in the literature, the most common being capsule-like (**Tri**^{2}**Di**^{3} and **Tet**^{2}**Di**^{4}), and tetrahedral (**Tri**^{4}**Di**^{6}) topologies. More exotic molecules have underlying topologies relating to cubes (**Tri**^{8}**Di**^{12}) and rhombic dodecahedra (**Tet**^{6}**Tri**^{8}). The larger topologies related to dodecahedra (**Tri**^{20}**Di**^{30}), cuboctahedra (**Tet**^{16}**Di**^{32}) and rhombicuboctahedra (**Tet**^{24}**Di**^{48}), remain more challenging synthetic targets. Even if the main obstacles in the synthesis are given by BB behaviours, solubility and shape persistence, the fact that such topologies could be realised for MOPs is encouraging. The main goal of this work was to provide a description of the four topological families, including their underlying symmetry, window sizes and relation to different shapes. The main idea was to clarify the ambivalence between geometrical shape and topology, whilst offering a new classification system for organic cages. This nomenclature system has the potential to be expanded and adapted beyond the topologies described and it could help removing confusion in the literature.

In Chapter 5 I am going to show the importance of computer simulation for the full prediction of structure, shape and shape-persistency. I am going to examine specific case studies and investigate to what extent computer simulations can assist in predicting the topology from knowledge of the BBs alone.

Chapter 5
Topological prediction of porous organic cages.

The work in this chapter led to the publication of:
Santolini V., Miklitz M., Berardo E. and Jelfs K.E., *Nanoscale*, 2017, 9, 5280-5298.



5.1. Introduction

In this chapter I will expand the discussion from Chapter 4 and describe how I used computational modelling to predict which topology is most likely to form among sets

of molecules reported in the literature, belonging to each topological family. The prediction is based on a computational pipeline that assembles and characterises organic cages and of which a schematic representation is shown in Figure 5.2. In the next few sections I will summarise each step of the pipeline, some of the methods involved in the characterisation of organic cages, and finally discuss the results obtained for the topology prediction of porous organic cages, to discuss in which cases the method was successful and in which it was not.

5.2. The prediction pipeline.

The prediction procedure in Figure 5.2 reproduces the process that led to the topological prediction of a **Tri**^{4}**Di**^{6} imine molecule from the **TriDi** topological family. It begins with step **A**, where a pair of precursors (ditopic and tritopic in this example) is selected and placed onto the vertices and edges (could be faces as well for other topology families) of a selection of polyhedra (**B**), as shown in Figure 5.1.* In order to generate the molecular structure, the pair of precursor molecules were generated based on their SMILES code and fed into RDKit software to generate precursor geometries, which are then optimised with the Universal Forcefield (UFF).[163] The precursors are then placed on the vertices and edges of the relevant polyhedra, the reactive end groups of the precursors are linked and the chemistry corrected so that the aldehyde and amine BBs form imine bonds. The assembled geometries are then minimised with the OPLS3 forcefield,[160] which, as previously discussed in the Methods chapter, was found to be an improvement of the previously used OPLS_2005.[98,215] In step **C**, conformer search methods are used to explore the PES around the initial guess to locate the lowest energy conformers of each molecule. If the lowest energy conformer is shape persistent ($\mathbf{C}_a$) the procedure moves on to the next step. By contrast, if the molecule is not shape persistent ($\mathbf{C}_b$), constrained MD can be used to locate higher energy open conformers,[215] which can then be analysed in step **D**. Only the geometries and energies of open conformations were compared, as these are likely to reproduce the structures that initially form in the reaction solution. In this latter step **D**, each open geometry is optimised with DFT methods to obtain a more accurate energetic ranking of the conformations.

*The code to assemble precursors into molecules was written in Python by Dr Enrico Berardo.

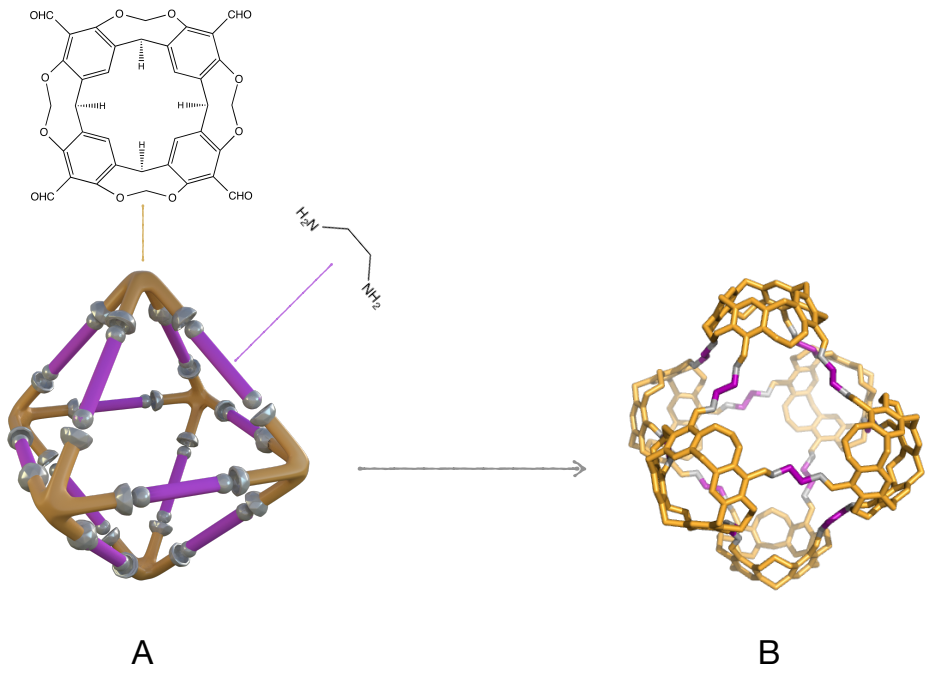


Figure 5.1.: Assembly of an octahedral molecule with **Tet**^{6}**Di**^{12} topology, starting from the tetraformyl cavitand and ethylene-1,2-diamine precursors. **A** is the underlying topology, **B** is the final assembled molecule.

In step **E**, relative internal energies per formula unit of all the optimised open conformers are compared. The use of DCC reactions allows one to assume that the molecule with lowest relative energy should be the one that evolves in greater concentration in term, and should therefore be the dominant product of the reaction.[74] When the comparison of relative energies is not possible due to molecules containing different species, formation energies per bond are calculated instead, with the assumption that the molecule with lowest formation energy is the most likely to form. Finally, if a crystal structure is available in the literature, an overlay between simulations and experiments is made in step **F**, and the RMSD value between the two structures is calculated to test the accuracy of the calculations.

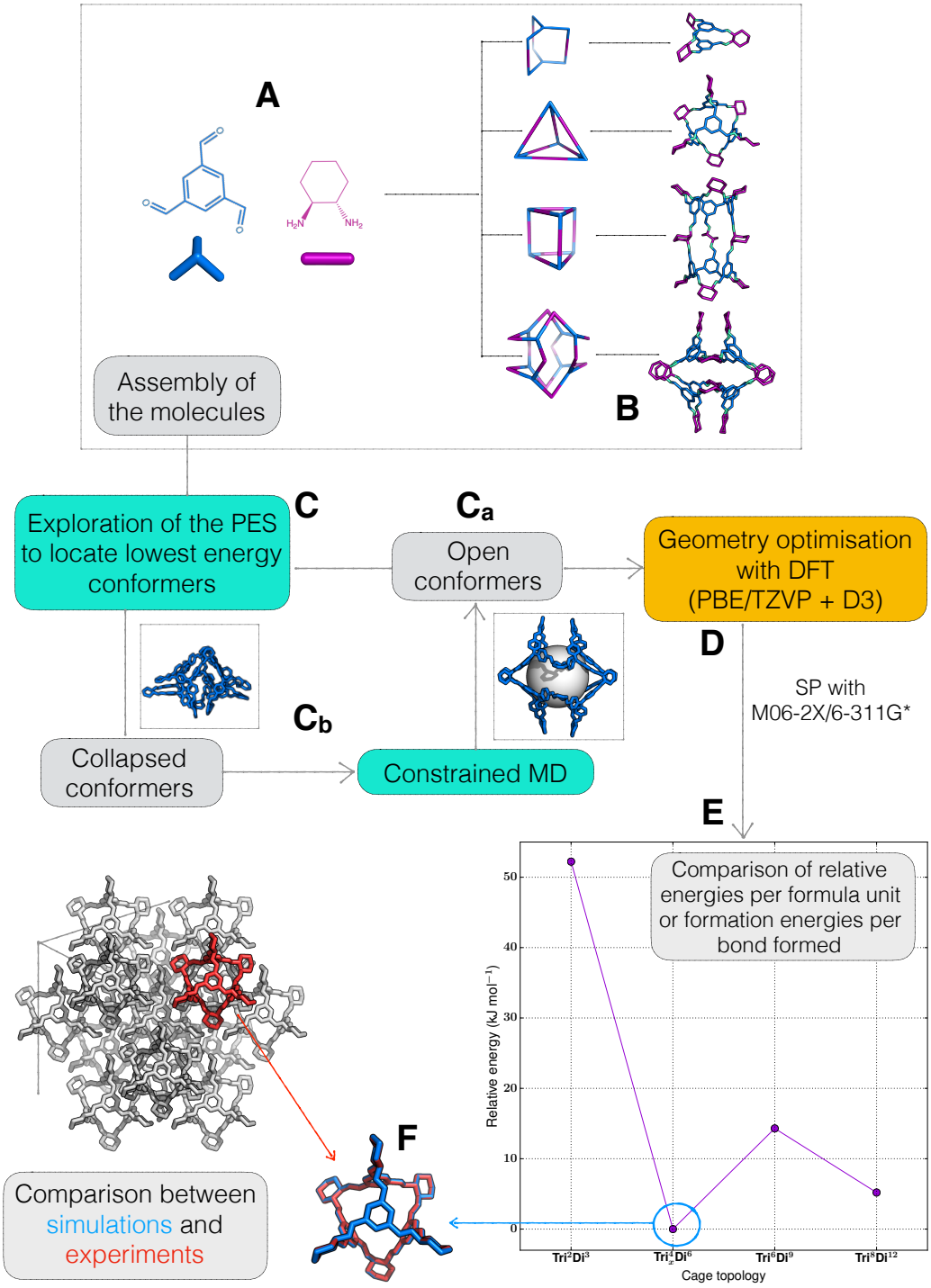


Figure 5.2.: Pipeline of the prediction approach.**A**: two precursors are chosen, in this example a ditopic (purple) and a tritopic (blue); **B**, precursors are assembled into molecular cages, according to a selection of possible topologies; **C**, Conformer search calculations are performed to locate the lowest energy conformer; if open ($\mathbf{C}_a$) the analysis proceeds to step **D**, if collapsed ($\mathbf{C}_b$) an open higher energy conformer is located with constrained MD; **D**: all the open conformers are analysed at the PBE and M06-2X level of theory; **E**: the relative internal energies per formula unit of all the cages, or their formation energies per bond, are compared to locate the lowest in energy, and therefore the topology which is most likely to form experimentally. **F**: if a crystal structure is available, the simulation and the crystal structure are overlaid

5.3. Methods involved in the prediction pipeline.

In the next few paragraphs I will discuss the principal techniques involved in this work and the reason why some methods were preferred over others. The simulation details adopted in this chapter are described for both conformer search methods and DFT calculations. Constrained MD is not described here, as it was already covered in Chapter 3.

5.3.1. Methods benchmarking

Sampling the PES with the low mode search

All the conformer search calculations were performed with the MacroModel package (version 10.3.015, Schrödinger, LLC, New York, NY, 2015-3) and the OPLS-AA force field,[160] which has been previously used for the structural and energetic description of organic cage molecules.[98,215] The PES was sampled with the low mode search method, which follows the low frequency eigenvectors of the molecule, as described in the Methods chapter. The quality of the sampling depends on the number of steps included in the calculation settings, on the maximum energy window allowed for sampling (i.e. difference in energy between the lowest and the highest energy conformers found), and the maximum/minimum movement the trajectory can undertake. Generally, the PES is considered sampled when increasing the number of steps for the chosen energy window leaves the number of unique conformers found unchanged. In order to check whether the PES was fairly well sampled, I wrote an analysis script in Python, which analysed the output file of the calculation and plotted the number of times each conformer was found (the higher the number, the better the sampling), and the number of lowest energy conformers found at each step of the calculation. Figure 5.3 shows the data obtained for an example case study on cage **CC1**, which was investigated with a simulation of 50,000 steps sampling an energy window of 30 kJ mol^{-1}, with a minimum move of 3 Å, and a maximum move of 10 Å. The simulation was then repeated with 100,000 steps. The former simulation identified 26 conformers within the chosen energy window (13 conformers plus their mirror images), although the sampling was quite scarce in some cases, as the crimson bar plot in Figure 5.3 shows. In my calculations, I have considered those cases in which each conformer was found more than a hundred times well sampled.

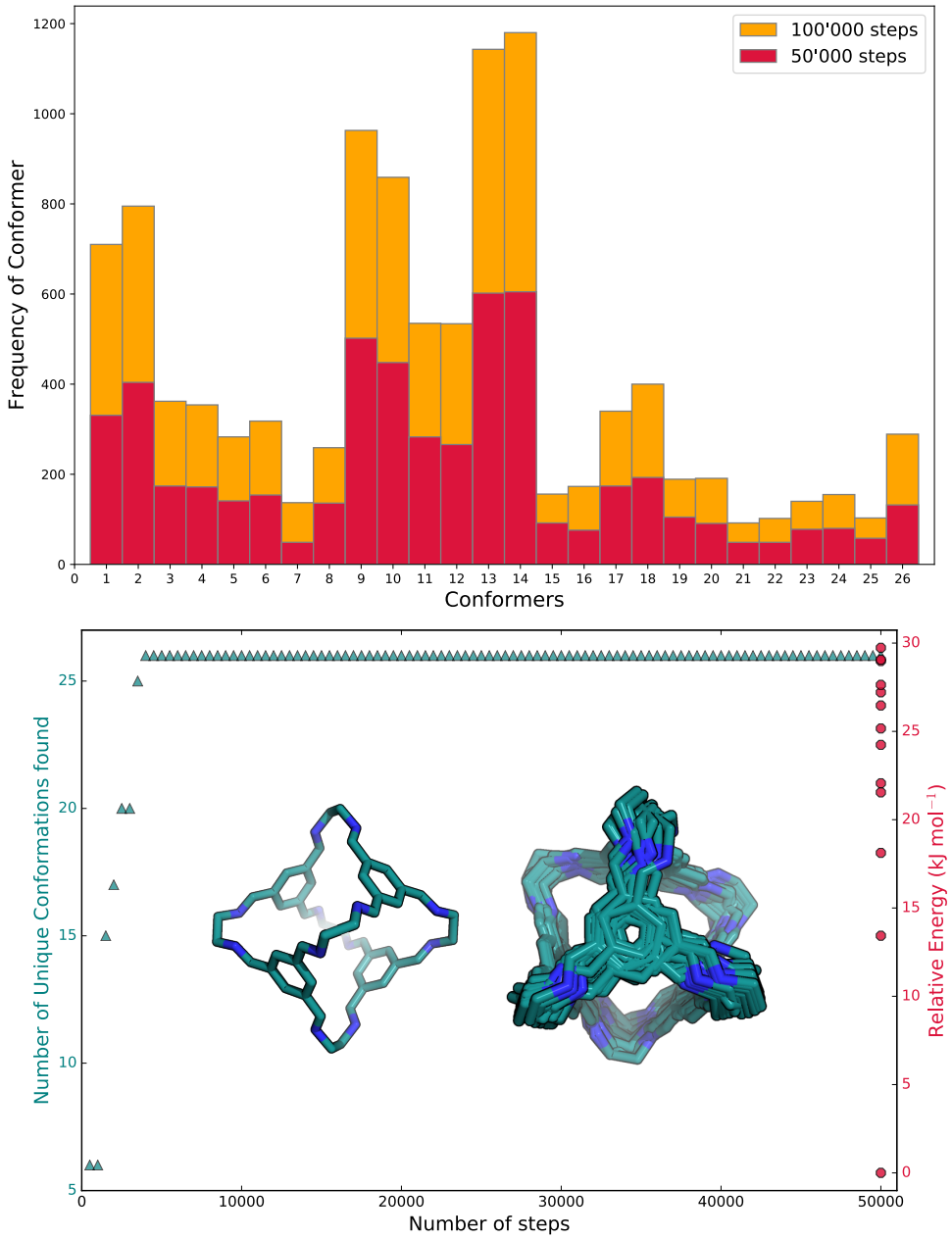


Figure 5.3.: Conformer search analysis for cage **CC1**. Top: bar plot calculating the number of times each conformer was found in a simulation of 50000 steps (crimson) and in one of 100000 steps (orange); bottom: plot showing the number of unique conformations found per number of steps (teal, left Y axis) and the relative energies of the final collection of conformers (red, right Y axis). In the middle of the plot, the geometries of all the 26 conformers are shown as an overlay on the right and the lowest energy conformer is shown (at a different angle) on the left.

Doubling the number of steps in the simulation led to better sampling of the surface, as each conformer was found hundreds of times, as it can be seen from

the orange bar plot in Figure 5.3. The number of unique conformations found was unvaried, a sign that no more conformations were available in that energy range. The plot at the bottom of Figure 5.3 shows that the number of conformers found reached a plateau quite early on in the simulation, a sign that no further sampling was needed (as the simulation run with 100,000 steps then confirmed). The second Y axes in the plot shows the energy of the conformers identified within 30 kJ mol^{-1} from the lowest energy one and of which the geometries are overlaid on the right hand side of the same plot. The lowest energy conformer is more stable than all the others by 13.4 kJ mol^{-1} and it is shown on the left hand side of the same plot. As a general rule of thumb, larger molecules require a larger number of steps to reach a plateau, as the number of bond stretches, angle bending and torsions to analyse is larger. Calculations on larger molecules and involving a large number of steps are considerably more computationally expensive. If a large number of simulations has to be performed and the number of atoms in the molecule is larger than 100, high temperature (500-1000 K) molecular dynamics simulations were generally preferred, as they allowed sampling of the surface effectively, and in less time than the low mode search.

Evaluation of the most suitable DFT functional for the description of relative and formation energy.

At the beginning of my PhD, I compared the performance of six different DFT functionals (and the HF wave function method) to find the functional and the level of theory that can most accurately reproduce the energy and relative stability of organic imino cages and that had the best computational cost/accuracy balance, given the size of the systems investigated. The methods were applied to study a fragment combining 1,3,5-triformylbenzene and 1,2-ethanediamine via imine condensation reaction, as shown within the bar plot in Figure 5.4. The formation energy for the fragment (on which cage **CC1** is based) was calculated, according to Equation 2.37 and with different methods. Geometry optimisations were performed on each reactant and product with the HF wave function method, the BLYP, B3LYP,[217] M06,[175] M06-2X,[175] PBE[174] functionals and MP2[218] approach. All methods were combined with triple zeta basis sets in the Gaussian quantum mechanics package,[219] the basis set being 6-311G** for all cases except for PBE, which was combined with the TZVP-MOLOPT[177] basis set. The convergence criteria were the default for Gaussian09, 10^{-8} a.u. $Å^{-1}$ for the self consistency field cycle, and 10^{-5} Å for the convergence of the geometry. The effect of including dispersion contributions to the energy was

studied by adding the DFT-D3 (D3)[176] correction to the BLYP, B3LYP and PBE functionals, whereas the long-range London interactions are intrinsically described by the M06, M06-2X and MP2 functionals.[220] The formation energies obtained are shown in the bar plot in Figure 5.4.

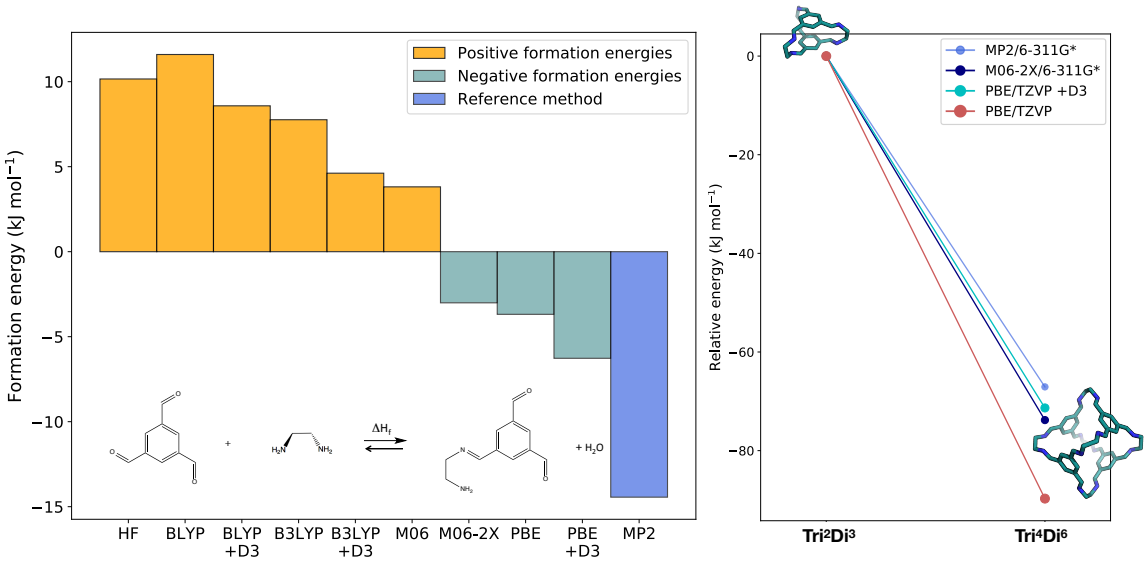


Figure 5.4.: Left: bar plot showing the formation energies per bond of the imine condensation between [1:1] triformyl-benzene and ethanediamine, calculated at different levels of theory. Right: relative formation energy per [2+3] formula unit calculated for the $\mathbf{Tri^2Di^3}$ and $\mathbf{Tri^4Di^6}$ topologies of cage **CC1** (shown in the plot, carbon is in teal, nitrogen in blue). All energies are calculated with respect to the $\mathbf{Tri^2Di^3}$ topology.

The MP2 method, although very accurate, can not be used to perform routine calculations on organic cages, as it would be computationally too expensive (the time required for MP2 calculations was one order of magnitude greater than that required for all the other methods). It was nonetheless used to calculate the formation energy of the fragment, and the result was treated as a reference, as it was the one calculated at the highest level of theory. MP2 predicts the formation energy of the reaction to be negative, therefore the energy of the products to be lower than that of the reactants. As one can see from Figure 5.4, among all the methods tested, only M06-2X and PBE predicted the formation energy to be negative, M06-2X predicting it to be -3.0 kJ mol^{-1}, PBE -3.7 kJ mol^{-1}, and PBE+D3 -6.3 kJ mol^{-1}. The choice of functionals to adopt for my research was therefore narrowed down to these two functionals, and their description of the structure's geometry and of the relative energies between porous organic cages of different sizes was further analysed.

Firstly, I compared the geometries optimised with PBE+D3 and M06-2X with

that optimised with MP2, in order to test the accuracy of the structural description. The fragments are overlaid in Figure 5.5 and, as one can see, both PBE+D3 and M06-2X match the MP2 description excellently, the RMSD being 0.05 Å between the PBE+D3 and MP2 optimised structures, and 0.03 Å between the M06-2X and MP2 optimised structures.

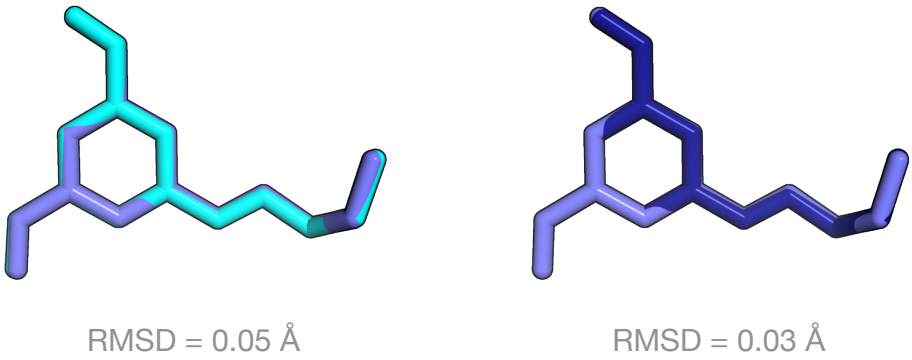


Figure 5.5.: Fragment of cage **CC1** investigated at different levels of theory. Left: overlay of the PBE+D3 (in cyan) and MP2 (in purple) optimised geometries; Right: overlay of the M06-2X (in dark blue) and MP2 (in purple) optimised geometries.

Having verified that both the M06-2X and PBE functionals describe well the geometries of the fragments investigated, I then had to verify whether they provided a suitable description of the relative energies per formula unit of two porous organic cages of different sizes, as the topological prediction of a reaction product in this thesis is based on this comparison. The lowest energy conformers of the $\mathbf{Tri}^2\mathbf{Di}^3$ and $\mathbf{Tri}^4\mathbf{Di}^6$ topologies of cage **CC1** were geometry optimised with both methods. Single point calculations were additionally performed on the PBE optimised geometry with the MP2 functional, so as to have a benchmark for the quality of the result. Given that the imine condensation reaction is usually under thermodynamic control and that therefore it should lead to the dominant product of a reaction, I assumed that the cage with the lowest relative energy per formula unit would have been the most likely to form. Results are shown in the plot on the right hand side of Figure 5.4. All methods predicted the cage with $\mathbf{Tri}^4\mathbf{Di}^6$ topology to be more than 60 kJ mol^{-1} per [2+3] formula unit lower in energy than the $\mathbf{Tri}^2\mathbf{Di}^3$ topology (confirming experimental results for cage **CC1**, which was found to have $\mathbf{Tri}^4\mathbf{Di}^6$ topology). Both M06-2X and PBE+D3 performed very well, as their description of the energy was not too distant from that of MP2 (respectively, 6.7 and 4.2 kJ mol^{-1} per formula unit lower than the MP2 energy). Excluding the contribution of the induced dipole

- induced dipole interactions made the energy description less accurate (22.6 kJ mol^{-1} lower than at the MP2 level), which confirmed, unsurprisingly, that the D3 correction had to be added when using the PBE functional.

Further on in my PhD, it became possible for me to run calculations in a faster and more efficient way on the UK's ARCHER supercomputer, therefore, due to the availability of the functionals on ARCHER, PBE calculations were run within the CP2K package[161] and M06-2X in the NWChem package.[221] Both methods were used to study organic cages. However, among these two, PBE was chosen over M06-2X when large numbers of calculations were involved (> 50) due to its optimal computational cost/accuracy balance and, in all calculations, it was combined with the D3 correction, as the latter contributes to the stabilisation of the molecule by lowering its energy. Both methods were already adopted in the literature for the description of porous organic cage systems.[98,164,165]

Comparison between force field and DFT accuracy

Finally, the results obtained from a conformer search simulation with the OPLS-AA force field were compared with those refined with the PBE/TZVP+D3 method. Force field methods can be used to sample the PES around the initial guess in order to locate the lowest energy conformers, however, they are not reliable for an accurate energetic description. For this reason, force fields were used to find the lowest energy geometries of the molecule investigated and DFT functionals were then applied to refine the description of the energy. This comparison was useful to understand: a) how accurate were the geometries obtained from conformer search simulations, b) how well force field methods described the energy. The conformer search simulation performed on **CC1** with 50,000 steps was re-analysed for this purpose. Results are reported in the plot in Figure 5.6. Force field calculations tend in general to underestimate the relative energy for conformers 2-13 with respect to 1, the biggest energy difference being 12.5 kJ mol^{-1} for conformer 12, the smallest being just 0.5 kJ mol^{-1} for conformer 4. The difference in energy between the two methods is 3.9 kJ mol^{-1} on average, with a standard deviation of 3.5 kJ mol^{-1}.

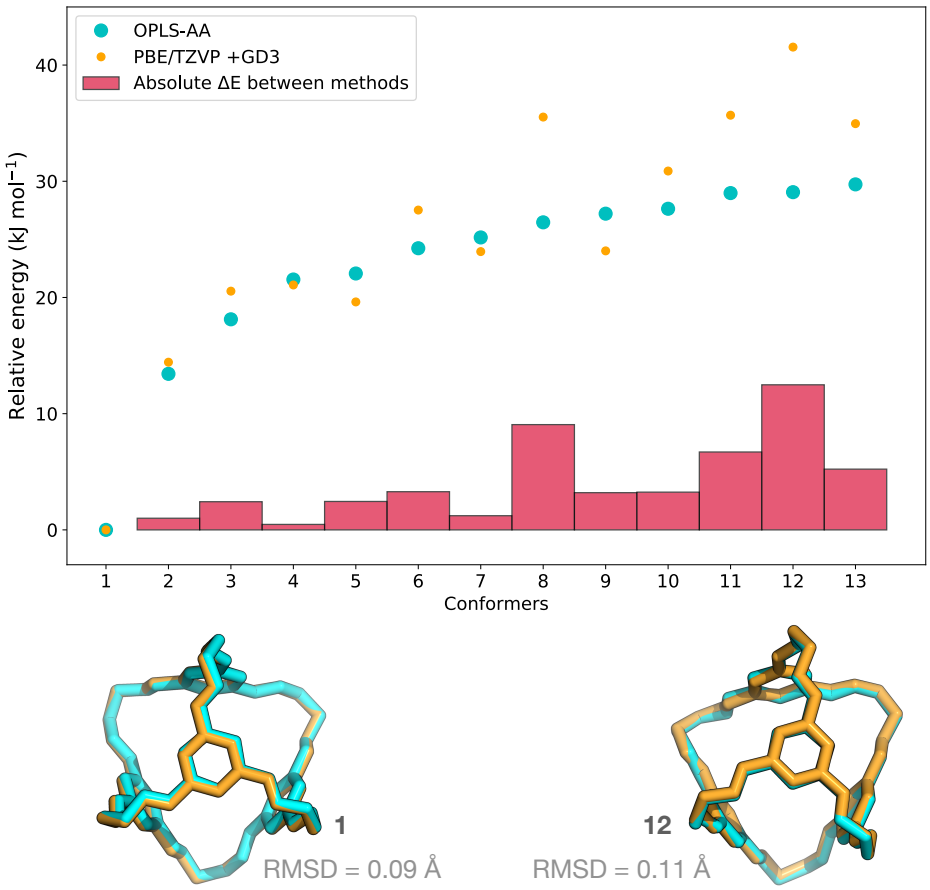


Figure 5.6.: Plot: Relative energies of the lowest energy conformers identified within a 30 kJ mol^{-1} with the OPLS-AA force field (in cyan) and the relative energies of the same conformers re-optimised at the PBE/TZVP+D3 level of theory (orange). The pink bars show the absolute energy difference between values obtained with the two methods. Bottom: overlays of the OPLS-AA (cyan) and PBE/TZVP +D3 (orange) geometries for conformers 1 (left), and 12 (right).

In the case of cage **CC1**, not many conformers are available in the selected energy window (13 excluding mirror images) and the energy gap between the lowest energy conformer and the next higher conformer is sufficiently large (13.4 kJ mol^{-1}) to consider the lowest energy conformer the preferred conformation. Nevertheless, in other cases, often when more flexible molecules are involved, conformational flexibility could increase the number of conformers found within the selected energy window and, bearing in mind that the error between the two methods can be as large as ∼12 kJ mol^{-1}, conformers found with OPLS-AA within a window of at least ∼15 kJ mol^{-1} should be refined with DFT methods. The refinement at a higher level of theory is to check whether any of the conformers that are higher in

energy at the OPLS level are then actually at a much lower energy with DFT. The geometries obtained at the OPLS level (in cyan in Figure 5.6) were found to be quite accurate when overlaid with those re-optimised at the PBE level (in orange), the difference in RMSD for the lowest energy conformer 1 (left hand side of Figure 5.6) being only 0.09 Å, and that for conformer 12 (of which the energies were described quite differently by OPLS and PBE) of only 0.11 Å. On this basis, unless otherwise specified, in this thesis a 15 kJ mol^{-1} confidence level was adopted.

5.3.2. Conformer search simulations

When the low search method was used, the number of steps (starting from a minimum of 20000) was increased until convergence was reached, *i.e.* the number of low energy conformers found reached a plateau. The energy window for saved conformers was set to 30 kJ mol^{-1} and the minimum and maximum moves as 3 Å and 10 Å, respectively. For the larger systems (all except the **TriDi** family), a high temperature Molecular Dynamics (MD) was run instead, as a more effective method to identify the lowest energy conformations. MD simulations were run with MacroModel, with the NVE ensemble, a time step of 1 fs, a total duration of 100 ns and sampling plus geometry optimising a structure every 10 ps. All calculations were carried out with the OPLS3 force field[160] and each sampled structure was minimised within the Macromodel package with the Polak-Ribier Conjugate Gradient method and a gradient convergence criteria of 0.05 kJ mol^{-1} $Å^{-1}$.This procedure was carried out at both 500 and 1000 K to ensure sampling of the PES. When the simulations reached convergence (according to the same criteria described for conformers search), the subset of conformers found was analysed, taking into account the structures that lay within an energy window of 20 kJ mol^{-1}. Duplicates were removed when the maximum atom deviation was equals or minus to 0.5 Å.

5.3.3. DFT methods

Once the lowest energy conformers were located for each molecule, DFT methods were applied to get a more accurate energetic ranking of the conformations. Firstly, the structures were DFT optimised with the PBE functional[174], the triple-zeta molecularly optimised basis sets[177], the GTH-type pseudopotential,[178] a plane wave grid cutoff of 350 Ry and the D3 dispersion correction in CP2K.[161] Secondly, single point calculations were performed on the PBE-optimised structures using the M06-2X functional[175] and 6-311G* basis set in NWChem.[221] The M06-2X functional

was successfully used in the past on porous organic cages,[164,165] and does not require additional dispersion corrections, as these are already embedded in the functional. All the energies discussed in this chapter are from the M06-2X/6-311G* calculations.

5.3.4. Comparison of the energies

Relative energies per formula unit and formation energies per bond formed are compared as discussed in the Methods chapter. The comparison of the relative internal energies of the hypothetical topologies for a given assembly of BBs gave an indication of the relative strain (or lack of it), should the BBs adopt those particular topologies. In the Introduction chapter, a large number of factors that could influence the topological outcome were outlined, some of which cannot be considered here, based on computational limitations that I will now discuss. Nonetheless, these kind of calculations remain important to uncover to what extent the relative internal energies can explain previous experimental observations.

5.3.5. Computational challenges for topology prediction.

Solvent effects.

Firstly, in terms of solvent effects, the "inflation" procedure reported in Chapter 1,[215] allows one to look at how solvent influences the conformations of cage molecules. The approach is useful to reproduce the scaffolding or "templating" effect of the solvent molecules, but does not explicitly model individual solvent interactions. To reproduce explicit solvent interactions, one would firstly have to greatly increase the number of atoms in the simulations by the necessity of including hundreds of solvent molecules.[96] Secondly, one would then need to sample the conformations of the molecules surrounding the cage to consider the dynamic nature of the solvation. Combined, these simulations are not currently tractable for these systems. This issue of solvent effect remains a big challenge in the field and efficient approaches need to be developed for the modelling of such complex systems.

Energy evaluation.

Ideally one should go beyond the relative internal energies from DFT calculations and compare free energies. This would include a consideration of the translational, rotational and vibrational entropy of the systems, which should tend to disfavour the formation of larger assemblies over smaller ones. Practically, this would require

a frequency analysis for each molecular conformation and this would be computationally intractable on a routine basis even for the smaller cage systems. Indeed, in a previous work, only fragments of a comparably small molecular cage could be considered for calculations of the free energy correction to guest binding energies.[155]

Catenanes as alternative reaction products.

Whilst not the focus of this work, one can also consider the use of simulations to predict preferential catenane formation over monomeric forms of the topologies discussed. Catenanes could be expected to form as thermodynamic products if a combination of solvent interactions, steric compatibility of the BBs involved in the reaction and the formation of stabilising interactions between the two monomeric units act as a thermodynamic driving force. The obvious starting point for simulations would be to calculate the binding energy for a catenane pair by comparing the total energy of the assembly to that of two gas space monomers. An unfavourable binding energy, for example resulting from steric clashes of the molecules in a catenated form, would suggest that the monomeric form would be preferred. However, favourable binding energies would need to be considered with caution, as these would indicate stabilisation of the catenane relative only to isolated gas phase monomers, without consideration of stabilisation of those monomers by either solvent interactions or intermolecular interactions gained via crystal packing. Thus, arguably, for a series of systems, binding energies might be best used to qualitatively rank a series of monomers in terms of likelihood of catenane preference over monomeric forms. I further note that for some systems there would be significant, and even potentially prohibitive, sampling requirements when looking for low energy catenated forms. These would arise from the requirement to consider both different mechanical interlocking arrangements and also configurational sampling for each interlocked form. Finally, reliable binding energies across a series of different systems are likely to require DFT calculations, rather than computationally cheaper FF calculations.

5.4. Results

Given all the previous considerations, I will now discuss the results I obtained for the prediction of molecules belonging to the four topology families.

5.4.1. The Tritopic + Ditopic family

I will begin the results discussion with systems belonging to the **TriDi** family. In the first subsection I will describe the five cages inspired by **CC3**, the **Tri**^{4}**Di**^{6} imine molecule synthesised by Gawronski and coworkers and then further characterised and used for applications by Cooper and coworkers.[7,75] In the second subsection, I will investigate a case based on the **CC5** and **CC8** imino cages,[72,96] that has been reported as an example of emergent behaviour, with the aim of rationalising the preferential formation of two different topologies with the additional of only a single CH2 unit to one of the precursors, with a consequential massive effect on the resultant properties.

CC3-like cages.

For the Tritopic + Ditopic family, I investigated the outcome of combining 1,3,5-triformylbenzene (**1** in Figure 5.7) and (*R*,*R*)-1,2-diaminocyclohexane (**2** in Figure 5.7), used in the synthesis of **CC3**, a **Tri**^{4}**Di**^{6} imine molecule.[7,75]

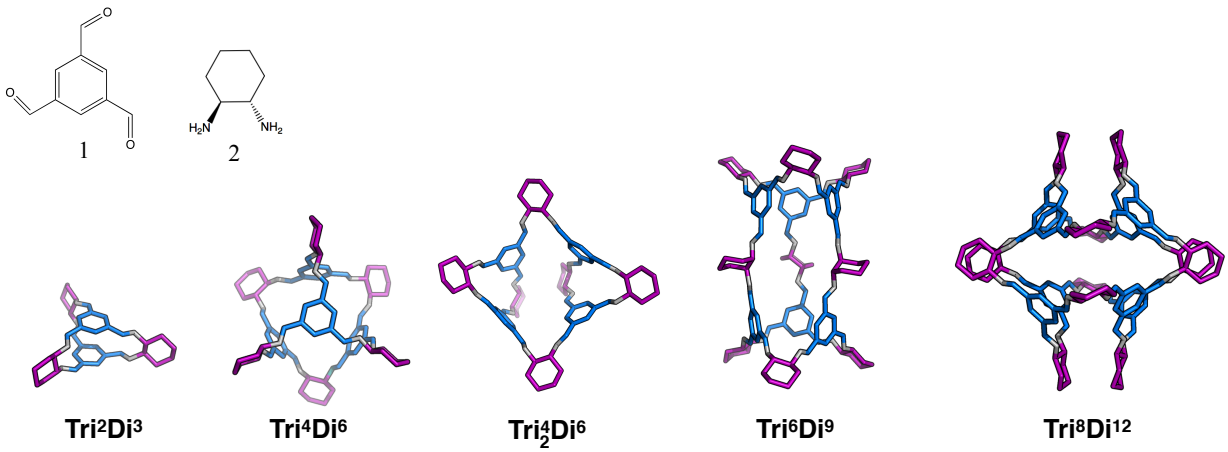


Figure 5.7.: Molecular structures of the lowest energy open conformations of the Tritopic + Ditopic topologies for the **CC3** imine cage system, formed from the reaction of 1,3,5-triformylbenzene (**1**) and (*R*,*R*)-1,2-diaminocyclohexane (**2**). Tritopic BBs are in blue, ditopic precursors in purple, the imine bonds formed in the reaction in grey.

The relative energies per [2+3] unit were compared for the lowest energy conformations of the **Tri**^{2}**Di**^{3}, **Tri**^{4}**Di**^{6}, **Tri**$_{2}^{4}$**Di**^{6}, **Tri**^{6}**Di**^{9} and **Tri**^{8}**Di**^{12} topologies. The structures of the lowest energy open conformations are shown in Figure 5.7, with the corresponding relative energies reported in Figure 5.8 and Table 5.1.

Table 5.1.: Information on the lowest energy open conformations of the **CC3**-based Tritopic + Ditopic topologies.

	Synthetically realised?	Energy relative to $\mathbf{Tri}^4\mathbf{Di}^6$ per [2+3] unit (kJ mol^{-1})	Shape persistent?	Void diameter (Å)
$\mathbf{Tri}^2\mathbf{Di}^3$	no	52	yes	1.3
$\mathbf{Tri}^4\mathbf{Di}^6$	yes	0	yes	5.5
$\mathbf{Tri}_2^4\mathbf{Di}^6$	no	39	yes	3.1
$\mathbf{Tri}^6\mathbf{Di}^9$	no	14	no	5.2
$\mathbf{Tri}^8\mathbf{Di}^{12}$	no	5	no	7.4

The $\mathbf{Tri}^2\mathbf{Di}^3$, $\mathbf{Tri}^4\mathbf{Di}^6$ and $\mathbf{Tri}_2^4\mathbf{Di}^6$ topologies showed shape persistency, whereas the larger $\mathbf{Tri}^6\mathbf{Di}^9$ and $\mathbf{Tri}^8\mathbf{Di}^{12}$ topologies displayed more flexibility and collapsed upon MD simulations, losing their internal cavities. For the latter two topologies, it was possible to find higher energy, open, metastable states with constrained MD simulations, 4 and 2 kJ mol^{-1} per [2+3] unit above the collapsed conformations for $\mathbf{Tri}^6\mathbf{Di}^9$ and $\mathbf{Tri}^8\mathbf{Di}^{12}$ respectively (details about the inflation of the latter two molecules are in Chapter 3). Only the relative energies of the open conformations were compared, as these are likely to reproduce the structures that initially form in the reaction solution; comparison between these geometries was considered more pertinent for determining the likely reaction outcome.

The $\mathbf{Tri}^4\mathbf{Di}^6$ molecule was identified as the lowest in energy and, it being the synthetically realised topology, could therefore be rationalised on the energy of the gas phase conformations alone. This was followed by the higher energy metastable inflated conformers of $\mathbf{Tri}^8\mathbf{Di}^{12}$ (5 kJ mol^{-1} per [2+3] unit) and $\mathbf{Tri}^6\mathbf{Di}^9$ (14 kJ mol^{-1} per [2+3] unit). An alternative topology was available with a [4+6] reaction, in which precursors are assembled into a doubly-bonded ring to give $\mathbf{Tri}^4{}_2\mathbf{Di}^6$; however this was located considerably higher in energy (39 kJ mol^{-1} per [2+3] unit) and therefore was considered less likely to form. The $\mathbf{Tri}^2\mathbf{Di}^3$, in which the precursors are assembled into a capsule, was the topology with the highest energy, 52 kJ mol^{-1} per [2+3] unit higher than $\mathbf{Tri}^4\mathbf{Di}^6$, probably due to the high geometric strain caused by arranging the precursors into a capsule. To investigate not only the topological prediction, but also the geometry prediction, the predicted structure of the $\mathbf{Tri}^4\mathbf{Di}^6$ **CC3** molecule was compared to that of the SCXRD structure.[7] A RMSD value of

0.11 Å was calculated, showing the conformation was well reproduced. The overlay of the two molecules is shown in Figure 5.9.

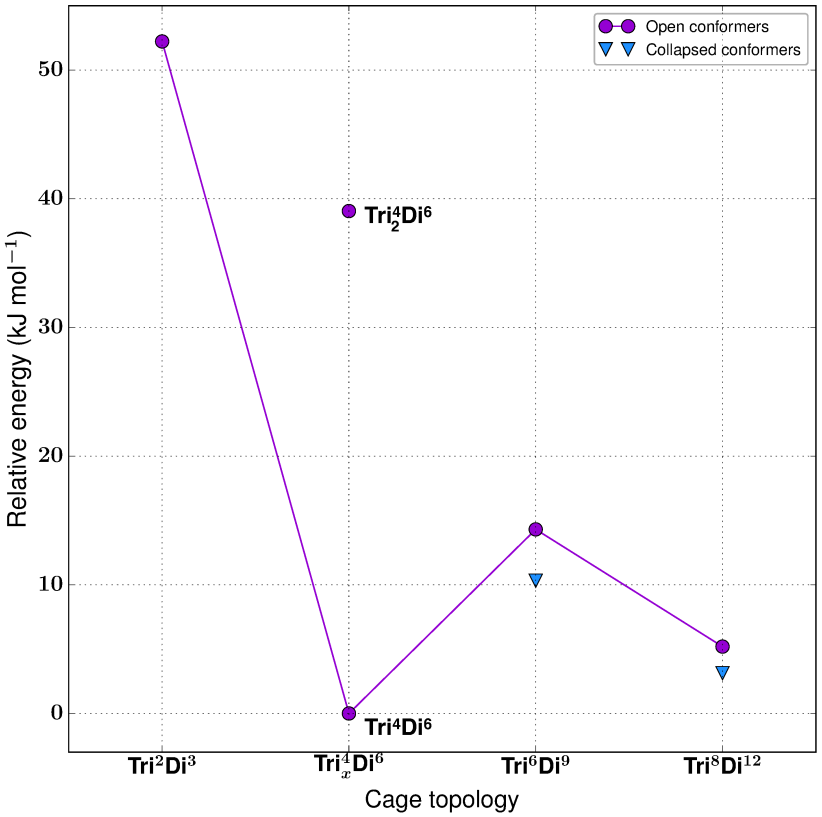


Figure 5.8.: The relative energies of the lowest energy conformations of the **CC3**-based Tritopic + Ditopic topologies, as calculated by M06-2X.

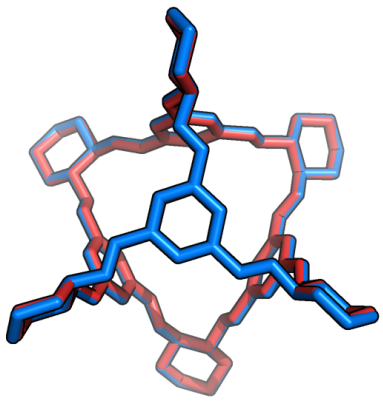


Figure 5.9.: Overlay of computed (blue) and crystal structure (red) of **Tri^{4}Di^6**.

CC5 and CC8-like cages.

Again for the Tritopic + Ditopic family, I attempted to analyse the preferential formation of the **CC5** and **CC8** imine cages, by comparing the energies of the cages that were experimentally found, and some alternative topologies. When tris(4-formylphenyl)amine (**3** in Figure 5.10) and (1*S*,2*S*)-cyclopentanediamine (**4** in Figure 5.10) are combined, they produce the porous **Tri**^{4}**Di**^{6} **CC5** molecule (with a very large BET surface area of 1333 m^2 g^{-1}),[72] but changing the diamine to (*R*,*R*)-1,2-diaminocyclohexene (**5** in Figure 5.10), adding only an additional CH_2 group, results in a **Tri**^{8}**Di**^{12} **CC8** molecule, which is non-porous and lacks shape persistence.[72,96]

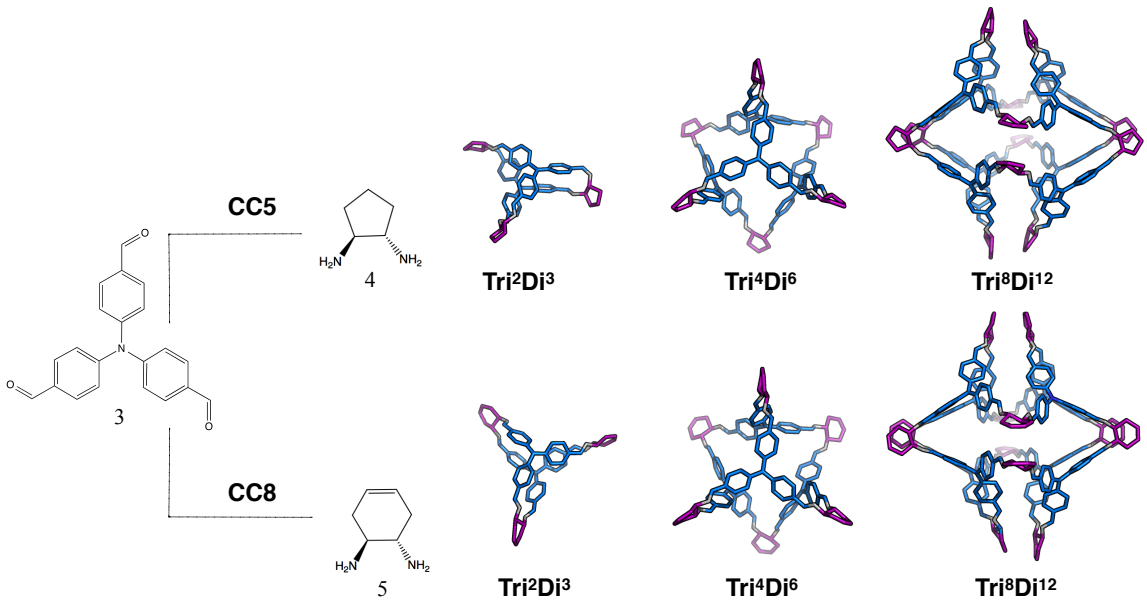


Figure 5.10.: Molecular structures of the lowest energy open conformations of the Ditopic + Tritopic topologies for the **CC5** and **CC8** imine cage system, formed from the reaction of 1,3,5-triformylbenzene (**3**) and (1*S*,2*S*)-cyclopentanediamine (**4**) for **CC5** and (*R*,*R*)-1,2-diaminocyclohexene (**5**) for **CC8**. Tritopic BBs are in blue, ditopic precursors in purple, the imine bonds formed in the reaction are in grey.

To try and understand why this behaviour occurs, relative energies per [2+3] unit and structures for the lowest energy conformations of the **Tri**^{2}**Di**^{3}, **Tri**^{4}**Di**^{6}, and **Tri**^{8}**Di**^{12} topologies were compared. The structures are show in Figure 5.10 and the energies in Figure 5.11 and Table 5.2.

Table 5.2.: Information on the lowest energy open conformations of the **CC5**- and **CC8**-based Tritopic + Ditopic topologies.

	Synthetically realised?	Energy relative to **Tri^{8}Di^{12}** per [2+3] unit (kJ mol^{-1})	Shape persistent?	Void diameter (Å)
CC5 Tri^{2}Di^3	no	87	-	0.4
CC5 Tri^{4}Di^6	yes	-1	yes	7.3
CC5 Tri^{8}Di^{12}	no	0	no	10.1
CC8 Tri^{2}Di^3	no	67	-	1.6
CC8 Tri^{4}Di^6	no	25	yes	6.5
CC8 Tri^{8}Di^{12}	yes	0	no	9.2

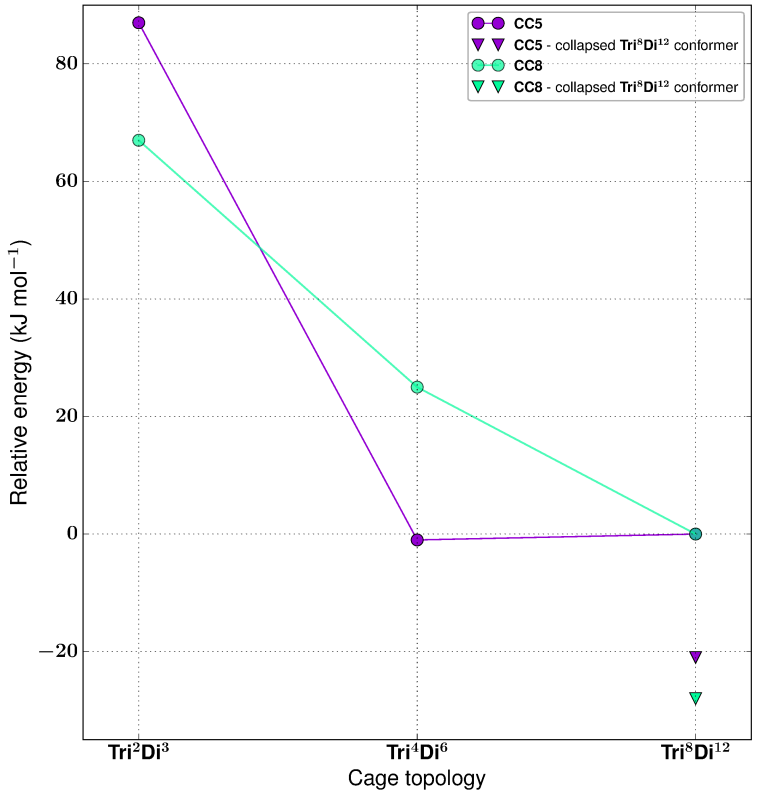


Figure 5.11.: The relative energies per [2+3] formula unit of the lowest energy conformations of the **CC5** and **CC8**-based Ditopic + Tritopic topologies, as calculated by M06-2X.

The **Tri**^{2}**Di**^{3} topology is too small to contain a cavity, whereas the **Tri**^{4}**Di**^{6} molecules are shape persistent with cavities ∼7 Å in diameter. However, both **Tri**^{8}**Di**^{12} structures collapse, as it is known experimentally for the **CC8** molecule, losing their internal cavities.[96] Constrained MD simulations on the **Tri**^{8}**Di**^{12} structures were used to find higher energy open metastable states, which were located 21 and 28 kJ mol^{-1} per [2+3] unit above the collapsed conformations for **CC5** and **CC8** respectively, with probe sizes of 4.5 Å. The open conformers of **CC5** and **CC8** have void diameters of 10.1 and 9.2 Å, respectively.

For both cage families, the **Tri**^{2}**Di**^{3} molecule, with only two carbon atoms bridging the tritopic BBs, was found to be quite strained and consequently high in energy compared to the other topological possibilities (by >40 kJ mol^{-1}). It is not surprising therefore that this topology was not experimentally observed. For the **CC8** molecule, the lowest energy conformation had the large **Tri**^{8}**Di**^{12} underlying topology, with an approximate octahedral symmetry. This corresponds to the experimental reports of this molecule forming with the diaminocyclohexene group.[96] As I previously reported in Chapter 3,[215] the computed open conformation for **CC8** well matched the geometry of a molecule extracted from SCXRD structure of the solvate (with a RMSD value of 1.67 Å, see Figure 3.5-A). The comparison between modelled and experimental structures improved slightly when relaxing the SCXRD chosen geometry (with the RMSD value moving to 0.63 Å, see Figure 3.5-B). The relaxation removed the "swelling" effect that solvent molecules can confer to the open geometry, causing in this case in a small contraction of the vertices.

Results for the **CC5** molecule differed from the ones obtained for the **CC8** molecule; no clear energetic preference was found between the **Tri**^{4}**Di**^{6} and **Tri**^{8}**Di**^{12} molecules, with a relative energy difference of only 1 kJ mol^{-1} per [2+3] unit. This molecule was in fact synthetically realised only as a **Tri**^{4}**Di**^{6} topology, with a BET surface area of 1333 m^{2} g^{-1}.[72] An overlay of the computed and the solvated SCXRD (see Figure 5.12) found a good match, with an RMSD of 0.71 Å. As in previous cases, this match improved when the solvate conformation was geometry optimised (with the RMSD value improved to 0.45 Å). The minimisation led to a contraction of the molecule, suggesting the presence of a scaffolding effect of the solvent that was lost in the calculations when the conformation was energy minimised after our artificial inflation procedure. In the case of **CC5**, the calculations could not successfully distinguish between the formation of a **Tri**^{4}**Di**^{6} or **Tri**^{8}**Di**^{12} topology. This could be attributed to the simulations not being able to include all the solvent effects (as evidenced by the overlays). Nonetheless, the scenario clearly differed from the **CC8**

topological energy landscape, where the largest topology was clearly preferred. It could be reasonably assumed, as discussed in the computational challenges section, that, if the entropic contribution to the free energies could be considered, this would be expected to lead to a destabilisation of the larger **Tri**^{8}**Di**^{12} topology with respect to the **Tri**^{4}**Di**^{6}.

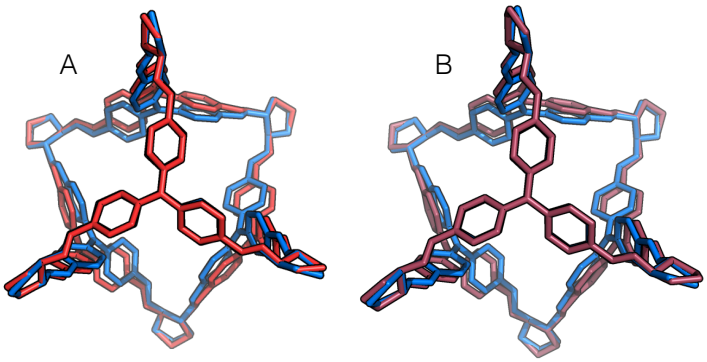


Figure 5.12.: CC5 - Tri^{4}Di^{6}, A) Overlay of computed (blue) and crystal structure (red), **B)** Overlay of computed (blue) and minimised crystal structures (dark red).

5.4.2. The Tritopic + Tritopic family

For the Tritopic + Tritopic family, the reaction between tris(2-aminoethyl)amine (**6** in Figure 5.13) and tris(4-formylphenyl)aldehyde (**7** in Figure 5.13) was investigated. It was used for the synthesis of **CC11**, a tetrapodal-shape **Tri**^{4}**Tri**^{4} molecule reported by Briggs and coworkers.[210]

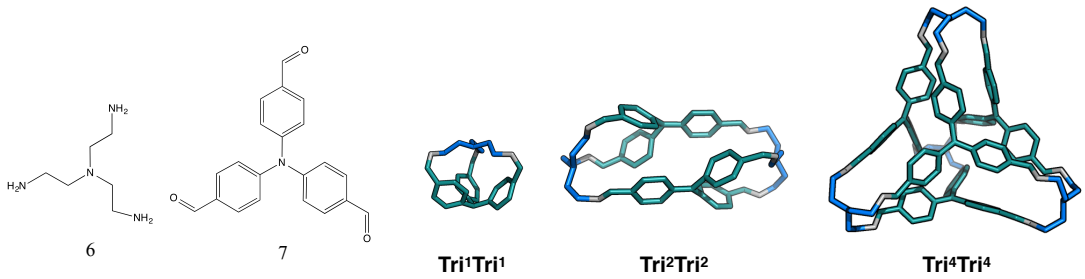


Figure 5.13.: Molecular structures of the lowest energy conformations of the Tritopic + Tritopic topologies for the **CC11** imine cage system, formed from the reaction of tris(2-aminoethyl)amine (**6**) and tris(4-formylphenyl)aldehyde (**7**). Tritopic BBs of **6** are in blue, tritopic BBs of **7** are in teal, the imine bonds formed in the reaction are in grey.

The relative energies per [1+1] unit were compared for the lowest energy conformations of the **Tri**^{1}**Tri**^{1}, **Tri**^{2}**Tri**^{2} and **Tri**^{4}**Tri**^{4} topologies. The geometries of these conformations are shown in Figure 5.13, with the relative energies shown in the plot in Figure 5.14 and Table 5.3.

Table 5.3.: Information on the lowest energy open conformations of the **CC11**-based Tritopic + Tritopic topologies.

	Synthetically realised?	Energy relative to **Tri**^{4}**Tri**^{4} per [1+1] unit (kJ mol^{-1})	Shape persistent?	Void diameter (Å)
Tri^1**Tri**^{1}	no	376	-	1.2
Tri$_2^2$**Tri**^{2}	no	123	-	0.4
Tri^4**Tri**^{4}	yes	0	yes	3.2

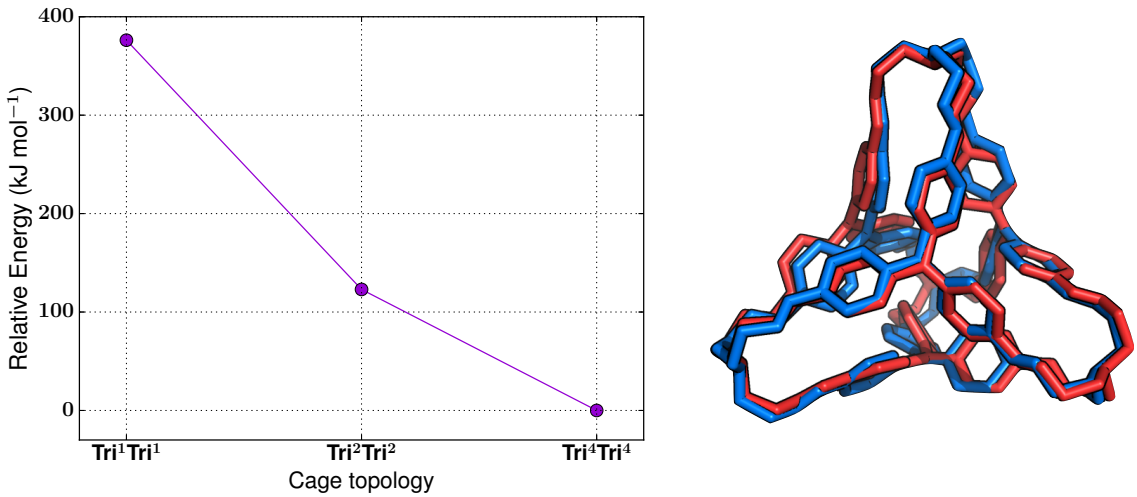


Figure 5.14.: Left hand side: the relative energies of the lowest energy conformations of the **CC11**-based Tritopic + Tritopic topologies, as calculated by M06-2X; Right hand side: Overlay of computed (blue) and crystal structure (red) of **Tri**^{4}**Tri**^{4}.

Both **Tri**^{1}**Tri**^{1} and **Tri**^{2}**Tri**^{2} were characterised by very strained geometries and therefore their energies were hundreds of kJ mol^{-1} higher than that of **Tri**^{4}**Tri**^{4} (376 and 123 kJ mol^{-1} per [1+1] unit, respectively). This was in agreement with the experimental result, which consist in the formation of a cage with **Tri**^{4}**Tri**^{4} topology, and with a tetrapodal geometric shape. The comparison with the single crystal X-ray diffraction led to an RMSD of 0.55 Å, with a good match between

the predicted structure and the experimentally reported geometry (see Figure 5.14, right hand side).

5.4.3. The Tetratopic + Ditopic family

For the Tetratopic + Ditopic family, I investigated the reaction of the tetraformyl-cavitand (**8** in Figure 5.15) and ethylene-1,2-diamine (**9** in Figure 5.15), reported by Warmuth and co-workers to produce a mixture of $\mathbf{Tet}_2^4\mathbf{Di}^8$, $\mathbf{Tet}^6\mathbf{Di}^{12}$ and $\mathbf{Tet}^8\mathbf{Di}^{16}$ topology molecules, dependent upon the solvent used.[99] Five different molecules were generated, $\mathbf{Tet}^2\mathbf{Di}^4$, $\mathbf{Tet}^3\mathbf{Di}^6$, $\mathbf{Tet}_2^4\mathbf{Di}^8$, $\mathbf{Tet}^6\mathbf{Di}^{12}$ and $\mathbf{Tet}^8\mathbf{Di}^{16}$. In the MD simulations, $\mathbf{Tet}^2\mathbf{Di}^4$, $\mathbf{Tet}_3^3\mathbf{Di}^6$ and $\mathbf{Tet}_2^4\mathbf{Di}^8$ were found to be shape persistent, whereas the larger $\mathbf{Tet}^6\mathbf{Di}^{12}$ and $\mathbf{Tet}^8\mathbf{Di}^{16}$ topologies collapsed.

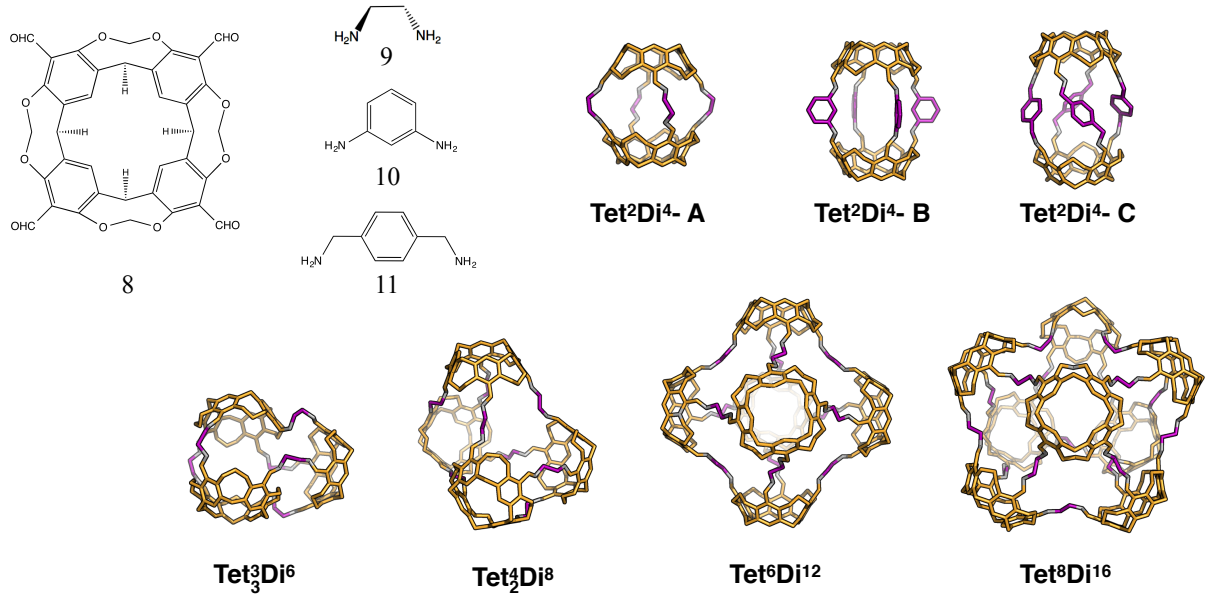


Figure 5.15.: Molecular structures of the lowest energy open conformations of the Tetratopic + Ditopic topologies, formed from the reaction of tetraformyl cavitand (**8**) and ethylene-1,2-diamine (**9**). Also shown with the $\mathbf{Tet}^2\mathbf{Di}^4$ topology, are the lowest energy conformation with *m*-phenylenediamine (**10**) and *p*-xylylenediamine (**11**) as the ditopic BBs. Tetratopic precursors are in yellow, ditopic precursors in purple, the imine bonds formed in the reaction are in grey.

Constrained MD simulations on the latter two structures were able to locate metastable, higher energy open conformations. The structures of the lowest energy open conformations are shown in Figure 5.15, with the relative energies shown in Figure 5.17 and reported in Table 5.4.

Table 5.4.: Information on the lowest energy open conformations of the Warmuth Tetratopic + Ditopic topologies. Values in brackets refer to formation energies per bond formed (kJ mol^{-1}), which were only calculated for cages **Tet^{2}Di^4** A, B, and C.

	Synthetically realised?	Energy relative to **Tet$_2^4$Di^8** per [1+2] unit (kJ mol^{-1})	Shape persistent?	Void diameter (Å)
Tet^{2}Di^4-A	no	27 (8)	yes	5.8
Tet^{2}Di^4-B	yes	(6)	yes	5.1
Tet^{2}Di^4-C	yes	(0)	yes	4.6
Tet^{3}Di^6	no	17	yes	3.1
Tet$_2^4$Di^8	yes	0	yes	6.3
Tet^{6}Di^{12}	yes	1	no	12.3
Tet^{8}Di^{16}	yes	from -21 to 17	no	8.4-10.4

The cage with the lowest energy was a partially collapsed conformer of **Tet^{8}Di^{16}**. The cage was not found to be shape persistent after high temperature MD, and the inflating procedure generated a very large number of higher energy metastable conformers, two of which are shown in Figure 5.16. Some conformers were partially open (like the lowest energy one on the left hand side of Figure 5.16), whereas others contained a larger pore size, the largest conformer presenting a void diameter of 10.4 Å.

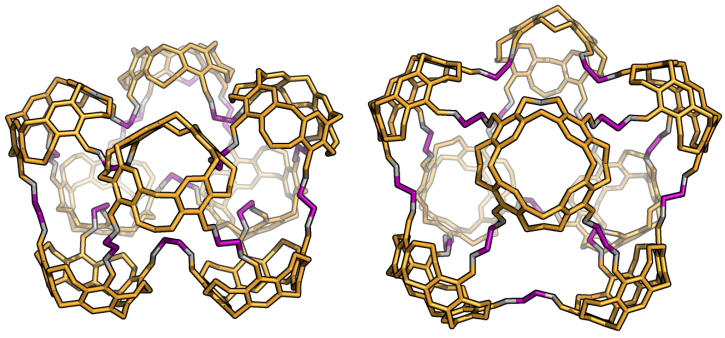


Figure 5.16.: Two different conformers of **Tet^{8}Di^{16}**, the less inflated and lower in energy on the left, and the most inflated and higher in energy on the right.

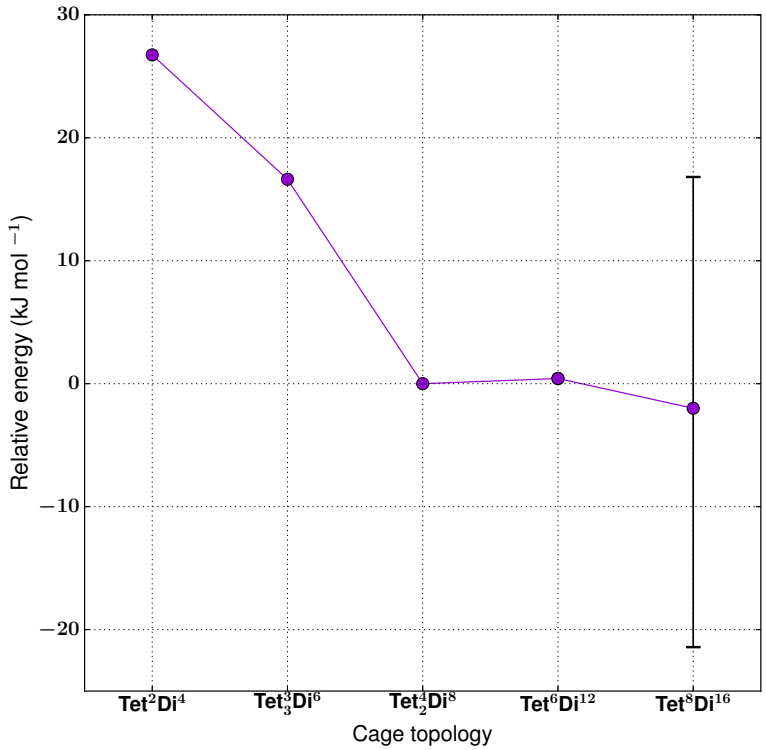


Figure 5.17.: The relative energies of the lowest energy conformations of the Warmuth Tetratopic + Ditopic topologies, as calculated by M06-2X. The error bar on point Tet^8Di^{16} shows the range of energies covered by conformations with differing degrees of inflation (see Figure 5.16, and Table 5.5), with more open conformers displaying higher energies.

As no crystal structure was available and it was not possible to be certain whether a single conformation represented the open structure, Figure 5.17 shows the full range of energies found for conformations with varying pore size, the conformation with the widest pore diameter shown on the right hand side of Figure 5.16. This larger and more open conformation was located 38 kJ mol^{-1} above the less inflated lower energy conformer, which also had a smaller pore diameter of 8.4 Å. The degree of strain of the structure increases with the degree of inflation.

Experimentally, the formation of $\mathbf{Tet^8Di^{16}}$ was obtained using CH_2Cl_2 as a solvent. We carried out a structural analysis on the resulting cage, in order to understand whether it was possible to relate the solvent used during the synthesis to the final topology. $\mathbf{Tet^8Di^{16}}$ has 10 windows, where the diameters of the circles enclosed in the windows range from 4 - 6 Å for the 8 triangular windows and from 6 - 9 Å for the 2 square windows, depending on the degree of inflation (see Table 5.5).[†]

[†]The window sizes were calculated with the pyWINDOW package[222] by Marcin Miklitz.

Table 5.5.: Information about the window sizes of $\mathbf{Tet}_2^4\mathbf{Di}^8$, $\mathbf{Tet}^6\mathbf{Di}^{12}$ and $\mathbf{Tet}^8\mathbf{Di}^{16}$.

	N^o of triangular windows (TW)	N^o of squared windows (SW)	Window Size Average (Å)
$\mathbf{Tet}_2^4\mathbf{Di}^8$	4	2	5.6 (TW), 1.8 (SW)
$\mathbf{Tet}^6\mathbf{Di}^{12}$	8	-	5.2 (TW)
$\mathbf{Tet}^8\mathbf{Di}^{16}$ - less inflated	8	2	4.4 (TW), 6.3 (SW)
$\mathbf{Tet}^8\mathbf{Di}^{16}$ - more inflated	8	2	4.5 (TW), 8.5 (SW)

About 22 and 21 kJ mol^{-1} above the lowest energy conformer of $\mathbf{Tet}^8\mathbf{Di}^{16}$ were $\mathbf{Tet}^6\mathbf{Di}^{12}$ and $\mathbf{Tet}_2^4\mathbf{Di}^8$, that were respectively found experimentally in THF and chloroform. Their windows diameters were analysed (an average of 5.2 Å for the 8 triangular windows of the octahedral $\mathbf{Tet}^6\mathbf{Di}^{12}$, and an average of 5.6 Å for the 4 biggest windows of $\mathbf{Tet}_2^4\mathbf{Di}^8$), but it was not possible to find any geometric correlation between the solvent used for the synthesis and the cage topologies. This particular aspect was not the object of the work, and will be left to further studies.

Other hypothetical topologies generated, $\mathbf{Tet}_3^3\mathbf{Di}^6$ and $\mathbf{Tet}^2\mathbf{Di}^4$, were more strained and thus higher in energy, justifying the fact that they were not experimentally observed. They were located respectively 17 and 27 kJ mol^{-1} higher than $\mathbf{Tet}_2^4\mathbf{Di}^8$.

Subsequently, I carried out an investigation to study the effect of changing the ditopic BB's geometry on the topological landscape. The study carried out by Warmuth and co-workers suggested that it was not possible to synthesise a cage with the shape of a $\mathbf{Tet}^2\mathbf{Di}^4$ capsule when using ethylene-1,2-diamine, which can be rationalised by the fact that the capsule lies 27 kJ mol^{-1} higher in energy per [1+2] unit than the observed $\mathbf{Tet}_2^4\mathbf{Di}^8$ topology.[99] However, they did find that the $\mathbf{Tet}^2\mathbf{Di}^4$ capsule could be formed by using different ditopic BBs; benzenes amino-functionalised in the *meta* or *para* position. Therefore, I assembled two more capsules in a [2+4] ratio, $\mathbf{Tet}^2\mathbf{Di}^4$**-B** and $\mathbf{Tet}^2\mathbf{Di}^4$**-C**, using *m*-Phenylenediamine (**10**) and *p*-Xylylenediamine (**11**) respectively. As these molecules contained different BBs, it was not possible to directly compare their relative energies per formula unit, thus I instead compared the relative formation energies of the three capsular $\mathbf{Tet}^2\mathbf{Di}^4$ molecules (see Table 5.4). In agreement with experimental results, both

cages **B** and **C** were lower in energy than **A**. Cage **Tet^{2}Di^4-C** in particular, was 6 kJ mol^{-1} per imine bond lower than **Tet^{2}Di^4-B**, and 8 kJ mol^{-1} lower than **Tet^{2}Di^4-A**. This is because kinked or twisted diamines allowed the two cavitands to align in a more stable and less hindered conformation, than when using BB **9**. The latter should adopt a very strained gauche conformation to permit the cavitands to align in the coplanar conformation required to reduce the strain.

5.4.4. The Tetratopic + Tritopic family

Finally, for the Tetratopic + Tritopic family, I studied the reaction of the tetraformyl-cavitand (**8** in Figure 5.18) and 1,3,5-tris(4-aminephenyl) benzene (**12** in Figure 5.18) to form a **Tet^{6}Tri^8** molecule, as previously reported by Warmuth and co-workers.[92]

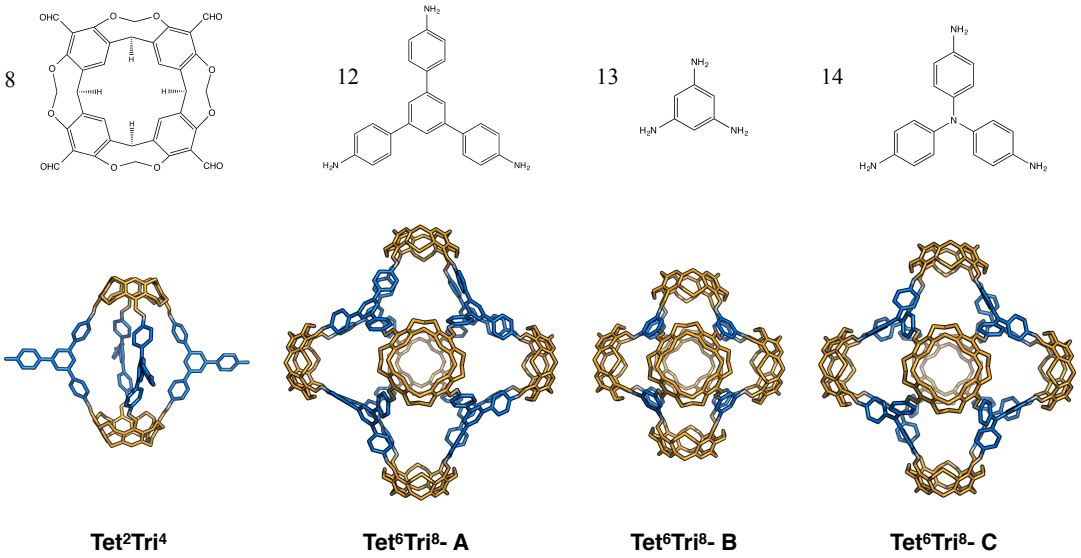


Figure 5.18.: Molecular structures of the lowest energy conformations of the Tetratopic + Tritopic topologies formed from the reaction of tetraformyl cavitand (**8**) and 1,3,5-tris(4-aminephenyl) benzene (**12**). Also shown with the **Tet^{6}Tri^8** topology are the lowest energy conformations with 1,3,5-triaminobenzene (**13**) and 4,4',4-triaminotriphenylamine (**14**) as the tritopic BBs. Tetratopic precursors are in yellow, tritopic precursors in blue, the imine bonds formed in the reaction are in grey.

The BBs were mixed in both a [6+8] ratio and into a [2+4] ratio to form two topologies, which were both observed depending upon whether a 3 aldehyde: 4 amine stoichiometric ratio was used experimentally (forming **Tet^{6}Tri^8-A**) or an

excess of the amine was used (forming a **Tet**^{2}**Tri**^{4} capsule with 3 unreacted amine groups). The lowest energy conformations are shown in Figure 5.18. The relative formation energies of the two topologies were compared (reported in Table 5.6) and found to be very similar (within 1 kJ mol^{-1}).

Table 5.6.: Information on the lowest energy open conformations of the Warmuth Tetratopic + Tritopic topologies.

	Synthetically realised?	Relative formation energy per bond formed (kJ mol^{-1})	Shape persistent?	Void diameter (Å)
Tet^2**Tri**^{4}	yes	3	yes	9.9
Tet^6**Tri**^{8}**-A**	yes	4	yes	17.8
Tet^6**Tri**^{8}**-B**	no	3	yes	12.2
Tet^6**Tri**^{8}**-C**	no	0	yes	15.5

Such a small difference in energy would explain why both topologies were experimentally observed in solution. Although the presence of the molecules in solutions was identified by Warmuth and coworkers with different techniques such as ^{1}H NMR, ^{13}C NMR and matrix-assisted laser desorption/ionization - time of flight (MALDI-TOF), the structures were never isolated from the solvent and neither solvated or desolvated crystal structures are currently available, therefore a direct geometric comparison between experimental and computed structures was not possible. Nevertheless, the calculations seem to suggest the lowest energy conformers of both the capsule and rhombic dodecahedron cage to be shape persistent and porous, therefore of synthetic interest. The void dimensions of both molecules were calculated and reported in Table 5.6. **Tet**^{2}**Tri**^{4} showed a void diameter 9.9 Å, and **Tet**^{6}**Tri**^{8} of 17.8 Å, which is very close to the 20.6 Å void diameter of the boronate cube synthesised by Mastalerz and coworkers, which currently holds the record of the molecule with the biggest intrinsic pore synthesised.

Considering the possible interest in a porous molecule of this cavity volume, the effect of the size of the triamine BB was investigated and two similar molecules with the **Tet**^{6}**Tri**^{8} topology were generated. Cavitand **8** was mixed with 1,3,5-triaminobenzene (**13**) to give **Tet**^{6}**Tri**^{8}**-B** and with 4,4',4-triaminotriphenylamine (**14**) to obtain **Tet**^{6}**Tri**^{8}**-C**. Both molecules were found to be shape persistent and therefore the formation energies per bond were calculated and compared to that of **Tet**^{6}**Tri**^{8}**-A** (see Table 5.6). **Tet**^{6}**Tri**^{8}**-B** was found to have a formation energy

$\sim$1 kJ mol^{-1} lower than **Tet^{6}Tri^{8}-A**, thus is potentially equally likely to form. **Tet^{6}Tri^{8}-C** showed a less strained geometry, which can be attributed to the central nitrogen atom on each tritopic BB allowing greater rotational freedom. This brought the formation energy 4 kJ mol^{-1} per imine bond lower than **Tet^{6}Tri^{8}-A**. This suggests that **Tet^{6}Tri^{8}-C** could be a particularly promising synthetic target. Both molecules are porous, showing respectively a shape persistent pore diameter of 12.2 Å and 15.5 Å. The synthesis of both molecules is currently being attempted by Cooper and coworkers.

5.4.5. Hypothetical larger molecules

Further to the structures analysed in the previous sections, I have constructed three larger topologies, **Tri^{20}Di^{30}**, **Tet^{16}Di^{32}** and **Tet^{24}Di^{48}** using, in the first case, the couple of BBs used for the **TriDi** family (1 and 2 in Figure 5.7), and for the latter two structures the couple of BBs chosen for the **TetDi** family (8 and 9 in Figure 5.15). These structures count 840, 1008 and 2016 atoms, respectively; they consist in large, extremely flexible molecules that are prone to collapse. Figure 5.19 shows these hypothetical structures as originally constructed, without minimisation or conformer searching, and thus retaining their open shape. I do not suggest any of these topologies as promising with these BBs, due to BB strain (**Tri^{20}Di^{30}**) or the accessibility of smaller collapsed unstrained structures, although their underlying topologies remain an interesting synthetic challenge.

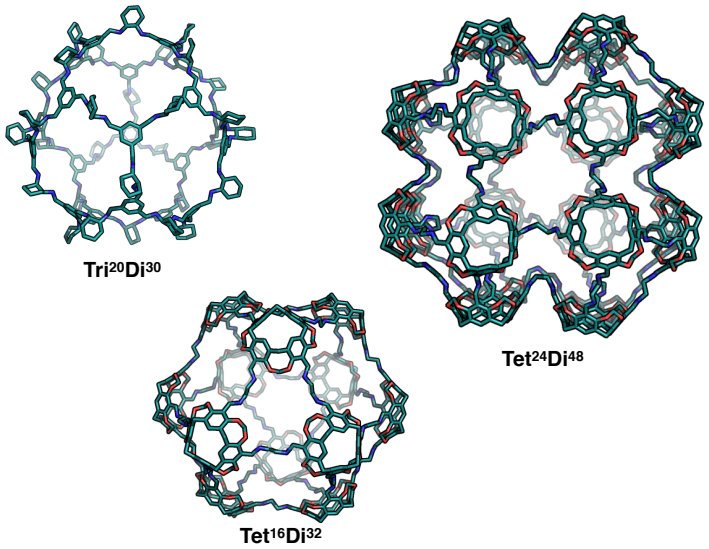


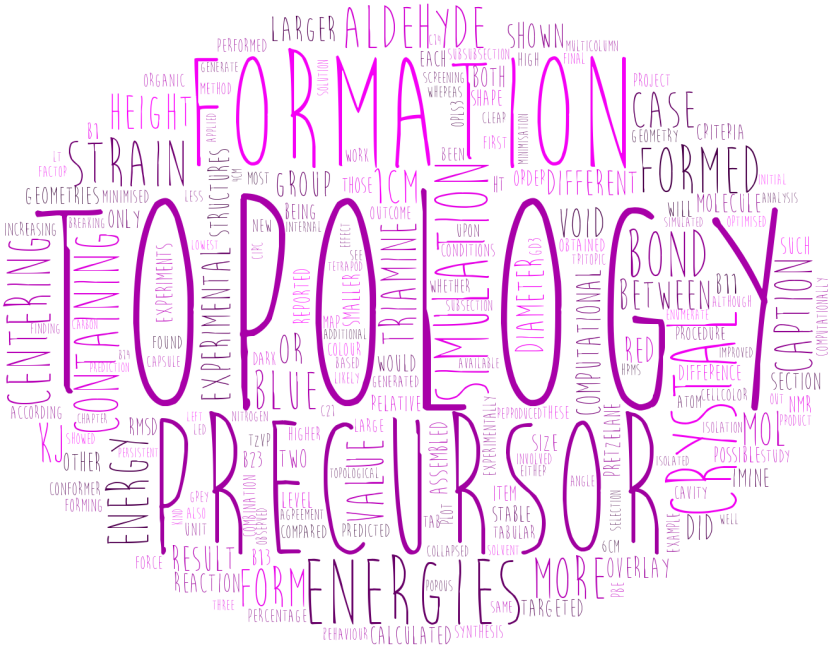
Figure 5.19.: Hypothetical structures of **Tri^{20}Di^{30}**, **Tet^{16}Di^{32}** and **Tet^{24}Di^{48}**, all of which lack shape persistence.

5.5. Conclusions

In this chapter I have tested to what extent calculations focusing on relative internal energies can assist in the prediction of topological outcome for a given DCC reaction that should reach the thermodynamic product. This approach was successful in the majority of the BB combinations examined, including for **CC3**, **CC5**, **CC11**, the **TetDi** family from Warmuth and the **TetTri** family from Warmuth. However, for the case of emergent behaviour in the **CC5** and **CC8** systems, whilst it was possible to correctly identify that the cyclohexane BB leads to a $\mathbf{Tri}^{8}\mathbf{Di}^{12}$ topology, for the cyclopentane BB, it was not possible to distinguish between the relative energies of $\mathbf{Tri}^{4}\mathbf{Di}^{6}$ and $\mathbf{Tri}^{8}\mathbf{Di}^{12}$. This means it was not possible to correctly identify the experimental outcome of a $\mathbf{Tri}^{4}\mathbf{Di}^{6}$ topology given that choice of precursors. This result is likely down to a combination of the influence of the solvent in disturbing the energy landscape and an entropic contribution disfavouring the larger topology, which could not be directly included in the simulations. The results presented in the next chapter will show that a larger energy gap ($\sim$30 kJ mol^{-1} at the PBE/TZVP+GD3 level) between two topologies generated from the same pair of precursors need to be calculated to consider a certain topology preferred over another. This means, therefore, that the level of confidence applied so far needs to be adjusted accordingly. For the Warmuth **TetDi** family, where the topology is known to be influenced by solvent choice, it was not possible to find any obvious geometric correlation between the structures and the successful solvent, thus this remains the subject of future research.

In general, this work highlights the design opportunity behind using simulations to consider the thermodynamic viability of a given target, which, together with chemical intuition, can provide new developments for this emerging field. The natural evolution of this project would be to explore the topological outcome for a much wider library of precursors with different topicity, length, angles, and rigidity, with the goal to rationalise which criteria would favour the formation of promising porous organic cages. However, the computational challenges encountered with modelling a relatively small set of organic precursors, suggest that brute force screening of large libraries is not currently computationally feasible. Instead, it would be better to focus on answering specific questions with smaller library subsets (<50), such as: How does variation in BB geometry influence topological outcome? Which BBs provide the most promising route to a desired topology? Which BB features are most critical to shape persistence in a desired topology?

Chapter 6
High-Throughput screening of porous organic cages



6.1. Introduction

In the previous chapters, I outlined the work carried out to classify, characterise and predict the formation of porous organic cage topologies. In this chapter, I will describe the application of the prediction strategy to a larger set of initial precursors,

aiming at the generation of a large collection of cages divided into three different topologies. The project was carried out in parallel with a high-throughput (HT) experimental screening project conducted by Dr. Becky Greenaway from the group of Prof. Andy Cooper (University of Liverpool), and simulations were performed to understand to what extent predictions could match experimental findings. The goal of the project was to accelerate discovery of families of new porous organic cages with different chemical components and topologies, which could then lead to the formation and characterisation of novel functional materials. The study involved mixing 3 differently functionalised triamines with 26 dialdehydes to generate 78 different combinations, characterised by 3 different topologies, namely $\mathbf{Tri}^2\mathbf{Di}^3$, a capsule, $\mathbf{Tri}^4\mathbf{Di}^6$, a tetrahedron, and $\mathbf{Tri}^4\mathbf{Tri}^4$, a tetrapod. The computational side of the work involved predicting the shape persistency of the full set of cages, and the topology of a selection of structures that, given the geometry of the starting precursors, could adopt more than one possible topology. In parallel, the project aimed at rationalising the geometric parameters (such as bond lengths, angles, and so on) that could lead to one topology rather than another, and to understand whether some chemical groups influenced the final outcome. Experimentally, a total of 33 new cages were discovered and analysed, and additionally, a serendipitously-discovered doubly bonded, triply interlocked species (which was renamed "pretzelane" due to its resemblance with a pretzel), was added to the computational study. Where possible, the structures of the cages obtained through X-ray crystal diffraction were compared with the conformations obtained with computational methods.

6.1.1. Overview of the project

In this section, I will mainly discuss the computational side of the project, but will draw upon the experimental findings where relevant for comparison with the computational results. Where topological outcomes in experiments are discussed, these are known from communication with Dr. Becky Greenaway, whose analysis is based upon the HT procedure described below, whereas the crystallographic results were provided by Dr. Marc Little (University of Liverpool). A manuscript in preparation will fully describe the experimental results.

Experimental overview

The reaction conditions for the synthesis of all the cages involved in the project needed to be carefully tested in order to obtain a robust workflow, which could

confidently be applied to the automated HT platform. The first concern was to be able to generate cages in solution and in reasonable yield without any crystallisation or precipitation occurring. This was important for two reasons: firstly, cages can sometimes be unstable to isolation, especially if they are not shape persistent, and avoiding precipitation would mean being able to carry out the HT solution analysis prior to isolation (to determine whether the cage has formed); secondly, it would avoid potential problems associated with handling suspensions on the automated platform.

The reaction conditions were first tested on one pair of precursors and then adopted for the automated procedure. This initial investigation showed that complete conversion to cage species was achieved using a small excess of triamine (one equivalent more per predicted cage stoichiometry, e.g. 5 for a reaction involving 4 triamines and 6 aldehydes), as using a larger excess of amine led to the formation of impurities. Alongside this, increasing the concentration of the reaction led to increasing amounts of precipitate forming, and, therefore, was avoided to make transfer to the automated platforms easier. The automated procedure that led to the cages' formation can be divided into a few steps:

1. A *Chemspeed SLT* robotic platform dispensed liquid stock solutions of the precursors into jacketed reactor vessels, which were heated for 3 days at 65 °C with vortexing. Deuterated chloroform was adopted as the solvent.

2. A solution analysis was performed prior to isolation to identify the formation of cages, and this was performed by a combination of techniques, such as ^{1}H NMR, high resolution mass spectroscopy (HRMS) and Fourier-transform infrared spectroscopy(FTIR).

3. Where clear cage formation was apparent, diffusion NMR was carried out to help identify the topology of the cage formed, by deducing the molecular mass based upon the solvation diameter. In some instances, although the ^{1}H NMR spectrum indicated clean cage formation, there was no clear mass ion in the HRMS corresponding to a cage species. Diffusion NMR was particularly useful in those cases too, as it provided an alternative method to determine the size, and hence topology, of the cages directly from the reaction mixtures. The average size and maximum diameters of the cages were calculated from simulations* and they were compared with the experimentally determined solvodynamic diameters. If there was agreement between the measured size and

*The diameters of the cages were calculated, in this case, by Marcin Miklitz.

the computed average size the combination of diffusion NMR and simulations provided a rapid and direct measure of the success of predictions.

4. For combinations that indicated cage formation, the reaction solutions were filtered, where necessary, to remove any insoluble precipitate and the solvent was removed.

5. The resulting isolated samples were analysed by PXRD and re-analysed by ^{1}H NMR spectroscopy to determine stability to isolation.

In some instances, the solid either did not fully re-dissolve, suggesting some instability to isolation, or decomposition was visible on comparing the spectra pre- and post-isolation. These factors led, in certain cases, to uncertainties about the final topology of the cage formed, and additional computational simulations were performed to calculate the relative energies of multiple possible topologies.

Computational overview

Initially, all the cages involved in the project were generated in the targeted topology (the criteria according to which precursors are targeted towards certain topologies are covered in the next few paragraphs) based on the building block structure, together with some additional alternative topologies for a selection of them. All the precursors, and the topologies they were targeted to are reported in Figure 6.1. Cages containing the ditopic linker 7 were omitted as the nitro-group on the aromatic ring of this precursor was poorly described by the OPLS3 forcefield. Precursor 10 was simplified by removing the branched alkyl substituents, as this external functionalisation was not important for consideration of the intrinsic strain in forming the cage topologies, and reducing the size and flexibility of the molecules reduced the computational cost. Aldehydes 1-10 were mixed with triamines A-C into a $\mathbf{Tri}^{2}\mathbf{Di}^{3}$ topology via imine condensation, aldehydes 11-21 into a $\mathbf{Tri}^{2}\mathbf{Di}^{3}$ topology, and aldehydes 22-26 into a $\mathbf{Tri}^{4}\mathbf{Tri}^{4}$ topology; all the cages were named after the combination of precursors they contain, so, for example, a cage obtained by mixing triamine "B" and dialdehyde "15" was named "B15".

The computational project started with the study of triamine B, already widely used for the synthesis of small cages,[179,223,224] which was mixed with terephthaldehyde (precursor 11) into four different topologies: $\mathbf{Tri}^{2}\mathbf{Di}^{3}$, $\mathbf{Tri}^{4}\mathbf{Di}^{6}$, $\mathbf{Tri}^{6}\mathbf{Di}^{3}$ and $\mathbf{Tri}^{8}\mathbf{Di}^{12}$. All the topologies were investigated with computational methods to predict the most likely reaction outcome and in order to test the timing and complexity

of the calculations that needed to be automated to be performed in parallel for all cages (details about the prediction are going to be covered in Section 6.3.1).

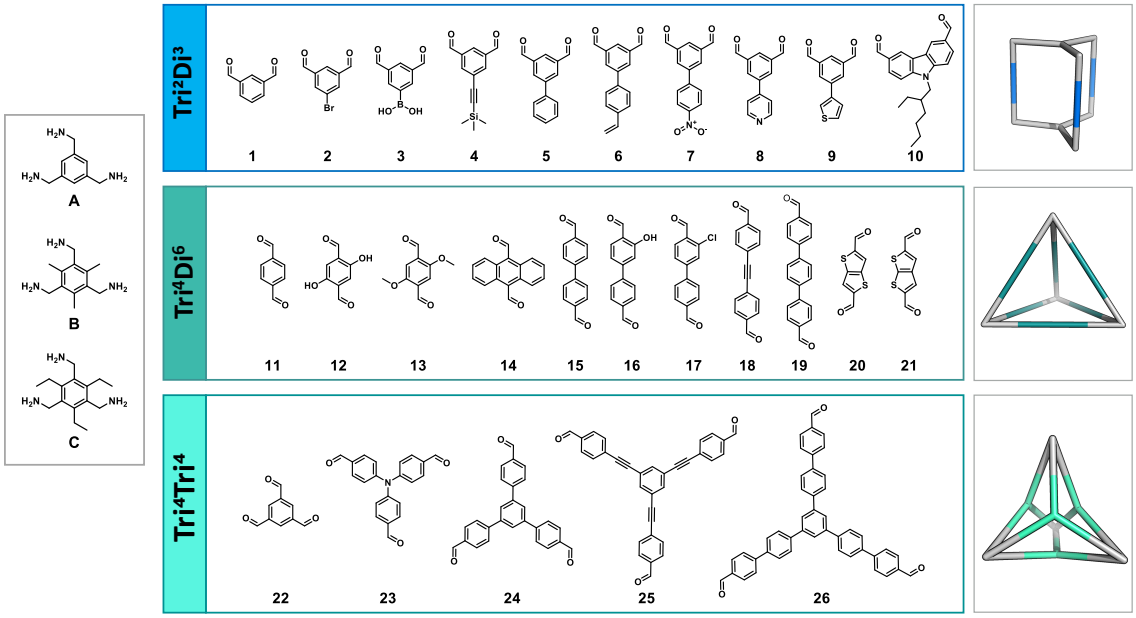


Figure 6.1.: Precursors involved in the HT screening and the presumed resultant topologies. The numbering is unrelated to that used in previous chapters.

Whilst the computational settings were optimised, the synthesis of B11 was attempted in order to calibrate the experimental conditions that would, then, be applied to the entire high-throughput screening. The reaction of these two precursors was previously reported to generate a complex precipitate;[225] however, in this study, lower concentration and heating led to the formation of a cage with **Tri**^{4}**Di**^{6} topology (matching the outcome of computational simulations - details will be provided in Section 6.3.1). Next, the selection of aldehyde precursors to be adopted for **Tri**^{4}**Di**^{6} topologies was expanded to a selection of longer and/or differently functionalised precursors. Then, the expansion to other cage topologies was considered, to confirm that the optimised conditions could be successfully applied to different topological families. The effect of decreasing the angle between the functional groups from 180° to 120° was studied and a second group of precursors was selected to generate cages with a **Tri**^{2}**Di**^{3} topology (1-10). Finally, a last group of tritopic aldehydes (precursors 22-26) was selected to generate **Tri**^{4}**Tri**^{4} cages via molecular panelling. Each group of aldehyde precursors was combined not only with triamine B, but also with the two A and C variants. To summarise, triamines A/B/C were mixed with precursors of three different kinds:

1. Ditopic precursors with aldehydes separated by an (aldehyde)-(centre of mass)-(aldehyde) angle of ∼120° with the aim of generating capsule-like molecules with $\mathbf{Tri}^2\mathbf{Di}^3$ topology (1-10).

2. Ditopic precursors with aldehydes separated by an (aldehyde)-(centre of mass)-(aldehyde) angle of ∼180° with the aim of generating capsule-like molecules with $\mathbf{Tri}^4\mathbf{Di}^6$ topology (11-21).

3. Tritopic precursors with the aldehydes arranged according to a C_3-fold symmetry to generate tetrapods with a $\mathbf{Tri}^4\mathbf{Tri}^4$ topology (22-26).

The above geometric criteria were chosen to reduce the strain of the precursors in the assembled cage and therefore to try and increase the chances of formation. More parameters were tested to understand whether some criteria could be applied to favour one topology over another. Some are listed below and illustrated in Figure 6.2:

a) Size of the precursors (meaning the distance from the centre of mass to the reactive ends in tritopic precursors).

b) Length of the precursors (meaning the distance between the two reactive ends in ditopic precursors).

c) Precursors with the same "core", but functionalised with different chemical groups.

d) Precursors that only differ by an atom, through either addition or substitution.

The precursors were assembled into the targeted topologies and according to the steps A-B described in Section 5.2. All the cages generated for the project are shown in Figures 6.3 (cages including triamine A), 6.4 (cages including triamine B) and 6.5 (cages including triamine C). The cages were characterised according to the procedure described in Section 6.2.

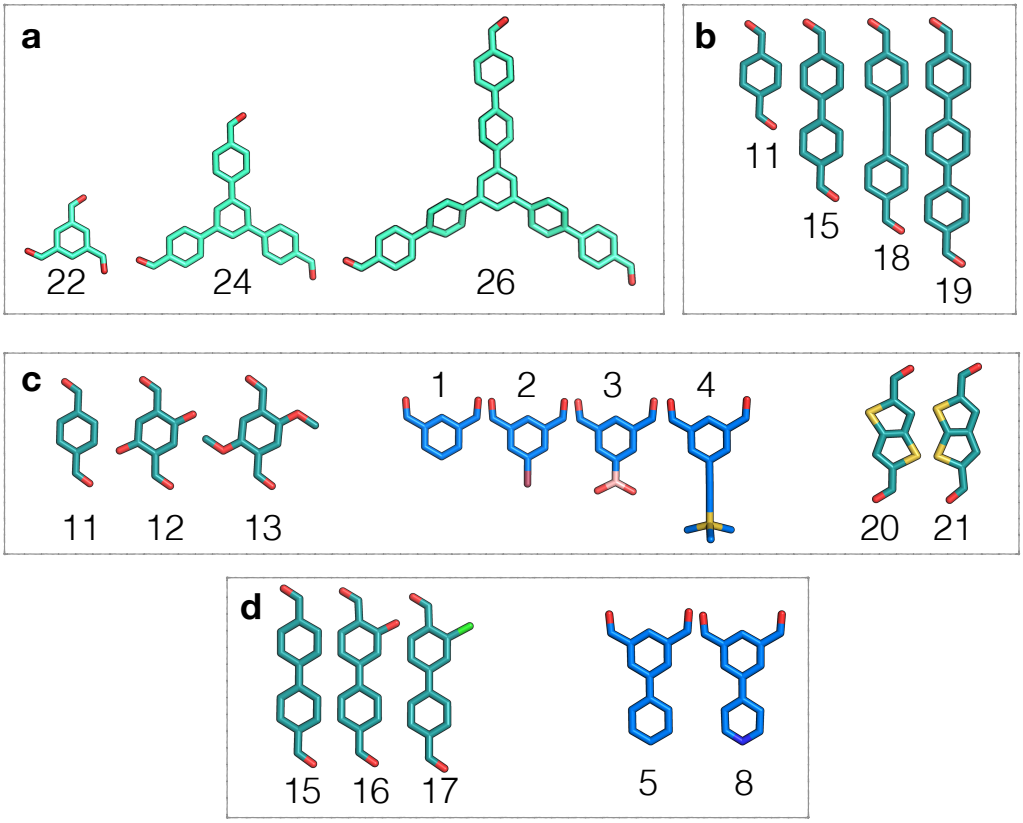


Figure 6.2.: Some of the criteria involved in the choice of the precursors; **a** is size of the precursors, **b** is length of the precursors, **c** precursors with different functionalities and **d** precursors that only differ by an atom. Carbon is in different shades of blue depending on which topology the precursors were targeted for, oxygen in red, nitrogen in blue, sulphur in yellow, boron in pink, bromine in dark red and chlorine in green.

A topology prediction was carried out for three cases (one for each topology studied in this work) in order to test each step of the procedure and whilst the synthesis of the same cages was attempted, B+1 to begin with, then B+11 and finally B+23. Firstly, triamine B and dialdehydes 1 and 11 were assembled not only in their planned topologies, a $\mathbf{Tri}^2\mathbf{Di}^3$ capsule and a $\mathbf{Tri}^4\mathbf{Di}^6$ tetrahedron, but also into the larger $\mathbf{Tri}^6\mathbf{Di}^9$ and $\mathbf{Tri}^8\mathbf{Di}^{12}$ topologies. Secondly, triamine B and dialdehyde 23 were assembled according to two additional topologies, $\mathbf{Tri}^1\mathbf{Tri}^1$ and $\mathbf{Tri}_2^2\mathbf{Tri}^2$. The relative energies of all cages, calculated at the DFT level, were compared to predict which topology was most likely to form, according to the same procedure defined in Section 2.4.1 and Chapter 5.

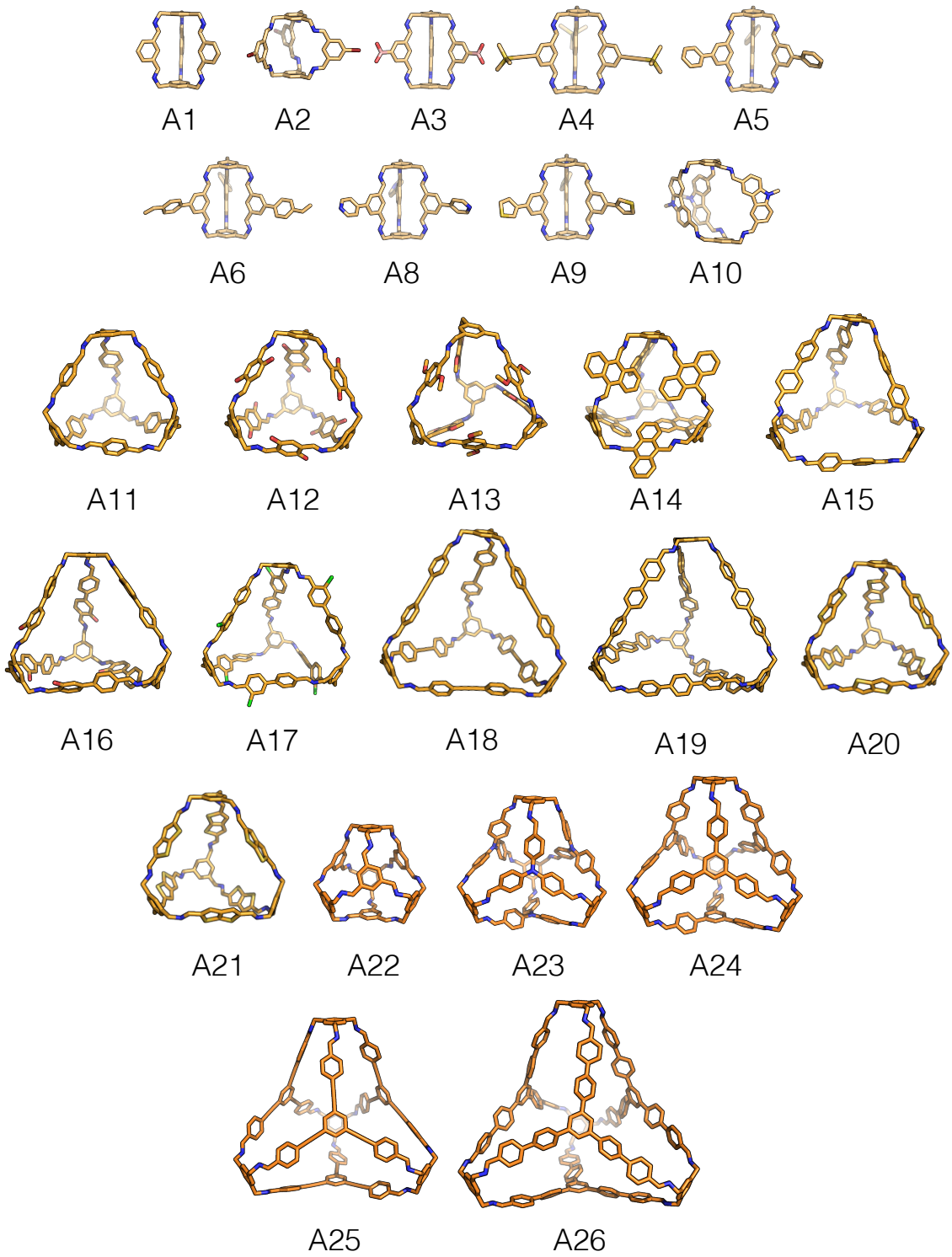


Figure 6.3.: PBE/TZVP+D3 optimised geometries of cages generated with triamine A. **Tri**^{2}**Di**^{3} cages (1-10) have carbon atoms in light yellow, **Tri**^{4}**Di**^{6} cages (11-21) in dark yellow and **Tri**^{4}**Tri**^{4} cages (22-26) in orange. Here, and throughout this chapter, nitrogens are in dark blue, oxygens in red, bromine in dark red, boron in pink, sulphur in yellow and chlorine in bright green. Cages A13, A14 and A17 collapsed after initial investigations, the starting open geometries being reported here. A7 is not shown as it was not included in the simulations.

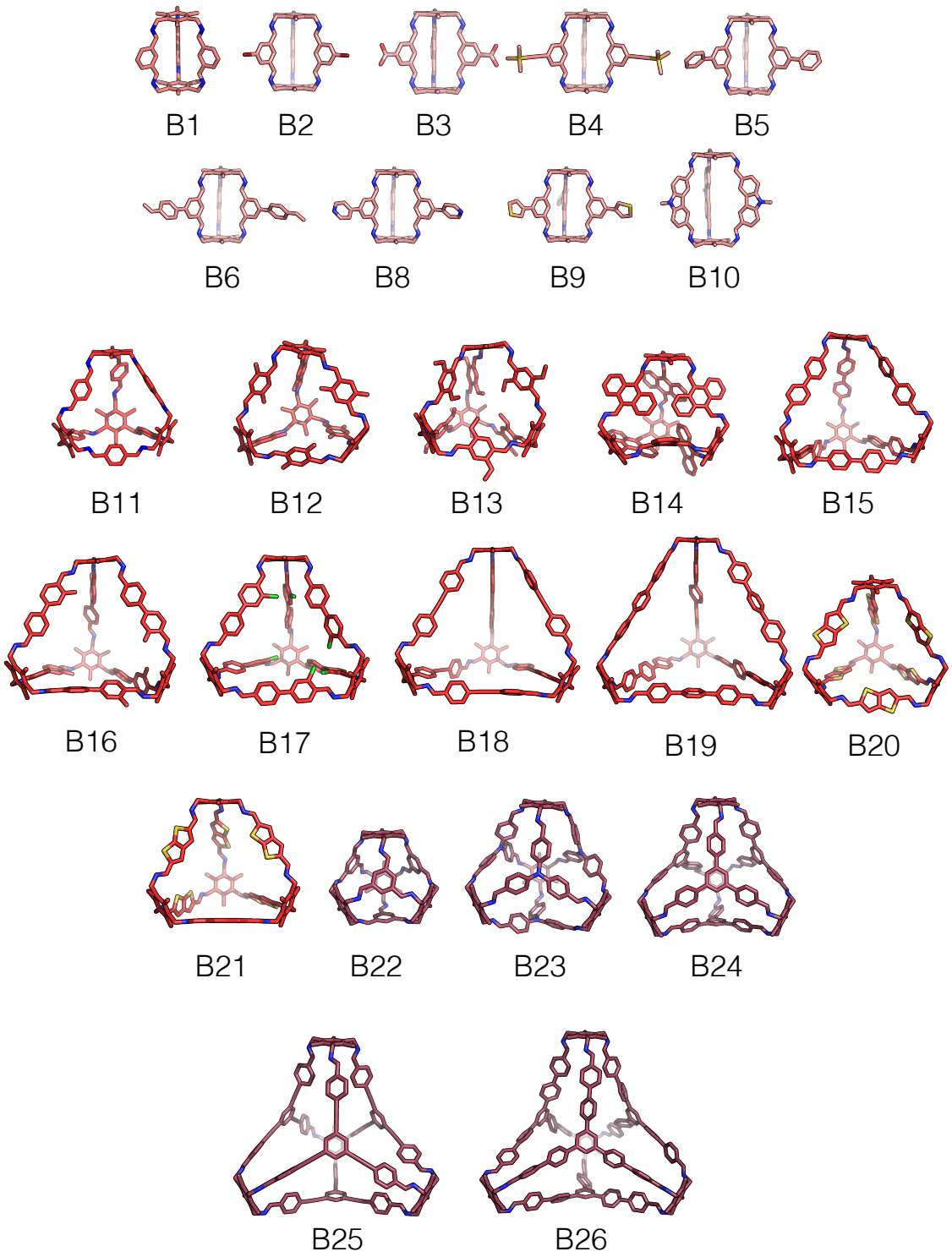


Figure 6.4.: PBE/TZVP+D3 optimised geometries of cages generated with triamine B. **Tri**^{2}**Di**^{3} cages (1-10) have carbon atoms in pink, **Tri**^{4}**Di**^{6} cages (11-21) in red and **Tri**^{4}**Tri**^{4} cages (22-26) in dark red. B7 is not shown as it was not included in the simulations.

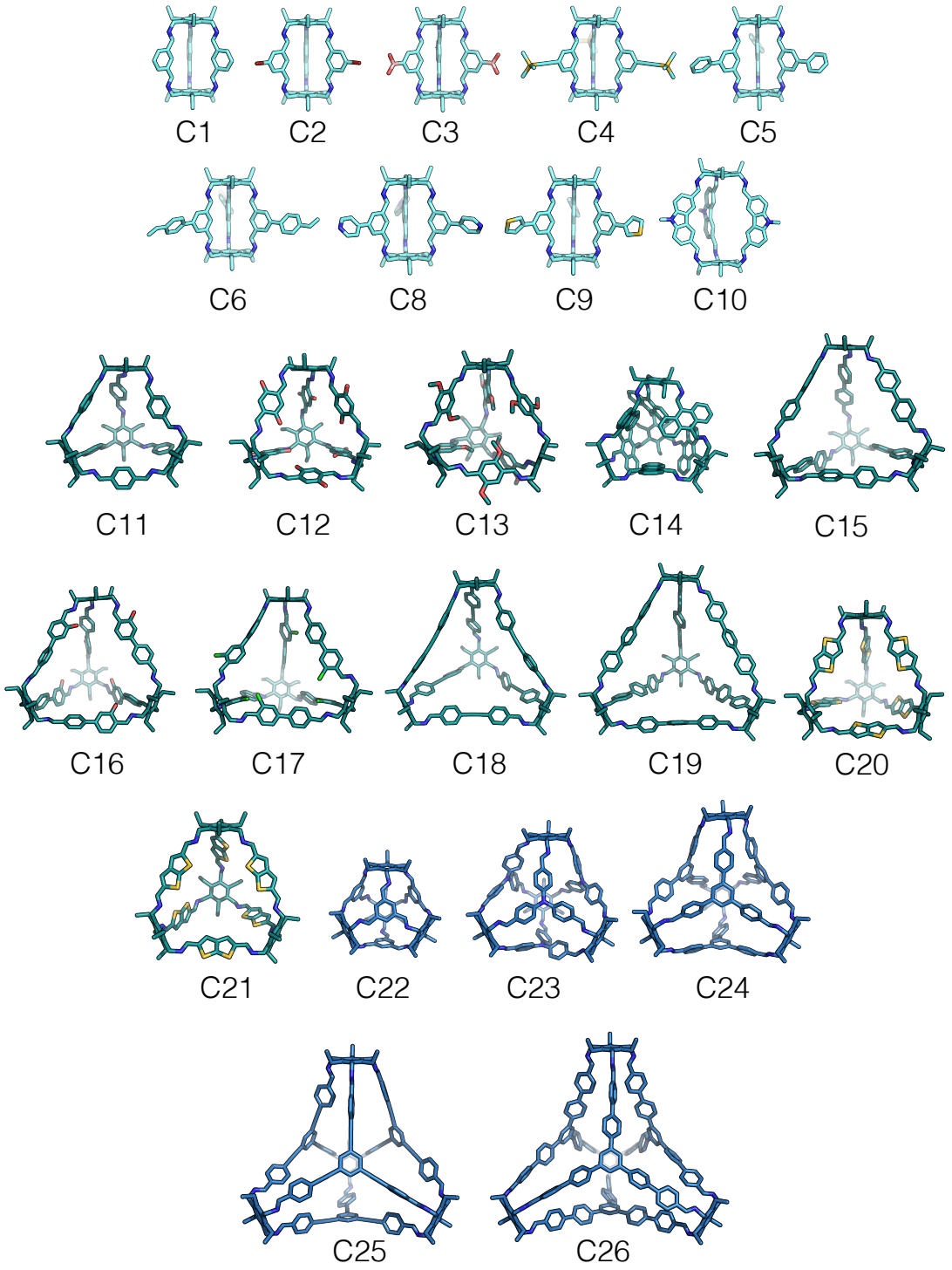


Figure 6.5.: PBE/TZVP+D3 optimised geometries of cages generated with triamine C. **Tri**^{2}**Di**^{3} cages (1-10) have carbon atoms in cyan, **Tri**^{4}**Di**^{6} cages (11-21) in teal and **Tri**^{4}**Tri**^{4} cages (22-26) in steel blue. C7 is not shown as it was not included in the simulations.

As the experimental characterisation of the cages progressed, it became clear that some precursors showed some unexpected behaviours, and that in those cases it was not clear whether the reactants assembled into larger tetrahedral topologies, in agreement with the initial plan, or whether they found a more stable assembly when forming a smaller capsule topology. This particular behaviour was observed for the combination of triamines A, B, and C with precursors 12, 14, 20 and 21. Therefore both the $\mathbf{Tri}^2\mathbf{Di}^3$ and the $\mathbf{Tri}^4\mathbf{Di}^6$ topologies were generated for this selection of cages, and the final internal energies were then compared to uncover which topology was predicted to form the most stable product. When it was serendipitously discovered that triamine B and C crystallised into a doubly bridged, triply interlocked novel "pretzelane" topology when combined with precursor 13, the new topologies were generated and their DFT energies compared with those of their tetrahedral equivalents.

6.2. Methods

The computational workflow for the screening can be described as follows:

1. All the cages were assembled as described in Figure 5.1. They were initially optimised with the universal force field (UFF) to produce feasible structures and were then geometry optimised with the OPLS3 force field using Macromodel (version 10.3.015, Schrödinger, LLC, New York, NY, 2015-3)

2. Each was then investigated with high temperature MD calculations in order to locate the lowest energy conformers, and simulations were repeated until no lower energy conformation could be found. Each run was performed at 1000 K for 100 ns, with a time step of 0.7 fs. All calculations were carried out with the OPLS3 force field,[160] and each sampled structure was minimised within the Macromodel package with the Polak-Ribier Conjugate Gradient method and a gradient convergence criteria of 0.05 kJ mol^{-1} $Å^{-1}$.

3. DFT calculations were performed on the lowest energy conformers in order to obtain a more accurate energetic ranking of the molecules, and they were run in CP2K[161] with the PBE potential,[174] combined with the TZVP basis set,[226] D3 dispersion corrections and a plane-wave cut-off of 350 Ry.

4. The formation energies of all the organic cages were calculated with Equation 2.37 and compared to rationalise which precursors were more likely to generate stable and shape persistent cages. Formation energies alone were not anticipated to be sufficient to predict whether or not a cage would cleanly form in a reaction, not least because formation energy alone does not include consideration of factors such as the translational, rotational and vibrational entropy of the systems in order to compare free energies, kinetic influences or external conditions. However, this combined experimental and computational large scale HT study provided the opportunity to test how informative formation energies were for this purpose.

In order to obtain more information about the stability of the imine cages, imine strain values were calculated with a script developed in the group.* This information was thought to be useful as evidence was found in previous works[227] that high imine strain in porous organic cages is often correlated with more unstable structures or low yields products. Furthermore, cages structurally deviating from the ideal geometry are characterised by higher strain and are therefore less likely to form. Initially, the script analysed the crystal structures of imine cages published in the literature and deposited in the Cambridge Structural Database (CSD) as of February 2016, which consisted of 24 cages.[21,22,33,228] For those cages, all the dihedral angles containing imine bonds were averaged; the value obtained was considered the optimal value for an imine bond in a porous organic cage and used as a reference. The optimal value averaged to 177.1°, with a standard deviation of 2.8° for the selection of 24 cages available. An example of a dihedral angle containing an imine bond (C^1-C^2=N^3-C^4) is shown in Figure 6.6. Subsequently, the script was applied to the cages being investigated in this project to calculate the average of all the dihedral angles containing imine bonds. These were then compared with the optimal value, with the deviation from it calculated as a percentage using the following equation:

$$Strain = 100\left|\left(\frac{\mathbf{Cage}_{OPT} - \mathbf{Cage}_{AVG}}{\mathbf{Cage}_{OPT}}\right)\right| \tag{6.1}$$

where $Cage_{AVG}$ is the average dihedral value for the cage analysed and $Cage_{OPT}$ is the optimal value. Based upon the data available at the time the script was written, cages with *Strain* percentage (considered as the percentage deviation from the ideal value) >2% were considered strained, and the same criteria will be adopted

*The script that calculates strain was written in Python by Dr Enrico Berardo. The crystal structures of recently published imine cages, necessary for the strain evaluation, were collected by Marcin Miklitz.

throughout this chapter.

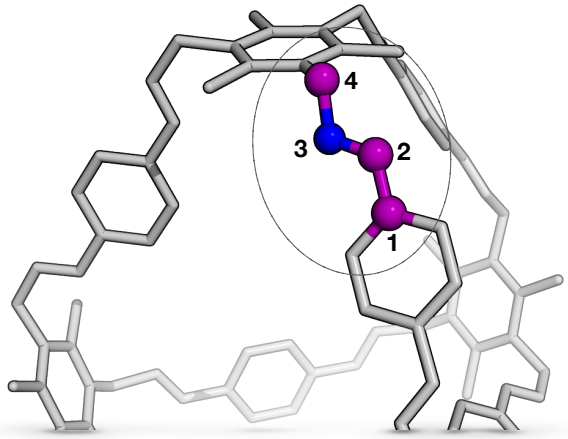


Figure 6.6.: Example of a dihedral angle (atoms 1 to 4) containing an imine bond. All atoms in the example cage (B11) are in grey; only atoms forming part of the dihedral are coloured, carbons in purple and nitrogen in blue.

6.3. Results

In this section, I will describe the results I obtained from the analysis of all the cages assembled. First I am going to discuss the three cases in which a full topology prediction was performed (B1, B11, B23). Then, I will discuss the results regarding all cages A/B/C, including those that displayed unexpected behaviour and that formed with a topology different from the initially targeted one. Finally, I will briefly discuss which criteria were considered desirable for the synthesis of novel porous organic cages, and the triply interlocked pretzelane topology.

6.3.1. Topology prediction for cages B1, B11, B23

The simulations on this selection of cages were performed to understand whether the topologies that were initially targeted were the most likely outcome of the reaction. A scheme representing the starting precursors for B1, B11 and B23, and some of the possible topologies arising from each combination are shown in Figure 6.7. In the next few subsections I will analyse each case separately.

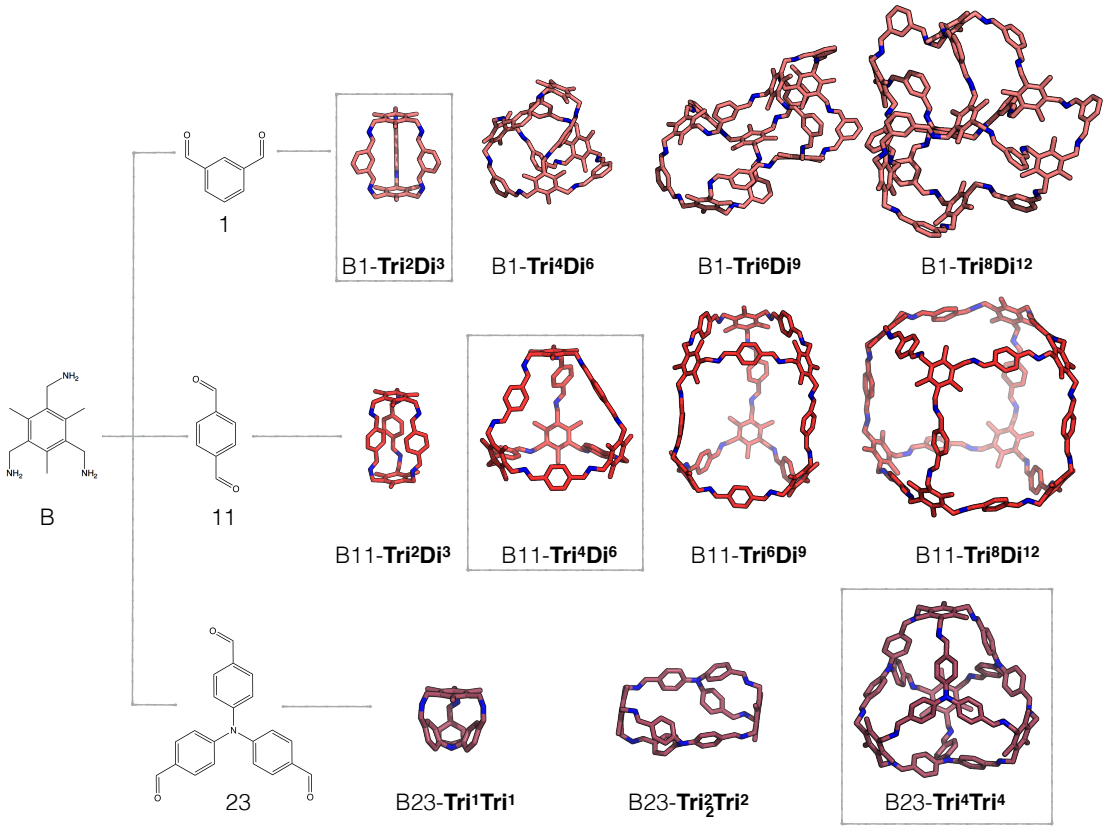


Figure 6.7.: Possible topologies arising from the combination of triamine B and aldehydes 1, 11 and 23. The topologies in the boxes are those targeted in the initial plan.

Cage B1

Precursors B and 1 were initially mixed into the $\mathbf{Tri^2Di^3}$, $\mathbf{Tri^4Di^6}$, $\mathbf{Tri^6Di^9}$ and $\mathbf{Tri^8Di^{12}}$ topologies, with the expectation that the smallest topology, $\mathbf{Tri^2Di^3}$, would form. After the initial formation step, and prior to any conformer search analysis, the four molecules were minimised with the OPLS3 force field. Only the smallest size was found to be shape persistent and all the other topologies collapsed immediately. The energies of the molecules were compared to study their relative stability and a plot summarising the results is shown in Figure 6.8.

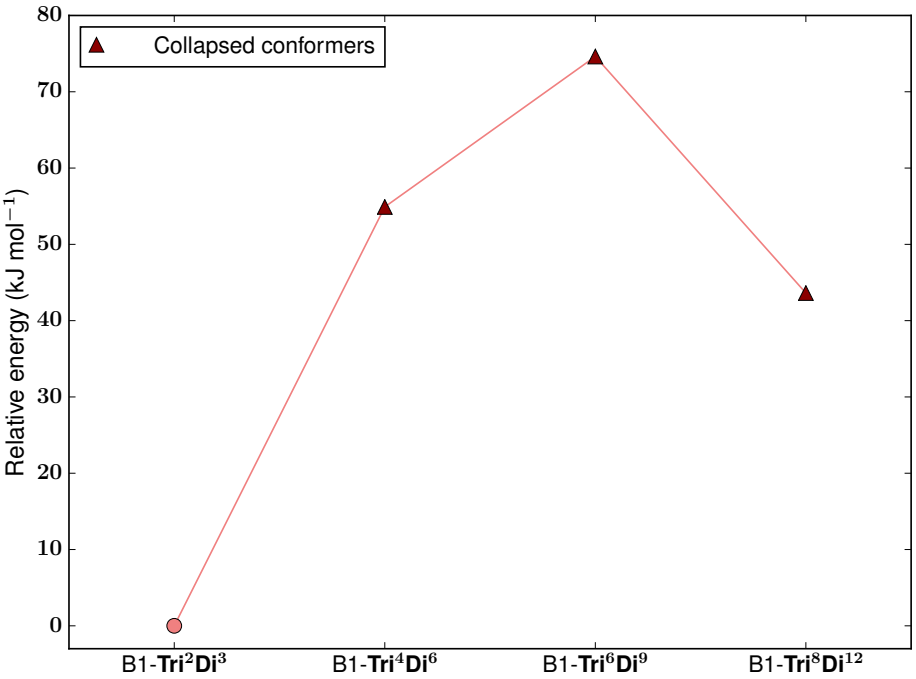


Figure 6.8.: B1: relative energies of the different topologies calculated at the PBE/TZVP+D3 level. All energies are calculated with respect to the **Tri^{2}Di^3** topology.

B1-**Tri^{2}Di^3** was the topology with the lowest relative energy and therefore predicted to be the most experimentally feasible. All the larger topologies collapsed and their relative energies were found to be vastly higher than the **Tri^{2}Di^3** ones (more than 40 kJ mol^{-1} in all cases). Open conformers of the collapsed topologies were not searched for, as their energies would have been even higher than their collapsed equivalents, and therefore not interesting for a prediction purpose. The finding that **Tri^{2}Di^3** was lowest in energy was in agreement with the experiments, where a cage formed by precursors B and 1 in a [2+3] ratio was identified by ^{1}H NMR and FTIR, and isolated. An X-ray diffraction crystal structure of cage B1 was obtained from the experiments, and an overlay of the two geometries is shown in Figure 6.9. The simulation accurately reproduced the reported crystal structure, with an RMSD value of 0.25 Å, and, with the same procedure adopted in the previous chapters, was then minimised with the OPLS3 force field to exclude the crystal packing effects that the simulations do not take into account, leading to a perfect match (RMSD = 0.00 Å).

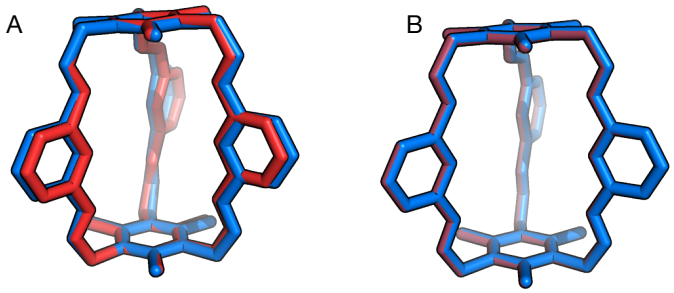


Figure 6.9.: B1-**Tri**^{2}**Di**^{3}: A) overlay of the simulation (blue) and the crystal structure (red); B) overlay of the simulation (blue) and the OPLS3-minimised crystal structure (dark red).

Cage B11

The equivalent set of simulations as for B1 were carried out for B11. Triamine B was mixed with aldehyde 11 into the **Tri**^{2}**Di**^{3}, **Tri**^{4}**Di**^{6}, **Tri**^{6}**Di**^{9} and **Tri**^{8}**Di**^{12} topologies, this time expecting the tetrahedral **Tri**^{4}**Di**^{6} topology to form. The relative energies of these four topologies are compared in Figure 6.10.

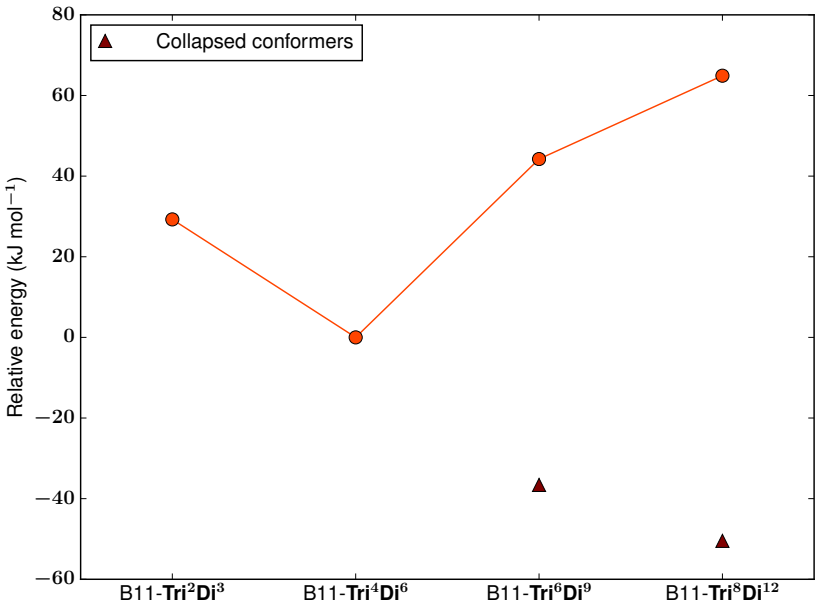


Figure 6.10.: B11: relative energies of the different topologies calculated at the PBE/TZVP+D3 level. All energies are calculated with respect to the **Tri**^{4}**Di**^{6} topology. Circles represent open conformers, whereas triangles collapsed conformers.

When conformer search methods were applied, the two smaller topologies, **Tri**^{2}**Di**^{3} and **Tri**^{4}**Di**^{6}, retained a void, whereas the two larger ones, **Tri**^{6}**Di**^{9} and **Tri**^{8}**Di**^{12}, collapsed via a vertex-folding mechanism, as shown in Figure 3.1. Nevertheless, it

was possible to locate two higher energy open conformers for both topologies after a preliminary minimisation of the initial guess structures, to optimise them with DFT methods and to compare their energy with those of the smaller topologies. As it can be deduced from Figure 6.10, $\mathbf{Tri}^4\mathbf{Di}^6$ was more than 29.3 kJ mol^{-1} lower in energy than all the other open conformers and, therefore, the most probable synthetic outcome. A result that, again, is in agreement with experimental findings. In this case as well, a crystal structure was obtained and could be compared with simulations; the overlays are shown in Figure 6.11. The match between the two geometries is reasonably good and the comparison results in an RMSD value of 1.07 Å. The differences between the two geometries are mainly given by the many different positions the teraphtaldeyde can adopt in the assembled structure, which provides cage B11 with a wide conformational freedom that was confirmed by the presence of large thermal ellipsoids in most non-hydrogen atoms within the crystal structure. Minimising the crystal structure left the RMSD unchanged.

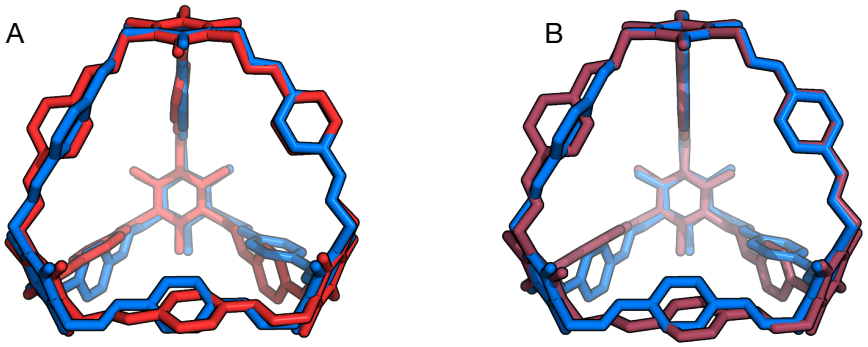


Figure 6.11.: B11-$\mathbf{Tri}^4\mathbf{Di}^6$: A) overlay of the simulation (blue); the crystal structure (red) and B) overlay of the simulation (blue) and the OPLS3-minimised crystal structure (dark red).

Cage B23

Finally, triamine B was mixed with the tritopic aldehyde 23 according to three different topologies, $\mathbf{Tri}^1\mathbf{Tri}^1$, $\mathbf{Tri}_2^2\mathbf{Tri}^2$ and $\mathbf{Tri}^4\mathbf{Tri}^4$, with the expectation of finding the last more stable. Relative energies calculated at the DFT level are compared in Figure 6.12. As predicted, $\mathbf{Tri}^4\mathbf{Tri}^4$ was found to be lowest in energy, in agreement with experimental findings that identified the successful formation of the cage. The large energy difference that separates the tetrapod from the other two topologies (>118 kJ mol^{-1} from the $\mathbf{Tri}^2\mathbf{Tri}^2$ topology and >350 kJ mol^{-1} from the smallest $\mathbf{Tri}^1\mathbf{Tri}^1$), in a behaviour similar to the one observed in Section 5.4.2, is most prob-

ably a result of the high level of strain that characterises both the **Tri**^{1}**Tri**^{1} and **Tri**^{2}**Tri**^{2} topologies, the geometries of which are shown in Figure 6.7.

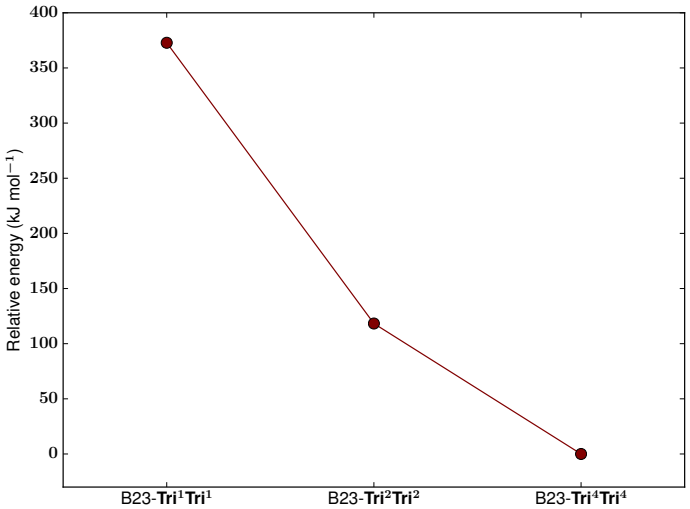


Figure 6.12.: B23: relative energies of the different topologies calculated at the PBE/TZVP+D3 level. All energies are calculated with respect to the **Tri**^{4}**Tri**^{4} topology.

Strain values were calculated for all cages and, whilst the largest tetrapod topology was characterised by a low level of strain (1.5%), B23-**Tri**^{1}**Tri**^{1} and B23-**Tri**^{2}**Tri**^{2} were assembled into very strained geometries associated with strain values of 5.5% and 7.0% respectively. A crystal structure was isolated for B23 too, so I overlaid it with my simulation in Figure 6.13 and identified a good match, with a RMSD value of 1.12 Å, which was improved to 0.94 Å upon force field minimisation of the crystal structure.

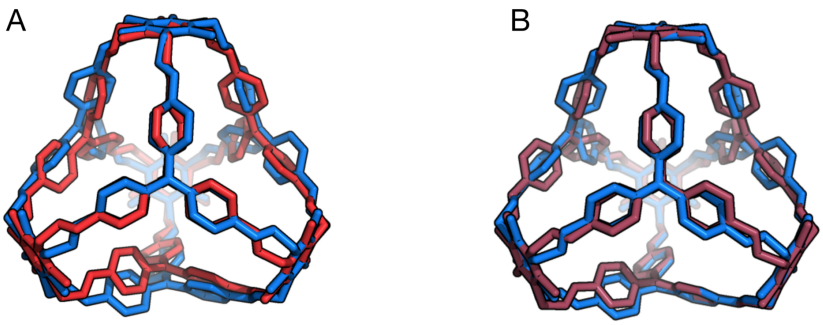


Figure 6.13.: B23-**Tri**^{4}**Tri**^{4}: A) overlay of the simulation (blue) and the crystal structure (red); B) overlay of the simulation (blue) and the OPLS3-minimised crystal structure (dark red).

6.3.2. HT screening - discussion of the results.

After the simulation procedure was tested on cages B1, B11 and B23, all the imine cages shown in Figures 6.3, 6.4, and 6.5 were assembled and computationally investigated with the methods discussed in Section 6.2. In this section I will discuss which cages collapsed and which retained a cavity, and whether this is an agreement with experimental results. In order to make the large amount of data generated clearer, results are summarised in the form of colour maps in Figure 6.14, with plot 1 showing the outcome of the synthesis, plot 2 reporting the formation energies per bond formed of the cages assembled into the targeted topologies and plot 3 showing the void diameters of the cages synthetically formed. The geometries of all the cages experimentally formed are shown in Figure 6.15.

Considerations based on the comparison between experimental and computational results

From plot 1 in Figure 6.14, it is possible to see that 33 new cages in total formed cleanly, according to ^{1}H NMR spectroscopy and HRMS. Additionally, there were 16 further cases where partial cage formation was observed and where, in addition to the cage, oligomeric products or unreacted aldehyde was present. In the cases in which the cages formed, there was good agreement between the targeted topology and those obtained experimentally, including the three cases in which simulations predicted cages A13, A14 and A17 to collapse, which could correspond to either collapse upon desolvation of the experimental structure,[96] or to a cage not forming at all. From a computational point of view, in all cases, the formation energy per imine bond was found to be negative, indicating that in principle cage formation could occur, however, some cages were noticeably less energetically favoured than others. With the exception of three of the $\mathbf{Tri}^{4}\mathbf{Di}^{6}$ cages (A13, A14 and A17), all of the hypothetical cages were predicted to be shape persistent.

At first glance, one can notice from plots 1 and 2 in Figure 6.14 that cages containing triamine A either did not form or formed with impurities (except for one case, A11), behaviour reflected by the formation energy values that are always slightly higher than those containing triamines B or C. Triamine A conferred increased flexibility to the final structures and did not offer the same degree of pre-organisation B and C did, due to the lack of additional methyl or ethyl groups, which, being more bulky, can affect the way molecules assemble within the cage.[229] Cages containing precursors B and C had a similar level of success in the HT screening, which could

imply that pre-configuration alone does not necessarily lead to cage formation and that rigidity of the precursors could offer a higher degree of success.

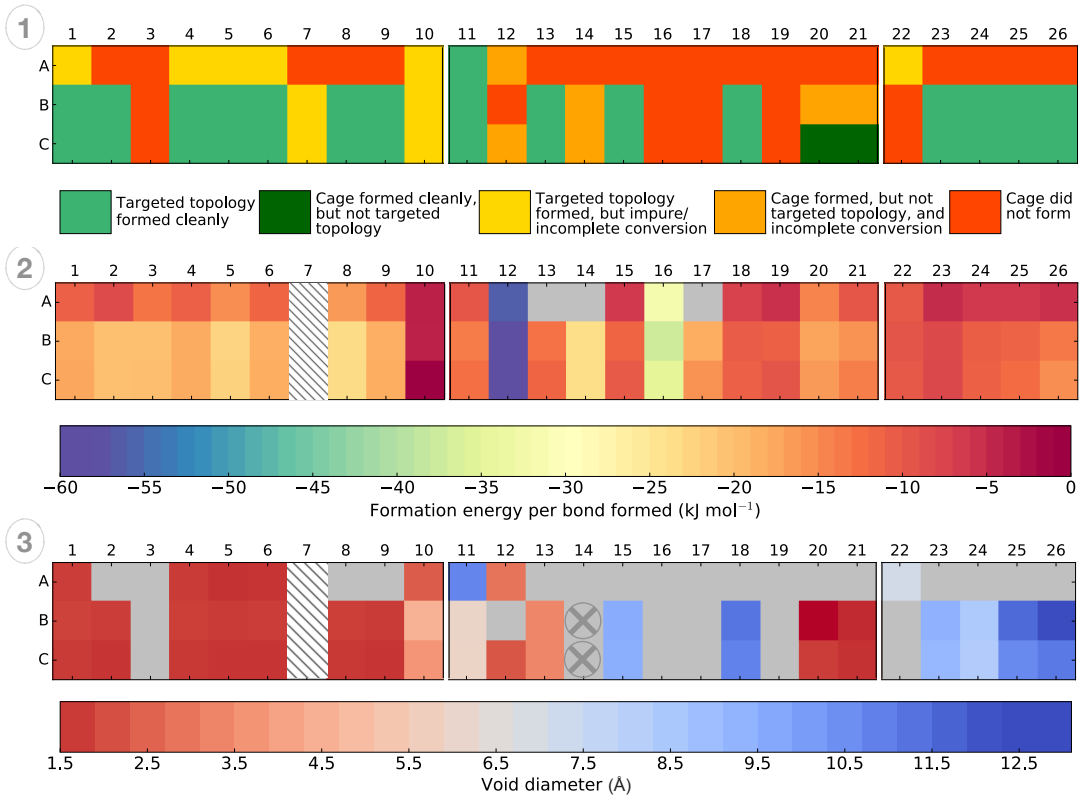


Figure 6.14.: Colour maps summarising the results obtained by the HT screening. 1) Description of the experimental outcome as per the legend below the colour map. 2) Colour map showing the formation energies per bond formed (calculated at the PBE/TZVP +D3 level) of each simulated cage assembled into the targeted topology; warmer colours represent more unstable cages, colder colours more stable cages, whereas grey squares are for cages that collapsed following high temperature MD and therefore their formation energy was not calculated. 3) Colour map showing the computationally calculated void diameters of the cages that successfully formed experimentally; shades of red represent smaller cavities, shades of blue larger cavities, grey squares cages that did not experimentally form, and crossed grey squares cages that formed but did not retain a cavity. In both plots 2) and 3), the squares corresponding to cages containing precursor 7 are barred, as simulations were not performed.

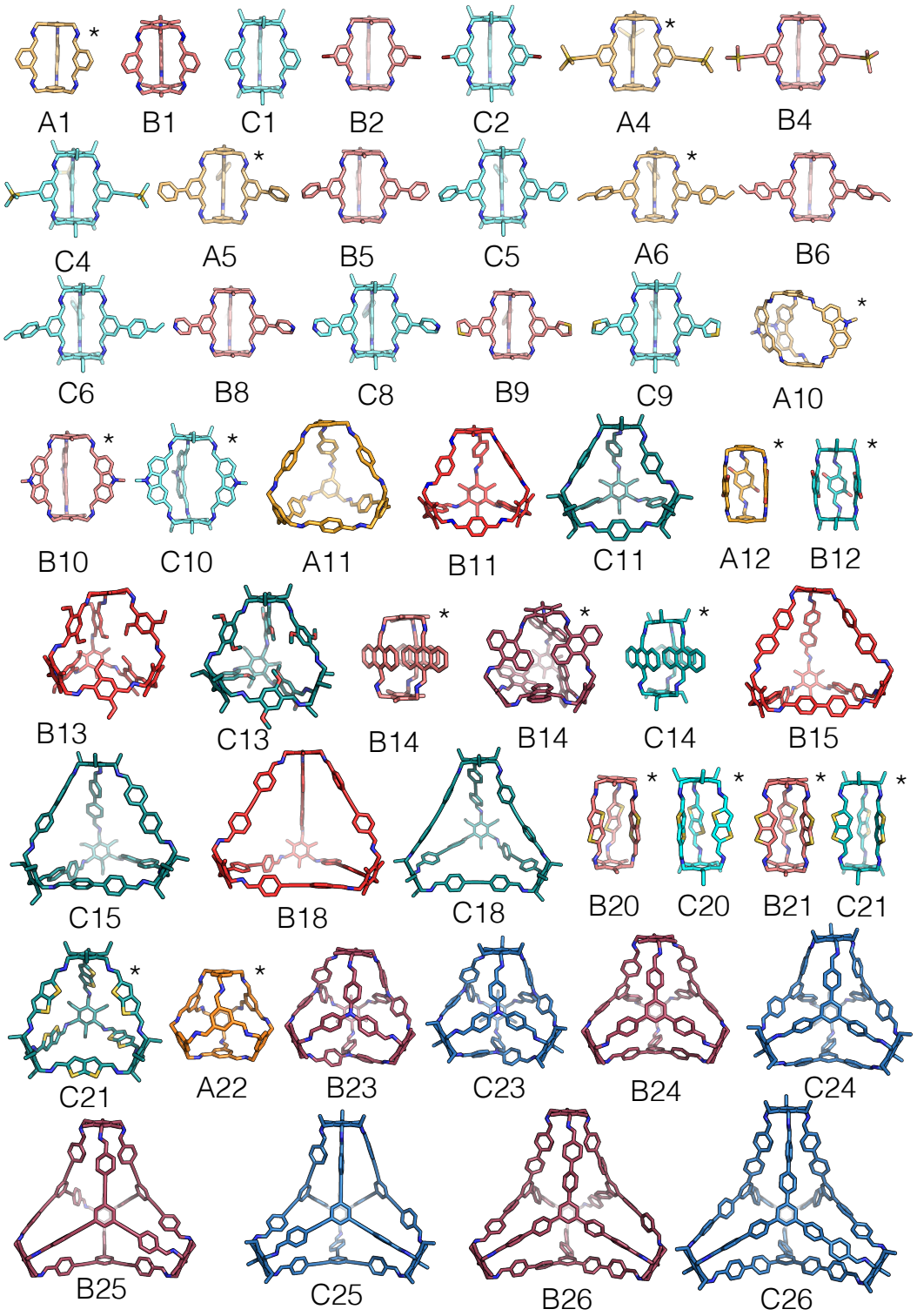


Figure 6.15.: Organic cages that formed as a result of the HT screening, optimised at the PBE/TZVP +D3 level. The colour code is as per Figures 6.3, 6.4 and 6.5. Cages marked with a star are those that formed partially, with the presence of residual aldehyde starting material or other oligomeric side-products.

6.3.3. Cages with Tri^2Di^3 topology

In general, cages with a $\mathbf{Tri^2Di^3}$ topology (containing precursors 1-10) were found to have similar formation energies with the exception of cages containing aldehyde 10, which were less energetically favoured. A smaller colour map containing only the formation energies of $\mathbf{Tri^2Di^3}$ cages was generated in order to better appreciate the difference in energy between cages (see Figure 6.16). Additionally, all the results regarding the values of the formation energies per bond, the strain, the void sizes, and information about experimental findings for $\mathbf{Tri^2Di^3}$ cages are reported in Table 6.1.

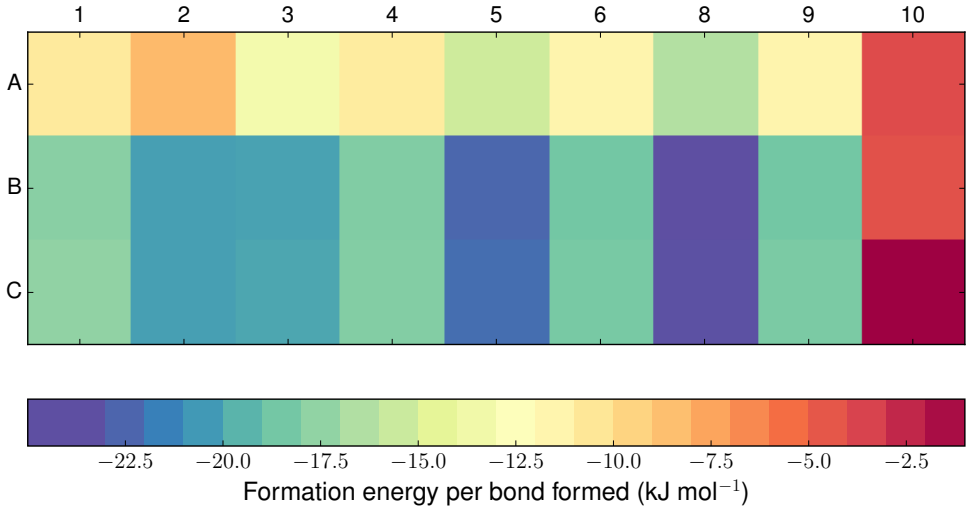


Figure 6.16.: Colour map containing the formation energies per bond formed calculated at the PBE/TZVP +D3 level, of cages with $\mathbf{Tri^2Di^3}$ topology.

As already mentioned, cages containing triamine A were less favourable than cages containing triamines B and C and therefore, cages of kind A were either considered less likely to form or harder to form as a pure product. This is in agreement with the experimental findings, which showed that cages of kind A either did not form at all or, if they did, contained some impurities or were part of a complex mixture. From a computational point of view only, it was possible to locate stable and symmetric conformers for all the cages with this topology, except for the cases of cages A10 and C10 that had very distorted, partially collapsed, geometries (A10, in particular, had a geometry more strained than all the others - as shown in Figure 6.15). This was mirrored by the experiments, in which difficulties were encountered to isolate pure products for cages containing aldehyde 10.

B and C cages containing precursors 1-4 were compared to understand the effect

of different functionalities (see box **c** in Figure 6.2). Computationally, the formation energies of these cages only differed by ∼2 kJ mol^{-1}, an indication that the peripheral modifications to the cage did not influence the formation energy of the final outcome. Only in one case did a disagreement between simulations and experiments occur, and it regarded cages containing precursor 3. Even though simulations predicted cages containing 3 to form symmetric shape-persistent conformations, and formation energies and strain values to be in the same range as cages 1-10 (which mostly formed), they did not form experimentally. This behaviour could be linked to experimental conditions I could not account for in my calculations, such as concentration, temperature, solvent, addition of a catalyst, removal of water during the reaction, rate and order of reagent addition.
As one can see from plot 3 in Figure 6.14, most $\mathbf{Tri^2Di^3}$ cages retained small cavities, although, because of their small elongated tubular shapes, all were below 1.8 Å in diameter, therefore too small to host any guests. Derivatisation of aldehyde 1 did not affect the cavity size, as the functionality was on the external periphery of the cage. Cages containing precursor 10 retained larger cavities, as a result of 10 having a larger carbazole as the central unit between the aldehyde substituents rather than a benzene ring as in 1.

Some crystal structures were available for comparison with the modelling. The crystal structure for cage B1 was already reported in Figure 6.9, however crystal structure geometries were also available for cages C1 (Figure 6.17), B2 and C2 (Figure 6.18). The simulation reproduced the crystal structure extremely well, matching the experiments with an RMSD of 0.12 Å, which did not change upon minimisation of the crystal structure, probably because of some disorder of the ethyl groups. The simulations of cages B2 and C2 reproduce crystal structures as well, with RMSD values of 0.32 Å (improved to 0.07 Å upon minimisation of the crystal structure geometry) and 0.21 Å (improved to 0.07 Å upon minimisation of the crystal structure geometry) respectively.

Table 6.1.: Formation energies per bond formed, strain values and void diameters of cages with **Tri^{2}Di^3** topology (A/B/C 1-10). Orange cells refer to simulated structures that were very distorted at the end of the computational procedure; cyan cells highlight structures with a strain percentage greater than 2.

Cage	Formation energy per bond formed (kJ mol^{-1})	Strain (%)	Void diameter (Å)
A1 - **Tri^{2}Di^3**	-11	0.1	1.7
B1 - **Tri^{2}Di^3**	-18	0.1	1.8
C1 - **Tri^{2}Di^3**	-18	0.1	1.7
A2 - **Tri^{2}Di^3**	-8	0.9	4.6
B2 - **Tri^{2}Di^3**	-20	0.1	1.7
C2 - **Tri^{2}Di^3**	-20	0.2	1.6
A3 - **Tri^{2}Di^3**	-13	0.1	1.6
B3 - **Tri^{2}Di^3**	-20	0.1	1.8
C3 - **Tri^{2}Di^3**	-20	0.2	1.7
A4 - **Tri^{2}Di^3**	-11	0.1	1.7
B4 - **Tri^{2}Di^3**	-18	0.5	1.8
C4 - **Tri^{2}Di^3**	-18	0.2	1.6
A5 - **Tri^{2}Di^3**	-15	0.2	1.6
B5 - **Tri^{2}Di^3**	-22	0.1	1.7
C5 - **Tri^{2}Di^3**	-22	0.2	1.6
A6 - **Tri^{2}Di^3**	-12	0.1	1.6
B6 - **Tri^{2}Di^3**	-19	0.3	1.7
C6 - **Tri^{2}Di^3**	-18	0.1	1.6
A8 - **Tri^{2}Di^3**	-16	0.1	1.5
B8 - **Tri^{2}Di^3**	-24	0.1	1.8
C8 - **Tri^{2}Di^3**	-23	0.1	1.6
A9 - **Tri^{2}Di^3**	-12	0.1	1.6
B9 - **Tri^{2}Di^3**	-19	0.1	1.7
C9 - **Tri^{2}Di^3**	-18	0.2	1.6
A10 - **Tri^{2}Di^3**	-4	2.4	2.3
B10 - **Tri^{2}Di^3**	-4	0.2	4.5
C10 - **Tri^{2}Di^3**	-1	0.3	3.7

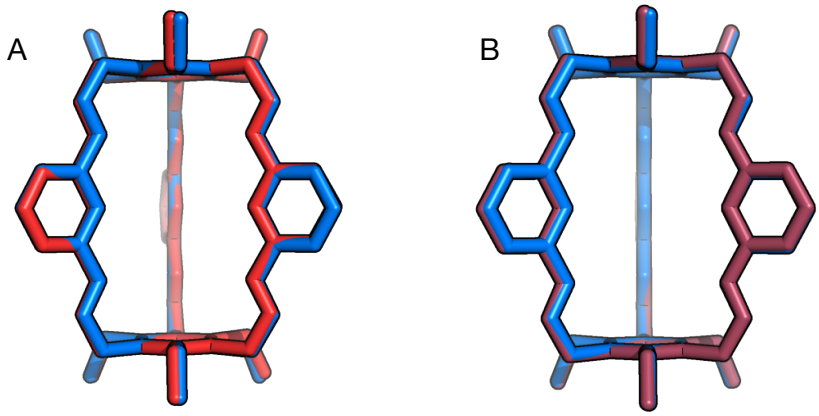


Figure 6.17.: C1: A) overlay of the simulation (blue) and the crystal structure (red); B) overlay of the simulation (blue) and the OPLS3-minimised crystal structure (dark red).

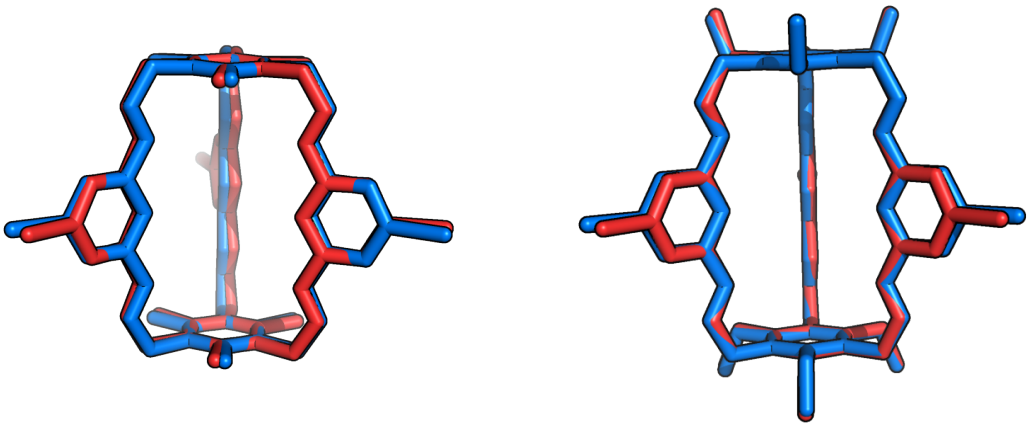


Figure 6.18.: Overlays of cages B2 (left) and C2 (right). The simulations are in blue, the crystal structures in red.

Among cages containing longer aldehyde groups (5-9), cages (B/C)5 and 8 seemed to be the most likely to form, with formation energies lower than -22 kJ mol^{-1}. Nonetheless, B and C cages containing precursors 6 and 9 were only ∼5 kJ mol^{-1} higher than the others, and were still able to form. In fact, a crystal structure was available for cages B9 and C9, which are overlaid with the simulations in Figure 6.19. The experiments were reproduced in both cases, and the reported RMSD values were 0.34 Å (reduced to 0.12 Å when minimising the crystal structure) and 0.52 Å (slightly improved to 0.48 Å when minimising the crystal structure) for B9 and C9 respectively.

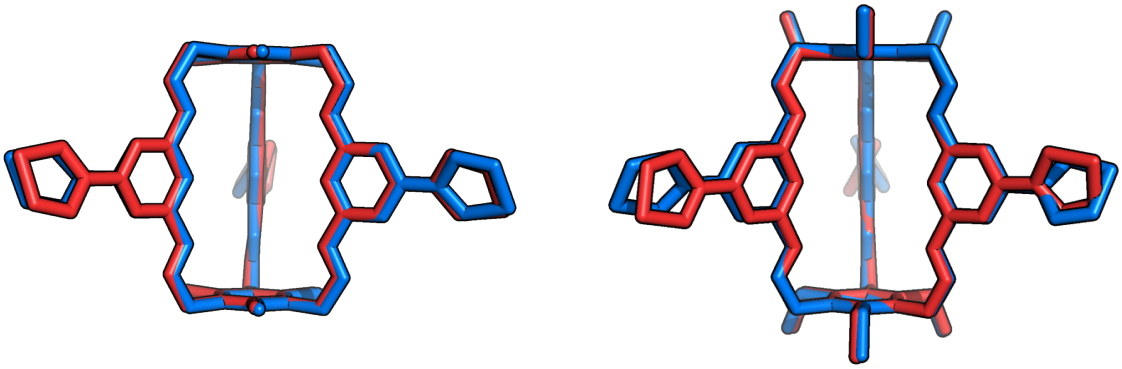


Figure 6.19.: Overlays of cages B9 (left) and C9 (right). The simulations are in blue, the crystal structures in red.

Summarising, cages with **Tri**^{2}**Di**^{3} topology were generally found to be the least strained and were therefore more frequently synthesised, with 14 out of 30 reactions resulting in clean cage formation, and 9 out of 30 reactions in partial formation alongside residual starting material or potential oligomers respectively. Overall, however, it was observed that the formation energies alone could not predict whether a cage would have been successfully obtained experimentally, as the energy gaps between cages were not large enough to suggest the formation of a specific outcome. For example, cages such as (A/B/C)10 were found to have the least favourable formation energies, with relatively high strain around the imine bonds in the cages, and yet they were all observed to form.

6.3.4. Cages with Tri^{4}Di^6 topology

Cages with larger topologies were on average less favourable than the smaller cages. In addition, cages with **Tri**^{4}**Di**^{6} topology showed additional complications that led to some cages (12, 14, 20, 21) forming according to the smaller **Tri**^{2}**Di**^{3} topology, making them the hardest to both investigate computationally and form/characterise experimentally. The values of the formation energies per bond formed, information about strain and void diameters are reported in Table 6.2. In order to make the analysis of cages with a **Tri**^{4}**Di**^{6} topology easier, I will divide the discussion into smaller sections and in each I will analyse the different criteria that might have contributed (positively or negatively) to the formation of such cages.

Table 6.2.: Formation energies per bond formed, strain values and void diameters of cages with **Tri**^{4}**Di**^{6} topology (A/B/C 11-19).

Cage	Formation energy per bond formed (kJ mol^{-1})	Strain (%)	Void diameter (Å)
A11 - **Tri**^{4}**Di**^{6}	-10	0.2	10.8
B11 - **Tri**^{4}**Di**^{6}	-14	0.9	6.0
C11 - **Tri**^{4}**Di**^{6}	-13	1.4	6.0
B13 - **Tri**^{4}**Di**^{6}	-13	1.1	3.4
C13 - **Tri**^{4}**Di**^{6}	-12	1.6	3.4
A15 - **Tri**^{4}**Di**^{6}	-7	0.6	11.5
B15 - **Tri**^{4}**Di**^{6}	-12	0.8	9.6
C15 - **Tri**^{4}**Di**^{6}	-10	1.4	9.5
A16 - **Tri**^{4}**Di**^{6}	-32	1.0	12.1
B16 - **Tri**^{4}**Di**^{6}	-37	1.0	9.7
C16 - **Tri**^{4}**Di**^{6}	-35	1.5	9.2
B17 - **Tri**^{4}**Di**^{6}	-18	0.6	7.7
C17 - **Tri**^{4}**Di**^{6}	-16	0.9	8.4
A18 - **Tri**^{4}**Di**^{6}	-7	0.4	14.8
B18 - **Tri**^{4}**Di**^{6}	-10	0.6	11.2
C18 - **Tri**^{4}**Di**^{6}	-11	0.9	10.8
A19 - **Tri**^{4}**Di**^{6}	-6	0.4	13.1
B19 - **Tri**^{4}**Di**^{6}	-11	0.6	12.6
C19 - **Tri**^{4}**Di**^{6}	-10	1.1	12.9

Cages containing precursors 11,15,18,19 - increasing the length of the aldehyde

As previously discussed, and as shown in box **b** of Figure 6.2, the effect of increasing the length of the aldehyde precursor was studied. As can be noticed from plot 3 in Figure 6.14, with increasing length of the ditopic aldehyde, the cavity size increased where cages were formed, as anticipated (with cage C18 retaining the largest cavity of 11.1 Å of diameter). However, there was a limit to this improvement, as no cage was formed for the largest precursor 19, nor did A15 or A18 in the case of the more flexible

triamine A. This suggests that increasing the size of the dialdehyde made cage formation increasingly difficult, which could be a result of increasing flexibility in the cage as it formed, and, in the case of 19, a mismatch in the preferred orientations due to the aromatic rings not being able to adopt the conformer required for cage formation, or premature precipitation of cage fragments due to poor solubility of the terphenyl containing fragments. This behaviour could not have been predicted by calculating the relative stability of the cages, nor their strain, as none of the cages lost their cavity according to simulations, the gap between the formation energies never being greater than 2 kJ mol^{-1}, and no cage showing a high level of strain.

Among the cages discussed in this section, two could be isolated and formed a crystal structure, B11 and B15. The overlays between crystal and simulated structure for B11 were shown already in Figure 6.11, whereas those for the latter are shown in Figure 6.20. The simulation well reproduced the shape of cage B15 and the shape of the cavity, with an RMSD of 1.23 Å, although the flexibility of the rings in precursor 15 does not lead to a more accurate geometrical match, as larger displacement ellipsoids show in the original crystal structure. The RMSD is slightly improved to 0.97 Å when the crystal structure is minimised.

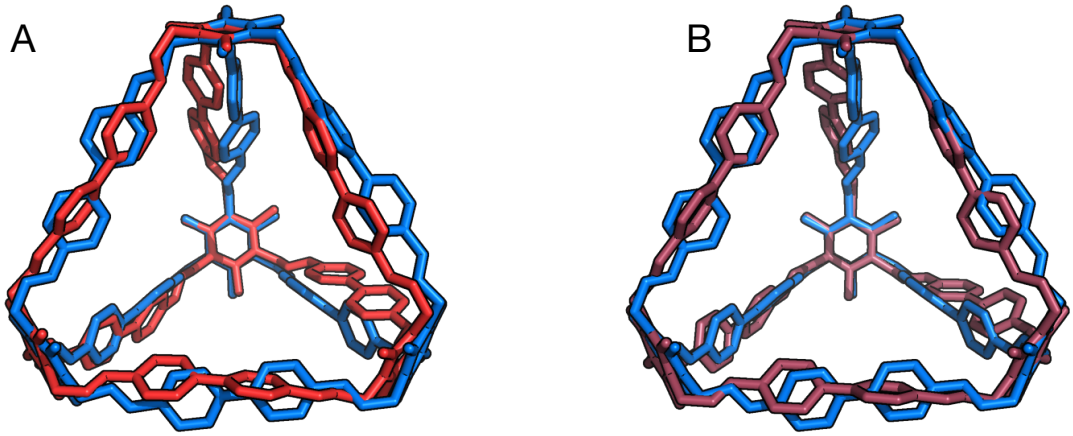


Figure 6.20.: B15: A) overlay of the simulation (blue) and the crystal structure (red); B) overlay of the simulation (blue) and the OPLS3-minimised crystal structure (dark red).

Cages containing precursors 11,12,13 - aldehydes with group addition/ substitution

The effect of substitution was interesting to investigate within cages with $\mathbf{Tri^4Di^6}$ topology. Cages containing precursors 11, 12 and 13 only differ by the way they are functionalised on the ortho position with respect to the aldehyde, 11 containing

just a hydrogen, 12 a hydroxyl group and 13 a methoxy group. Formation energies show that cages containing precursor 12 are particularly stable, almost certainly because of the formation of hydrogen bonds (H-bonds) between the hydrogen of the hydroxyl group and the imine nitrogen centre. Interestingly, the cages containing the bulkier substituents formed an alternative topology, (A/C)12 into a $\mathbf{Tri}^2\mathbf{Di}^3$ topology (which will be discussed in detail later in the chapter) and (B/C)13 according to a $\mathbf{Tri}^4\mathbf{Di}^6$ topology prior to isolation (as identified by ^{1}H NMR and HRMS) and then re-equilibrated to form a triply interlocked $\mathbf{Tri}^8\mathbf{Di}^{12}$ topology ("pretzelane" which will be discussed in Section 6.3.7), during recrystallisation studies. The reasoning behind these two mechanism is still partially unclear and the discussion will be further expanded in Section 6.3.7. Experimental analysis conducted prior to isolation showed that cages B12 and A13 did not form.

Atom substitution - cages containing precursors 15,16,17

Cages containing precursors 15, 16 and 17 differed only by the atom in the ortho position with respect to one of the aldehyde groups (H in 15, OH in 16 and Cl in 17). Looking at the formation energies, cages containing precursor 16 seem to be particularly stable, the energies of the other cages being more than 20 kJ mol^{-1} higher (except for cage A17, which collapses). This is probably caused by the formation of stabilising H-bonds between the hydroxy and imino groups. Despite this, only cages (B/C)15 formed, none of the other cages assembling during the synthesis. This cannot be rationalised computationally and is probably related to experimental conditions, such as those mentioned earlier, which cannot be reproduced in my simulations.

Formation of alternative topologies - cages containing precursors 12, 14, 20, 21

Contrary to the initial plan, some combination of precursors did not assemble according to the targeted topology, but into a smaller $\mathbf{Tri}^2\mathbf{Di}^3$ one. This involved cages (A/C)12, (B/C)14, (B/C)20 and (B/C)21, of which both targeted and final topologies are shown in Figure 6.21. Computationally, four additional $\mathbf{Tri}^2\mathbf{Di}^3$ cages were assembled and investigated, in order to understand why one topology formed over another. The relative energy difference, per [3+2] unit, between $\mathbf{Tri}^2\mathbf{Di}^3$ and $\mathbf{Tri}^4\mathbf{Di}^6$ topologies are listed next to each example in Figure 6.21, where negative differences are associated with more stable $\mathbf{Tri}^4\mathbf{Di}^6$ topologies and positive differ-

ences vice versa. The formation energies per bond formed, the strain percentages and the void diameters for the same cages are reported in Table 6.3.

Cages containing precursor 12

A and C cages containing precursor 12 are illustrated in the top left box of Figure 6.21, the $\mathbf{Tri}^2\mathbf{Di}^3$ topology on the left hand side, the $\mathbf{Tri}^4\mathbf{Di}^6$ on the right. ^{1}H NMR identified the formation of cages A12 and C12, then HRMS and diffusion NMR confirmed the assembly in solution of cages with a $\mathbf{Tri}^2\mathbf{Di}^3$ topology. Computationally, simulations were run on both topologies for each case; the molecules remained shape persistent after undergoing repeated MD simulations and retained cavities with an average diameter of ∼2.5 Å for the smaller topologies, and of 10.6 and 6.9 Å for A12 and C12 respectively. Comparing their internal relative energies shows that both cages A and C 12 were more stable in their $\mathbf{Tri}^4\mathbf{Di}^6$ topology, by 5 kJ mol^{-1} in the first case and 13 kJ mol^{-1} in the second. Therefore, calculating the internal energy does not help understand why the targeted topologies did not assemble. I then considered the strain percentage (although a high level of strain in the geometry normally reflects higher energies), but could not justify the formation of the smaller topologies, the values for the $\mathbf{Tri}^2\mathbf{Di}^3$ topologies (∼4.5%) being almost double those for the $\mathbf{Tri}^4\mathbf{Di}^6$ topologies.

The formation of cages A12 and C12 in the $\mathbf{Tri}^2\mathbf{Di}^3$ topology could have been driven by the favourable formation of H-bonds between the hydroxyl group of precursor 12 and the nitrogen of the imine bond, as illustrated by Figure 6.22. The presence of the intramolecular H-bond could have forced the direction of the imine bond, generating a geometrical asset more favourable for the smaller $\mathbf{Tri}^2\mathbf{Di}^3$ topology to form. The results of the DFT calculations do not allow me to prove this point, however, a possible way of obtaining more information would involve comparing the energies of the cages one step before the product was formed, rather than of the final product. That would allow one to compare the cost of forming a bond, rather than just internal energies.

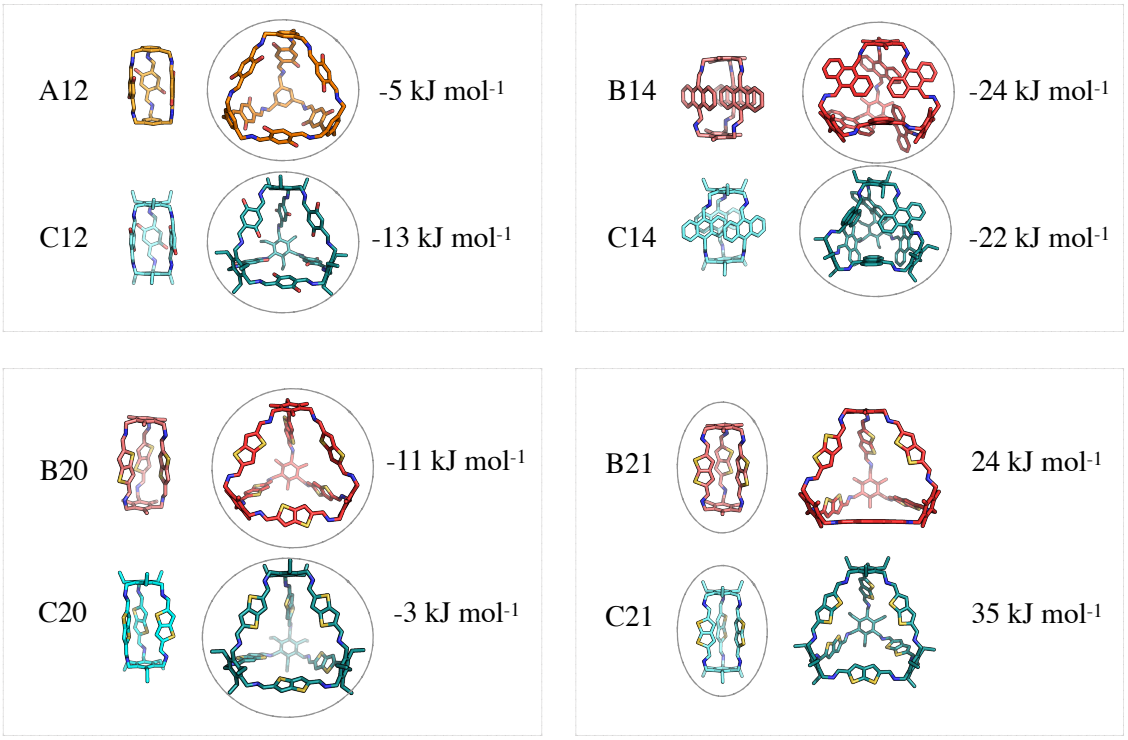


Figure 6.21.: Preference for **Tri**^{4}**Di**^{6} over **Tri**^{2}**Di**^{3} topology for cages (A/C)12, (B/C)14, (B/C)20 and (B/C)21 where a negative value indicates an energetic preference for **Tri**^{4}**Di**^{6} and a positive value an energetic preference for **Tri**^{2}**Di**^{3} topology (the preferred topologies according to the simulations are circled in grey for clarity). All were experimentally found to result in a **Tri**^{2}**Di**^{3}, rather than the planned **Tri**^{4}**Di**^{6} topology, with the exception of B14 and C21, in which both topologies were observed.

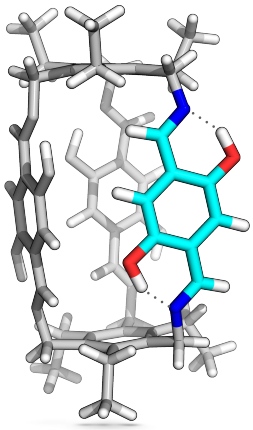


Figure 6.22.: Example of two intramolecular hydrogen-bonds forming in cage C12-**Tri**^{2}**Di**^{3} between the hydroxyl group in aldehyde 12 and the nitrogen of the imine bond. All atoms are in grey, except for the atoms involved in the hydrogen-bond (carbon is in cyan, oxygen in red and nitrogen in blue).

Table 6.3.: Formation energies, strain values and void diameters for the **Tri**^{2}**Di**^{3} and **Tri**^{4}**Di**^{6} topologies of cages (A/C)12, (B/C)14, (B/C)20 and (B/C)21. Cells in cyan highlight cages with a strain percentage value >2.

Cage	Formation energy per bond formed (kJ mol^{-1})	Strain (%)	Void diameter (Å)
A12 - **Tri**^{2}**Di**^{3}	-55	4.4	2.9
A12 - **Tri**^{4}**Di**^{6}	-56	0.2	10.6
C12 - **Tri**^{2}**Di**^{3}	-56	4.6	2.2
C12 - **Tri**^{4}**Di**^{6}	-58	2.1	6.9
B14 - **Tri**^{2}**Di**^{3}	-20	3.9	-
B14 - **Tri**^{4}**Di**^{6}	-24	1.8	3.1
C14 - **Tri**^{2}**Di**^{3}	-20	4.2	-
C14 - **Tri**^{4}**Di**^{6}	-24	1.2	3.3
B20 - **Tri**^{2}**Di**^{3}	-16	3.1	1.0
B20 - **Tri**^{4}**Di**^{6}	-17	1.0	8.7
C20 - **Tri**^{2}**Di**^{3}	-16	3.5	1.7
C20 - **Tri**^{4}**Di**^{6}	-16	1.4	8.6
B21 - **Tri**^{2}**Di**^{3}	-16	1.9	1.4
B21 - **Tri**^{4}**Di**^{6}	-16	1.0	8.5
C21 - **Tri**^{2}**Di**^{3}	-21	0.6	1.6
C21 - **Tri**^{4}**Di**^{6}	-14	1.4	8.2

The difference in energy between the two topologies is small enough to assume that other factors that I could not take into account might have influenced the outcome of the reaction, such as the formation of alternative outcomes, the influence of solvent interactions and many other experimental conditions. If two different topologies are predicted to be close in energy (less than ∼25 kJ mol^{-1}), then it could mean that the driving force towards a specific topology is not strong enough not be competitive with external experimental factors, and either topologies could be able to form. If the difference in energy is predicted to be larger, then there could be too strong a driving force for the topology landscape to be shifted to another outcome by external factors.

Cages containing precursor 14

Cages containing precursor 14 were originally planned for the synthesis of **Tri**^{4}**Di**^{6}

molecules, mainly because of the bulky core of the aldehyde that was thought not to fit into a smaller topology. Despite this reasoning, experimental investigations showed that cages B14 and C14 assembled into $\mathbf{Tri}^2\mathbf{Di}^3$ cages, the reaction pot of B14 in particular presenting a mixture of the $\mathbf{Tri}^2\mathbf{Di}^3$ and $\mathbf{Tri}^4\mathbf{Di}^6$ topologies, according to HRMS and diffusion NMR. Again, I analysed the formation energies and the strain of both topologies of cages B14 and C14 to try and explain the experimental results, all the values being reported in Table 6.3. In this case too, both the relative internal energies and the strain values would computationally predict the formation of the larger topology (which is favoured by 24 kJ mol^{-1} per [3+2] formula unit in the case of cage B, and by 22 kJ mol^{-1} per [3+2] formula unit in that of cage C), therefore it is again not intuitively clear why the smaller topology not only assembled, but they could also assemble in a crystalline phase when the larger could not. Looking at the geometries of the cages, which are shown in the top right box of Figure 6.21, the smaller cages' experimental formation could be driven by the π-stacking interactions forming between the anthracene units of the aldehyde, as shown in more detail in Figure 6.23, although the stabilising interaction was not reflected in more stable internal energies.

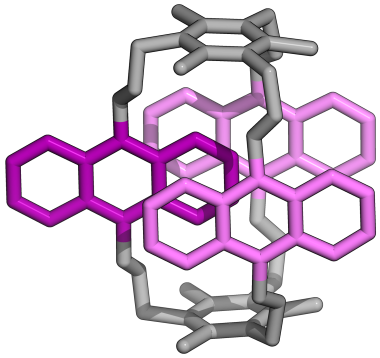


Figure 6.23.: π-stacking between the anthracene units of cage B14-$\mathbf{Tri}^2\mathbf{Di}^3$. All atoms are in grey, except for the aromatic rings involved in the π-stacking, which are in two different shades of purple.

The crystal structure of cage C14-$\mathbf{Tri}^2\mathbf{Di}^3$ was available for comparison, therefore experimental and simulated geometries were overlaid to see whether the experiments were well reproduced. The superposition is reported in Figure 6.24, where it is possible to see that the match between the two structures is reasonably good. The RMSD value between simulated and crystal structure geometries summed up to 1.74 Å (Figure 6.24-A), and it was further improved to 0.48 Å when the crystal structure was minimised to exclude the crystal packing effects neglected by the simulations

(Figure 6.24-B). Given the steric hindrance caused by the anthracene units, none of the cages containing precursor 14 retained a large cavity, the largest being that of cage C14-**Tri**^{4}**Di**^{6}, of 4.2 Å in diameter.

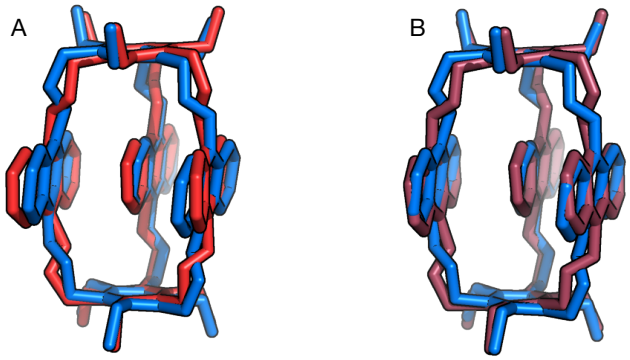


Figure 6.24.: C14-**Tri**^{2}**Di**^{3}: A) overlay of the simulation (blue) and the crystal structure (red); B) overlay of the simulation (blue) and the OPLS3-minimised crystal structure (dark red).

Cages containing precursor 20

Cages containing precursor 20 unexpectedly assembled into a **Tri**^{2}**Di**^{3} topology, although they were targeted for the larger **Tri**^{4}**Di**^{6} topologies. Simulations were run on both topologies for the two cages and the resulting lowest energy geometries are shown in the bottom left box in Figure 6.21.

All the cages showed shape persistency, the two larger molecules retaining a void of 8.7 Å (B20) and 8.6 Å (C20) in diameter. The smaller topologies did not have very large spherical voids, as they were characterised by a tubular-shaped void given by their intrinsic elongated shape. Computational simulations predicted the larger **Tri**^{4}**Di**^{6} topology to form, although the difference in relative energy per [3+2] unit is quite small in both cases, precisely -11 kJ mol^{-1} for B20 and -3 kJ mol^{-1} for C20. The strain for the smaller cages was quite large, but nonetheless the molecules were able to form in solution. That could have been because of an electronic delocalisation effect going on between the conjungated double bonds adjacent to the sulphur, that could have had an effect on the directionality of the imine bond.

Cages containing precursor 21

Cages containing precursor 21 showed a different behaviour again. Cages of kind A did not form, so only cages B and C have been analysed here, and their geometries and relative energy difference per [3+2] unit can be seen in Figure 6.21, whereas formation energies, strain values and void diameters are reported in Table 6.3. Ex-

perimentally, B21 only assembled into a **Tri**^{2}**Di**^{3} topology, whereas, interestingly, C21 formed a **Tri**^{2}**Di**^{3} capsule in the HT screen; then on scale-up and direct recrystallization of the reaction solution, a crystal structure of both a **Tri**^{2}**Di**^{3} capsule and a **Tri**^{4}**Di**^{6} tetrahedron was obtained (the first using CH_2Cl_2/MeCN, the second CH_2Cl_2/Et_2O). Edge-to-face π-stacking interactions were observed in the crystal structure of the **Tri**^{2}**Di**^{3} product; when the reaction was carried out in chloroform at reflux in the HT screen, the formation of the **Tri**^{2}**Di**^{3} capsule dominated. In contrast, in dichloromethane at reflux, a mixture of the **Tri**^{2}**Di**^{3} capsule and **Tri**^{4}**Di**^{6} tetrahedron was indicated by a more complex ^{1}H NMR spectra and determined by HRMS. The formation of different cage topologies from the same precursors by simply changing conditions has been observed previously and is another indication of the potential sensitivity of the topology to the experimental conditions. The crystal structures of both topologies of C21 have been overlaid with the simulated geometries, as shown in Figure 6.25, the RMSD being 0.75 Å for the smaller topology and 0.91 Å for the larger (the first reduced to 0.58 Å upon minimisation of the crystal structure geometry, the second to 0.90 Å). In this case, computationally, the study of the relative internal energies suggests a strong preference for the **Tri**^{2}**Di**^{3} topologies and therefore agree with experiments for the case of B21 (in which the smaller topology is preferred by 24 kJ mol^{-1} per [3+2] formula unit) and only partially for the case of C21, as such a big difference in energy would not suggest the formation of the larger tetrahedron (35 kJ mol^{-1} per [3+2] formula unit).

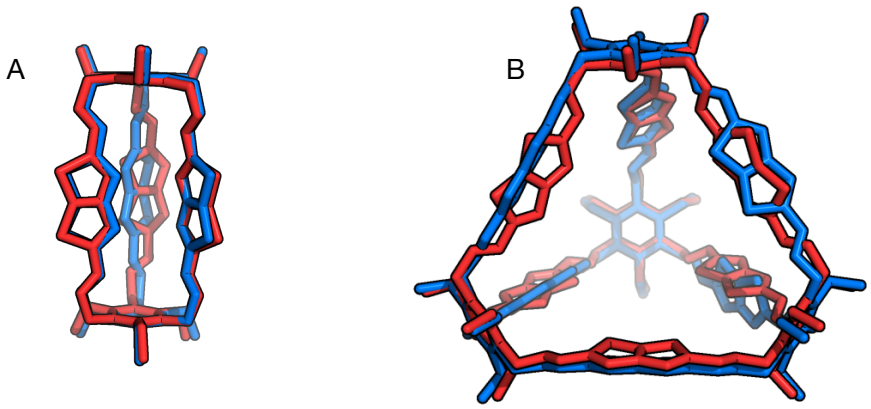


Figure 6.25.: A) overlay of the simulation (blue) and the crystal structure (red) of C21-**Tri**^{2}**Di**^{3}; B) overlay of the simulation (blue) and the crystal structure (red) of C21-**Tri**^{4}**Di**^{6}.

In general, all the cases where an alternative topology was formed were those where the $\mathbf{Tri}^4\mathbf{Di}^6$ topology was planned, which, one could conclude, had the smallest thermodynamic driving force for formation in the example system. Of the 33 reactions predicted to form $\mathbf{Tri}^4\mathbf{Di}^6$ tetrahedrons, 16 did not form cages, 9 formed $\mathbf{Tri}^4\mathbf{Di}^6$ tetrahedrons, 7 formed $\mathbf{Tri}^2\mathbf{Di}^3$ capsules and 1 formed a $\mathbf{Tri}^2\mathbf{Di}^3/\mathbf{Tri}^4\mathbf{Di}^6$ mixture. These cages, therefore, were experimentally found to be the most susceptible to other factors influencing the topological driving force, such as solvent choice, or modifications to the precursors. The lack of a strong driving force for a particular topology might also be a contributing factor as to why so many more of the $\mathbf{Tri}^4\mathbf{Di}^6$ targeted reactions failed to produce a cage compared to the $\mathbf{Tri}^2\mathbf{Di}^3$ and, as can be seen from the next section, $\mathbf{Tri}^4\mathbf{Tri}^4$ targeted reactions.

In all the cases that "break the pattern" and in which an alternative, unplanned, topology formed, it was not possible to computationally rationalise the causes upon the internal energies/ strain of the isolated cage molecules alone. In these cases other factors that I do not consider, such as the translational, rotational and vibrational entropy of the systems in order to compare free energies, kinetic influences, or external conditions (solvent interactions or the synthetic method used), can influence the topological outcome. This suggests that, in the computational design of a desired topology, an energetic preference of over 30 kJ mol^{-1} for every repeat unit in the structure should be the objective.

6.3.5. Cages with $\mathbf{Tri}^4\mathbf{Tri}^4$ topology

Triamines A/B/C were mixed with tritopic aldehydes 22-26 into $\mathbf{Tri}^4\mathbf{Tri}^4$ topologies. All the information about the formation energies per bond, the strain and the void sizes are listed in Table 6.4, whereas an additional colour map with the formation energies per bond formed has been added to better appreciate the energy differences in Figure 6.26.

In this case, the study is based on mixing the triamines with tritopic precursors of increasing size, to study the possible formation of tetrapods of different sizes, that could eventually be applied for molecular encapsulation. Simulations predicted cages (A/B/C)22 to be equally likely to form (their formation energies being -10 kJ mol^{-1}, -9 kJ mol^{-1}, -10 kJ mol^{-1} respectively), to be shape persistent and not particularly strained, and to all have a void diameter of $\sim$7 Å. Experiments were not in complete agreement with the simulations and saw the formation of only

cage A22, although clear cage formation could not be identified by ^{1}H NMR, and only HRMS and diffusion NMR could see the formation of a cage with Tri^{4}Tri^{4} topology. It is not clear why cages containing B and C did not form; it is possible that some repulsive interactions occurred between the precursors, which in this topology would be assembled into a tighter geometry in respect to the other larger tritopic precursors adopted, although the repulsion would have been reflected by less favourable formation energies. Again, the answer could be related to experimental factors I could not include in my calculations.

Table 6.4.: Formation energies per bond formed, strain values and void diameters of cages with **Tri**^{4}**Tri**^{4} topology (A/B/C 22-26).

Cage	Formation energy per bond formed (kJ mol^{-1})	Strain (%)	Void diameter (Å)
A22 - **Tri**^{4}**Tri**^{4}	-10	1.6	7.1
B22 - **Tri**^{4}**Tri**^{4}	-10	1.4	7.3
C22 - **Tri**^{4}**Tri**^{4}	-10	1.1	7.3
A23 - **Tri**^{4}**Tri**^{4}	-5	0.8	8.8
B23 - **Tri**^{4}**Tri**^{4}	-8	1.5	9.3
C23 - **Tri**^{4}**Tri**^{4}	-8	0.9	9.1
A24 - **Tri**^{4}**Tri**^{4}	-6	0.8	10.7
B24 - **Tri**^{4}**Tri**^{4}	-11	1.2	8.4
C24 - **Tri**^{4}**Tri**^{4}	-11	1.1	8.3
A25 - **Tri**^{4}**Tri**^{4}	-7	0.3	13.2
B25 - **Tri**^{4}**Tri**^{4}	-11	0.3	11.5
C25 - **Tri**^{4}**Tri**^{4}	-12	0.4	10.6
A26 - **Tri**^{4}**Tri**^{4}	-6	1.0	13.1
B26 - **Tri**^{4}**Tri**^{4}	-14	0.4	12.3
C26 - **Tri**^{4}**Tri**^{4}	-15	0.4	11.1

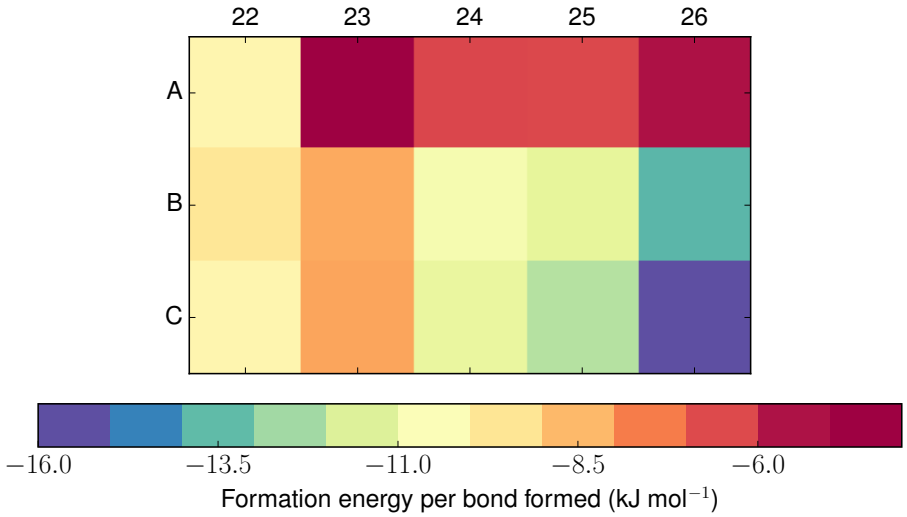


Figure 6.26.: Colour map containing the formation energies per bond formed calculated at PBE/TZVP+D3 level of theory for cages with **Tri**^{4}**Tri**^{4} topology.

From the colour map, it is possible to notice a pattern forming for cages containing precursors 22-26: cages containing triamine A are less stable than cages containing triamines B and C, with the difference in formation energy between cages of kind A and C increasing with the size of aldehydes 22-26 from 0.01 to 9.9 kJ mol^{-1}, as shown in the plot reported in Figure 6.27. This means that the presence of triamine C (and B, which led to cages with similar stabilities) progressively increased the stability of tetrapods with growing aldehyde sizes. This was in agreement with the experiments, which saw that cages A(23-26) did not form, but that reported that their B and C equivalents did form, based upon ^{1}H NMR, HRMS, diffusion NMR and, in the cases of (B/C) 23, B24 and (B/C) 26, PXRD analysis.

Crystal structures were determined for cages C23, B24, B26, and C26, and an overlay of the first three with my simulations are reported in Figure 6.28. RMSD values are slightly higher for these larger structures, which is mainly due to a larger conformational flexibility of the tritopic precursors involved. C23 is fairly well reproduced with an RMSD of 1.10 Å, that is very slightly improved to 1.06 Å upon minimisation of the crystal structure; similarly the simulation of B24 reproduced the experiments with a RMSD of 1.11 Å, which the minimisation of the crystal structure reduced to 1.03 Å. Marginally higher are the RMSD values for the larger B26 and C26 cages, which correspond to 1.24 Å and 1.76 Å respectively, both improved by minimising the crystal structure to 1.09 Å and 1.62 Å.

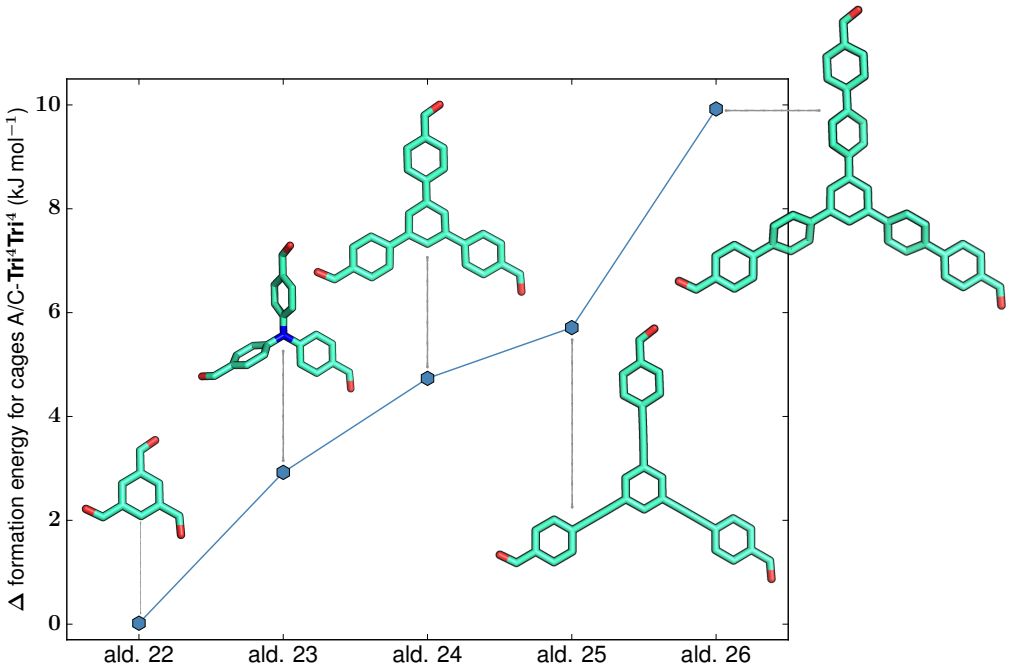


Figure 6.27.: Formation energy difference between cages A and C (22-26) calculated at the PBE/TZVP +D3 level. A positive difference in formation energy corresponds to C cages being more stable than A cages.

The family of tetrapod molecules was the one that produced the largest predicted cavities across the entire screen. As expected, the cavity size increased with increasing aldehyde size (for example, 26 (12Å on average) > 25 (11.3Å on average) > 24 (8.6Å on average)). The largest potential cavity would have been for A26, with a diameter of 13.1 Å, but this cage did not form; instead, the largest cavity of a cage that successfully formed was for B26, which had a cavity of 12.3 Å. The strain percentage values were low for all cages.

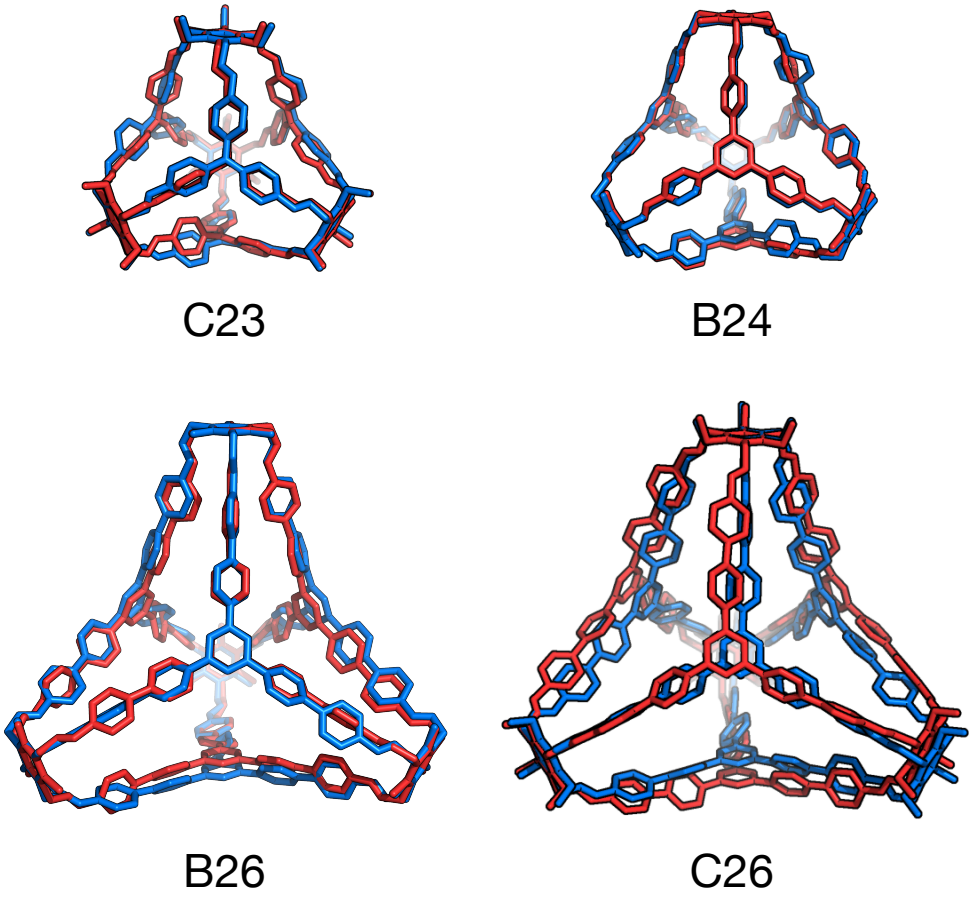


Figure 6.28.: Overlays of cages C23, B24 and B26. The simulations are in blue, the crystal structures in red.

6.3.6. Criteria for the synthesis of promising candidates

Examining the results, some criteria to approach the synthesis of highly porous molecules were defined. In order to obtain large voids, one should probably aim at the synthesis of a **Tri**^{4}**Tri**^{4} cage, with substituent groups providing pre-configuration of the tritopic precursors' conformation, and a maximised angle between reactive end groups on the precursor (i.e., 120° for tritopic and 180° for ditopic precursors). Preliminary studies into the sorption capabilities of B11 and B23 following solvent exchange experiments indicated that the cages can indeed be porous "as-made" with comparable porosity to other previously reported POCs in the literature, with both demonstrating uptake of H_2 and N_2: B11, N_2 8.18 mmol g^{-1}, H_2 3.15 mmol g^{-1} (77 K, 1 bar); B23, N_2 2.55 mmol g^{-1}, H_2 2.31 mmol g^{-1} (77 K, 1 bar). Looking at the void colour map, if one assumes that a diameter of more than 2.89 Å(the kinetic diameter of H_2)[194] is required for the cages to have the potential to host any guest

at all, then from the entire screen we find 21 cages that meet this criteria. There was therefore a 27% chance of getting a porous cage from our HT screen.

6.3.7. A serendipitous discovery, the pretzelane topology

Whilst attempting the characterisation of cage B13 (see Figure 6.4), a cage formed by 8 triamines of kind B and 12 precursors of kind 13 was isolated. It was found to be a covalently bonded triply interlocked topology reminiscent of a pretzel, of which the geometry is shown in Figure 6.29. The mechanism of formation of the pretzelane molecule based upon the current experimental settings is not clear yet. Amine B and aldehyde 13 formed a cage with **Tri**^{4}**Di**^{6} topology in solution; however, they then crystallised in the pretzelane larger topology, the same behaviour being observed for cage C13 (of which, however, a final crystal structure is not yet available).

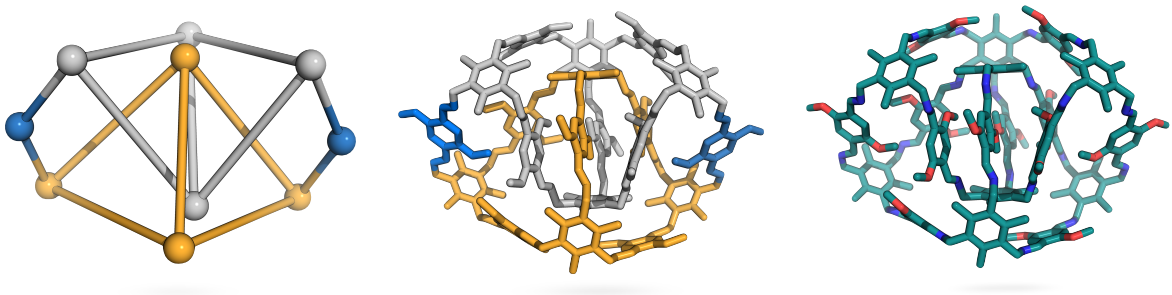


Figure 6.29.: Left: topology scheme of B13-pretzelane shows two interlocked units (in yellow and grey) bridged in two positions (blue); Middle: B13-pretzelane coloured in grey, yellow and blue to make the connectivity more understandable; Right: geometry of B13 pretzelene, with carbons in teal, oxygens in red and nitrogens in blue.

Even though the pretzelane topology was not originally included in the computational study, simulations were performed on both B13-**Tri**^{4}**Di**^{6} and B13-pretzelane to understand whether a substantial difference in formation energy could have driven the crystallisation of the larger topology. Both molecules' lowest energy conformers showed shape persistency, meaning that the structures did not collapse, although no large void was observed; in the case of B13-**Tri**^{4}**Di**^{6}, this was due to the position of the methoxy groups that were oriented towards the centre of the cavity; by contrast, the topology of B13-pretzelane was intrinsically non-porous. The formation energy, the strain and the void diameters of both molecules are reported in Table 6.5.

Table 6.5.: Formation energies, strain values and void diameters for B13-**Tri**^{4}**Di**^{6} and B13-pretzelane

Cage	Formation energy per bond formed (kJ mol^{-1})	Strain (%)	Void diameter (Å)
B13 - **Tri**^{4}**Di**^{6}	-13	1.1	3.4
B13 - pretzelane	-31	1.5	-

The formation energy per bond of the pretzelane molecule was more than 18 kJ mol^{-1} stable than the one of B13-**Tri**^{4}**Di**^{6}, the energetic difference being large enough to explain why the first was the structure that could be isolated experimentally. Such a great stabilisation could be caused by the formation of stabilising intramolecular interactions (such as π-stacking) within the pretzelane topology. Given that the crystal structure of B13-pretzelane was available, the RMSD between the crystal structure and the simulated structure was calculated to be 1.19 Å, showing that the experiments were well reproduced.

6.4. Conclusions

This work highlights how combining computational modelling with state-of-the-art automation and high throughput characterisation techniques can lead to the discovery of a large range of new organic cages. By investigating 78 building block combinations, 33 new cages were synthesised cleanly, with a number of other combinations also showing evidence of cage formation. Recrystallization studies led to 18 different crystal structures consisting of 15 different precursor combinations, alongside the serendipitous discovery of a new di-covalently connected triply interlocked cage species. The percentage of success of the experiments for each topological family is shown in Figure 6.30. Cages belonging to either the **Tri**^{2}**Di**^{3} or **Tri**^{4}**Tri**^{4} family performed better, 76.8% cages forming in the first case (46.5 % cleanly) and 59.6% in the second (53.5% cleanly), probably because the geometry of the starting precursor could lead to stable and reasonably rigid structures.

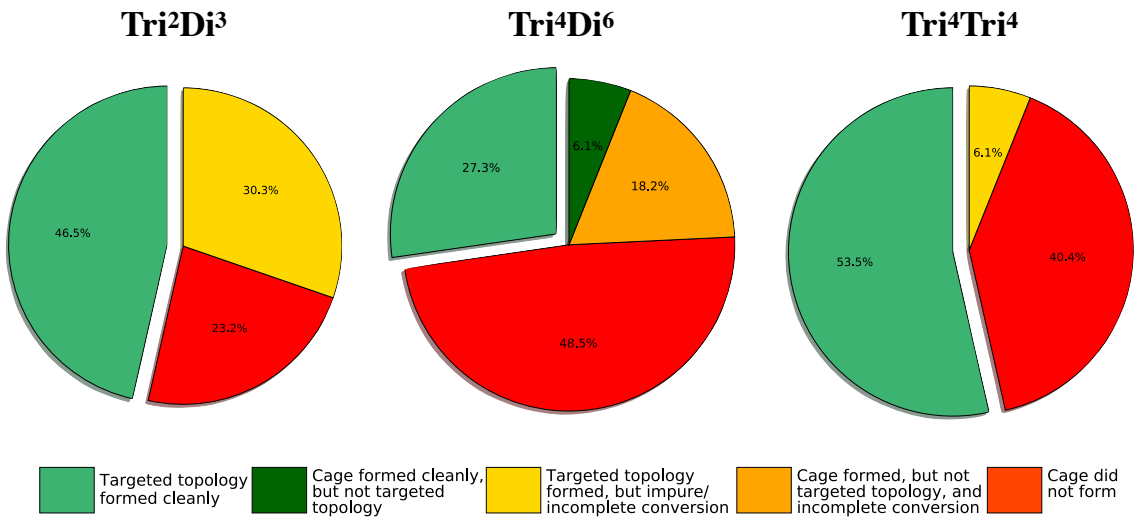


Figure 6.30.: Pie charts representing the percentages of success of the experiments for each topological family.

From a computational point of view, a strong topological energetic preference towards a desired cage is needed (∼30 kJ mol^{-1}), as this both increases the chance of the cage being formed and ensures the desired topology can be achieved experimentally. With the complexity of the topological possibilities in this set, the computation alone would not have helped in the selection of reactions screened, as it is clear from this investigation that it is complex to predict the most promising organic cages - as evidenced by solvent influence, topology emergent behaviour and other as yet unclear factors that affect which molecules actually form. However, using computational modelling alongside the experimental HT screening proved highly beneficial and the simulations can be used to determine which of the HT "hits" are the most promising to follow up for further characterisation, preventing wasted time on the scale-up, isolation and characterisation of the least promising molecules. Additionally, if the targeted topology of a cage is known, computational techniques can aid the experiments in determining whether a cage is shape persistent, and whether its cavity would be large enough, before conducting additional investigations.

This kind of study is also required as a base for further machine learning studies (where both successful and failed reactions would make up the training set), although more precursors (100-1000) would be needed for a sensible training set. In relation to organic cages, this work flow still proved successful as it provided a new collection of cages that notably expanded the selection of those currently in the literature and also provided a potential route for investigating other combinations.

Applications of these new cages are currently under investigation; for example,

some of the hits are already demonstrating porosity (B11 and B23) and therefore could have possible applications in gas uptake and separations. Furthermore, they might also have applications as tectons in other assemblies, such as cage co-crystallisations or hydrogen-bonded frameworks; or, for those with differing external functionalisation, they could be used as linkers in extended frameworks, such as cage polymers, cage MOFs or cage covalent organic frameworks.

Chapter 7
Conclusions and future work

The focus of my PhD was to develop a strategy for the prediction of the relative stability and topology of porous organic cages, in order to direct the synthesis towards promising candidates. In this context, Chapter 3 discussed the development of a new approach that aimed at finding higher energy open metastable conformations of partially or densely collapsed molecules, by using an "inflation" technique based on constrained molecular dynamics. That allowed me to investigate the scaffolding effect of the solvent on molecular geometries and thus to be able to study the properties of the open structures, without needing to take into account host-guest chemical interactions. The procedure was successfully applied to two imino cages from the literature, **CC8** and Mastalerz endo-functionalised cage, two hypothetical imino-cages, the [6+9] and [8+12] sizes of cage **CC3**, two cucurbitutils, CB6 and CB10-ns, and finally Cryptophane A, all shown in Figure 7.1. The method can potentially be applied in the future for the prediction of the solvated structure of a broad range of systems.

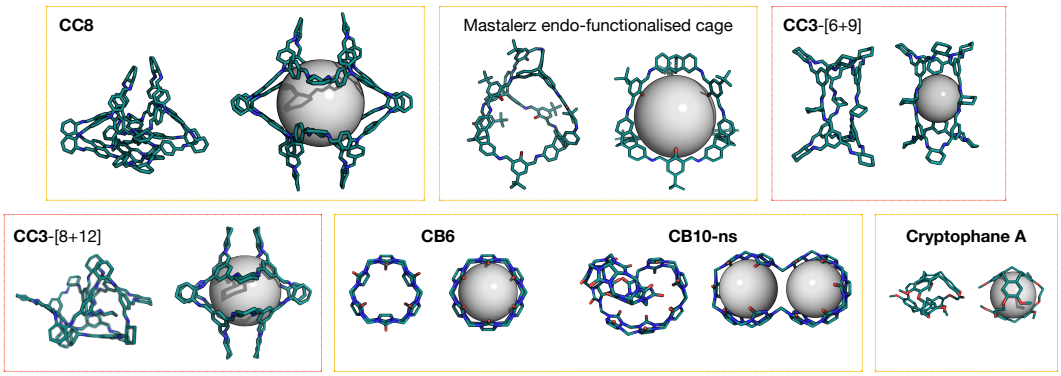


Figure 7.1.: Molecules inflated with constrained MD in Chapter 3, in each box the molecule on the left is collapsed, whereas that on the right is open, with a grey sphere highlighting the cavity. Orange boxes surround cages reported in the literature, red boxes hypothetical cages.

In Chapters 4 and 5, I approached the topological prediction of organic cage molecules. In Chapter 4 I discussed how I classified porous organic cages into four different topological families (**TriDi**, **TriTri**, **TetDi**, **TetDi**, shown in Figure 7.2) depending on the number of reactive ends of the precursors they contained. A new nomenclature was defined for 20 topologies that could be accessed via the DCC synthesis of organic cage molecules, 12 of which were already available in the literature. The underlying symmetry of the topologies, their window sizes and relation with different shapes were described too, with the main goal of clarifying the ambivalence between the ideas of geometrical shape and topology.

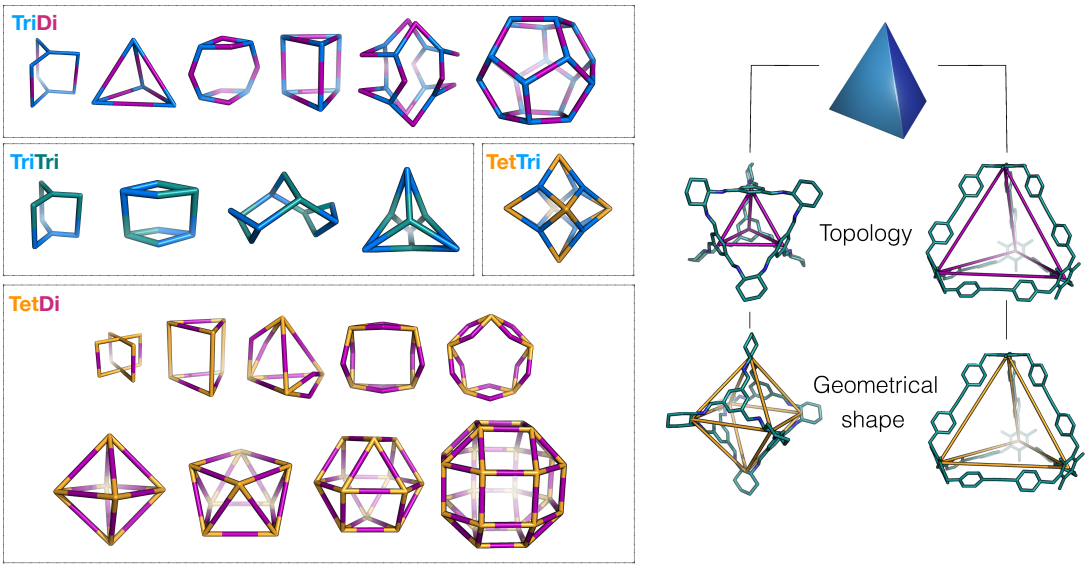


Figure 7.2.: The four topological families classified in Chapter 4, and a scheme showing the difference between topology and geometrical shape.

In Chapter 5, I tested to what extent the calculation of relative and formation energy can assist the prediction of the topological outcome for a selection of DCC reactions. The prediction pipeline was applied to a few organic cages from the literature, namely **CC3**, **CC5**, and **CC8** for the **TriDi** family, **CC11** for the **TriTri** family, and two cages from Warmuth and co-workers for the **TetDi** and **TetTri** family (all shown in Figure 7.3). The prediction of the final topology was successful in all cases, except for cage **CC5**. It was, in fact, not possible to correctly predict the formation of a **Tri**^{4}**Di**^{6} topology over a **Tri**^{8}**Di**^{12} topology, as the difference in energy between the two structures was not large enough. The correct result can potentially be simulated by: i) including the effect of the solvent on the energy landscape, ii) calculating the entropic contributions that would disfavour the larger topology and iii) considering alternative reaction pathways, or relative stabilities of intermediate

compounds. The object of future research would be to successfully predict the effect of different solvents on cage topology, as, as seen for the Warmuth **TetDi** family, it was not possible to find any obvious geometric correlation between the structures and the solvent that led to their formation with the simulation procedure I used during my PhD. Nonetheless, the work showed how simulations and calculating the relative stability of molecules with different topologies is a useful tool to obtain information about the thermodynamic feasibility of a given target.

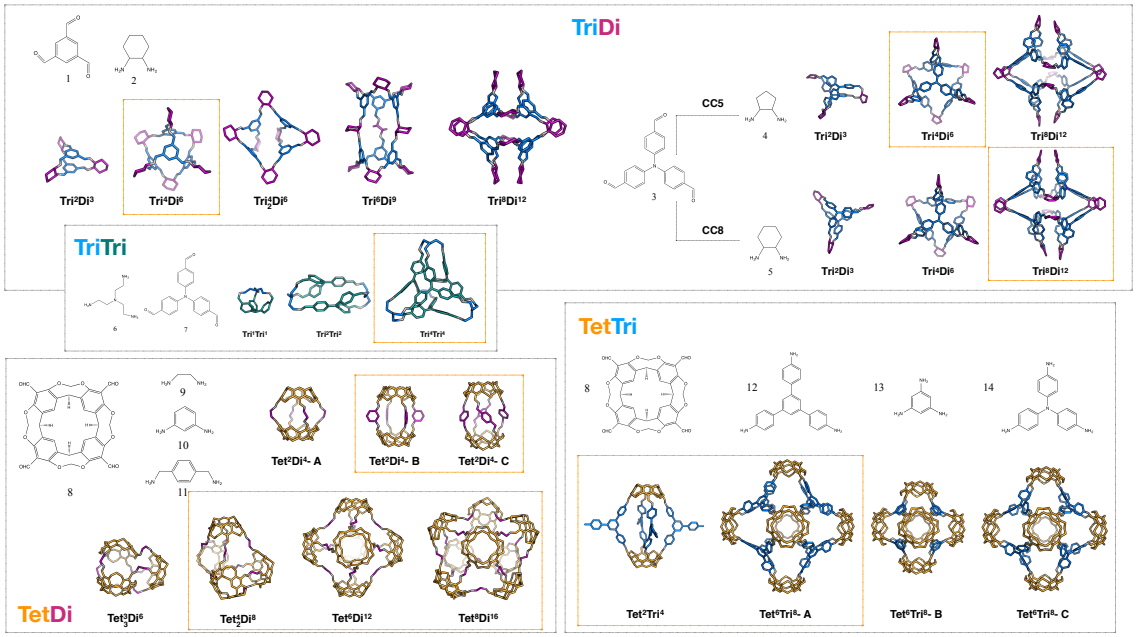


Figure 7.3.: The molecules investigated in Chapter 5. Orange boxes highlight the molecules reported in the literature.

In the last chapter, Chapter 6, computational modelling was combined with state-of-the-art automation in the context of a high-throughput synthesis of porous organic cages. 26 building blocks were combined into 78 cages (shown in Figure 7.4), 33 of which formed cleanly and 16 more with impurities. Recrystallization experiments led to 18 crystal structures generated from 15 different precursor combinations, alongside the serendipitous discovery of a new double-bridged triply interlocked cage species. All cages were computationally studied, to both predict the topology and shape-persistency of the final product. Studying such a large collection of cages was very important to realise the advantages and limits of the prediction technique. The study highlighted that, when comparing the relative formation energies per bond formed of a collection of cages, a strong topological energetic preference towards a desired cage is needed ($\sim$30 kJ mol^{-1}) to be confident that the targeted topology

can be achieved experimentally. Factors such as solvent influence, thermodynamic events leading to emergent behaviour and other experimental variables have a large effect on the final outcome. Often it was not possible to include the description of those factors in the simulations. Computation alone would have ruled out cages containing triamine A, as they had higher formation energies and in some cases collapsed, but would not have helped in the selection of the remaining reactions considered. Nonetheless, simulations were largely helpful to determine: i) which of the cages formed in solution were the most promising to follow up for further characterisation; ii) the shape-persistency of cage; iii) whether the cavity of a cage is large enough to selectively host guests. Machine learning studies to predict which BBs would lead to novel porous organic materials, on a training set containing 100-1000 precursors, would be the natural evolution of this project.

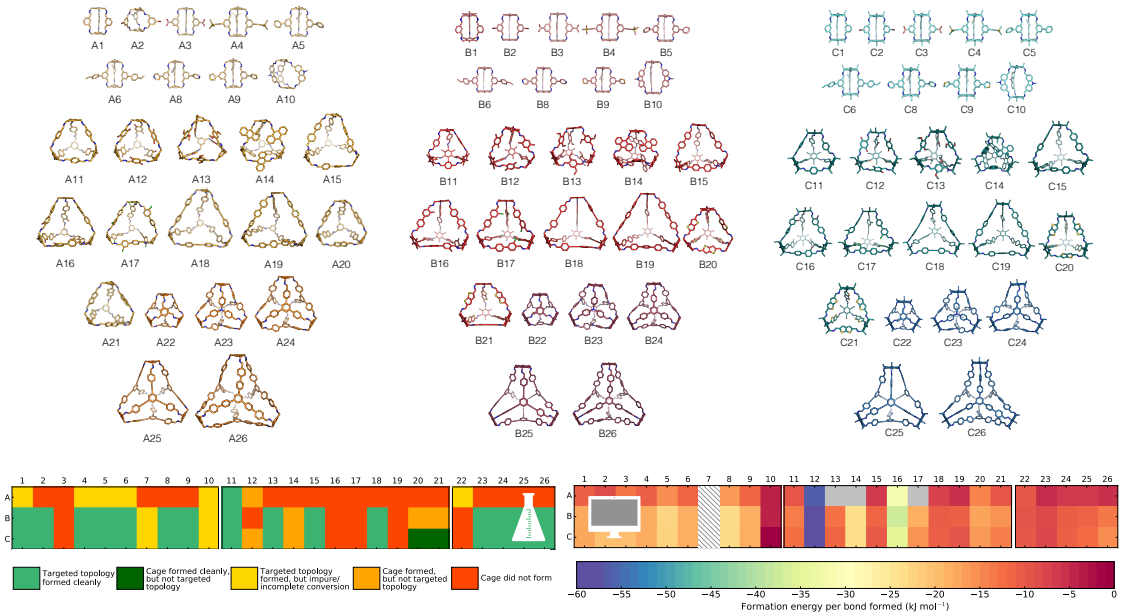


Figure 7.4.: Cages investigated in Chapter 6, together with the plots reporting results from experiments (left), and simulations (right).

Whilst I was reaching the end of my PhD and this thesis was already partially completed, I was told by the group's collaborators that a system of which I predicted the feasibility at the very beginning of my first year was found to be a promising candidate for a new porous material. Results are known from Dr. Angeles Pulido (Cambridge Crystallographic Data Centre) for the part regarding crystal structure prediction and from Dr. Becky Greenaway and Dr. Marc Little (Cooper's group, University of Liverpool) for the part about synthesis and crystallisation of materials.

Becoming aware of this result was quite exciting for me, as it completed the picture outlined in Figure 1.2 and that I am showing again in Figure 7.5.

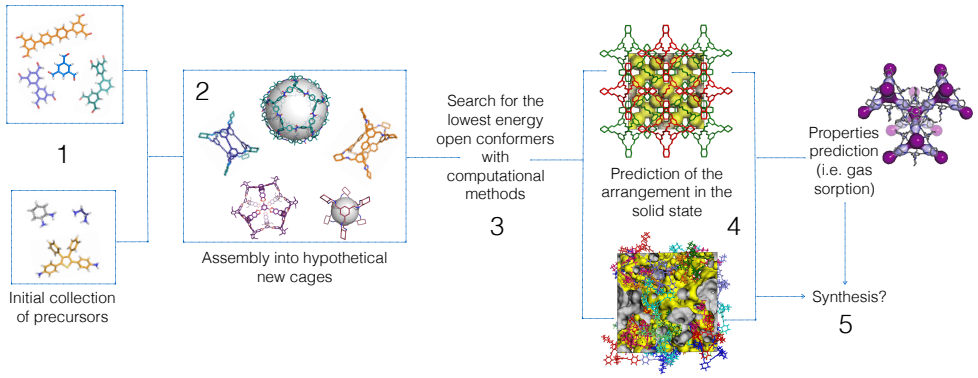


Figure 7.5.: Workflow integrating computational and experimental methods for the successful discovery of new porous organic materials. Steps 1-2-3: computational investigation of hypothetical single cages, assembled from a pair of organic precursors; Step 4: promising candidates from step 3 are assembled into the solid state with crystal structure prediction methods; Step 5: promising candidates (of which properties can be studied with computational simulations) from step 4 are suggested for the synthesis.

The molecules I designed (according to Steps 1-2), were three-component imine cages obtained combining three ditopic amines with three tritopic aldehydes and capping the unreacted ends with a tritopic amine, as shown in Figure 7.6. They were named "cage pots" (**CP**) due to their resemblance with a pot. **CP1** was obtained by combining tris(2-aminoethyl)amine with the precursors contained in cage **CC1**, **CP2** with those contained in cage **CC3**, and **CP4** with those contained in cage **CC4**.

According to step 3 of Figure 7.5, the cage pots were investigated with conformer search methods, the lowest energy conformers were identified and their energies and structures were refined at the PBE/TZVP +D3 level of theory. As one can see from the values reported in Figure 7.6, the energy difference between the cage pots and their parent two-components imine cages was not sufficiently large to to think they were unlikely to form. The one with lowest formation energy, **CP3**, was investigated with crystal structure prediction methods by Dr. Pulido, who, as a preliminary result, identified three possible stable crystal packing arrangements. The first combines *R*-**CP3**, the second **CP3** molecules with opposite chirality, whereas the third is a co-crystal of [1:4] *R*-**CC3** and *S*-**CP3**. The molecules are predicted to pack in layers in the first case and window-to-window in the latter two cases, as shown in Figure 7.7.

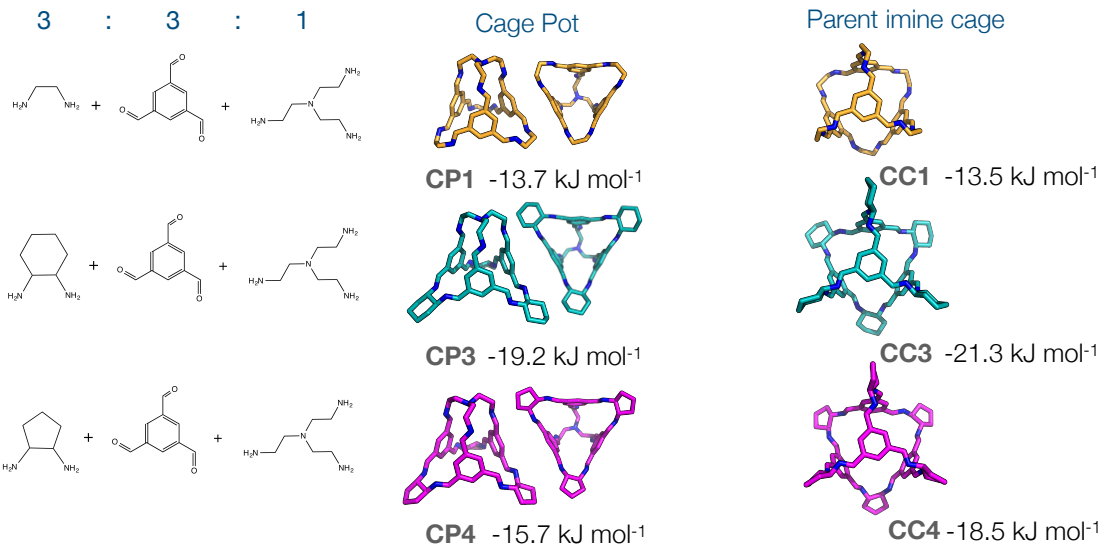


Figure 7.6.: **CP1**, **CP3**, and **CP4**, and their parent imine cages, **CC1**, **CC3**, and **CC4**. The numbers below each cage refer to the formation energy per bond formed calculated at the PBE/TZVP +D3 level of theory.

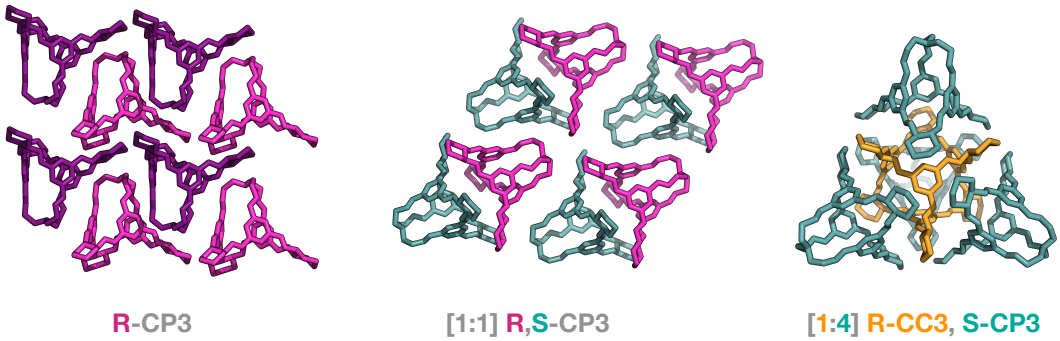


Figure 7.7.: Clusters showing interactions between **CP** molecules in the solid state, as predicted by crystal structure prediction. **CP3** molecules are in different colours for clarity: left) magenta and purple, middle) the *R* enantiomer is in magenta and the *S* in teal, right) *R*-**CC3** is in orange and *S*-**CP3** in teal.

The synthesis of **CP3** was attempted and the formation of cage pot molecules was identified by high-performance liquid chromatography and ^{1}H NMR. Recrystallisation studies led to the formation of two crystal structures, a racemic crystal of *R,S*-**CP3s** and a co-crystal of [1:4] *R*-**CC3** and *S*-**CP3**. Preliminary analysis confirms that window-recognition occurs between molecules of different chirality. More cage pots obtained by substitution or functionalisation of the precursors involved in this project are under current experimental investigation. These preliminary successful results are a promising starting point for a new integrated and more efficient workflow that, starting from a selection of building blocks and via the simulation of

assembled single cages, can aim at proposing promising candidates for the synthesis of porous organic materials.

In conclusion, this thesis provided an advance in the classification and computational characterisation of porous organic molecules. The projects I carried out during my PhD progressively combined all the features needed to build a computational prediction pipeline that can be integrated with the solid state analysis and the synthesis of organic cage molecules. The pipeline was thoroughly tested, firstly, on systems already available in the literature and then on a larger collection of cages. This last step was fundamental to become aware of both the advantages and the limits of the procedure, and of the parts that will need to be improved in the future, in order to be able to design more efficiently and with the aid of automated processes, novel porous organic materials.

Appendix A
Constrained MD input files

A.1. Input files for Constrained MD calculations

A.1.1. Spherical probe

Below an example input file for a constrained MD simulation is reported (the input file is meant for the use of DL_POLY2.20[230] combined with the PLUMED2[171] plug-in). In the first part the centre of mass of the system is defined (**COM**). The second part calculates the distance between the COM and each atom of the molecule. The distance is calculated with the equation reported above. The final section defines the restraint (KAPPA) and its position(AT0/AT1) at the time defined by STEP. In the case of constrained MD it is going to be applied between the centre of mass and any atom. The restraint is gradually increased from 0 to 50000 kJ mol^{-1} nm^{-2} over the first 20000 steps (from STEP0 to STEP1). Some testing was required on the VERSE keyword, which tells PLUMED whether the restraint should act for collective variables larger (U) or smaller (L) than the restraint, or both (B). The default value for the keyword VERSE is B, but in our case the keyword needs to be set to (L), so that the restraint is only applied to distances lower than the position of the restraint (which is placed at 0.3 nm in the example).

```
# This should be the COM of whole molecule
COM ATOMS=1-488 LABEL=com
#The difference between the largest and smallest distance
DISTANCES GROUPA=com GROUPB=1-488 MIN=BETA=10 LABEL=d1
MOVINGRESTRAINT ...
ARG=d1.min
STEP0=0 AT0=0.3 KAPPA0=0.0
```

```
STEP1=20000 AT1=0.3 KAPPA1=50000.0
VERSE=L
...  MOVINGRESTRAINT
```

A.1.2. Cylindrical probe

The input file for calculations involving cylindrical probes with PLUMED2 is as shown below:

```
# com1 is the COM of the top face of the cylinder, com2 of the other.
# The orientation is given by the vector from com1 to com2
COM ATOMS=3,25,39 LABEL=com1
COM ATOMS=49,69,81 LABEL=com2
# inplane distance between com1 and com2
d1:  INPLANEDISTANCES VECTORSTART=com1 VECTOREND=com2
GROUP=1-162 MIN=BETA=5

MOVINGRESTRAINT ...
ARG=d1.min
STEP0=0 AT0=0.5 KAPPA0=0.0
STEP1=20000 AT1=0.5 KAPPA1=50000.0
LABEL=m1
VERSE=L
...  MOVINGRESTRAINT
```

Bibliography

[1] J. R. Holst, A. Trewin and A. Cooper, *Nat. Chem.*, 2010, **2**, 915–920.

[2] S. Kitagawa, *Angew. Chem. Int. Ed.*, 2015, **54**, 10686–10687.

[3] M. Kondo, T. Yoshitomi, H. Matsuzaka, S. Kitagawa and K. Seki, *Angew. Chem. Int. Ed.*, 1997, **36**, 1725–1727.

[4] A. K. Cheetham, G. Ferey and T. Loiseau, *Angew. Chem. Int. Ed.*, 1999, **38**, 3268–3292.

[5] H. Furukawa, K. E. Cordova, M. O'Keeffe and O. M. Yaghi, *Science*, 2013, **341**, 1230444–1,1230444–12.

[6] A. P. Côté, A. Benin, N. W. Ockwig, M. O'Keefe, A. J. Matzger and O. M. Yaghi, *Science*, 2005, **310**, 1166–1170.

[7] T. Tozawa, J. T. A. Jones, S. I. Swamy, S. Jiang, D. J. Adams, S. Shakespeare, D. Bradshaw, T. Hasell, S. Y. Chong, C. Tang, S. Thompson, J. Parker, A. Trewin, J. Bacsa, A. M. Z. Slawin, A. Steiner and A. Cooper, *Nat. Mater.*, 2009, **8**, 973–978.

[8] C. S. Cundy and P. A. Cox, *Chem. Rev.*, 2003, **103**, 663.

[9] C. S. Cundy and P. A. Cox, *Microporous Mesoporous Mater.*, 2005, **82**, 1–78.

[10] O. M. Yaghi and H. Li, *J. Am. Chem. Soc.*, 1995, **117**, 10401–10402.

[11] S. S. Chui, S. M. Lo, J. P. Charmant, A. G. Orpen and I. D. Williams, *Science*, 1999, **283**, 1148–1150.

[12] O. M. Yaghi, H. Li, M. Eddaoudi and M. O'Keeffe, *Nature*, 1999, **402**, 276–279.

[13] H. Deng, S. Grunder, K. E. Cordova, C. Valente, H. Furukawa, M. Hmadeh, F. Gándara, A. C. Whalley, Z. Liu, S. Asahina, H. Kazumori, M. O'Keeffe, O. Terasaki, J. F. Stoddart and O. M. Yaghi, *Science*, 2012, **336**, 1018–1023.

[14] C. S. Diercks and O. M. Yaghi, *Science*, 2017, **355**, 1–8.

[15] N. L. Rosi, J. Eckert, M. Eddaoudi, D. T. Vodak, J. Kim, M. O'Keeffe and O. M. Yaghi, *Science*, 2003, **300**, 1127–1129.

[16] S.-T. Yang, J. Kim, H.-Y. Cho, S. Kim and W.-S. Ahn, *RSC Adv.*, 2012, **2**, 10179–10181.

[17] J. Evans, D. M. Huang, M. Haranczyk, A. W. Thornton, C. J. Sumby and C. J. Doonan, *CrystEngComm*, 2016, 1–11.

[18] A. Cooper, *Angew. Chem. Int. Ed.*, 2012, **51**, 7892–7894.

[19] M. Mastalerz, *Chem. - Eur. J.*, 2012, **18**, 10082–10091.

[20] J. Tian, P. K. Thallapally and B. P. McGrail, *CrystEngComm*, 2012, **14**, 1909–1919.

[21] A. G. Slater and A. Cooper, *Science*, 2015, **348**, aaa8075–1–aaa8075–10.

[22] T. Hasell and A. Cooper, *Nat. Rev. Mater.*, 2016, **1**, 1–15.

[23] L. Barbour, *Chem. Commun.*, 2006, 1163–1168.

[24] O. M. Yaghi, M. O'Keefe and M. Kanatzidis, *J. Solid State Chem.*, 2000, **152**, 1–2.

[25] M. O'Keefe and O. M. Yaghi, *Chem. Rev.*, 2012, **112**, 675–702.

[26] O. Delgado-Friedrichs, M. D. Foster, M. O'Keefe, D. M. Proserpio, M. M. J. Treacy and O. M. Yaghi, *J. Solid State Chem.*, 2005, **178**, 2533–2554.

[27] M. O'Keefe, M. A. Peskov, S. J. Ramsden and O. M. Yaghi, *Acc. Chem. Res.*, 2008, **41**, 1782–1789.

[28] E. V. Alexandrov, V. A. Blatov, A. V. Kochetkov and D. M. Proserpio, *CrystEngComm*, 2011, **13**, 3947–3958.

[29] C. E. Wilmer, M. Leaf, C. Y. Lee, O. K. Farha, B. G. Hauser, J. T. Hupp and R. Q. Snurr, *Nat. Chem.*, 2011, **4**, 83–89.

[30] O. M. Yaghi, M. O'Keefe, N. W. Ockwig, H. K. Chae, M. Eddaoudi and J. Kim, *Nat. Commun.*, 2003, **423**, 705–714.

[31] R. S. Forgan, J.-P. Sauvage and J. F. Stoddart, *Chem. Rev.*, 2011, **111**, 5434–5464.

[32] R. Chakrabarty, P. S. Mukherjee and P. J. Stang, *Chem. Rev.*, 2011, **111**, 6810–6918.

[33] G. Zhang and M. Mastalerz, *Chem. Soc. Rev.*, 2014, **43**, 1934–1947.

[34] S. Chong and A. Cooper, *Comprehensive Supramolecular Chemistry {II}*, Elsevier, Oxford, 2017, pp. 139 – 197.

[35] A. I. Joseph, S. H. Lapidus, C. M. Kane and K. T. Holman, *Angew. Chem. Int. Ed.*, 2014, **54**, 1471–1475.

[36] M. J. Bojdys, M. E. Briggs, J. T. A. Jones, D. J. Adams, S. Y. Chong, M. Schmidtmann and A. Cooper, *J. Am. Chem. Soc.*, 2011, **133**, 16566–16571.

[37] T. Hasell, S. Y. Chong, M. Schmidtmann, D. J. Adams and A. Cooper, *Angew. Chem., Int. Ed.*, 2012, **51**, 7154–7157.

[38] T. Hasell, M. Schmidtmann, C. A. Stone, M. W. Smith and A. Cooper, *Chem. Commun.*, 2012, **48**, 4689–4691.

[39] M. Liu, M. A. Little, K. E. Jelfs, J. T. A. Jones, M. Schmidtmann, S. Y. Chong, T. Hasell and A. Cooper, *J. Am. Chem. Soc.*, 2014, **136**, 7583–7586.

[40] Q. Song, S. Jiang, T. Hasell, M. Liu, S. Sun, A. K. Cheetham, E. Sivaniah and A. Cooper, *Adv. Mater.*, 2016, **28**, 2629–2637.

[41] T. Hasell, H. Zhang and A. Cooper, *Adv. Mater.*, 2012, **24**, 5732–5737.

[42] P. E. Eaton and T. W. Cole, *J. Am. Chem. Soc.*, 1964, **86**, 3157–3158.

[43] R. M. Barrer and V. SHANSON, *J. Chem. Soc., Chem. Commun.*, 1976, **0**, 333–334.

[44] H. R. Allcock and L. A. Siegel, *J. Am. Chem. Soc.*, 1964, **86**, 5140–5144.

[45] P. Sozzani, S. Bracco, A. Comotti, L. Ferretti and R. Simonutti, *Angew. Chem. Int. Ed.*, 2005, **44**, 1816–1820.

[46] G. Couderc, T. Hertzsch, N. Behrnd, K. Kramer and J. Hulliger, *Microporous Mesoporous Mater.*

[47] K. J. Msayib, D. Book, P. M. Budd, N. Chaukura, K. D. M. Harris, M. Helliwell, S. Tedds, A. Walton, M. Warren, John Eand Xu and N. B. McKeown, *Angew. Chem. Int. Ed.*, 2009, **48**, 3273–3277.

[48] M. Mastalerz and I. M. Oppel, *Angew. Chem. Int. Ed.*, 2012, **51**, 5252–5255.

[49] A. Pulido, L. Chen, T. Kaczorowski, D. Holden, M. A. Little, S. Y. Chong, B. J. Slater, D. P. McMahon, B. Bonillo, C. J. Stackhouse, A. Stephenson, C. M. Kane, R. Clowes, T. Hasell, A. I. Cooper and G. M. Day, *Nature*, 2017, **543**, 657–664.

[50] Y. Liu, X. Liu and R. Warmuth, *Chem. - Eur. J.*, 2007, **13**, 8953–8959.

[51] G. D. Andreetti, R. Ungaro and A. Pochini, *J. Chem. Soc., Chem. Commun.*, 1979, 1005–1007.

[52] S. J. Dalgarno, P. K. Thallapally, L. Barbour and J. L. Atwood, *Chem. Soc. Rev.*, 2007, **36**, 236–245.

[53] H. Kudo, H. Kudo, R. Hayashi, R. Hayashi, K. Mitani, K. Mitani, T. Yokozawa, T. Yokozawa, N. C. Kasuga, N. C. Kasuga, T. Nishikubo and T. Nishikubo, *Angew. Chem. Int. Ed.*, 2006, **45**, 7948–7952.

[54] P. Stutte, W. Kiggen and F. Vögtle, *Tetrahedron*, 1987, **43**, 2065–2074.

[55] D. J. Cram, S. Karbach, Y. H. Kim, L. Baczynskyj and G. W. Kallemeyn, *J. Am. Chem. Soc.*, 1985, **107**, 2575–2576.

[56] J.-M. Lehn, *Science*, 1985, **227**, 849–856.

[57] M. Quan and D. J. Cram, *J. Am. Chem. Soc.*, 1991, **113**, 2754–2755.

[58] J. Lehn, *Chem. - Eur. J.*, 1999, **5**, 2455–2463.

[59] H. M. El-Kaderi, J. R. Hunt, J. L. Mendoza-Cortes, A. P. Côté, R. E. Taylor, M. O'Keefe and O. M. Yaghi, *Science*, 2007, **316**, 268–272.

[60] C. Férey, C. Mellot-Draznieks, C. Serre, F. Millange, J. Dutour, S. Surblé and I. Margiolaki, *Science*, 2005, **309**, 2040–2042.

[61] M. Mastalerz, *Angew. Chem.*, 2010, **49**, 5042–5053.

[62] A. Avellaneda, P. Valente, A. Burgun, J. D. Evans, A. W. Markwell-Heys, D. Rankine, D. J. Nielsen, M. R. Hill, C. J. Sumby and C. J. Doonan, *Angew. Chem. Int. Ed.*, 2013, **52**, 3746–3749.

[63] W. Kiggen and F. Vögtle, *Angew. Chem. Intern. Ed.*, 1984, **23**, 714–715.

[64] T. J. Mcmurry, S. J. Rodgers and K. N. Raymond, *J. Am. Chem. Soc.*, 1987, **109**, 3451–3453.

[65] W. Maison, J. V. Frangioni and N. Pannier, *Org. Lett.*, 2004, **6**, 4567–4569.

[66] T. J. Ryan, G. Lecollinet, T. Velasco and A. P. Davis, *Proc. Natl. Acad. Sci.*, 2002, **99**, 4863–4866.

[67] J. L. Katz, K. J. Selby and R. R. Conry, *Org. Lett.*, 2005, **7**, 3505–3507.

[68] Z. Wu, S. Lee and J. S. Moore, *J. Am. Chem. Soc.*, 1992, **114**, 8730–8732.

[69] K. E. Jelfs, F. Schiffmann, J. T. A. Jones, B. Slater, F. Cora and A. Cooper, *Phys. Chem. Chem. Phys.*, 2011, **13**, 20081–20085.

[70] T. Hasell, X. Wu, J. T. A. Jones, J. Bacsa, A. Steiner, T. Mitra, A. Trewin, D. J. Adams and A. Cooper, *Nat. Chem.*, 2010, **2**, 750–755.

[71] T. Hasell, M. A. Little, S. Y. Chong, M. Schmidtmann, M. E. Briggs, V. Santolini, K. E. Jelfs and A. I. Cooper, *Nanoscale*, 2017, **9**, 6783–6790.

[72] J. T. A. Jones, T. Hasell, X. Wu, J. Bacsa, K. E. Jelfs, M. Schmidtmann, S. Y. Chong, D. J. Adams, A. Trewin, F. Schiffman, F. Cora, B. Slater, A. Steiner, G. M. Day and A. Cooper, *Nat. Commun.*, 2011, **474**, 367–371.

[73] T. Hasell, J. L. Culshaw, S. Y. Chong, M. Schmidtmann, M. A. Little, K. E. Jelfs, E. O. Pyzer-Knapp, H. Shepherd, D. J. Adams, G. M. Day and A. Cooper, 2014, **136**, 1438–1448.

[74] D. P. Lydon, D. P. Lydon, N. L. Campbell, D. J. Adams and A. Cooper, *Synth. Commun.*, 2011, **41**, 2146–2151.

[75] P. Skowronek and J. Gawronski, *Org. Lett.*, 2008, **10**, 4755–4758.

[76] S. M. Elbert, F. Rominger and M. Mastalerz, *Chem. - Eur. J.*, 2014, **20**, 16707–16720.

[77] M. Mastalerz, *Chem. Commun.*, 2008, 4756–4758.

[78] M. Mastalerz, M. W. Schneider, I. M. Oppel and O. Presly, *Angew. Chem. Int. Ed.*, 2011, **50**, 1046–1051.

[79] M. W. Schneider, I. M. Oppel, H. Ott, L. G. Lechner, H.-J. S. Hauswald, R. Stoll and M. Mastalerz, *Chem - Eur. J.*, 2012, **18**, 836–847.

[80] J. Sun, J. Bennett, R. Emge and R. Warmuth, *J. Am. Chem. Soc.*, 2011, **133**, 3268–3271.

[81] Z. Lin, J. Sun, B. Efremovska and R. Warmuth, *Chem - Eur. J.*, 2012, **18**, 12864–12872.

[82] N. M. Rue, J. Sun and R. Warmuth, *Isr. J. Chem.*, 2011, **51**, 743–768.

[83] K. E. Jelfs and A. Cooper, *Curr. Opin. Solid State Mater. Sci.*, 2013, **17**, 19–30.

[84] G. Zhang, O. Presly, F. White, I. M. Oppel and M. Mastalerz, *Angew. Chem. Int. Ed.*, 2014, **53**, 1516–1520.

[85] M. Thommes, K. Kaneko, A. V. Neimark, J. P. Olivier, F. Rodríguez-Reinoso, J. Rouquerol and K. S. W. Sing, *Inf. Bull. - I. U. P. A. C., Tech. Rep.*, 2015.

[86] S. Klotzbach, T. Scherpf and F. Beuerle, *Chem. Commun.*, 2014, **50**, 12454–12457.

[87] S. Klotzbach and F. Beuerle, *Angew. Chem.*, 2015, **127**, 10497–10502.

[88] A. Dhara and F. Beuerle, *Chem. - Eur. J.*, 2015, **21**, 17391–17396.

[89] C. Zhang, Q. Wang, H. Long and W. Zhang, *J. Am. Chem. Soc.*, 2011, **133**, 20995–21001.

[90] S. L. Diemer, M. Kristensen, B. Rasmussen, S. R. Beeren and M. Pittelkow, *Int. J. Mol. Sci.*, 2015, **16**, 21858–21872.

[91] K. D. Okochi, G. S. Han, I. M. Aldridge, Y. Liu and W. Zhang, *Org. Lett.*, 2013, **15**, 4296–4299.

[92] Y. Liu, X. Liu, and R. Warmuth, *Chem. - Eur. J.*, 2007, **13**, 8953–8959.

[93] K. Crane, U. Pinkall and P. Schröder, *ACM Trans. Graph.*, 2013, **32**, year.

[94] S. Lee, A. Yang and T. P. Moneypenny, *J. Am. Chem. Soc.*, 2016, **138**, 2182–2185.

[95] D. Xu and R. Warmuth, *J. Am. Chem. Soc.*, 2008, **130**, 7520–7521.

[96] K. E. Jelfs, X. Wu, M. Schmidtmann, J. T. A. Jones, J. E. Warren, D. J. Adams and A. Cooper, *Angew. Chem., Int. Ed.*, 2011, **50**, 10653–10656.

[97] Q. F. Sun, J. Iwasa, D. Ogawa, Y. Ishido, S. Sato, T. Ozeki, Y. Sei, K. Yamaguchi and M. Fujita, *Science*, 2010, **328**, 1144–1147.

[98] K. E. Jelfs, E. Eden, J. L. Culshaw, S. Shakespeare, E. O. Pyzer-Knapp, H. P. G. Thompson, J. Bacsa, G. M. Day and D. J. Adams, *J. Am. Chem. Soc.*, 2013, **135**, 9307–9310.

[99] X. Liu and R. Warmuth, *J. Am. Chem. Soc.*, 2006, **128**, 14120–14127.

[100] G. Zhang, O. Presly, F. White, I. M. Oppel and M. Mastalerz, *Angew. Chem. Int. Ed.*, 2014, **53**, 5126–5130.

[101] G. O. Brunner, *Zeitschrift Fur Kristallographie*, 1981, **156**, 295–303.

[102] D. J. Tranchemontagne, Z. Ni, M. O'Keefe and O. M. Yaghi, *Angew. Chem. Int. Ed.*, 2008, **47**, 5136–5147.

[103] B. Olenyuk, M. D. Levin, J. A. Whiteford, J. E. Shield and P. J. Stang, *J. Am. Chem. Soc.*, 1999, **121**, 10434–10435.

[104] N. J. Young and B. P. Hay, *Chem. Commun.*, 2013, **49**, 1354–1379.

[105] D. B. Amabilino, *Template Strategies in Self-Assembly*, John Wiley & Sons, Ltd, Chichester, UK, 2012.

[106] M. Rekharsky and Y. Inoue, *Solvation Effects in Supramolecular Recognition*, John Wiley & Sons, Ltd, Chichester, UK, 2012.

[107] Y. Jin, A. Jin, R. Mccaffrey, H. Long and W. Zhang, *J. Org. Chem.*, 2012, **77**, 7392–7400.

[108] S. Mecozzi and J. Rebek, *Chem. - Eur. J.*, 1998, **4**, 1016–1022.

[109] R. T. S. Lam, *Science*, 2005, **308**, 667–669.

[110] J. W. Kamplain and C. W. Bielawski, *Chem. Commun.*, 2006, 1727–1733.

[111] B. Brisig, J. Sanders and S. Otto, *Angew. Chem. Int. Ed.*, 2003, **42**, 1270–1273.

[112] M. E. Briggs, A. G. Slater, N. Lunt, S. Jiang, M. A. Little, R. L. Greenaway, T. Hasell, C. Battilocchio, S. V. Ley and A. I. Cooper, *Chem. Commun.*, 2015, **51**, 17390–17393.

[113] M. Kitchin, K. Konstas, C. J. Sumby, M. L. Czyz, P. Valente, M. R. Hill, A. Polyzos and C. J. Doonan, *Chem. Commun.*, 2015, **51**, 14231–14234.

[114] B. Içli, N. Christinat, J. Tönnemann, C. Schüttler, R. Scopelliti and K. Severin, *J. Am. Chem. Soc.*, 2009, **131**, 3154–3155.

[115] J. T. A. Jones, D. Holden, T. Mitra, T. Hasell, D. J. Adams, K. E. Jelfs, A. Trewin, D. J. Willock, G. M. Day, J. Bacsa, A. Steiner and A. Cooper, *Angew. Chem.*, 2010, **123**, 775–779.

[116] T. Düren and R. Q. Snurr, *J. Phys. Chem. B*, 2004, **108**, 15703–15708.

[117] H. Furukawa, Y. B. Go, N. Ko, Y. K. Park, F. J. Uribe-Romo, J. Kim, M. O'Keeffe and O. M. Yaghi, *Inorg. Chem.*, 2011, **50**, 9147–9152.

[118] M. A. Little, S. Y. Chong, M. Schmidtmann, T. Hasell and A. Cooper, *Chem. Commun.*, 2014, **50**, 9465–9468.

[119] T. Hasell, S. Y. Chong, K. E. Jelfs, D. J. Adams and A. Cooper, *J. Am. Chem. Soc.*, 2012, **134**, 588–598.

[120] A. G. Slater, M. A. Little, A. Pulido, S. Y. Chong, D. Holden, L. Chen, C. Morgan, X. Wu, G. Cheng, R. Clowes, M. E. Briggs, T. Hasell, K. E. Jelfs, G. M. Day and A. I. Cooper, *Nat. Chem.*, 2017, **9**, 17–25.

[121] N. O'Reilly, N. Giri and S. L. James, *Chem - Eur. J.*, 2007, **13**, 3020–3025.

[122] S. L. James, *Adv. Mater.*, 2016, **28**, 5712–5716.

[123] G. Melaugh, N. Giri, C. E. Davidson, S. L. James and M. G. Del Pópolo, *Phys. Chem. Chem. Phys.*, 2014, **16**, 9422–9431.

[124] N. Giri, C. E. Davidson, G. Melaugh, M. G. Del Pópolo, J. T. A. Jones, T. Hasell, A. Cooper, P. N. Horton, M. B. Hursthouse and S. L. James, 2012.

[125] N. Giri, M. G. Del Pópolo, G. Melaugh, R. L. Greenaway, K. Rätzke, T. Koschine, L. Pison, M. F. C. Gomes, A. Cooper and S. L. James, *Nat. Commun.*, 2015, **527**, 216–220.

[126] F. Zhang, F. Yang, J. Huang, B. G. Sumpter and R. Qiao, *J. Phys. Chem. B*, 2016, **120**, 7195–7200.

[127] S. L. Price, *Acc. Chem. Res.*, 2009, **42**, 117–126.

[128] G. M. Day, T. G. Cooper, A. J. Cruz-Cabeza, K. E. Hejczyk, H. L. Ammon, S. Boerrigter, J. S. Tan, R. G. Della Valle, E. Venuti, J. Jose, S. R. Gadre, G. Desiraju, T. S. Thakur, B. P. van Eijck, J. C. Facelli, V. E. Bazterra, M. B. Ferraro, D. W. M. Hofmann, M. A. Neumann, F. J. J. Leusen, J. Kendrick, S. L. Price, A. J. Misquitta, P. G. Karamertzanis, G. W. A. Welch, H. A. Scheraga, Y. Arnautova, M. U. Schmidt, J. van de Streek, A. K. Wolf and B. Schweizer, *Acta Crystallogr. Sect. B: Struct. Sci.*, 2009, **65**, 107–125.

[129] S. L. Price, *Chem. Soc. Rev.*, 2014, **43**, 2098–2111.

[130] A. M. Reilly, R. I. Cooper, C. S. Adjiman, S. Bhattacharya, A. D. Boese, J. G. Brandenburg, P. J. Bygrave, R. Bylsma, J. E. Campbell, R. Car, D. H. Case, R. Chadha, J. C. Cole, K. Cosburn, H. M. Cuppen, F. Curtis, G. M. Day, R. A. DiStasio, A. Dzyabchenko, B. P. Van Eijck, D. M. Elking, J. A. Van Den Ende, J. C. Facelli, M. B. Ferraro, L. Fusti-Molnar, C. A. Gatsiou, T. S. Gee, R. De Gelder, L. M. Ghiringhelli, H. Goto, S. Grimme, R. Guo, D. W. Hofmann, J. Hoja, R. K. Hylton, L. Iuzzolino, W. Jankiewicz, D. T. De Jong, J. Kendrick, N. J. De Klerk, H. Y. Ko, L. N. Kuleshova, X. Li, S. Lohani, F. J. Leusen, A. M. Lund, J. Lv, Y. Ma, N. Marom, A. E. Masunov, P. McCabe, D. P. McMahon, H. Meekes, M. P. Metz, A. J. Misquitta, S. Mohamed, B. Monserrat, R. J. Needs, M. A. Neumann, J. Nyman, S. Obata, H. Oberhofer, A. R. Oganov, A. M. Orendt, G. I. Pagola, C. C. Pantelides, C. J. Pickard, R. Podeszwa, L. S. Price, S. L. Price, A. Pulido, M. G. Read, K. Reuter, E. Schneider,

C. Schober, G. P. Shields, P. Singh, I. J. Sugden, K. Szalewicz, C. R. Taylor, A. Tkatchenko, M. E. Tuckerman, F. Vacarro, M. Vasileiadis, A. Vazquez-Mayagoitia, L. Vogt, Y. Wang, R. E. Watson, G. A. De Wijs, J. Yang, Q. Zhu and C. R. Groom, *Acta Crystallogr., Sect. B: Struct. Sci., Cryst. Eng. Mater.*, 2016, **72**, 439–459.

[131] E. O. Pyzer-Knapp, H. P. G. Thompson, F. Schiffmann, K. E. Jelfs, A. Cooper and G. M. Day, *Chem. Sci.*, 2014, **5**, 2235–2245.

[132] M. A. Little, M. E. Briggs, J. T. A. Jones, M. Schmidtmann, T. Hasell, S. Y. Chong, K. E. Jelfs, L. Chen and A. Cooper, *Nat. Chem.*, 2015, **7**, 153 –159.

[133] J. Pikunic, C. Clinard, N. Cohaut, K. Gubbins, J. Guet, R. Pellenq, I. Rannou and J. Rouzaud, 2003, **19**, 8565–8582.

[134] S. Jiang, J. T. A. Jones, T. Hasell, D. Bléger, C. E. Blythe, D. J. Adams, A. Trewin and A. Cooper, *Nat. Commun.*, 2011, **2**, 1–7.

[135] L. J. Abbott and C. M. Colina, *Macromolecules*, 2011, **44**, 4511–4519.

[136] G. S. Larsen, P. Lin, K. E. Hart and C. Colina, *Macromolecules*, 2011, **44**, 6944–6951.

[137] L. J. Abbott, A. G. McDermott, A. Del Regno, R. G. D. Taylor, C. G. Bezzu, K. J. Msayib, N. B. McKeown, F. R. Siperstein, J. Runt and C. M. Colina, *J. Phys. Chem. B*, 2013, **117**, 355–364.

[138] S. Jiang, K. E. Jelfs, D. Holden, T. Hasell, S. Y. Chong, M. Haranczyk, A. Trewin and A. I. Cooper, *J. Am. Chem. Soc.*, 2013, **135**, 17818–17830.

[139] J. D. Evans, D. M. Huang, M. R. Hill, C. J. Sumby, D. S. Sholl, A. W. Thornton and C. J. Doonan, *J. Phys. Chem. C*, 2015, **119**, 7746–7754.

[140] D. Holden, S. Y. Chong, L. Chen, K. E. Jelfs, T. Hasell and A. Cooper, *Chem. Sci.*, 2016, 1–6.

[141] L. Chen, P. S. Reiss, S. Y. Chong, D. Holden, K. E. Jelfs, T. Hasell, M. a. Little, A. Kewley, M. E. Briggs, A. Stephenson, K. M. Thomas, J. a. Armstrong, J. Bell, J. Busto, R. Noel, J. Liu, D. M. Strachan, P. K. Thallapally and A. I. Cooper, *Nat. Mater.*, 2014, **13**, 954–960.

[142] R. Manurung, D. Holden, M. Miklitz, L. Chen, T. Hasell, S. Y. Chong, M. Haranczyk, A. I. Cooper and K. E. Jelfs, *J. Phys. Chem. C*, 2015, **119**, 22577–22586.

[143] D. Holden, K. E. Jelfs, A. Cooper, A. Trewin, A. Trewin and D. J. Willock, *J. Phys. Chem. C*, 2012, **116**, 16639–16651.

[144] A. Kewley, A. Stephenson, L. Chen, M. E. Briggs, T. Hasell and A. Cooper, *Chem. Mater.*, 2015, **27**, 3207–3210.

[145] T. Mitra, K. E. Jelfs, M. Schmidtmann, A. Ahmed, S. Y. Chong, D. J. Adams and A. Cooper, *Nat. Chem.*, 2013, **5**, 276–281.

[146] T. Hasell, M. Miklitz, A. Stephenson, M. A. Little, S. Y. Chong, R. Clowes, L. Chen, D. Holden, G. A. Tribello, K. E. Jelfs and A. Cooper, *J. Am. Chem. Soc.*, 2016, **138**, 1653–1659.

[147] J. Camp and D. S. Sholl, *J. Phys. Chem. C*, 2015, 1110–1120.

[148] A. Kewley, A. Stephenson, L. Chen, M. E. Briggs, T. Hasell and A. Cooper, *Chem. Mater.*, 2015, **27**, 3207–3210.

[149] J. L. Daschbach, X. Sun, P. K. Thallapally, B. P. McGrail and L. X. Dang, *J. Phys. Chem. B*, 2010, **114**, 5764–5768.

[150] S. Alavi, T. K. Woo, A. Sirjoosingh, S. Lang, I. Moudrakovski and J. A. Ripmeester, *Chem–Eur. J.*, 2010, **16**, 11689–11696.

[151] Y. J. Colon and R. Q. Snurr, *Chem. Soc. Rev.*, 2014, **43**, 5735–5749.

[152] B. J. Sikora, C. E. Wilmer, M. L. Greenfield and R. Q. Snurr, *Chem. Sci.*, 2012, **3**, 2217–2223.

[153] C. E. Wilmer, O. K. Farha, Y.-S. Bae, J. T. Hupp and R. Q. Snurr, *Energy Environ. Sci.*, 2012, **5**, 9849–9856.

[154] C. E. Wilmer, M. Leaf, C. Y. Lee, O. K. Farha, B. G. Hauser, J. T. Hupp and R. Q. Snurr, 2011, **4**, 83–89.

[155] M. A. Zwijnenburg, E. Berardo, W. J. Peveler and K. E. Jelfs, *J. Phys. Chem. B*, 2016, **120**, 5063–5072.

[156] T. Hasell, M. A. Little, S. Y. Chong, M. Schmidtmann, M. E. Briggs, V. Santolini, K. E. Jelfs and A. I. Cooper, *Nanoscale*, 2017, **9**, 6783–6790.

[157] A. Pal and K. Vanka, *Chem. Commun.*, 2011, **47**, 11417–11419.

[158] G. Zhu, Y. Liu, L. Flores, Z. R. Lee, C. W. Jones, D. A. Dixon, D. S. Sholl and R. P. Lively, *Chem. Mater.*, 2018, **30**, 262–272.

[159] E. Polak and G. Ribiere, *ESAIM: Mathematical Modelling and Numerical Analysis*, 1969, **3**, 35–43.

[160] E. Harder, W. Damm, J. Maple, C. Wu, M. Reboul, J. Y. Xiang, L. Wang, D. Lupyan, M. K. Dahlgren, J. L. Knight, J. W. Kaus, D. S. Cerutti, G. Krilov, W. L. Jorgensen, R. Abel and R. A. Friesner, *J. Chem. Theory Comput.*, 2016, **12**, 281–296.

[161] J. VandeVondele, M. Krack, F. Mohamed, M. Parrinello, T. Chassaing and J. Hutter, *Comput. Phys. Commun.*, 2005, **167**, 103–128.

[162] D. F. Shanno, *Mathematics of Computation*, 1970, **24**, 647–647.

[163] A. Rappe, C. J. Casewit, K. Colwell, W. GoddardIII and W. M. Skiff, *J. Am. Chem. Soc.*, 1992, **114**, 10024–10035.

[164] A. Burgun, P. Valente, J. D. Evans, D. M. Huang, C. J. Sumby and C. J. Doonan, *Chem. Commun.*, 2016, **52**, 8850–8853.

[165] M. Saleh, H. M. Lee, K. C. Kemp and K. S. Kim, *ACS Appl. Mater. Interfaces*, 2014, **6**, 7325–7333.

[166] W. Jorgensen, D. Maxwell and J. TiradoRives, *J. Am. Chem. Soc.*, 1996, **118**, 11225–11236.

[167] W. Jorgensen and J. TiradoRives, 1988, **110**, 1657–1666.

[168] W. Cornell, P. Cieplak, C. Bayly, I. Gould, K. Merz, D. Ferguson, D. Spellmeyer, T. Fox, J. Caldwell and P. Kollman, 1995, **117**, 5179–5197.

[169] I. Kolossvary and W. Guida, *J. Am. Chem. Soc.*, 1996, **118**, 5011–5019.

[170] L. Verlet, *Phys. Rev.*, 1967, **159**, 98–103.

[171] G. A. Tribello, M. Bonomi, D. Branduardi, C. Camilloni and G. Bussi, *Comput. Phys. Commun.*, 2014, **185**, 604–613.

[172] P. Hohenberg and W. Kohn, *Phys. Rev.*, 1964, **136**, B864–B872.

[173] W. Kohn and L. J. Sham, *Phys. Rev.*, 1965, **140**, A1133–A1138.

[174] J. Perdew, K. Burke and M. Ernzerhof, *Phys. Rev. Lett.*, 1996, **77**, 3865–3868.

[175] Y. Zhao and D. G. Truhlar, *Theor. Chem. Acc.*, 2007, **120**, 215–241.

[176] S. Grimme, J. Antony, S. Ehrlich and H. Krieg, *J. Chem. Phys.*, 2010, **132**, 154104–1–154104–18.

[177] J. VandeVondele and J. Hutter, *J. Chem. Phys.*, 2007, **127**, 114105–1,114105–10.

[178] S. Goedecker, M. Teter and J. Hutter, *Phys. Rev. B*, 1996, **54**, 1703–1710.

[179] K. Acharyya and P. S. Mukherjee, *Chem. Commun.*, 2014, **50**, 15788–15791.

[180] A. J. Cruz-Cabeza, S. Karki, L. Fábián, T. Friščić, G. M. Day and W. Jones, *Chem. Commun.*, 2010, **46**, 2224–2226.

[181] D. Lewis, D. Willock, R. Catlow, J. Thomas and G. Hutchings, *Nat. Commun.*, 1996, **382**, 604–606.

[182] R. Pophale, F. Daeyaert and M. W. Deem, *J. Mater. Chem. A*, 2013, **1**, 6750–6760.

[183] Y. Liu, J.-H. Her, A. Dailly, A. J. Ramirez-Cuesta, D. A. Neumann and C. M. Brown, *J. Am. Chem. Soc.*, 2008, **130**, 11813–11818.

[184] F. X. Coudert, C. Mellot-Draznieks, A. H. Fuchs and A. Boutin, *J. Am. Chem. Soc.*, 2009, **131**, 3442–3443.

[185] L. Bouëssel du Bourg, A. U. Ortiz, A. Boutin and F.-X. Coudert, *APL Materials*, 2014, **2**, 124110–1–124110–10.

[186] L. Chen, J. P. S. Mowat, D. Fairen-Jimenez, C. A. Morrison, S. P. Thompson, P. A. Wright and T. Dueren, *J. Am. Chem. Soc.*, 2013, **135**, 15763–15773.

[187] K. I. Assaf and W. M. Nau, *Chem. Soc. Rev.*, 2014, **44**, 394–418.

[188] T. Brotin and J.-P. Dutasta, *Chem. Rev.*, 2009, **109**, 88–130.

[189] F. R. Robert Behrend, Eberhard Meyer, *Justus Liebig Annalen der Chemie*, 1905.

[190] W.-H. Huang, S. Liu, P. Y. Zavalij and L. Isaacs, *J. Am. Chem. Soc.*, 2006, **128**, 14744–14745.

[191] O. Taratula, P. A. Hill, N. S. Khan, P. J. Carroll and I. J. Dmochowski, *Nat. Commun.*, 2010, **1**, 148–157.

[192] T. P. Grubmuller H., Heymann B., *Science*, 1996, **271**, 997–999.

[193] P. J. Steinhardt, D. R. Nelson and M. Ronchetti, *Phys. Rev. B*, 1983, **28**, 784–805.

[194] L. M. Robeson, *J. Membr. Sci.*, 1991, **62**, 165–185.

[195] L. Isaacs, S.-K. Park, S. Liu, Y. H. Ko, N. Selvapalam, Y. Kim, H. Kim, P. Y. Zavalij, G.-H. Kim, H.-S. Lee and K. Kim, *J. Am. Chem. Soc.*, 2005, **127**, 18000–18001.

[196] L. Garel, J.-P. Dutasta and A. Collet, *Angew. Chem. Int. Ed.*, 1993, **32**, 1169–1171.

[197] K. Bartik, M. Luhmer, J.-P. Dutasta, A. Collet and J. Reisse, *J. Am. Chem. Soc.*, 1998, **120**, 784–791.

[198] G. Huber, T. Brotin, L. Dubois, H. Desvaux, J.-P. Dutasta and P. Berthault, *J. Am. Chem. Soc.*, 2006, **128**, 6239–6246.

[199] T. Brotin and J.-P. Dutasta, *Eur. J. Org. Chem.*, 2003, **2003**, 973–984.

[200] D. Barth, O. David, F. Quessette and V. Reinhard, *Lecture notes in Computer Science*, 2015, **9125**, 235–246.

[201] N. Christinat, R. Scopelliti and K. Severin, *Angew. Chem.*, 2008, **120**, 1874–1878.

[202] P. Kieryk, J. Janczak, J. Panek, M. Miklitz and J. Lisowski, *Org. Lett.*, 2015, **15**, 12–15.

[203] H. Takahagi, S. Fujibe and N. Iwasawa, *Chem. - Eur. J.*, 2009, **15**, 13327–13330.

[204] M. W. Schneider, I. M. Oppel and M. Mastalerz, *Chem. - Eur. J.*, 2012, **18**, 4156–4160.

[205] H. Ding, Y. Yang, B. Li, F. Pan, G. Zhu, M. Zeller, D. Yuan and C. Wang, *Chem. Commun.*, 2015, **51**, 1976–1979.

[206] P. Skowronek, U. Rychlewska and J. Gawronski, *Chem. Commun.*, 2013, **49**, 2524–2526.

[207] R. Warmuth, E. F. Maverick, C. B. Knobler and D. J. Cram, *J. Org. Chem.*, 2003, **68**, 2077–2088.

[208] V. Barba and I. Betanzos, *J. Organomet. Chem.*, 2007, **692**, 4903–4908.

[209] K. Kataoka, T. D. James and Y. Kubo, *J. Am. Chem. Soc.*, 2007, **129**, 15126–15127.

[210] M. E. Briggs, K. E. Jelfs, S. Y. Chong and C. Lester, *Cryst. Growth Des.*, 2013, **13**, 4993–5000.

[211] S. Hong, M. R. Rohman, J. Jia, Y. Kim, D. Moon, Y. Kim, Y. H. Ko, E. Lee and K. Kim, *Angew. Chem. Int. Ed.*, 2015, **54**, 13241–13244.

[212] Q. Wang, C. Zhang, B. C. Noll, H. Long, Y. Jin and W. Zhang, *Angew. Chem. Int. Ed.*, 2014, **53**, 10663–10667.

[213] Y. Jin, B. A. Voss, R. D. Noble and W. Zhang, *Angew. Chem., Int. Ed.*, 2010, **49**, 6348–6351.

[214] M. Brutschy, M. W. Schneider, M. Mastalerz and S. R. Waldvogel, *Adv. Mater.*, 2012, **24**, 6049–6052.

[215] V. Santolini, G. A. Tribello and K. E. Jelfs, *Chem. Commun.*, 2015, **51**, 15542–15545.

[216] M. Fujita, K. Umemoto, M. Yoshizawa, N. Fujita, T. Kusukawa and K. Biradha, *Chem. Commun.*, 2001, 509–518.

[217] A. Becke, *J. Chem. Phys.*, 1993, **98**, 5648–5652.

[218] C. Moller and M. Plesset, *Phys. Rev.*, 1934, **46**, 618–622.

[219] M. J. Frisch, G. W. Trucks, 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 and D. J. Fox, *Gaussian 09 Revision D.01*, 2010, Gaussian Inc. Wallingford CT.

[220] S. Grimme, *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 2011, **1**, 211–228.

[221] M. Valiev, E. J. Bylaska, N. Govind, K. Kowalski, T. P. Straatsma, H. J. J. Van Dam, D. Wang, J. Nieplocha, E. Apra, T. L. Windus and W. A. De Jong, *Comput. Phys. Commun.*, 2010, **181**, 1477–1489.

[222] M. Miklitz and K. E. Jelfs, *Manuscript in preparation*, 2018.

[223] M. Arunachalam, I. Ravikumar and P. Ghosh, *J. Org. Chem.*, 2008, **73**, 9144–9147.

[224] K. Acharyya, S. Mukherjee and P. S. Mukherjee, *J. Am. Chem. Soc.*, 2013, **135**, 554–557.

[225] N. D. Rycke, J. Marrot, F. Couty and O. R. David, *Tetrahedron Letters*, 2010, **51**, 6521 – 6525.

[226] R. Z. Khaliullin, J. VandeVondele and J. Hutter, *J. Chem. Theory Comput.*, 2013, **9**, 4421–4427.

[227] V. Santolini, M. Miklitz, E. Berardo and K. E. Jelfs, *Nanoscale*, 2017, **9**, 5280–5298.

[228] J. D. Evans, C. J. Sumby and C. J. Doonan, *Chemistry Letters*, 2015, **44**, 582–588.

[229] J. C. Lauer, W.-S. Zhang, F. Rominger, R. R. Schröder and M. Mastalerz, *Chem. - Eur. Jour.*, 2018, **24**, 1816–1820.

[230] W. Smith, C. W. Yong and P. M. Rodger, *Mol. Simul.*, 2002, **28**, 385–471.