Discovery of Synthesisable Organic Materials

by

Steven Bennett

Supervisors:
Prof. Kim E. Jelfs
Dr. Rebecca L. Greenaway

A document submitted in partial fulfilment of the requirements for the degree of
Doctor of Philosophy
at
Imperial College London

Declaration of Originality

I, Steven Bennett, certify that the work in this thesis, which I present for the Doctor of Philosophy award, is my original work unless otherwise referenced or acknowledged.

Copyright Decleration

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-Non Commercial 4.0 International Licence (CC BY-NC). Under this licence, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work. This is on the condition that: you credit the author and do not use it, or any derivative works, for a commercial purpose. When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes. Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

Abstract

Materials development is often synonymous with technological advancement, accompanying the progression of humanity, and can help address many contemporary issues humans face, including climate change, pollution, and resource scarcity. With their unique structural properties and vast surface areas, porous materials have emerged as promising materials for addressing some of these challenges, with applications in gas separation, storage, catalysis, and sensing. Porous organic cages are a sub-class of porous materials that have gathered significant attention in recent years due to their advantages over porous framework materials and the diverse range of geometric shapes that cages can form. However, the discovery of porous organic cages often relies on incremental modifications to known precursors or serendipitous discoveries. Computational screening can assist but is often hindered by the inability to experimentally realise many of the computationally predicted candidates.

In this thesis, I present a computational workflow to bridge the gap between computational prediction and synthetic realisation of porous organic cages. I developed a machine learning model to mimic the decision-making process of expert chemists when selecting synthesisable cage precursors. I used this model to filter a large library of cage precursors and used computational simulations to determine whether each pair of precursors would form a cage with a permanent internal cavity.

Next, I used the predictions from this computational screening workflow to guide the creation of a precursor library of diamines and trialdehydes, which were combined using high-throughput experimentation. To analyse the outcome of the 144 reactions performed, I developed an automated reaction analysis workflow to determine whether a cage formed. By combining this workflow with computational simulations, I was able to predict which precursor combinations would form shape persistent cages, identifying candidates that were expected to be porous. The automated reaction analysis workflow in my thesis provides a foundation to expand the number of reactions that can be performed and analysed in parallel using high-throughput experimentation, which could, in time, lead to the prediction of cage assembly using data-driven techniques.

Contents

Abstract 4

Acknowledgements 9

List of Publications 11

List of Figures 12

List of Tables 19

List of Abbreviations 21

1 Introduction 24
1.1 Porous Materials . . . 24
1.2 Porous Molecular Materials . . . 25
1.2.1 History of Porous Molecular Materials . . . 25
1.2.2 Porous Organic Cages . . . 26
1.2.3 Properties of Porous Materials . . . 29
1.3 Computational Modelling of Porous Molecular Materials . . . 30
1.3.1 Structure Prediction . . . 31
1.3.2 Disordered Materials . . . 35
1.3.3 Property Prediction . . . 35
1.4 Applications of Artificial Intelligence in Porous Molecular Materials . . . 36
1.4.1 Encoding Porous Materials . . . 37
1.5 Considering the Synthetic Accessibility of Porous Molecular Materials . . . 39
1.5.1 Exploring Materials Space Algorithmically . . . 42
1.6 Experimental Automation . . . 45
1.6.1 Closed-Loop Materials Discovery . . . 47
1.7 Thesis Structure . . . 48

2 Computational Methods 50
2.1 Introduction to Computational Chemistry . . . 50
2.1.1 Force Fields . . . 51
2.1.2 Geometry Optimisation . . . 53
2.1.3 Molecular Dynamics . . . 54
2.1.4 Density Functional Theory . . . 56
2.2 Machine Learning . . . 60
2.2.1 Decision Trees . . . 61
2.2.2 Performance Metrics . . . 62
2.2.3 Chemical Descriptors . . . 64
2.3 Quantifying Synthetic Difficulty . . . 65

3 Development and Comparison of Synthetic Difficulty Scoring Methods 68
3.1 Introduction . . . 68
3.2 Methods . . . 69
3.2.1 Datasets . . . 69
3.2.2 Obtaining Chemist Training Data . . . 70
3.2.3 Model Selection and Training . . . 73
3.3 Results and Discussion . . . 75
3.3.1 Labelled Training Data . . . 75
3.3.2 Materials Precursor Score (MPScore) Training and Evaluation . . . 78
3.3.3 Comparison with Existing Synthetic Accessibility Scores . . . 84
3.3.4 Conclusion . . . 88

4 High-Throughput Screening for Porous Organic Cages 89
4.1 Introduction . . . 89
4.2 Methods . . . 90
4.2.1 Computational Screening Workflow . . . 90
4.2.2 Identifying Synthesisable Precursors . . . 92
4.2.3 Cage Construction and Conformational Searching . . . 92
4.2.4 Porosity Evaluation . . . 93
4.3 Results and Discussion . . . 94
4.3.1 Precursor Database and Synthesisable Precursors . . . 94
4.3.2 Analysis of Shape Persistent POCs . . . 99
4.4 Conclusions . . . 104

5 Developing an Automated Experimental Workflow for Cage Discovery 106
5.1 Introduction . . . 106
5.1.1 Precursor Library . . . 107
5.1.2 Synthesis of Trialdehyde Precursors . . . 110
5.1.3 Experimental Conditions . . . 110
5.2 Computational Methods . . . 114
5.2.1 Automated Reaction Analysis . . . 114
5.2.2 ^{1}H NMR Spectroscopy . . . 115
5.2.3 HRMS . . . 117
5.2.4 Categorisation of High-Throughput Results . . . 121
5.2.5 Computational Prediction Workflow . . . 123
5.3 Results and Discussion . . . 124
5.3.1 High-Throughput Screening Results . . . 125
5.3.2 Clean Cage Hits . . . 125
5.3.3 Formation of a Single Cage Topologies . . . 130
5.3.4 Formation of a Mixture of Topologies . . . 135
5.3.5 Unsuccessful Combinations . . . 141
5.3.6 Considering Kinetic Intermediates . . . 143
5.3.7 Precursor Trends . . . 144
5.4 Conclusions . . . 153

6 Conclusions and Future Work 155
6.1 Future Work . . . 157

7 Experimental Data 159
7.1 General Synthetic and Analytical Methods . . . 159
7.1.1 Synthesis . . . 159
7.1.2 HRMS . . . 159
7.1.3 NMR . . . 159
7.1.4 IR . . . 160
7.1.5 High-Throughput Experimental Details . . . 160
7.2 Chapter 6 Experimental . . . 161
7.2.1 Synthesis of Trialdehydes . . . 161
7.2.2 High-Throughput Cage Screen . . . 165
7.2.3 High-Throughput Results . . . 174

Bibliography 208

A Appendix 220
A.1 Chapter 3 Supplementary Data 220
A.2 Chapter 5 Supplementary Data 246
A.2.1 Stacked NMR Spectra 246
A.2.2 High-Throughput HRMS Data 255
A.3 Copyright Permission Documents 294

Acknowledgements

I would first like to express my sincerest gratitude to all those that have contributed, both scientifically and personally, to the completion of this thesis.

First and foremost, I would like to extend my deepest appreciation to my supervisors, Prof. Kim Jelfs and Dr. Rebecca Greenaway. Their constant support, insights, and guidance have been invaluable in shaping this thesis, and I have felt extremely fortunate to have had them as my supervisors throughout the PhD journey. Not only has Kim's guidance been instrumental in developing my ability as a researcher, but she has provided me with countless opportunities for professional development, which will no doubt shape my future immensely. To Becky, who trusted a computational chemist enough to enter the lab, I am especially thankful for the opportunity to have had such an exciting project, her experimental insight, and her support throughout the PhD. I could not have asked for better supervisors to guide me through this challenging and rewarding journey.

I am also grateful to all the members of the Jelfs group, both past and present, for providing such a welcoming environment to conduct research, share ideas, and learn from each other, with whom many I hope to be lasting friends. In particular, I would like to thank (Dr.!) Irene Bechis, who I shared the majority of my PhD journey with, and whose afternoon chats and Benugo coffee breaks I will remember fondly; Dr. Sophie Finnigan for the cycling adventures; and Dr. Andrew Tarzia for always welcoming a discussion about many spontaneous ideas. In addition, I would also like to thank Dr. Lukas Turcani for his time teaching me the art of programming; Dr. Filip Szczypiński for his guidance, which has shaped my ability to perform research; Dr. Austin Mroz, whose incredible writing has helped shape my own; Dr. Emma Wolpert, whose incredible passion for science is a constant source of motivation; and Charlotte Breakwell, for her endless positivity. Also, I would like to thank Harry Hagan, Dr. James Pegg, Dr. Victor Posligua, Jiajun Zhou, and Dr. Julia Schmidt, who have each played a part in this journey.

I would also like to thank all the members of the Greenaway group for welcoming me so warmly into the lab, despite the uncertainty of having just come out of lockdown. In particular, I would like to thank Dr. Ben Egleston for introducing me to the lab, his wizardry with NMR, the many cycling adventures, and the many excellent samosas. I would also like to thank Ruby Morel, Zarik Yusaf, and Callum John, who have each made the lab a wonderful place to work.

Last but certainly not least from the Jelfs and Greenaway groups, I would like to extend a very large thanks to Annabel Basford. Together, we navigated the challenges of cage synthesis, and whose assistance in the lab throughout my PhD was essential in its completion. Her boundless enthusiasm and positive energy make the office and lab richer places to work.

Outside of my direct research groups, I would also like to thank the past and present members of Lab 109, particularly Dr. Florian Song, with whom countless coffee breaks and adventures were shared, both in person and virtually. The friendships I have formed with you and Irene have profoundly impacted my PhD journey, and I have no doubt that they will persevere beyond the geographical distances that may now separate us. In addition, I would also like to thank Oli, Léonie, and Francesca, who each made Lab 109 and Imperial a better place to be. I would also like to extend my thanks to my friends outside of Imperial, Rob and Amy, for the many visits to London, and to my friends from UCL, including Mira, Seb, Akhil, Krishan, Theo, James, and many others, whose long-lasting friendships I value dearly.

Importantly, I would also like to thank my fiancée, Sophie, with whom I have lived in three different houses and travelled to four different continents during this PhD. Together, we started our respective degrees, and together we are embarking on the next step of our lives. Thank you for your unwavering support, love, and motivation over the course of this PhD. I could not have completed this journey without you.

Finally, I would like to acknowledge my parents, Keith and Suzanne, for their endless support and guidance throughout my entire educational life, and my sister, Samantha; I would not be in the position I am today without you. Thank you.

List of Publications

The following published and unpublished work have contributed to the content of this thesis:

Journal Articles

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

S. Bennett, F. T. Szczypiński, L. Turcani, M. E. Briggs, R. L. Greenaway and K. E. Jelfs, Materials Precursor Score: Modeling Chemists' Intuition for the Synthetic Accessibility of Porous Organic Cage Precursors, *J. Chem. Inf. Model.*, 2021, 61, 4342–4356.

A. Basford, S. Bennett, R. L. Greenaway, and K. E. Jelfs, Streamlining the Automated Discovery of Porous Organic Cages, *unpublished work*

Book Chapters

S. Bennett, A. Tarzia, M. A. Zwijnenburg and K. E. Jelfs, in *Machine Learning in Chemistry: The Impact of Artificial Intelligence*, ed. H. M. Cartwright, Royal Society of Chemistry, Cambridge, 2020, pp. 280–310.

S. Bennett and K. E. Jelfs, in *AI-Guided Design and Property Prediction for Zeolites and Nanoporous Materials*, John Wiley & Sons, Ltd, 2023, pp. 251–282.

List of Figures

1.1 Examples of porous molecular materials . . . 26
1.2 The synthesis of **CC3** . . . 28
1.3 Examples of cages formed using irreversible bond formation . . . 29
1.4 Post-synthetic modification of **CC3**, forming **RCC3** and **FT-RCC3** . . . 30
1.5 The origin of intrinsic and extrinsic porosity in **CC2** . . . 31
1.6 Construction of different topologies using *stk* . . . 32
1.7 The energy-structure-functions maps of three porous molecular materials . . . 34
1.8 Overview of molecular descriptors . . . 39
1.9 Overview of methods used to quantify synthetic difficulty . . . 41
1.10 Examples of algorithms for generative molecular design . . . 43
1.11 Results of a high-throughput experimental screen . . . 46

2.1 Example of navigation across a potential energy surface . . . 54
2.2 Example confusion matrix for a binary classification problem . . . 63
2.3 Overview of the Extended-Connectivity Fingerprint algorithm . . . 65

3.1 Training dataset scaffold analysis . . . 71
3.2 Training data labelling interface . . . 72
3.3 Molecules where chemists disagreed . . . 77
3.4 Precision-recall curve . . . 80
3.5 Calibration curve . . . 81
3.6 Model feature importances . . . 82
3.7 Representative molecules from each class . . . 83
3.8 Comparison between synthetic difficulty scores . . . 85
3.9 Change in synthetic difficulty values during the synthesis of Polmacoxib . . . 86

4.1 Computational pipeline for the high-throughput computational screen . . . 91
4.2 Distribution of synthetic difficulty scores of the precursor library . . . 95
4.3 Distriubtion of cavity sizes for shape persistent POCs . . . 98
4.4 Structural properties of POC precursors . . . 100

4.5 Most frequently appearing precursors 102
4.6 The six largest POCs from the high-throughput computational screen 103

5.1 Precursor library for the high-throughput experimental screen 108
5.2 Stacked ^{1}H NMR spectra showing the effect of changing precursor equivalents 113
5.3 Stacked ^{1}H NMR spectra showing the effect of reaction time 114
5.4 HRMS of the initial condition screening of **G16** (**CC3**) 115
5.5 Example ^{1}H NMR spectra of the different outcomes 117
5.6 Example of the high-throughput mass ion detection workflow 120
5.7 Overview of the results from the high-throughput experimental screen 126
5.8 Sankey diagram illustrating different outcomes of the high-throughput experimental screen 127
5.9 Precursor combinations in category 1 128
5.10 Stacked ^{1}H NMR spectra of **CC3**, **S16**, and **I16** 129
5.11 Relative energy per formula unit plots of **I16** and **S16** 130
5.12 Precursor combinations in category 2 132
5.13 Relative energies per formula unit plot of **S28** 132
5.14 Precursor combinations in categories 5 and 6 134
5.15 Relative energy per formula unit plot for **G16** 135
5.16 Precursor combinations in category 3 137
5.17 Precursor combinations in category 4 138
5.18 Relative energy per formula unit plots of **G27**, **I28**, **I27**, and **K16** 140
5.19 Precursor combinations in categories 7 and 8 142
5.20 Relative energies per formula unit plots of **I15** and **J32** 143
5.21 The number of times a precursor appeared in an experimentally observed cage . 145
5.22 Examples of cages formed from flexible diamines 146
5.23 Examples of cages formed from rigid diamines 147
5.24 Colour map showing the formation energy per imine bond for each precursor combination in four different topologies 148
5.25 Colour map showing the cavity diameter of experimentally formed cages 151
5.26 Experimentally observed cages predicted to be shape persistent 152

7.1 ^{1}H NMR spectrum ($CDCl_3$) of trialdehyde **R**. 162
7.2 ^{13}C NMR spectrum ($CDCl_3$) of trialdehyde **R**. 162
7.3 ^{1}H NMR spectrum ($(CD_3)_2SO$) of trialdehyde **U**. 164
7.4 ^{13}C NMR spectrum ($(CD_3)_2SO$) of trialdehyde **U**. 164
7.5 ^{1}H-^{13}C HSQC spectrum ($(CD_3)_2SO$) of trialdehyde **U**. 165

7.6 Crude ^{1}H NMR spectrum ($CDCl_3$) for **G29**$_{[2+3]}$ taken directly from the high-throughput screen. . . . 174
7.7 HRMS spectra for crude reactions taken from the high-throughput screen for cage **G29**$_{[2+3]}$ . . . 175
7.8 Crude ^{1}H NMR spectrum ($CDCl_3$) for **H25**$_{[2+3]}$ taken directly from the high-throughput screen. . . . 176
7.9 HRMS spectra for crude reactions taken from the high-throughout screen for cage **H25**$_{[2+3]}$ . . . 177
7.10 Crude ^{1}H NMR spectrum ($CDCl_3$) for **U10**$_{[2+3]}$ taken directly from the high-throughput screen. . . . 178
7.11 HRMS spectra for crude reactions taken from the high-throughput screen for cage **U10**$_{[2+3]}$ . . . 179
7.12 Crude ^{1}H NMR spectrum ($CDCl_3$) for **U30**$_{[2+3]}$ taken directly from the high-throughput screen. . . . 180
7.13 HRMS spectra for crude reactions taken from the high-throughput screen for cage **U30**$_{[2+3]}$ . . . 181
7.14 Crude ^{1}H NMR spectrum ($CDCl_3$) for **I16**$_{[4+6]}$ taken directly from the high-throughput screen. . . . 182
7.15 HRMS spectra for crude reactions taken from the high-throughout screen for cage **I16**$_{[4+6]}$ . . . 183
7.16 Crude ^{1}H NMR spectrum ($CDCl_3$) for **I25**$_{[2+3]}$ taken directly from the high-throughput screen. . . . 184
7.17 HRMS spectra for crude reactions taken from the high-throughput screen for cage **I25**$_{[2+3]}$ . . . 185
7.18 Crude ^{1}H NMR spectrum ($CDCl_3$) for **I29**$_{[2+3]}$ taken directly from the high-throughput screen. . . . 186
7.19 HRMS spectra for crude reactions taken from the high-throughput screen for cage **I29**$_{[2+3]}$ . . . 187
7.20 Crude ^{1}H NMR spectrum ($CDCl_3$) for **I31**$_{[2+3]}$ taken directly from the high-throughput screen. . . . 188
7.21 HRMS spectra for crude reactions taken from the high-throughput screen for cage **I31**$_{[2+3]}$ . . . 189
7.22 Crude ^{1}H NMR spectrum ($CDCl_3$) for **K10**$_{[2+3]}$ taken directly from the high-throughput screen. . . . 190
7.23 HRMS spectra for crude reactions taken from the high-throughout screen for cage **K10**$_{[2+3]}$ . . . 191

7.24 Crude ^{1}H NMR spectrum ($CDCl_3$) for **R25**$_{[2+3]}$ taken directly from the high-throughput screen. . . . 192
7.25 HRMS spectra for crude reactions taken from the high-throughput screen for cage **R25**$_{[2+3]}$ . . . 193
7.26 Crude ^{1}H NMR spectrum ($CDCl_3$) for **R28** taken directly from the high-throughput screen. . . . 194
7.27 HRMS spectra for crude reactions taken from the high-throughput screen for cage **R28**$_{[2+3]}$ . . . 195
7.28 Crude ^{1}H NMR spectrum ($CDCl_3$) for **S10**$_{[2+3]}$ taken directly from the high-throughput screen. . . . 196
7.29 HRMS spectra for crude reactions taken from the high-throughput screen for cage **S10**$_{[2+3]}$ . . . 197
7.30 Crude ^{1}H NMR spectrum ($CDCl_3$) for **S16**$_{[4+6]}$ taken directly from the high-throughput screen. . . . 198
7.31 HRMS spectra for crude reactions taken from the high-throughout screen for cage **S16**$_{[4+6]}$ . . . 199
7.32 Crude ^{1}H NMR spectrum ($CDCl_3$) for **S25**$_{[2+3]}$ taken directly from the high-throughput screen. . . . 200
7.33 HRMS spectra for crude reactions taken from the high-throughout screen for cage **S25**$_{[2+3]}$ . . . 201
7.34 Crude ^{1}H NMR spectrum ($CDCl_3$) for **S26**$_{[2+3]}$ taken directly from the high-throughput screen. . . . 202
7.35 HRMS spectra for crude reactions taken from the high-throughput screen for cage **S26**$_{[2+3]}$ . . . 203
7.36 Crude ^{1}H NMR spectrum ($CDCl_3$) for **S30**$_{[2+3]}$ taken directly from the high-throughput screen. . . . 204
7.37 HRMS spectra for crude reactions taken from the high-throughput screen for cage **S30**$_{[2+3]}$ . . . 205
7.38 Crude ^{1}H NMR spectrum ($CDCl_3$) for **S32**$_{[2+3]}$ taken directly from the high-throughput screen. . . . 206
7.39 HRMS spectra for crude reactions taken from the high-throughput screen for cage **S32**$_{[2+3]}$ . . . 207

A.1 Molecules provided to R.L.G for the blind classification task . . . 220
A.2 Stacked ^{1}H NMR spectra ($CDCl_3$) of the crude reactions of diamine **10** with trialdehdyes **G-U**, taken directly from the high-throughput screen. . . . 247

A.3 Stacked ^{1}H NMR spectra ($CDCl_3$) of the crude reactions of diamine **11** with trialdehdyes **G-U**, taken directly from the high-throughput screen. 247
A.4 Stacked ^{1}H NMR spectra ($CDCl_3$) of the crude reactions of diamine **12** with trialdehdyes **G-U**, taken directly from the high-throughput screen. 248
A.5 Stacked ^{1}H NMR spectra ($CDCl_3$) of the crude reactions of diamine **14** with trialdehdyes **G-U**, taken directly from the high-throughput screen. 248
A.6 Stacked ^{1}H NMR spectra ($CDCl_3$) of the crude reactions of diamine **15** with trialdehdyes **G-U**, taken directly from the high-throughput screen. 249
A.7 Stacked ^{1}H NMR spectra ($CDCl_3$) of the crude reactions of diamine **16** with trialdehdyes **G-U**, taken directly from the high-throughput screen. 249
A.8 Stacked ^{1}H NMR spectra ($CDCl_3$) of the crude reactions of diamine **25** with trialdehdyes **G-U**, taken directly from the high-throughput screen. 250
A.9 Stacked ^{1}H NMR spectra ($CDCl_3$) of the crude reactions of diamine **26** with trialdehdyes **G-U**, taken directly from the high-throughput screen. 250
A.10 Stacked ^{1}H NMR spectra ($CDCl_3$) of the crude reactions of diamine **27** with trialdehdyes **G-U**, taken directly from the high-throughput screen. 251
A.11 Stacked ^{1}H NMR spectra ($CDCl_3$) of the crude reactions of diamine **28** with trialdehdyes **G-U**, taken directly from the high-throughput screen. 251
A.12 Stacked ^{1}H NMR spectra ($CDCl_3$) of the crude reactions of diamine **29** with trialdehdyes **G-U**, taken directly from the high-throughput screen. 252
A.13 Stacked ^{1}H NMR spectra ($CDCl_3$) of the crude reactions of diamine **30** with trialdehdyes **G-U**, taken directly from the high-throughput screen. 252
A.14 Stacked ^{1}H NMR spectra ($CDCl_3$) of the crude reactions of diamine **31** with trialdehdyes **G-U**, taken directly from the high-throughput screen. 253
A.15 Stacked ^{1}H NMR spectra ($CDCl_3$) of the crude reactions of diamine **32** with trialdehdyes **G-U**, taken directly from the high-throughput screen. 253
A.16 Stacked ^{1}H NMR spectra ($CDCl_3$) of the crude reactions of diamine **33** with trialdehdyes **G-U**, taken directly from the high-throughput screen. 254
A.17 Stacked ^{1}H NMR spectra ($CDCl_3$) of the crude reactions of diamine **34** with trialdehdyes **G-U**, taken directly from the high-throughput screen. 254
A.18 HRMS spectra for crude reactions taken from the high-throughput screen for cage $\mathbf{H26}_{[2+3]}$ 255
A.19 HRMS spectra for crude reactions taken from the high-throughput screen for cage $\mathbf{S28}_{[4+6]}$ 255
A.20 HRMS spectra for crude reactions taken from the high-throughput screen for cage $\mathbf{U28}_{[2+3]}$ 256

A.21 HRMS spectra for crude reactions taken from the high-throughput screen for cage **S31**$_{[2+3]}$ . . . 256
A.22 HRMS spectra for crude reactions taken from the high-throughput screen for cage **H31**$_{[2+3]}$ . . . 257
A.23 HRMS spectra for crude reactions taken from the high-throughput screen for cage **K25**$_{[2+3]}$ . . . 257
A.24 HRMS spectra for crude reactions taken from the high-throughput screen for cage **T25**$_{[2+3]}$ . . . 258
A.25 HRMS spectra for crude reactions taken from the high-throughput screen for cage **U26**$_{[2+3]}$ . . . 258
A.26 HRMS spectra for crude reactions taken from the high-throughput screen for cage **T26**$_{[2+3]}$ . . . 259
A.27 HRMS spectra for crude reactions taken from the high-throughput screen for cage **T10**$_{[2+3]}$ . . . 259
A.28 HRMS spectra for crude reactions taken from the high-throughput screen for cage **I28**$_{[4+6],[6+9]}$ . . . 260
A.29 HRMS spectra for crude reactions taken from the high-throughput screen for cage **G27**$_{[2+3],[4+6]}$ . . . 261
A.30 HRMS spectra for crude reactions taken from the high-throughput screen for cage **G31**$_{[2+3],[4+6]}$ . . . 262
A.31 HRMS spectra for crude reactions taken from the high-throughput screen for cage **J31**$_{[2+3],[4+6],[6+9]}$ . . . 263
A.32 HRMS spectra for crude reactions taken from the high-throughput screen for cage **G30**$_{[2+3],[4+6]}$ . . . 264
A.33 HRMS spectra for crude reactions taken from the high-throughput screen for cage **I30**$_{[2+3],[4+6]}$ . . . 265
A.34 HRMS spectra for crude reactions taken from the high-throughput screen for cage **G25**$_{[2+3],[4+6]}$ . . . 266
A.35 HRMS spectra for crude reactions taken from the high-throughput screen for cage **I26**$_{[2+3],[4+6]}$ . . . 267
A.36 HRMS spectra for crude reactions taken from the high-throughput screen for cage **K29**$_{[2+3],[4+6]}$ . . . 268
A.37 HRMS spectra for crude reactions taken from the high-throughput screen for cage **U25**$_{[2+3],[4+6]}$ . . . 269
A.38 HRMS spectra for crude reactions taken from the high-throughput screen for cage **U31**$_{[2+3],[4+6]}$ . . . 269

A.39 HRMS spectra for crude reactions taken from the high-throughput screen for cage **K16**$_{[4+6],\,[8+12]}$ 270
A.40 HRMS spectra for crude reactions taken from the high-throughput screen for cage **I27**$_{[2+3],\,[4+6]}$ 271
A.41 HRMS spectra for crude reactions taken from the high-throughput screen for cage **K31**$_{[4+6],\,[6+9],\,[8+12]}$ 272
A.42 HRMS spectra for crude reactions taken from the high-throughput screen for cage **K30**$_{[2+3],\,[4+6]}$ 273
A.43 HRMS spectra for crude reactions taken from the high-throughput screen for cage **H30**$_{[2+3],\,[4+6]}$ 274
A.44 HRMS spectra for crude reactions taken from the high-throughput screen for cage **K26**$_{[2+3],\,[4+6]}$ 275
A.45 HRMS spectra for crude reactions taken from the high-throughput screen for cage **G16**$_{[4+6]}$ 276
A.46 HRMS spectra for crude reactions taken from the high-throughput screen for cage **H15**$_{[2+3]}$ 276
A.47 HRMS spectra for crude reactions taken from the high-throughput screen for cage **R34**$_{[2+3]}$ 277
A.48 HRMS spectra for crude reactions taken from the high-throughput screen for cage **T12**$_{[2+3]}$ 277
A.49 HRMS spectra for crude reactions taken from the high-throughput screen for cage **J15**$_{[2+3]}$ 278
A.50 HRMS spectra for crude reactions taken from the high-throughput screen for cage **J26**$_{[2+3]}$ 278
A.51 HRMS spectra for crude reactions taken from the high-throughput screen for cage **K32**$_{[2+3]}$ 279
A.52 HRMS spectra for crude reactions taken from the high-throughput screen for cage **R15**$_{[2+3]}$ 279
A.53 HRMS spectra for crude reactions taken from the high-throughput screen for cage **R27**$_{[2+3]}$ 280
A.54 HRMS spectra for crude reactions taken from the high-throughput screen for cage **R31**$_{[2+3]}$ 280
A.55 HRMS spectra for crude reactions taken from the high-throughput screen for cage **R32**$_{[2+3]}$ 281
A.56 HRMS spectra for crude reactions taken from the high-throughput screen for cage **H10**$_{[2+3]}$ 282

A.57 HRMS spectra for crude reactions taken from the high-throughput screen for cage **J10**$_{[2+3]}$ 282
A.58 HRMS spectra for crude reactions taken from the high-throughput screen for cage **K15**$_{[2+3]}$ 283
A.59 HRMS spectra for crude reactions taken from the high-throughput screen for cage **R10**$_{[2+3]}$ 283
A.60 HRMS spectra for crude reactions taken from the high-throughput screen for cage **U15**$_{[2+3]}$ 284
A.61 HRMS spectra for crude reactions taken from the high-throughput screen for cage **I10**$_{[2+3],\,[4+6]}$ 285
A.62 HRMS spectra for crude reactions taken from the high-throughput screen for cage **I15**$_{[4+6],\,[6+9],\,[8+12]}$ 286
A.63 HRMS spectra for crude reactions taken from the high-throughput screen for cage **J25**$_{[2+3],\,[4+6]}$ 287
A.64 HRMS spectra for crude reactions taken from the high-throughput screen for cage **J32**$_{[2+3],\,[4+6]}$ 288
A.65 HRMS spectra for crude reactions taken from the high-throughput screen for cage **R29**$_{[2+3],\,[4+6]}$ 289
A.66 HRMS spectra for crude reactions taken from the high-throughput screen for cage **I32**$_{[2+3],\,[4+6]}$ 290
A.67 HRMS spectra for crude reactions taken from the high-throughput screen for cage **T29**$_{[2+3],\,[4+6]}$ 291
A.68 HRMS spectra for crude reactions taken from the high-throughput screen for cage **T30**$_{[2+3],\,[4+6]}$ 292
A.69 HRMS spectra for crude reactions taken from the high-throughput screen for cage **T31**$_{[2+3],\,[4+6]}$ 293
A.70 The copyright permission document permitting re-use of content in reference (39). 294
A.71 The copyright permission document permitting re-use of content in reference (40). 294

List of Tables

3.1 Breakdown of all molecules in the training dataset by their functional group . . 71
3.2 Hyperparameter options explored in the random forest model 74
3.3 The number of molecules labelled by each chemist 76
3.4 A summary of the molecules labelled by each synthetic materials chemist 77
3.5 Chemists' answers to the question, "Can you make 1 g of this compound in under 5 steps?" 78
3.6 Evaluation metrics for the MPScore model 78
3.7 Cumulative metrics across each of the 5-folds in the cross-validation procedure . 79
3.8 Performance metrics for the blind validation task used to assess the performance of the SAScore, SCScore and MPScore 87

4.1 Number of overlapping precursors scored as easy-to-synthesise 96
4.2 The number and percentage of shape persistent cages formed from precursors selected by each synthetic difficulty model 97
4.3 The six largest shape persistent POCs are presented with their corresponding precursor pair synthetic difficulty scores and calculated cavity diameters. 104

5.1 Categorisation criteria of high-throughput screening results 121
5.2 The breakdown of precursor combinations where kinetic intermediate was observed to form by HRMS 144

7.1 Precursor stock solutions used in the high-throughput screen 166
7.2 Reaction volumes used for each stock solution in the high-throughput screen . 167
7.3 Precursor combination categorisations 171

A.1 Synthetic difficulty scores of molecules provided to R.L.G for the blind classification task 242

List of Abbreviations

MOF Metal-Organic Framework

COF Covalent Organic Framework

PMM Porous Molecular Material

SOF Supramolecular Organic Framework

POC Porous Organic Cage

DCC Dynamic Covalent Chemistry

FF Force Field

ETKDG Experimental-Torsion-Knowledge Distance Geometry

UFF Universal Force Field

CSP Crystal Structure Prediction

DFT Density Functional Theory

MD Molecular Dynamics

PL Porous Liquid

AI Artificial Intelligence

ML Machine Learning

RF Random Forest

NN Neural Network

DL Deep Learning

ECFP Extended-Connectivity Fingerprint

SOAP Smooth Overlap of Atomic Positions

CSD Cambridge Structural Database

SVD Singular Value Decomposition

GNN Graph Neural Network

RNN Recurrent Neural Network

MOC Metal-Organic Cage

QSAR Quantitative Structure–Activity Relationship

BO Bayesian Optimisation

HTE High-Throughput Experimentation

LDA Local Density Approximation

HEG Homogenous Electron Gas

GGA Generalized Gradient Approximation

HF Hartree-Fock

CART Classification and Regression Tree

RMSE Root Mean Square Error

MSE Mean Square Error

ROC Receiver Operating Characteristic

HRMS High-Resolution Mass Spectrometry

NMR Nuclear Magnetic Resonance

IR Infrared

GIAO Gauge Invariant Atomic Orbitals

ADAP Automated Data Analysis Pipeline

HPLC High-Performance Liquid Chromatography

QTOF Quadrupole Time-of-Flight

PXRD Powder X-ray Diffraction

Chemical Abbreviations

TPP Tris(o-phenylenedioxy) phosphonitrile trimer

TTEB 3,3,4,4-Tetrakis(trimethylsilylethynyl)biphenyl

TFB Triformylbenzene

CHDA Cyclohexanediamine

Tp 1,3,5-Triformylphloroglucinol

TEA Triethylamine

1 Introduction

1.1 Porous Materials

Porous materials have a rich history spanning thousands of years. Dating back to ancient times when activated charcoal was used for various purposes to modern applications of naturally occurring or engineered zeolites, humans have been utilising these materials throughout history.[1] In today's world, porous materials have a diverse range of industrial applications, including washing detergents,[2] gas sorbents,[3] and molecular separations.[4] Current thermal separation processes for separating mixtures consume 10-15% of the world's energy.[5] However, these methods can be substituted with porous material membranes that are tailored for separation, resulting in significant energy savings of up to 100 million tonnes of CO_2 per year, making them an appealing alternative material for industrial scale separation processes.[6]

Many top performing porous materials currently are based on inorganic or hybrid organic-inorganic materials.[7] In each of these materials, an extended network of stronger covalent or coordination bonding creates a permanent rigid network of pores. Zeolites, metal-organic frameworks (MOFs), and covalent organic frameworks (COFs) each fit into this category; with our expanding ability to understand structure on the macro-, meso- and nanoscopic scale using computational models and experimental techniques, we are now better able to design porosity in these materials. Porosity in these materials can be measured by the specific surface area that is accessible to gas molecules in the solid state.

Some of the earliest examples of inorganic porous materials were zeolites, where early examples typically achieved pore sizes below 8 Å.[8] Now, zeolites are used commercially, such as zeolite A, where millions of tonnes are used per year on applications from catalysis to washing detergents.[9] Of the zeolites that have found commercial use, Zeolite A is one of the earliest discovered and most well-studied zeolites,[10] with a calculated pore volume of 0.508 $cm^3\ g^{-1}$.[11] For MOFs meanwhile, the highest porosity achieved is 5.02 $cm^3\ g^{-1}$,[12] an order of magnitude greater than the specific pore volume of zeolite A,[11] and it is believed to approach the possible upper limit of solid-state porous materials.[13]

While network materials attain ever-higher porosities, they suffer from several drawbacks. Generally, network materials are insoluble in a wide range of solvents, making the fabrication of mem-

branes challenging. The same lack of solubility makes the separation of impurities challenging and reduces the ability to post-functionalise the material to alter its properties. Inorganic materials are also limited to a few specific elements, lack the rich diversity of organic materials, and lack porosity originating from the dynamic behaviour of flexible organic molecules.[14] Due to these limitations, there has been increased interest in the discovery of new porous materials comprised of discrete or polymeric organic molecules.

1.2 Porous Molecular Materials

Porous molecular materials (PMMs) are an emerging subclass of porous materials, which do not feature an extended network of strong intermolecular bonding, and are often defined by the absence of directional covalent and coordination bonding.[15] Specifically, PMMs are defined as porous solids or liquids with only discrete molecular subunits connected only through weaker intermolecular bonding interactions, such as hydrogen bonding, van der Waals, and dipole interactions. Both organic and metal-organic molecules have been shown to produce PMMs; however, my thesis notably excluded metal-organic PMMs, instead focussing on their organic counterparts.[15] Their organic nature offers several advantages over their network counterparts, as outlined in the previous section. Indeed, their organic nature has generated great interest as chemists can now apply a catalogue of synthetic organic transformations to create a new material with entirely new properties. Despite their inherent advantages, PMMs are rare, and their discovery is increasingly the product of combined experimental and computational efforts.[16]

1.2.1 History of Porous Molecular Materials

Interest in PMMs originated with the discovery of Dianin's molecule (Figure 1.1), initially synthesised in 1914; its porosity was later demonstrated in 1976.[17] Featuring a relatively simple molecular structure, inefficient packing in the crystal structure led to the formation of hourglass-shaped columns running along a single axis.[18] Notably, these columns did not collapse upon desolvation, showcasing the structural rigidity of organic molecules and their ability to adsorb gas molecules in the solid state.[19] Over 100 years after the discovery, the host-guest interactions of this material are still being studied.[20]

Since this initial discovery, a plethora of notable examples of porous molecular crystals have been discovered. Following Dianin's compound was the discovery tris(o-phenylenedioxy) phosphonitrile trimer (**TPP**),[21]; despite possessing only weakly directional intermolecular interactions, the steric bulk of substituents encouraged lower density packing arrangements. Similar lower density packing afforded the porous material 3,3,4,4-tetrakis(trimethylsilylethynyl)biphenyl (**TTEB**).[22]

Stronger, more directional intermolecular interactions such as π-π stacking and hydrogen bonding have been exploited in the hydrogen-bonded organic framework (HOF) and supramolecular organic framework (SOF) porous materials 9,10-bis(4-((3,5-dicyano-2,6-dipyridyl) dihydropyridyl)phenyl) anthracene (**SOF-1**),[23] and 4,4,4,4-tetra(4,6-diamino-s-triazin-2-yl) tetraphenylmethane (**HOF-1**),[24] and several crystaline dipeptides,[25] each shown in Figure 1.1. In each of these examples, PMMs possessed surface areas of around 200-300 m^2 g^{-1}, substantially lower than their network counterparts due to the attractive intermolecular interactions that dominate.

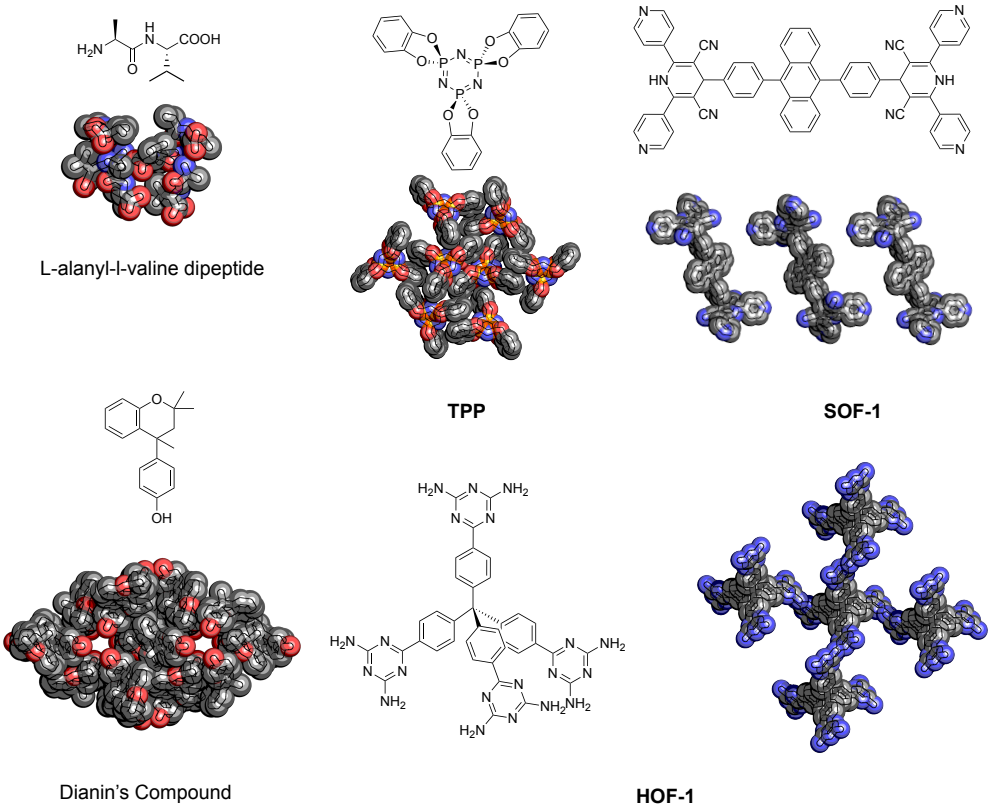


Figure 1.1: Examples of porous molecular crystals extracted from the unit cell in their crystal structure. Oxygen is shown in red, nitrogen in blue, and carbon is shown in black. All hydrogen atoms have been omitted for clarity, and solvent molecules have been removed from the **HOF-1** crystal structure.

1.2.2 Porous Organic Cages

Porous organic cages (POCs) are discrete molecules that possess a permanent internal cavity, which is accessible to guest molecules by windows surrounding the internal cavity. A molecule with more than two windows is termed a cage.[26] These windows can hinder diffusion out of and into the internal cavity, giving rise to potential applications, including selective guest adsorption and sep-

arations. Recently, these types of materials have been able to achieve porosity values comparable to network materials, including achieving surface areas of 3,758 m^2g^{-1}.[27]

In the literature, POC formation is often limited to a few robust reversible dynamic covalent chemistry (DCC) reactions. While imine condensation is the most popular,[28] alkyne metathesis,[29] and boronic ester condensation have also been used in the literature.[27] Other reversible reactions have been employed for cage synthesis, although less frequently, including disulfide formation and boroxine formation.[30]

POCs have been observed to form highly symmetrical structures based on archimedean and platonic solids, which are denoted the topology of the cage.[31] The thermodynamic control of DCC reactions allows the formation of the most energetically stable product, which can result in the formation of these highly symmetrical structures. Cage formation typically relies on the assumption that, given enough time, the thermodynamic product will dominate over shorter-lived kinetic intermediates. Santolini *et al.* defined a naming convention for the topology of POCs, describing the topology in terms of the topicity of the precursors that form the cage and the number of precursor molecules that make up the discrete unit of the cage.[31] For example, **Tri**^{4}**Di**^{6} describes the topology of a cage consisting of four tritopic precursors and six ditopic precursors, thus describing a POC with the underlying topology of an octahedron. The imine condensation used to form a **Tri**^{4}**Di**^{6} topology can also be described by the notation [4 + 6], representing the reaction between four aldehydes with six amines.

Since the initial discovery of covalent cages,[32,33] a plethora of POCs have been discovered with a diverse range of topologies, including the notable discoveries of a series of triptycene-based cages,[27] cage capsules,[34], and cavitands.[35] A prominent organic cage in literature, covalent cage 3 (**CC3**), is shown in Figure 1.2. **CC3** is formed from the imine condensation of 4 equivalents of triformylbenzene (TFB) with 6 equivalents cyclohexanediamine (CHDA) to form a POC which maintains a permanent internal cavity upon desolvation and has been extensively studied.[36] The flexibility of imine condensation can also result in the formation of exotic structures, such as a dumbbell-shaped cage and an organic cage pot.[37] Of the other types of DCC reactions, boronate-ester condensation was used to synthesise a [8 + 12] cubic structure with an exceptionally large pore volume of 1.84 cm^3g^{-1}.[27]

While each of these examples exploits the reversible nature of the DCC reaction, irreversible reaction chemistries have also been used to form POCs. These include "click" chemistry (Figure 1.3a),[39] more general nucleophilic substitution reactions (Figure 1.3b),[40] and amide bond formation (Figure 1.3c).[41] Delecluse *et al.* exploited irreversible bond formation to form a small hemicryptophane cage using urea bond formation, enhancing anion binding due to the presence of urea moieties and the stability of the structure to decomposition.[42] However, the successful application of irreversible bond formation heavily relies on pre-organised precursors with com-

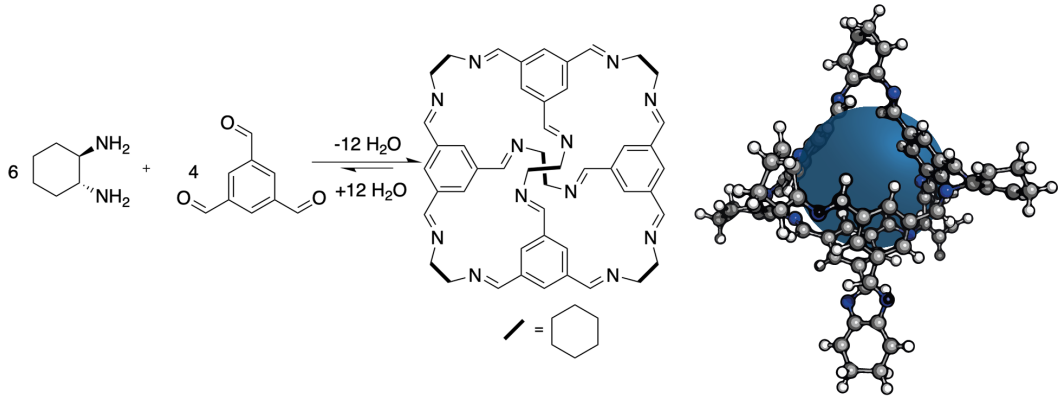


Figure 1.2: The imine condensation of 4 equivalents of triformylbenzene (TFB) with 6 equivalents cyclohexanediamine (CHDA) to form covalent cage 3 (**CC3**). **CC3** has been shown to maintain a permanent internal cavity upon desolvation, shown by the blue sphere, which is accessible to guest molecules.[38]

plementary binding angles to achieve the targeted topology. For instance, a [2 + 3] imide POC was formed using a flexible triamine and a linear dianhydride, but in a low yield.[43] Although irreversible bond formation can lead to increased stability and predictable topological outcomes, their low yields and lack of error-correcting mechanisms limit their widespread application for POC synthesis.

Instead of using typically low-yielding irreversible chemistries, post-synthetic modification of POCs can achieve similar enhancements in stability with the benefit of high-yielding DCC reactions for initial synthesis.[44] For imine POCs, reducing the imine bond to irreversible amine bonds can enhance stability but result in a collapsed pore structure, as exemplified by the reduction of **CC3** to **RCC3**, shown in Figure 1.4.[45] Indeed, the pore structure of **RCC3** was shown to collapse due to the reduced rigidity of the amine bond, facilitating the collapse of the internal cavity. To regain the internal pore structure, shape persistence can be re-introduced by animal formation, 'tying' the amine groups together with carbonyl-containing molecules, forming **FT-RCC3**. Alternatively, the imine group can be oxidised into an amide group, enhancing stability across a wide pH range while maintaining the cavity.[46] The range of possible post-synthetic modifications of POCs is diverse: POCs can dynamically convert to COFs and become coordination ligands in metal-organic frameworks.[44]

The landscape of POCs is rapidly expanding and has been extensively covered in a range of review articles.[28,32,38,47] However, despite the rapid developments across the field, we still cannot reliably predict what combination of precursors will assemble to form a POC. Even now, new systems are the product of small iterative changes to precursors known to form POCs and are often blind experimental successes rather than rational design performed using *in silico* predictions.[48]

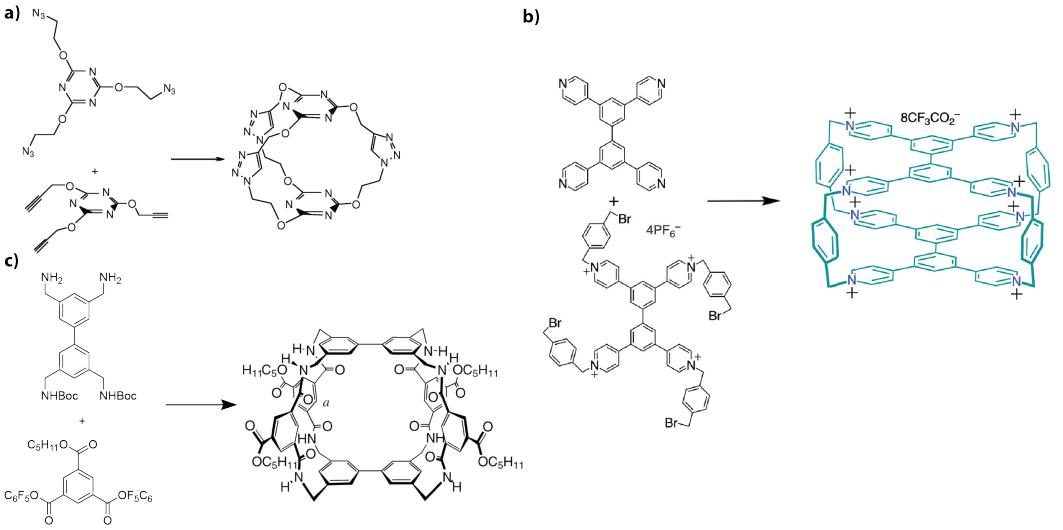


Figure 1.3: The synthesis of molecular cages using irreversible bond formation, such as click chemistry (a),[39] nucleophilic substitution reactions (b),[40] and amide bond formation (c).[41] Additional reagents for each reaction have been omitted. Reproduced with permission from references (39) (a), copyright 2008 American Chemical Society; (40) (b), copyright 2020 American Chemical Society; and (41) (c), copyright 2002 National Academy of Science.

Combining computational prediction and experimental validation has resulted in the successful discovery of POCs. When combined with experimental automation, computation is now able to achieve accelerated POC discovery.[48]

1.2.3 Properties of Porous Materials

The discrete molecular nature of POCs, and PMMs more broadly, gives rise to both *intrinsic* and *extrinsic* void spaces. Extrinsic porosity arises due to inefficient packing, resulting in void spaces between discrete molecules. In smaller molecules which lack a central internal cavity, such as **HOF-1**, **SOF-1** and **TPP**, porosity is predominantly extrinsic and arises due to the formation of energetically favourable low-density polymorphs. Achieving such porosity in discrete molecular systems is challenging since molecules must overcome predominantly attractive intermolecular bonding, leading to more compact packing arrangements with reduced void space. However, there has been recent progress in designing extrinsically porous molecular crystals by using molecules with 'awkward' shapes and possessing hydrogen bond-forming moieties to form strong directional packing arrangements.[49]

Meanwhile, intrinsic porosity arises from the well-defined internal cavity in the discrete molecular unit. This type of porosity is present in many macrocyclic compounds, such as cyclodextrins, calixarenes, cucurbiturils, and POCs.[28] In each of these molecules, the presence of windows

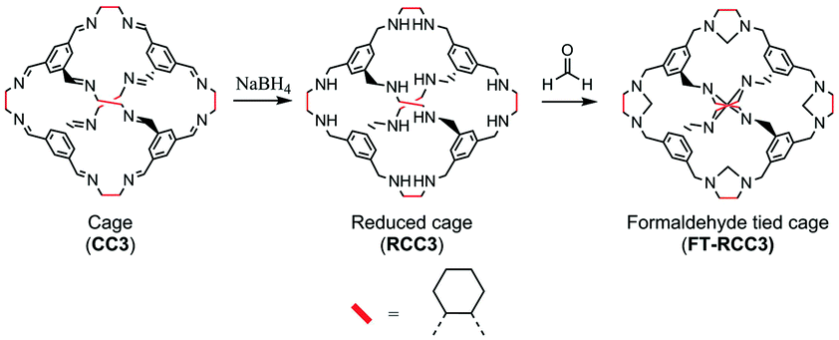


Figure 1.4: The reduction of the imine bonds in **CC3** to amine bonds, forming **RCC3**, which results in a collapse of the central internal cavity. Amine bonds can be 'tied' using aminal formation, forming **FT-RCC3** and regaining the internal cavity.[45] Reproduced from reference (45) with permission from the Royal Society of Chemistry and licensed under CC BY-SA 3.0.

surrounding the internal cavity allows for the selective diffusion of guest molecules through the internal cavity. For POCs, the term shape persistent describes a cage with a cavity accessible to a guest molecule upon removal of solvent molecules. In contrast, a collapsed cage would instead not possess any accessible cavity or windows.

Although the cavity of shape persistent POCs is permanent, they can form non-porous solids, indicating that intrinsic porosity alone is not sufficient for a material to be porous. In addition to intrinsic void spaces, a porous material must also possess an interconnected void network that allows guest molecules to navigate through it successfully. Such requirements are evident in the interconversion of different polymorphs of **CC1**.[28] Some porosity can only be accounted for by considering the influence of the guest molecule on the host network. Holden *et al.* were only able to explain the difference in porosity between H_2 and CO_2 in **CC2** by accounting for the effects of the guest on the host network, shown in Figure 1.5.[50] The complex origin of porosity hinders the rational design of new porous materials. While intrinsic porosity is typically easier to engineer in molecules using computational methods and chemist intuition, there is no certainty that those molecules will pack in a way that porosity is maintained.

1.3 Computational Modelling of Porous Molecular Materials

The objective of computational modelling is to provide atomistic insight into chemical phenomena on a range of length- and time-scales. An atomistic understanding can often lead to the derivation of structure-property relationships, creating a predictive function of a chemical structure to a property of interest. Properties of porous molecules, canvassed in Section 1.2.3, can be dictated

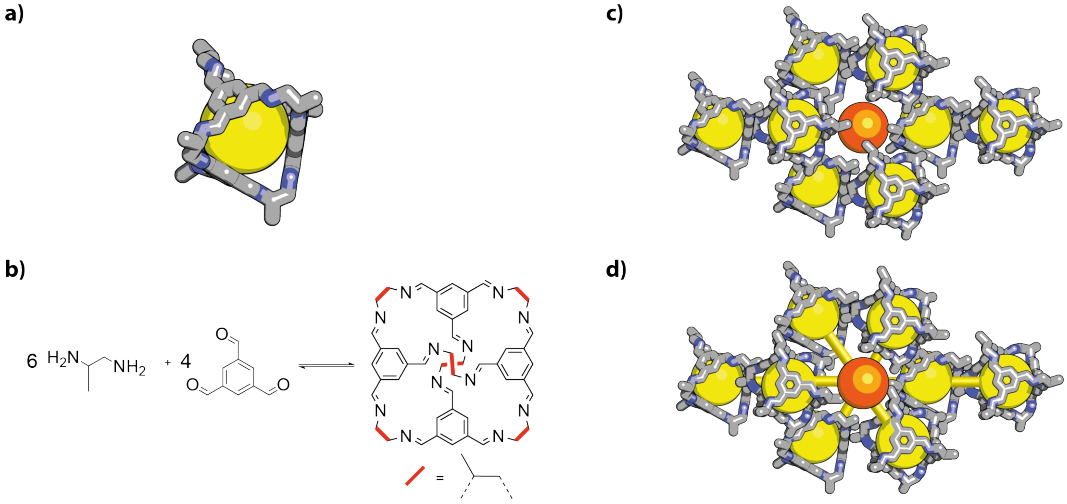


Figure 1.5: The forms of porosity in **CC2** (b), which possess a permanent internal cavity (a), shown in yellow. One-dimensional channels arise from the window-to-arene packing, however, the packing arrangement leaves these cavities disconnected in the solid state (c). Porosity can only be explained by the influence of guest molecules on the host work (d), allowing dynamic expansion of the interconnected voids.

by both the discrete molecular unit and the manner in which these units pack in the solid state. Therefore, a wide range of simulations, performed at different scales and at different levels of theory are desirable for the analysis of PMMs.

While properties typically arise from the bulk structure of a molecule, the properties of the isolated molecular unit can often be extrapolated to those of the extended structure.[51] Indeed, even simplistic geometric descriptions of molecular porosity correlate well with overall material porosity.[52] This allows for an analysis of the discrete units that constitute PMMs using more rigorous computational methods, with can provide insight into how these units assemble in the crystal structure. If the single molecular unit cannot accurately describe the porosity of the bulk structure, less accurate computational methods, such as force field (FF) or coarse-grained methods, can instead be used to probe the non-covalent interactions that dictate properties. High-throughput screening is also often restricted to these types of methods, relying on the evaluation of thousands of candidates.[53]

1.3.1 Structure Prediction

Most computational modelling workflows begin with an initial 'guess' molecular structure from atom or bond information of a molecule of interest. For small molecules, this information is contained in the SMILES or InChI line notation, which directly encodes the molecular connectivity graph. A reasonable three-dimensional structure can then be derived using rule-based approaches,

such as the experimental-torsion-knowledge distance geometry (ETKDG) algorithm.[54] The large number of atoms and complex topologies POCs can form complicates computational structure generation. The supramolecular toolkit (*stk*) is a Python package for the automated construction of three-dimensional POCs by providing only two-dimensional precursor representations, and has also been applied to the generation of polymers, metal-organic polyhedra, and COFs.[55,56] A typical construction workflow for different cage topologies is outlined in Figure 1.6. Initially, precursors are placed on pre-defined vertices corresponding to a specific topology, and unphysical long bonds are created between defined functional groups (shown as dashed red lines in Figure 1.6). An optimisation workflow, as will be discussed in the next subsection, can be used to identify a plausible structure from the underlying structure generated by *stk*. Similar construction capabilities have been achieved for metallocages and host-guest systems using *cgbind*.[57]

Identifying low energy molecular conformations is a key aspect of many computational screening workflows, and is routinely performed for small drug-like molecules in the pharmaceutical industry.[58] The typically large size and high flexibility of molecules that constitute PMMs require robust conformational search tools to effectively sample the potential energy surface. Combined

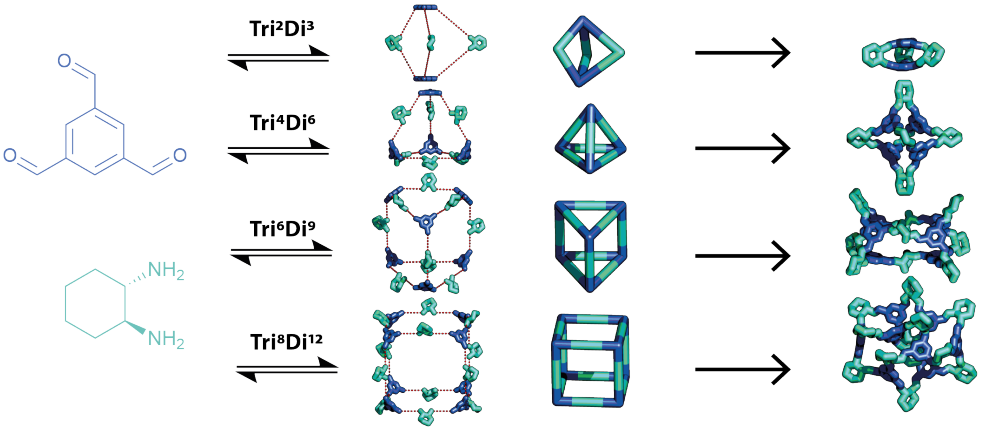


Figure 1.6: Possible topological outcomes of the reaction between CHDA (green) and TFB (blue). Using the supramolecular toolkit (*stk*), each topology can be constructed by placing each precursor in a pre-defined position and creating bonds (red dashed line) between precursors on vertices (blue) and edges (green). Once constructed, an optimisation procedure eliminates unphysical bond distances, and can be used to determine whether a cage will retain its permanent internal cavity. Hydrogens have been omitted for clarity.

with FF or electronic structure methods, molecular mechanics can help achieve this and has found widespread use in many PMM workflows.

Most conformational generation methods rely on an energetic ranking of conformations using electronic structure or, more commonly, FFs for large numbers of conformations. Often tested on a small subset of materials, it is challenging to develop a truly transferable FF that is applicable to multiple material types. For example, the universal force field (UFF)[59] and COMPASS[60] fail to accurately reproduce the experimentally observed bond angles of imine bonds, a bond type present in many POCs. Parameterising existing FFs with additional parameters can increase transferability; however, the OPLS all-atom FF can accurately reproduce the POC structure and energetic trends.[61] For organic molecules, OPLS can accurately reproduce the energies of solvation, protein-ligand binding,[62] and dimer interaction energies, and is increasingly being used to model POCs.

Automated construction of POCs has accelerated the rate at which hypothetical structures of organic molecules can be generated, enabling high-throughput virtual screening of these systems and providing the tools to probe the formation of different structural isomers of POCs.[63] Although the generation of initial structures is now possible, predicting realistic conformations for these systems is non-trivial, and while FFs can accurately describe the structure of the single molecule, it is more challenging to predict the structure of the molecular crystal.

Crystal Structure Prediction

For many systems, analysis of the single molecule is insufficient to describe the bulk properties. Crystal structure prediction (CSP) can instead be used to probe the weak intermolecular forces between discrete PMM units that give rise to bulk properties.[64] CSP is a computational method that aims to predict the most stable crystal structure of a given molecule, which in turn can be used to investigate the bulk properties of materials. Due to the vast number of possible crystal structures, CSP is typically performed using an iterative filtering procedure: i) many initial crystal structures are generated with large deviations in their conformations using Monte-Carlo (MC) or quasi-random sampling; ii) lattice energies of each crystal structure are calculated, typically using computationally inexpensive FFs, and finally; iii) high energy or structurally similar conformations are removed. Increasing computational power has allowed for the use of more accurate density functional theory (DFT) ranking of lowest energy polymorphs, where the inclusion of dispersion-corrected functionals can differentiate between low-lying energy minima.[65]

Computational prediction of crystal structures is a long-standing challenge and of significant interest in the pharmaceutical industry for the discovery of new polymorphs. Typically conducted on small rigid molecules, flexibility and weak intermolecular forces between molecules in PMMs can create a large number of energetically similar crystal structures.[66] Despite the challenges, CSP

has been successfully applied to PMMs, including POCs.[67] Once constructed, a visual mapping of the structure to properties, entitled energy-structure-function (ESF) maps, can be created to provide a global view of the relationship between the crystal structure and their corresponding properties, aiding in the identification of promising materials for specific applications. These visual maps can be used to identify promising polymorphs of porous materials given by low energy, low density 'spikes' on the landscape. Using ESF maps, Zhao *et al.* probed the structure-energy landscape of six HOFs (shown in Figure 1.7), identify promising targets in which stronger hydrogen and π-π directed packing to lower density states.[66] ESF maps have not only been applied to the investigation of new molecules but can be applied to identify porous polymorphs of well-studied molecules such as trimesic acid and adamantane-1,3,5,7-tetracarboxylic acid.[68] Unlike molecules without an internal cavity, many more structures on the landscape of POCs are predicted to be porous. Although many PMMs form crystalline materials, they have also been reported beyond the crystalline phase as amorphous solids and porous liquids.

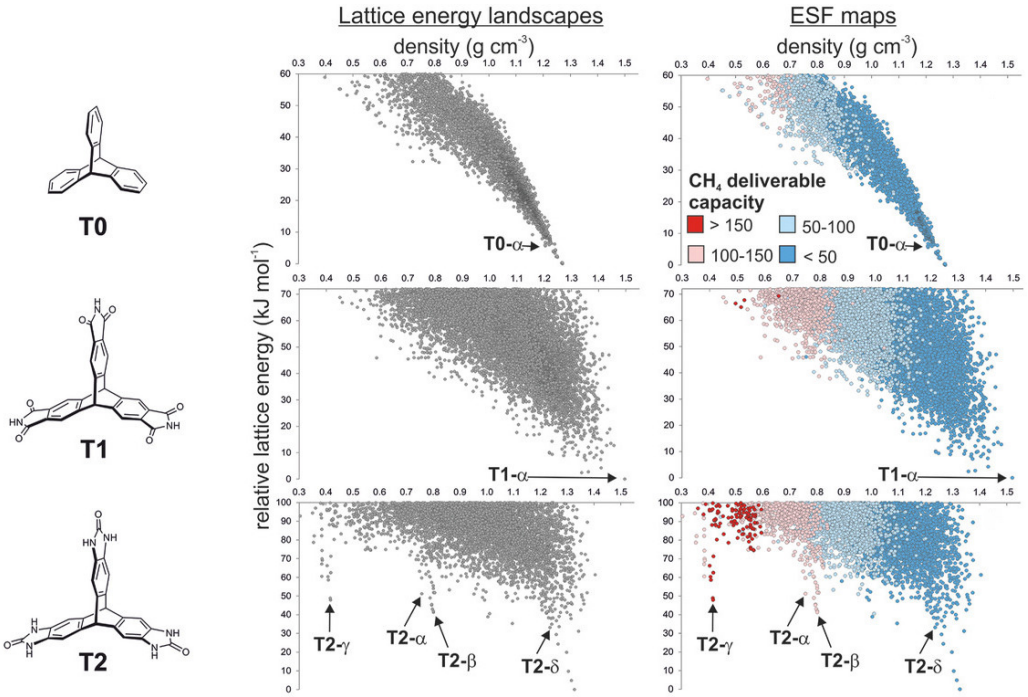


Figure 1.7: The energy-structure-function maps of three porous molecular materials, reported by Zhao *et al.*, which were used to successfully predict the existence of four experimentally stable polymorphs of **T2**. By visually displaying the relationship between stability and properties, these maps can facilitate the identification of metastable polymorphs with favourable properties. Figure adapted from reference (66) with permission from Nature Portfolio and licensed under CC BY-SA 4.0.

1.3.2 Disordered Materials

For PMMs that form amorphous solids or liquids, the inherent disorder adds additional modelling challenges. The absence of long-range order makes identifying the cause of porosity challenging, and atomistic simulations are necessary to gain insight into the underlying pore network. Computational workflows based on molecular dynamics (MD) have been developed to simulate the amorphisation process to study the pore network of amorphous polymers,[69] organic molecules of intrinsic microporosity,[70] and POCs.[71] In these workflows, MD is performed to model the dynamics of atoms and molecules over time under given thermodynamic conditions, allowing simulation of the dynamic behaviour of PMMs in the amorphous phase and enabling identification of structural features responsible for their porosity.

Through simulations, it has been possible to elucidate design rules to increase the porosity of organic molecules in the amorphous phase: the addition of bulky end groups, increased rigidity, and decreased planarity have been shown to theoretically enhance porosity.[72] Similar relationships have been observed for amorphous POCs, with the addition that a large internal cavity is likely to result in a highly porous amorphous structure, which can often exceed the porosity of the crystalline phase.[71] Indeed, the amorphous phase is of interest because molecules have been shown to exhibit greater porosity in this state than their crystalline counterparts. In amorphous materials, irregular packing may allow access to void spaces for guest molecules which may be inaccessible in their crystalline counterparts.

The internal cavity of POCs allows this porosity to be maintained in the liquid state to form permanently porous liquids (PLs).[73] PLs can be a neat liquid (type I), dissolved (type II), dispersed (type III) in a solvent whose size prohibits entry to an internal cavity, or neat meltable framework materials (type IV), each relying on intrinsic porosity.[7] Similar to amorphous materials, developing an accurate computational model of PLs is a challenging task. Explicit solvent models in MD simulations can indicate whether a solvent is likely to diffuse through a pore. To design POCs for use in PLs, flexible alkyl appendages can be attached to the periphery of the cage, reducing the melting point and increasing the solubility.[74] A combination of umbrella sampling and weighted histogram analysis methods can be used to analyse the free energy profile of the solvent diffusion pathways, providing insight into whether the internal cavity of the cage will be occluded by external functionalities.[74]

1.3.3 Property Prediction

For individual molecules that constitute PMMs, an analysis of the internal cavity using computational software can determine whether a molecule is likely to possess intrinsic porosity. Prior to property analysis, it is important that a structure representative of the experimentally observed

one is identified, using the methods described in the previous sub-sections. Porosity calculated from high energy unfeasible structures is unlikely to reproduce experimentally observed porosities; only when an adequate structure is proposed can the porosity arising from the structure of the internal cavity be analysed.

pyWindow is a Python software package designed to analyse the internal pore structure of discrete molecules. To achieve this, *pyWindow* uses vectors that connect the centre of mass to a sphere of sampling points surrounding the molecule. Using this approach, cage windows, cavity sizes, and the number of accessible windows can be calculated.[51] The average diameter of a molecule, calculated as the mean distance between the centre of mass and the van der Waals surface of atoms, agrees well with experimental solvodynamic diameters.[48] However, each of these methods assumes a monoatomic spherical guest and does not account for the dynamic flexibility of the host, which is often the origin of porosity.[50] Kravchenko *et al.* expanded this to multiatomic guests with arbitrary shapes, predicting whether a host and guest will form a cage complex.[75]

1.4 Applications of Artificial Intelligence in Porous Molecular Materials

In recent decades, artificial intelligence (AI) has seen a rapid increase in popularity with high-profile cases such as AlphaFold[76] and large-language models such as GPT-3,[77] and applications in chemistry are no exception to this trend. Machine learning (ML) is a subfield of AI, and is broadly defined as a model able to make predictions given a set of training data. Typically, models are provided with a set of inputs and outputs (supervised learning) and iteratively update functions to create a better mapping of input to output. However, models can also be developed based on limited labelled training data (semi-supervised learning), or even no labelled training data (unsupervised learning). Linear regression, decision trees, random forests (RFs), and neural networks (NNs) all fit into the former supervised category, creating a learnt mapping from input descriptors to output properties. In the unsupervised category, clustering algorithms such as k-means, t-SNE, and DBSCAN can be used to group data points by similarity. In materials discovery, each of these methods has found widespread use, from predicting the properties of hypothetical materials to proposing a plausible synthetic route.[78]

More recently, deep learning (DL) methods have emerged as a powerful tool for capturing complex non-linear relationships between input and output data.[79] Typically making use of NNs with multiple hidden layers of interconnected nodes, DL methods have exceeded human performance in fields such as natural language processing, computer vision, and speech recognition. Since their original conception in 1943,[80] the popularity of NNs has grown with the amount of accessible computing power and the ability to train them efficiently. Based on biological synaptic connec-

tions, NNs apply a series of nonlinear transformations to the input data to create a predictive function of the output. By calculating the gradient of a user-defined loss function with respect to the weights of the network's nodes, the NN can be trained to minimise the loss function in a process known as backpropagation. Many forward and backward propagations are typically required to achieve optimal predictive performance; however, the flexibility and complexity of the trained model can be superior to other ML methods, especially combined with large amounts of training data.[81]

Many fields have seen the application of DL, and chemistry is no exception.[82] Large databases of organic reactions, biological activities, and pharmacokinetic properties have enabled the use of DL techniques in the field of drug design.[83] Although the absence of large datasets required for an effective DL model has hindered the widespread application of DL in the PMM field,[84] there have been several recent adoptions of generative DL, including the notable work of Kim *et al.* to propose plausible new zeolite structures.[85] Although many DL applications are limited by an inadequate quantity and quality of training data, ML algorithms have been successfully applied to the field of porous materials,[86] including the property prediction of porous molecular crystals,[87] MOFs,[88] and POCs.[89]

1.4.1 Encoding Porous Materials

Once the data has been collected for a predictive task, the next step is to transform the data into numeric vectors that can be used as input to the ML model. For property prediction, the goal of the descriptor is to capture sufficient information into a numerical vector, matrix, or tensor that can establish a relationship between molecule structure and property, and are typically engineered according to the specific task. An overview of the different types of descriptors that are commonly used for encoding porous materials is shown in Figure 1.8.

Alternatively, vector encodings of the underlying atom and bond connectivity, termed molecular fingerprints, can be used to directly capture molecular structure, and can also be used for substructure searching and structure comparison.[90] Two-dimensional molecular fingerprints of POC precursors have been used to predict if a POC will remain shape persistent,[89] and can be used to predict crystalline properties.[91] While various algorithms exist for generating molecular fingerprints, including the popular Extended-Connectivity Fingerprints (ECFPs) and Topological Torsion Fingerprints, these descriptions cannot capture the complex intermolecular interactions that frequently dictate the properties of porous materials.

For porous materials, where the representation of the underlying molecular graph is often insufficient to describe porosity, descriptors based on the *local* environment or *global* extended material structure can instead be used (Figure 1.8b). For the former description, the smooth overlap of atomic positions (SOAP) kernel quantises the local environment information for every atom

in a molecule.[92] Using this description, it is possible to distinguish between the properties of polymorphs, which, despite being isostructual, exhibit vastly different properties.[66] Using the SOAP-REMatch kernel, an extension to the original SOAP descriptor which performs an averaging over each local similarity kernel, the charge mobility and energy could be probed for different polymorphs of pentacene and azapentacene. Combined with clustering techniques, atomic environment descriptors are able to capture key structural motifs indicative of different stacking arrangements. Such structural characterisations can be used for the identification of new polymorphs on ESF maps; Zhao *et al.* were able to identify 'landmark' structures of six porous materials with directional hydrogen bonding.[66]

Porosity in PMMs can arise in two ways. First, it can result from small changes in the local atomic environment, which can lead to pore formation, as seen with the internal cavities of POCs. Second, it can occur due to the interconnected network of the molecular structure of the material, such as when pores are formed by the long-range packing of PMMs or the network structure of COFs. To encode larger structural features of materials, persistent homology can describe long-range geometries, including macroscopic pores and interstitial voids, encapsulating topological descriptions in a unique barcode.[93] Persistent homology is a mathematical framework used to study topological features, which have gained increased attention in the field of porous materials due to its ability to capture long-range connectivity and geometry of pore structure (Figure 1.8c).[94] Using this description, Lee *et al.* were able to distinguish between materials with different pore shapes better than the SOAP kernel.[94] Using kernel ridge regression, Moosavi *et al.* were able to predict the methane storage capacities of three different triptycene-based molecules with different crystal structures, achieving a mean absolute error (MAE) of 7 v STP v^{-1} compared with simulations and only requiring 600 computationally-derived crystal structures.[87] The persistent homology descriptor has been shown to differentiate between cages and non-cages within the Cambridge Structural Database (CSD) combined with a RF classifier trained on unique barcodes, resulting in the identification of 1,839 different cages.[95]

Despite the ability of both local and global structural descriptors to differentiate between molecular structures, each relies on the identification of plausible polymorphs from experimental crystallographic data or CSP. As such, these methods are primarily suited to describing the structural characteristics of rigid molecules such as those used for organic photovoltaics.[96] Alternatively, the intrinsic cavity structure of cages can be directly encoded into a lower dimensional vector using singular value decomposition (SVD). Sturluson *et al.* performed SVD on three-dimensional images of 74 experimentally validated cages, using t-SNE to cluster cages according to their salient structural features.[97]

The choice of descriptor for porous materials is vast, however, information about the material is often lost during the encoding process. Instead, DL methods can be used to directly learn

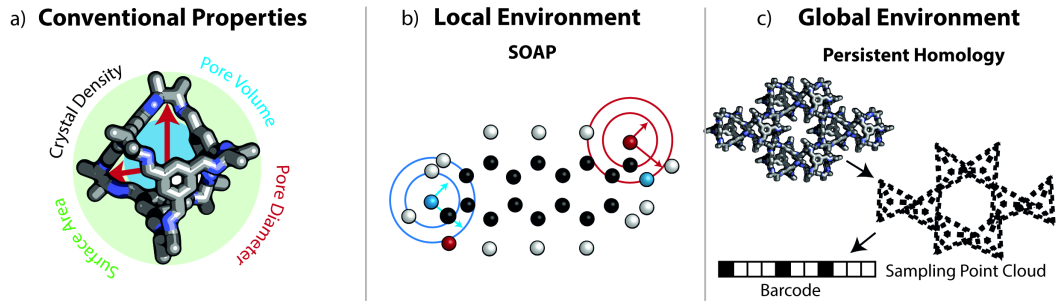


Figure 1.8: An overview of the different descriptors commonly employed for encoding porous materials. Traditional descriptors (a) approximate porosity through various calculated measures, but may not fully capture the entire geometric features of the void space. Alternatively, the smooth overlap of atomic positions (SOAP) kernel (b) has emerged as an effective method for encoding *local* atomic environments. In cases where porosity originates from long-range order, persistent homology (c) captures topological information on the scale of the *global* molecular structure.

a relationship between a raw material structure and its properties. To achieve this, graph neural networks (GNNs) and recurrent neural networks (RNNs) have been used to encode chemical structures.[98] GNNs are derived from the field of computer vision and learn a mapping from a molecular graph, with nodes (atoms) and edges (bonds), to a continuous vector. Meanwhile, RNNs are derived from the natural language processing field and use sequence-based representations of molecules, such as SMILES strings, as input. It is challenging to represent porous materials in a sequence-based representation due to their large size and periodic nature (in the case of MOFs and COFs), which hinders the application of RNNs in the field of porous materials. GNNs meanwhile have been applied to predict the band gaps of MOFs and, more recently, to predict shape persistent POCs more accurately than a combination of two-dimensional molecular fingerprints and a RF model.[99]

1.5 Considering the Synthetic Accessibility of Porous Molecular Materials

The initial stage of any materials discovery workflow is the precursor selection process. For organic materials, precursor synthesis is often a time- and resource-intensive aspect of a material's discovery process.[100] Indeed, even slight modifications in precursor structure can lead to significant changes in material properties. For organic and metal-organic materials, such as metal-organic cages (MOCs), POCs, and even MOFs, there are millions of commercially available compounds that can be used, and a nearly-infinite number of hypothetical precursors that can be used to synthesise a material, even when the precursors are restricted to a select few functional groups. De-

spite providing a vast number of potential new materials, each with its own individual properties, it is likely that only a small subset of hypothetical precursors are synthesisable. Therefore, the consideration of synthetic accessibility in computational workflows can be used to guide experimental screening.

Synthetic Accessibility of Small Organic Molecules

Several attempts have been made to computationally assess the synthetic accessibility of organic molecules, albeit primarily focused on the synthetic requirements of drug discovery programs. These methods broadly fit into four categories (Figure 1.9), which are: (i) identifying structurally complex features that could be indicative of a challenging synthesis, (ii) data-driven approaches that make predictions from extensive reaction datasets, (iii) capturing chemists' intuition using statistical methods, and (d) identifying synthetic routes to commercially available compounds using computer-assisted retrosynthesis.

Early attempts to quantify synthetic accessibility used a quantitative analysis of complex structural motifs present in molecules, such as heteroatoms, chiral centres, and large macrocyclic rings (Figure 1.9a).[101] However, complexity does not necessarily correlate with synthetic difficulty; for example, natural products are typically complex, but can be readily accessible. Indeed, complexity has been shown to fluctuate throughout a series of successive synthetic steps to a target molecule.[102] To address this, fragments that frequently appear in chemical compound databases could indicate that they are more accessible, which is the approach used by Ertl and Schuffenhauer in their synthetic accessibility score (SAScore).[103] Alternatively, chemists' intuition can be modelled using quantitative structure–activity relationships (QSARs) to identify molecular features that chemists may avoid due to their challenging synthesis. Despite not necessarily agreeing on what could make a molecule difficult-to-synthesise, aggregated chemists' scores can create a filter to eliminate molecules that chemists themselves would avoid in a drug or materials discovery program.[104]

Recently, the development of large reaction databases has facilitated the development of data-driven synthetic accessibility predictions (Figure 1.9b).[105] In particular, Coley *et al.* developed the synthetic complexity score (SCScore) that used ML to predict the position a molecule is likely to lie in a reaction pathway.[102] Using the large Reaxys database,[106] a NN was trained to predict the likely position a molecule was likely to lie in a reaction pathway, ranking molecules more likely to be located toward the end of a synthetic pathway as more synthetically inaccessible.

While each of the aforementioned methods provides a continuous measure of synthetic accessibility, recent progress has been made in applying ML to computer-assisted retrosynthesis (Figure 1.9d), enabling the fast computational prediction of synthetic routes to a target molecule. Segler *et al.* combined the Monte Carlo tree search algorithm with a series of NNs, which could propose viable reaction transformations, predict whether suggested reactions are feasible, and

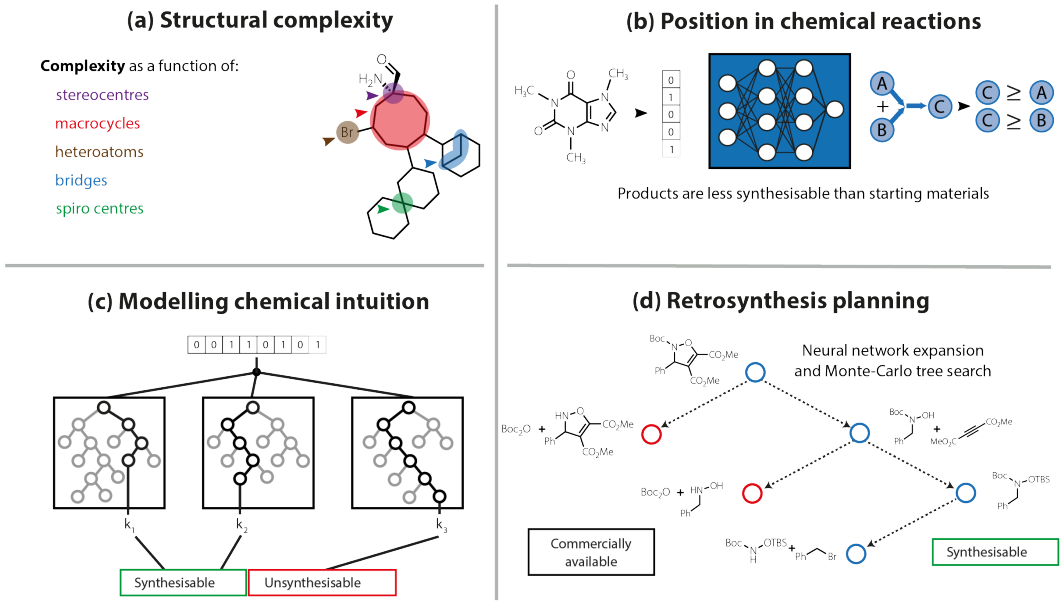


Figure 1.9: An overview of four methods used to calculate the synthetic difficulty of small organic molecules. The approaches to calculate synthetic difficulty are as follows: (a) a function of the number of complex structural motifs present within a molecule; (b) the use of data-driven approaches trained on extensive reaction databases; (c) modelling the decision-making process of expert chemists; and (d) identifying reaction pathways from commercially available compound databases using automated retrosynthesis planning methods.

predict the best node (molecule) to expand the synthetic pathways further.[107] Combined with tree search algorithms, DL methods from natural language processing have also been applied to computer-assisted retrosynthesis.[108] Using the popular transformer DL architecture, retrosynthesis was framed as a sequence-to-sequence translation task, where, given a target molecule, the model was tasked with predicting reactants. Current state-of-the-art computer-assisted retrosynthesis planners are able to readily predict the correct reagents for a target molecule 50% of the time, although non-ML approaches such as hard-coding reaction rules or identifying reagents to similar molecules have also proven successful.[109]

Predicting Reaction Outcome

While the structures cages adopt are well-defined, it is challenging to predict the topological outcome of a single set of precursors. The outcome of DCC reactions can lead to mixtures of topologies, polymers, or even oligomeric products can form. Precursors can be designed to direct towards a specific topology over others, however, even small modifications can lead to dramatic changes in the topology that forms and, subsequently, the properties of the POC. Altering the alkyl chain

length of a precursor by a single carbon atom can result in different topological outcomes, thought to be an 'odd-even' effect on cage topology exemplified by Jelfs *et al.*[110] However, this effect is not universal; Bera *et al.* showed how altering the chain length had no effect on topology when replacing triformylbenzene (TFB) with 1,3,5-triformylphloroglucinol (Tp).[111] Even when using the same set of precursors, reaction conditions such as solvent choice, concentration, and temperature can result in different topological outcomes, complicating reaction outcome prediction.[48] The assumption that the DCC reaction will result in the formation of the thermodynamic product is not always true. Kinetic pathways can result in the formation of highly stable intermediates, and can result in stabilising templating effects, such as π-π stacking, which direct reaction outcome.[112]

It is challenging to predict the outcome of a reaction using conventional computational methods. While formation energies can identify the most stable thermodynamic product, kinetic pathways can influence reaction outcome and cannot predict the formation of polymers or oligomers. The presence of solvent can result in high-energy crystal structures, such as those experimentally observed for **CC1**.[67] Only when these solvent molecules are explicitly included in computational models using Monte Carlo methods can the observed structure be rationalised. Solvent can also act as a scaffold in the cage cavity, keeping the cage cavity open in solution. Santolini *et al.* showed how a spherical probe can be used to artificially inflate the cage cavity, revealing an open formation upon subsequent constrained MD optimisations.[61]

1.5.1 Exploring Materials Space Algorithmically

Materials discovery is often serendipitous. New discoveries are increasingly the product of incremental changes to known materials (or materials precursors), prioritising candidates with a high probability of success. This confines the search space to well-explored regions of chemical space, neglecting unexplored regions where an ideal material, or a specific material structure, may lie. To tackle this exploratory bias, optimisation algorithms are becoming increasingly popular for materials discovery, and have been applied to the discovery of porous materials.[113] An efficient optimisation algorithm should identify the global maximum in the fewest number of iterations, finding the solution to a well-defined, typically multi-objective, function. In this section, I will discuss the application of optimisation algorithms to the discovery of porous materials, how they have shone a light on unexplored regions of chemical space, and the challenges of translating these theoretical candidates into experimental successes. An overview of the different types of generative models that have been applied to materials discovery is outlined in Figure 1.10.

Of all the optimisation algorithms applied to chemistry, evolutionary algorithms (EAs) (Figure 1.10c) have enjoyed the most widespread use, from developing new light-absorbing molecules to identifying new MOF adsorbates.[113,114] EAs abstract the structure of molecules into genetic equivalents; in POCs, for example, Berardo *et al.* translated a POC into a chromosome by as-

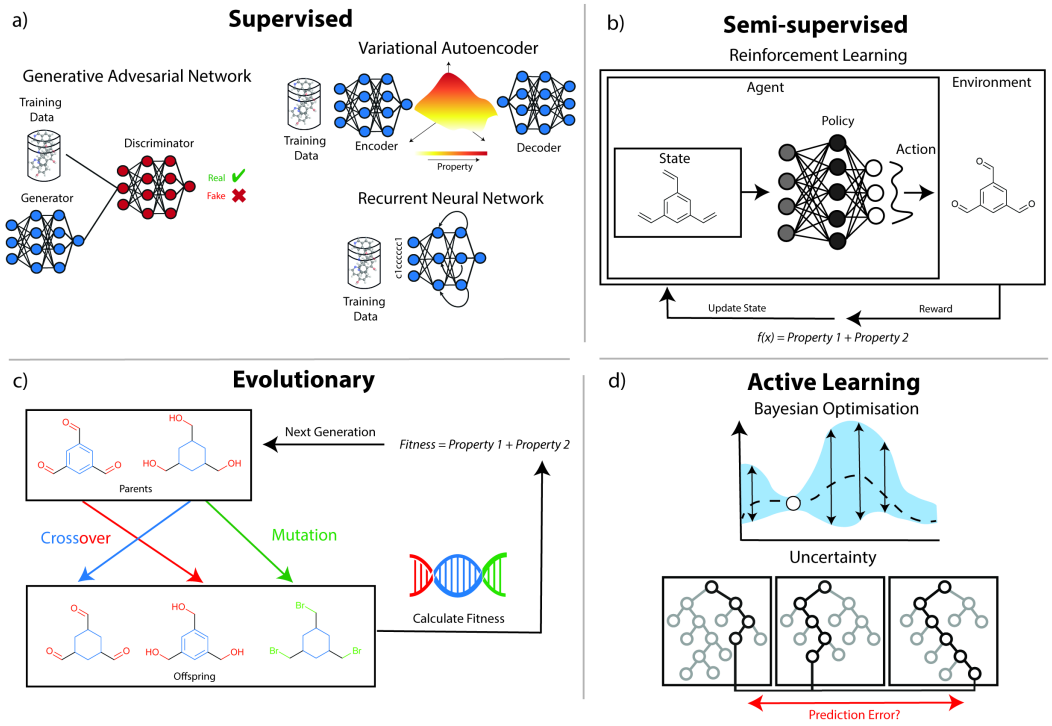


Figure 1.10: An overview of generative algorithms. Supervised generative algorithms (a) can be trained using extensive databases of molecules, often using recurrent neural networks to encode molecular structure. Alternatively, reinforcement learning (b) can be used for molecular generation, where the objective is to train an agent to take actions that result in new molecules with desirable properties. When generating training data is expensive or limited, evolutionary approaches (c) can be applied, using concepts from natural selection to converge on a potential minimum for a given fitness function. Also, when calculations are costly, active learning (d) can be implemented to optimise the training data generation process and maximise model performance while minimising the number of data points required.

cribing precursors and topology to genes.[115] For this work, the fitness function comprised of the variables POC pore size, window size, and a measure of shape asymmetry. Using a crossover process, where precursors between two POCs are interchanged, and mutation, where a precursor is replaced with another, the algorithm was able to identify shape persistent POCs with high symmetry and cavity sizes in excess of 16 Å. Indeed, fitness functions can also be designed to identify materials for specific applications. Miklitz *et al.* identified several POCs that were able to encapsulate fullerene, using the binding affinity between the host cage and guest fullerene as part of the fitness function.[116] For the former example, the initial database of 153 trialdehydes and 39,203 diamines would result in nearly 6 million possible POCs, even when restricting the search space to a single topology.[115] Fortunately, EAs are effective at traversing the enormous search spaces, avoiding becoming trapped in local minima on the fitness landscape.[117]

While EAs are effective global optimisation tools, fitness function evaluations can be prohibitively expensive. Bayesian optimisation (BO) (Figure 1.10d) is an alternative optimisation strategy that aims to identify the global maximum in the fewest evaluations. Frequently used to find the optimal hyperparameters of computationally expensive ML models, BO has also been applied to the design of reaction optimisation, chemical design, and porous material discovery.[118] Comprised of surrogate and acquisition functions, the two work in tandem to balance the exploration of uncertain regions of the function landscape with regions already known to be promising (the exploration versus exploitation trade-off).[119] This is in contrast to active learning methods, which choose the next sampling point based on identifying the region with maximum uncertainty. Gaussian processes (GPs) are typically employed as surrogate functions to mimic the expensive objective function due to their ability to quantify uncertainty. The purpose of the acquisition function is to select the next sample to maximise the expected improvement over the current best solution.

This method has been used to identify COFs from a database of 70,000 hypothetical structures. Using BO, the authors were able to minimise the number of computationally expensive grand canonical Monte Carlo simulations performed.[118] This approach was compared to random sampling, identifying the best candidate in 110 simulations using BO compared to 500 simulations to find only the 170th best with random sampling. To reduce the number of computationally expensive data points required for CSP, Pyzer-Knapp *et al.* used BO to minimise the number of lattice energy and methane uptake calculations to discover promising new porous molecular crystal structures.[120] The development of a novel parallelised acquisition function also enabled scalable function evaluations, accelerating the construction of ESF maps.

On the Synthetic Accessibility of Generative Models

With the advent of deep generative models and powerful optimisation tools, computational materials generation is now achievable. However, despite being able to suggest candidates that maximise objective functions, progress is stymied by proposed molecules being synthetically inaccessible.[121] Despite suggesting POCs that were shape persistent, many precursors exhibited complex external functionalities that would make them difficult to synthesise.[115] Gao and Coley analysed the synthetic accessibility of generative models, using their ASKCOS retrosynthesis tool to propose plausible synthetic routes and comparing with the methods of evaluating synthetic accessibility.[121] Methods that optimised specific objective functions, such as EAs, proposed the most synthetically unfeasible molecules, whereas deep generative models, which interpolate between the distribution of chemical space on which they were trained, proposed more synthetically accessible molecules.

Despite computational generative models and virtual screening workflows often proposing synthetically unfeasible candidates, a number of strategies can be used to bias generative tools towards more synthesisable molecules: (i) using a preliminary filter to eliminate synthetically in-

accessible molecules in databases for virtual screening workflows or generative models; or (ii) this filter can be applied following the generation process to eliminate unsynthesisable molecules *post hoc*; or (iii) a heuristic measure of synthetic accessibility can be used in the objective function of generative models to guide towards more synthesisable molecules.[121] Filtering molecules *a priori* guarantees that all molecules that undergo expensive property calculations can be synthesised, but limits the size of the chemical spaces explored. In generative models, restricting the training set of molecules would skew the distribution of generated molecules towards those that are likely to be synthesised. For models unlikely to propose any or very few synthesisable molecules, applying these filters following the generative process would hinder the performance of the generative model, likely wasting the computational resources used to analyse unsynthesisable molecules. Finally, using a quantitative measure of synthetic accessibility in the objective function could miss candidates that are slightly more challenging to synthesise but have exceptional properties.

1.6 Experimental Automation

Experimental chemists naturally gravitate towards choosing reactions or precursors that they are familiar with or are similar to precursors known to form materials with promising properties. Using unfamiliar precursors carries a high risk of failure, making it a time-consuming and expensive process to explore new precursors with vastly different structures. Recently, the growing accessibility of high-throughput automated synthesis techniques has allowed a much larger number of candidates to be tested systematically and reproducibly, accelerating serendipitous discovery.

Although automation is currently limited to non-complex reactions, current techniques excel at performing a single reaction type with a throughput many times greater than that of a single chemist, referred to as high-throughput experimentation (HTE). Robotic platforms can screen arrays of reactions, which can be used to find the optimal conditions for a single reaction, or combine many different precursors combinatorially using a small number of reaction conditions.[122] Indeed, HTE has accelerated the discovery of new pharmaceuticals[123] and materials,[124] with PMMs being no exception to this trend.[48] While HTE and automation can struggle to replicate the specialised nature of many organic chemistry reactions, it can be used productively when investigating a single reaction type, uncovering new insights into catalysis,[125] even when performing reactions on a nanomolar scale.[126]

For POCs, where experimental discovery is typically slow, HTE has been shown to accelerate discovery.[48] Greenaway *et al.* used the Chemspeed Accelerator SLT-100 robotic platform to react 78 precursor combinations using 3 amines and 26 aldehydes, gaining insight into how the structural flexibility of the precursor and the position of the reactive functional group can influence the topological outcome. In total, 33 POCs were synthesised (Figure 1.11), 18 single crystal

structures were determined, and a new cage topology was discovered, a covalently bonded cage catenane. Using computation, formation energies of each of the cages were calculated using DFT and each candidate was evaluated for shape persistency. Hypothetical cages with a more negative formation energies correlated well with cages that were experimentally observed, albeit with several exceptions. These exceptions could be due to complexities introduced in the self-assembly process by the formation of stable kinetic intermediates or the effect of experimental conditions not factored into the simulation.[112]

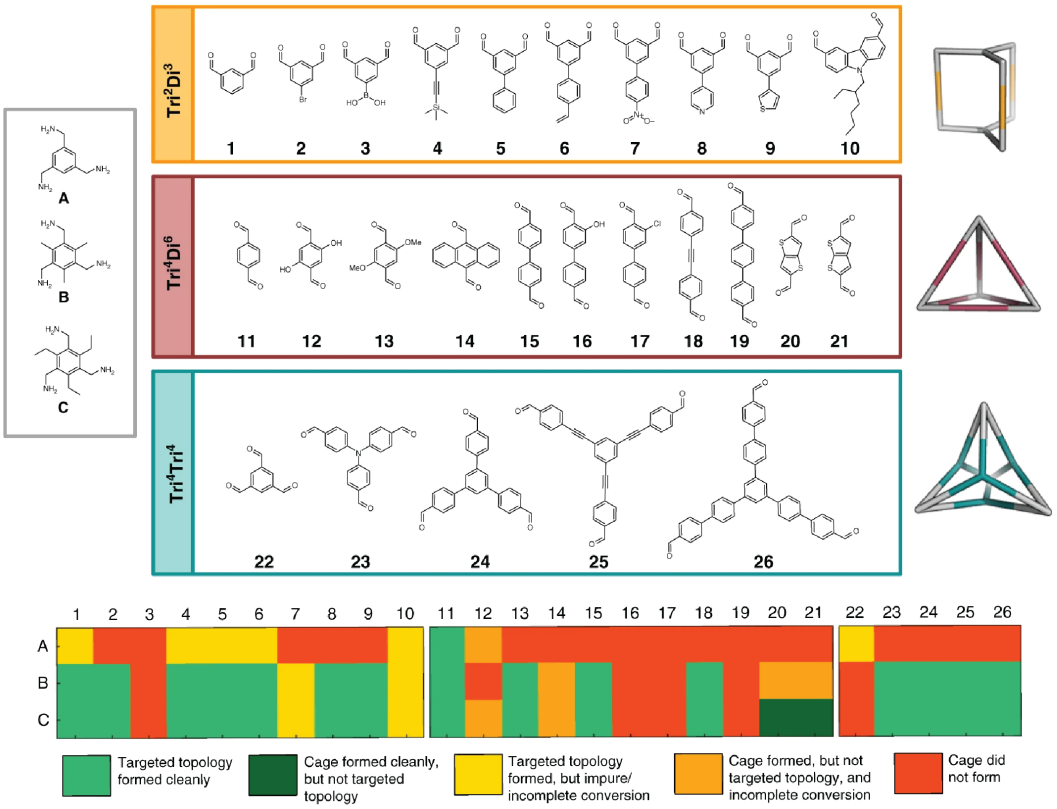


Figure 1.11: The 3 diamines and 26 trialdehydes used in the high-throughput experimental screen by Greenaway *et al.*, alongside the targeted topology for each precursor combination. The outcome of each precursor combination is shown below, where light green represents the clean formation of a targetted topology, dark green represents the formation of an untargeted topology, yellow represents partial conversion to a targetted topology, orange represents partial conversion to an untargeted topology, and red represents cases where no cage formation was observed. In total, 33 cages were observed to form cleanly. Adapted with permission from reference (48), and licensed under CC BY-SA 4.0..

A similar automated approach was used to investigate the self-sorting mechanism of three-component cages using combinations of tri- and ditopic amines and tritopic aldehydes.[37] This systematic exploration resulted in the discovery of **OCP3**, an organic cage pot with windows re-

sembling **CC3**. Although socially self-sorted cages can form, a statistical distribution of cages with different precursors located on different vertices can also be present, creating a 'scrambled' mixture. By exploiting the scrambling behaviour of cages, Kearsey *et al.* used HTE to automate the discovery of porous liquids. Using the diamine and trialdehyde used to form **CC13** and an additional diamine in different ratios, the solubility of the resulting mixture was tested in 6 different solvents, resulting in the discovery of 29 porous liquids.[127] Each of these studies takes advantage of automation, facilitating systematic navigation across a wide range of different precursors or reaction conditions. However, for POCs, the time and synthesis of precursors required to conduct the experiment still remain a bottleneck for many autonomous applications.

1.6.1 Closed-Loop Materials Discovery

By combining recent developments in both automated experimentation and global optimisation algorithms, it is possible to envisage a 'closed-loop' discovery process, by which an interface between experiment and algorithm allows for immediate feedback cycles. Although automated experimentation has been used for PMM discovery, it is challenging to link synthesis, analysis, and model training in a single workflow. Often, complex tasks such as data analysis require the insight of human experts to assess data quality and infer conclusions.[128]

Despite the many challenges, there have been a number of successful applications of closed-loop discovery. Christensen *et al.* integrated the Chemspeed SWING platform with a high-throughput experiment scheduler in Python, and Agilent HPLC-UV to optimise the yield of Suzuki-Miyaura cross-coupling reactions. Combined with a BO algorithm,[130] a 65% yield and a 1.6:1 selectivity were achieved for the desired *E* isomeric product by exploring different phosphine ligands and varying concentrations.[129] For the same reaction type, Angello *et al.* used the uncertainty of the BO model to guide the experiment selection process. The authors found that the algorithm explored both high and low-yielding regions of the search space equally, in contrast to literature-reported yields, which biased towards high-yielding regions.[131] For metal-organic systems, Porwol *et al.* explored combinations of four azides, two aminoalkynes, and two aldehydes with two possible metal coordination ions.[132] For each of the 56 possible ligands, the authors tested different reaction times, temperatures, and relative quantities, for a total of 10 billion possible reactions: a task that would be impossible for a single chemist to perform. Using mass spectroscopy, pH, and UV-Vis to determine reaction outcomes, the algorithm navigated toward regions of more promising reactivity by determining the degree of change between starting materials and products. While closed-loop discovery has not been applied to the discovery of PMMs, it has found application in the discovery of organic photovoltaics,[133] organic hole transport materials,[134] and the discovery of new reactivities.

In each of these aforementioned approaches, each search space is curated using chemist intuition and, therefore, subject to human bias. Bias can hinder the discovery of new materials: from limiting the structural space of precursors explored to limited variation in reaction conditions, chemists' choices are not always more likely to result in a successful reaction.[135] This could be overcome by using computation to define the search space, but is currently limited by instrument capabilities and current optimisation algorithms requiring a defined search space. While the on-demand synthesis of precursors is far from realisation, the ability to perform a higher throughput of reactions and receive immediate feedback from computation can reduce the inherent bias in experiment selection.

The combination of computational and experimental methods has proven to be effective in the discovery of new POCs. Computational approaches can predictively identify promising potential precursors, while HTE systematically combines these precursors *in situ*. By leveraging these approaches, researchers can explore the limitations of computational models for simulations and obtain a larger amount of experimental data, leading to the discovery of more POC precursors through efficient experimental designs.

1.7 Thesis Structure

In the introduction, I provided a relatively broad overview of the different aspects and challenges of POC design and discovery, all the way from the computational prediction of synthetically accessible precursors, to the experimental realisation of these materials. The discussion also highlighted the recent accomplishments in employing ML techniques in POC discovery, outlining their benefits in streamlining computational screening workflows and generating new hypothetical candidate molecules for potential synthesis. With recent development in both computational screening and high-throughput experimentation, it is clear a synergistic combination of these two approaches can work towards the broad goal of accelerated materials discovery.

Chapter 2 I will begin with an overview of the computational methods used in my thesis. I will canvas computational chemistry methods such as FF, MD, and DFT, in addition to some of the ML methods that I used. I will also cover some approaches to quantify the performance of ML models, including addressing imbalanced datasets, in addition to the computational methods of quantifying the synthetic accessibility of small organic molecules.

Chapter 3 Next, I will cover the initial stages of any materials design workflow: precursor selection. Here, I present computational methods of evaluating the synthetic accessibility of cage precursors highlighting how pre-existing methods in the literature can be applied to screen POC

precursors and the limitations of these methods that are typically applied to drug discovery. Using ML, I develop a model able to replicate the decision-making process of expert chemists, evaluating the performance of this model to select synthesisable POC precursors.

Chapter 4 Next, I present how the synthetic accessibility models can be applied in a high-throughput computational screening workflow to identify promising hypothetical POCs. In this chapter, I elucidate several design rules for POC precursors from cages predicted to be shape persistent, and use these rules to identify promising precursors for experimental synthesis and validation.

Chapter 5 Following the precursor selection process, I outline a combined high-throughput experimentation and computational screening workflow that can be applied to tackle the challenge of predicting cage topology. Using a library of cage precursors, I combine them *in situ* using experimental automation techniques and develop an automated workflow for evaluating the reaction outcome. Here, computation was used to rationalise the observed topological outcome, working towards the goal of predicting POC formation.

2 Computational Methods

In this chapter, I aim to provide a theoretical background behind the computational methods employed in this thesis. First, I will canvas methods based on classical mechanics, such as FF potentials and MD, which were primarily employed in high-throughput computational screening applications for POCs in Chapter 4. I will also introduce DFT, which was employed to rationalise observed experimental outcomes in Chapter 5. Furthermore, I will discuss the ML techniques that were employed to predict materials precursors easy-to-synthesise for POC, covering the evaluation metrics of the ML model, and methods for creating a numerical representation of molecules, which were used in Chapter 3. Finally, I will present an overview of alternate methods for quantifying synthetic accessibility, which were evaluated in this thesis along with the trained ML model.

2.1 Introduction to Computational Chemistry

The goal of computational chemistry is to provide a mathematical description of chemical systems on a range of length- and timescales, with the aim of creating an explicit function to derive the system's properties and energy. By providing a more detailed atomistic view of molecular systems, computational analysis can reveal insights that are unattainable through experimentation alone. These methods typically complement experimental approaches, providing a deeper understanding of the underlying chemical interactions. One crucial application of this atomistic understanding is in materials design, where methods that reproduce experimental properties can predict the properties of previously undiscovered materials. Computational methods vary in accuracy and time requirements, but are generally less expensive and more efficient than experiments. As a result, they are often used to guide experimental chemists toward more promising candidates for study.

Broadly, computational methods can be categorised into electronic structure and molecular mechanics methods. The former, electronic structure, aims to solve the Schrödinger equation to determine the electronic structure of molecules, providing an insight into properties such as bond energies, and spectra prediction. Examples of electronic structure methods include DFT, Hartree-Fock (HF), and post-HF methods, all termed *ab initio* methods. Meanwhile, molecular mechanics

methods use empirically-derived potentials to describe the interaction between atoms, allowing the potential energy landscape to be investigated using these cheaper potentials. Examples include MD and Monte Carlo simulations, which both simulate the motion and interactions of atoms. These methods can be used synergistically; the combination of electronic structure and molecular mechanics can lead to a more accurate prediction of molecular properties, since the former takes into account electronic effects, while the latter is more computationally efficient.

More recently, cheminformatics methods have become popular additions to traditional computational chemistry tools that make use of ML for property prediction. Cheminformatics involves the application of algorithms to analyse, store, and derive new QSARs, which create a mathematical link between the structure of a molecule and its properties. ML algorithms, such as NNs and decision trees, are commonly used in cheminformatics to extract patterns and insights from large chemical datasets.

In this thesis, I make use of each of these methods to predict the structure and properties of POCs at different stages of a POC screening workflow. Firstly, cheminformatic tools are employed to analyse large libraries of cage precursors, facilitating the identification of promising candidate molecules. Next, molecular mechanics methods are used to explore the complex potential energy landscape of POCs, enabling the identification of low energy conformations. Finally, electronic structure methods are used to rationalise the experimentally-observed outcomes following a high-throughput computational screen, providing insight into the thermodynamic driving forces of POC formation.

2.1.1 Force Fields

To navigate the potential energy surface, it is necessary to compute the energy of that system based on a given set of atomic coordinates. FFs are mathematical functions that describe the interactions between atoms in a system, providing an explicit function for calculating the potential energy of a system given a set of atomic coordinates. Instead of using the Schrödinger equation to describe bonding interactions, FFs define bonding explicitly, comprising of several terms that capture the interaction between atoms. These include bonded terms that describe covalent bonds and angle bends, as well as non-bonded terms that represent van der Waals and electrostatic interactions. Although the functional form of the FF energy can differ between specific implementations, the potential energy function generally follows the format:

$$E_{\mathrm{FF}} = E_{\mathrm{str}} + E_{\mathrm{bend}} + E_{\mathrm{tors}} + E_{\mathrm{vdw}} + E_{\mathrm{el}} \tag{2.1}$$

In this section, I will provide an overview of the various components that make up a FF, using the OPLS3 FF as an example, which is the specific FF used in this thesis.[136] The first two terms,

E_{str} and E_{bend}, describe the energy required to stretch a bond and deform the bond angle between two pairs of bonds from the equilibrium position. These terms are defined by the harmonic approximations:

$$E_{\mathrm{str}} = \sum_{\text{bonds}} K_r (r - r_{\mathrm{eq}})^2 \tag{2.2}$$

$$E_{\mathrm{bend}} = \sum_{\text{angles}} K_\theta (\theta - \theta_{\mathrm{eq}})^2 \tag{2.3}$$

where for the E_{str} term, r is the bond length, r_{eq} is the equilibrium bond length, and K_r is the force constant of the bond, each summed over all the bonds in the system. The E_{bend} component is determined by the angle θ formed from three adjacent atoms, θ_{eq} being the equilibrium angle, and K_θ the angle bending parameter, each summed over all the bond angles in the system. Generally, bonds are unlikely to deviate significantly from their equilibrium positions, making the simple harmonic approximation suitable for most cases. However, higher-order terms can be included to account for more complex interactions.

A truncated Fourier series defines the torsional term:

$$E_{\mathrm{tors}} = \sum_{\text{dihedrals}} \left[\frac{V_1}{2}(1 + \cos\varphi) + \frac{V_2}{2}(1 - \cos 2\varphi) + \frac{V_3}{2}(1 + \cos 3\varphi) + \frac{V_4}{2}(1 - \cos 4\varphi) \right] \tag{2.4}$$

Here, φ denotes the torsional angle and V_{1-4} represents the size of the energy barriers associated with four periodic rotations, 360°, 180°, 120°, and 90°. The Fourier series ensures the correct number of energy maxima and minima when describing the complete rotation of a torsional angle from 0° to 360°. In some molecular systems, specific symmetries can result in a lack of torsional rotation energy barriers, which would normally create energy maxima and minima around the rotation. When this occurs, the coefficients of certain terms in the Fourier series may be equal to zero, simplifying the energy landscape of the torsional rotation.

The non-bonded terms, E_{vdw} and E_{el}, represent the pairwise interactions between charged sites in the molecule and determine the intermolecular bonding behaviour. The Coulomb potential is used to calculate E_{el}, which is given by the sum over all pairs of charged sites i and j:

$$E_{\mathrm{el}} = \sum_{i<j} \left(\frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \right) f_{ij} \tag{2.5}$$

Here, q_i and q_j denote the charges of the interacting sites separated by a distance r_{ij}, and ϵ_0 is the permittivity of free space. In OPLS3, the term f_{ij} is used to scale the intermolecular energies of

atoms separated by three or fewer bonds, and applies to van der Waals and electronic functions.[136] The Lennard-Jones potential is used to describe van der Waals interactions, which is given by:

$$E_{\text{vdw}} = \sum_{i<j} 4\varepsilon_{ij} \left(\frac{\sigma_{ij}^{12}}{r_{ij}^{12}} - \frac{\sigma_{ij}^{6}}{r_{ij}^{6}} \right) f_{ij} \tag{2.6}$$

Here, ε_{ij} and σ_{ij} denote the Lennard-Jones parameters for the interacting sites separated by a distance r_{ij}.

Once the functional form of the FF is defined, the next step is to determine the parameters for each term. To assign parameters, molecules are divided into structural units, called atom types, which consist of specific atoms and bonds. Fortunately, identical atom types present in different molecules behave similarly and can be assigned the same parameters. This enables FFs to be widely applicable to different molecules as long as there is sufficient coverage of atom types. Stretching, bending, and torsional parameters are typically obtained by fitting quantum chemical measurements or crystallographic data; OPLS3 used DFT calculations at the B3LYP/6-31G* level to derive 1,187 stretching terms and 15,236 bending terms. Furthermore, the MP2 localised methodology was used to derive 48,142 torsional terms.[136]

Determining parameters for non-bonded terms, which govern intermolecular interactions, is a more challenging task. Assignment of atomic charges in the electronic term can be achieved by fitting them to the electrostatic potential generated by electronic structure methods. OPLS3 uses the CM1A-BCC charge model, which includes virtual charge sites to accommodate the behaviour of aromatic nitrogen-containing rings and aryl halides. However, the van der Waals interaction is the most difficult to compute, as DFT methods are unreliable in modelling dispersion. Therefore, experimental data are often used to fit these parameters. The parameters in OPLS3 are derived from liquid simulations fitted to experimental thermodynamic properties, resulting in 124 atom types. With these parameters, OPLS3 can predict protein-ligand binding energies with an accuracy of 1 kcal mol^{-1} and predict conformational energy with an accuracy of 1.2 kcal mol^{-1}, as compared to quantum mechanical methods.[136]

2.1.2 Geometry Optimisation

Geometry optimisation is an important concept in computational chemistry that involves finding the optimal geometry of atomic systems that corresponds to a minimum on the potential energy surface. The geometry optimisation process involves iteratively adjusting the positions of atoms to minimise the forces on each atom, until the forces on each atom are nearly zero, and a stationary point on the potential energy surface is reached. The resulting optimised geometry can provide insight into the chemical properties and reactivity of molecules, as well as serve as a starting point for further computational studies. Figure 2.1 illustrates the differences in exploring the potential

energy surface (PES) between gradient descent and other local optimisation methods, as well as MD, a global optimisation algorithm.

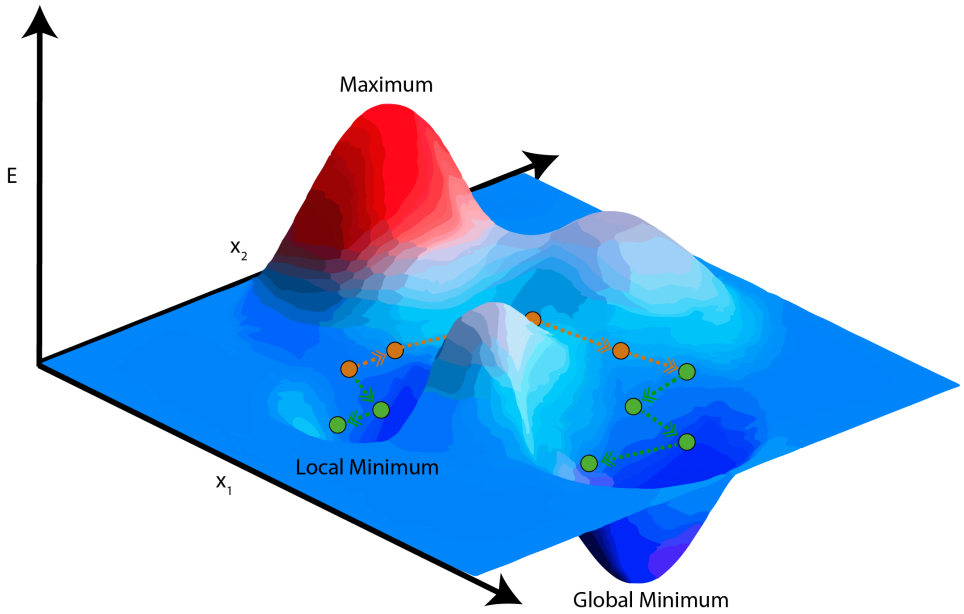


Figure 2.1: An example of a two-dimensional potential energy surface, where the green arrow depicts gradient-based methods that converge towards the nearest minima, while the orange arrow represents methods that can sample between neighbouring minima, like molecular dynamics (MD).

2.1.3 Molecular Dynamics

To explore the PES of atoms and molecules and gain insight into their motion on short timescales, MD can be used. Unlike gradient descent optimisation, which only minimises forces acting on the system, MD is a physical simulation that reproduces the physical movement of particles. By approximating the behaviour of atoms as classical particles, Newton's laws of motion can be used to simulate their dynamics over a period of time that is computationally feasible. Since the PES can be extremely complex, the parameters of MD simulations must be carefully chosen to efficiently traverse this landscape.

Starting from a set of atomic coordinates, MD iteratively calculates the new position of each atom on the basis of the forces acting on it from its current position. Forces can be expressed in differential form as:

$$-\frac{dV}{d\mathbf{r}} = m\frac{d^2\mathbf{r}}{dt^2} \tag{2.7}$$

where V is the potential energy at the atomic positions $\boldsymbol{r}$. In Cartesian coordinates, $\boldsymbol{r}$ is a vector of length $3N_{atoms}$ that contains the coordinates of all particles in the system. The left-hand side is the negative of the energy gradient (the force), and the right-hand side is the product of the mass m and the second derivative of the atomic positions (acceleration).

The velocity Verlet algorithm is a widely-used method for numerical integration of Newton's laws of motion in MD simulations.[137] This algorithm extends the Verlet equations and allows for a prediction of future particle positions based on the current position and velocity. However, the Verlet algorithm suffers from truncation errors due to finite precision and does not explicitly include velocities. These drawbacks are remedied by the leap-frog algorithm, which has better numerical accuracy and includes velocities, although velocities remain out-of-phase with positions by half a time step. The velocity Verlet algorithm overcomes this limitation, allowing computation of both positions and velocities at the same time step.[138] It can be expressed as:

$$\mathbf{r}_{i+1} = \mathbf{r}_i + \mathbf{v}_i \Delta t + \frac{1}{2}\mathbf{a}_i \Delta t^2 \tag{2.8}$$

$$\mathbf{v}_{i+1} = \mathbf{v}_i + \frac{1}{2}(\mathbf{a}_i + \mathbf{a}_{i+1}) \Delta t \tag{2.9}$$

where, at times at time $i+1$ and time i respectively, $\boldsymbol{r}_{i+1}$ and $\boldsymbol{r}_i$ are position vectors, $\boldsymbol{v}_i$ and $\boldsymbol{v}_{i+1}$ are velocity vectors, $\boldsymbol{a}_i$ and $\boldsymbol{a}_{i+1}$ are acceleration vectors of a particle, and Δt is the time step.

The required forces for MD are typically derived from the atomic FF and are used to compute the acceleration of the system at given positions, which can then be used to propagate the atomic coordinates across time. For accurate simulation results, the time step used must be sufficiently small, typically on the order of femtoseconds and at least one order of magnitude less than the fastest process in the system, which are often atomic vibrations. However, decreasing the time step increases computational expense significantly; for example, simulating 1 nanosecond requires evaluating 10^6 time steps of 1 femtosecond. While a nanosecond-long simulation captures atomic vibrations, it is insufficient for modelling longer processes, such as chemical reactions. The initial atomic coordinates greatly affect the trajectory of the system and are commonly generated using FF optimisations or chemist intuition. Short MD simulations are often used to equilibrate the system and achieve a reasonable starting geometry, as an unphysical starting geometry is unlikely to reproduce experimental dynamics.

In MD simulations, the microcanonical ensemble (NVE) is often used, which conserves the number of particles, volume, and energy of the isolated system throughout the simulation. However, canonical ensembles (NVT) can also be used, where the temperature is maintained by allowing the system to exchange energy with a thermal reservoir via a thermostat, which is typically

applied to maintain the combined energy of the system and the reservoir. For larger molecules with many degrees of freedom, such as POCs, a high simulation temperature of 700 K is often employed in MD simulations to overcome high energy barriers and sufficiently sample the PES. After the entire trajectory has been computed, gradient descent optimisation is used to minimise the forces acting on particles and obtain low energy conformations. In this thesis, Schrödinger's MacroModel software is used to perform MD conformational searches and subsequent FF optimisations, using the OPLS3 FF.[136,139]

2.1.4 Density Functional Theory

Although FFs are widely used for calculating the potential energy of molecular systems due to their computational efficiency, they have several limitations. FFs rely on empirical parameters derived from experimental data, making them less reliable for systems outside of their training dataset. Additionally, FFs based on classical mechanics do not account for electronic effects, which can be important in types of systems such as those containing transition metals or strong covalent bonds. Instead, DFT is a quantum mechanical method that has been shown to produce accurate electronic structure calculations for a wide range of systems. Unlike FFs, DFT often does not rely on empirical parameters and can be used to study a wide range of chemical phenomena.

In quantum mechanics, the Schrödinger equation describes the time evolution of the wave function of a system. Although solving the Schrödinger equation would provide a wavefunction of the system, allowing the quantised energies of the system to be calculated, solving the equation for a many-body system with more than a few electrons is intractable due to the inability to solve the partial differential equation of a function with $3N$ variables, where N is the number of particles. In DFT, the electron density of the system is instead used, circumventing the need to solve the many-body problem by replacing the problem with a functional of density, which is instead a function of only three variables.

The origins of DFT can be traced back to the Thomas-Fermi model,[140] which produced relatively coarse results by taking only kinetic energy into account while treating nuclear-electron and electron-electron interactions classically. Modern DFT methods are derived from the two main Hohenburg-Kohn theorems.[141] The first theorem states that the external potential acting on a system of electrons is determined by the electron density of the system, allowing the individual components of the ground-state energy to be written as a functional of the ground-state electron density. The second variational theorem for density states that the functional that determines the ground-state energy of a system of electrons is minimised when the input density is the true ground-state density. These two theorems show that DFT can be used to compute the total energy of a system, but does not indicate which approximation should be used for the unknown functionals.

A strategy was developed to develop a possible solution for unknown functionals using the Kohn-Sham methodology.[142] In their approach, a series of orbitals were used to describe a non-interacting reference, allowing most of the kinetic energy to be calculated. The exact kinetic energy term is derived from a reference system comprised of non-interacting electrons, expressed as a Slater determinant of Kohn-Sham orbitals, where:

$$T_S = \sum_{i=1}^{N_{\text{elec}}} \langle \phi_i | -\frac{1}{2}\nabla^2 | \varphi_i \rangle \tag{2.10}$$

The additional correction term is collected in an additional exchange-correlation term, which accounts for the kinetic energy due to the interacting electrons, allowing the total energy of the system calculated by DFT to be written as:

$$E_{\text{DFT}}[\rho] = T_S[\rho] + E_{\text{ne}}[\rho] + J[\rho] + E_{\text{xc}}[\rho] \tag{2.11}$$

where $T_S[\rho]$ is the kinetic energy of non-interacting electrons, $E_{\text{ne}}[\rho]$ is the interaction energy between the electrons and the nucleus, and the final two terms represent the electron-electron interaction energies, where $J[\rho]$ is the classical Coulomb interaction between the electrons and $E_{\text{xc}}[\rho]$ is the exchange-correlation term. Each of these terms can be written in an explicit form, with the exception of $E_{\text{xc}}[\rho]$, where the exact functional is unknown. When the DFT energy is compared to the exact energy, $E_{\text{xc}}[\rho]$ can be expressed as individual components. That is,

$$E_{xc}[\rho] = (T[\rho] - T_S[\rho]) + (E_{ee}[\rho] - J[\rho]) \tag{2.12}$$

where the first set of parentheses describes the difference in kinetic energy between the reference non-interacting electron system and an interacting system, and the second set of parentheses represents the difference between the quantum mechanical electron-electron interaction energy and the Coulombic repulsion between electrons.

In order to obtain the total energy of the system using DFT, a set of Kohn-Sham orbitals that minimise the energy must be obtained, subject to orthonormality constraints, obtaining the eigenfunction-analogous equation:

$$h_i^{\text{KS}} \phi_i = \epsilon_i \phi_i \tag{2.13}$$

Initially, a guess for the electron density is made. The effective potential is calculated (Equation 2.14), which includes the external potential as well as the electron-electron repulsion term.

$$\mathbf{v}_{\text{eff}}(\mathbf{r}) = \mathbf{V}_{\text{ne}}(\mathbf{r}) + \int \frac{\rho(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} d\mathbf{r}' + \mathbf{V}_{\text{xc}}(\mathbf{r}) \tag{2.14}$$

Solving the Kohn-Sham equation using this potential will produce a new set of orbitals from which the electron density can be calculated. This step is iterative; at each stage, the new electron density is compared with the initial density. If these densities are the same, self-consistency is reached and the total energy can be obtained. However, despite having a pathway to obtain the total density, the exact exchange-correlation functional is unknown, and so modern DFT efforts are tailored toward the identification of suitable exchange-correlation functionals.[143]

Exchange-Correlation Functionals

As the exact form of $E_{\mathrm{xc}}[\rho]$ is unknown, an approximation for the unknown exchange-correlation term must be used. The Local Density Approximation (LDA) was an early functional developed for this purpose, which assumes that the local density can be described as a homogeneous electron gas (HEG). In this approximation, $E_{\mathrm{xc}}[\rho]$ is decomposed into exchange terms ($E_{\mathrm{x}}[\rho]$) and correlation ($E_{\mathrm{c}}[\rho]$) terms, with the analytical form of the HEG being used for the exchange term. Although analytical expressions for the correlation term are available for both high- and low-density systems, intermediate density values require expressions derived from quantum Monte Carlo simulations.[144] LDA approximations work well for materials with relatively uniform electron density, such as semiconductors; however, they struggle to provide accurate results for systems with significantly varying electron densities.

Improvements over the LDA were made in the Generalized Gradient Approximation (GGA), which included the derivative of the electron density as a variable, accounting for the non-homogeneity of the electron density. There are many flavours of GGA functionals, including the Perdew-Burke-Ernzerhof (PBE) functional,[145] the Becke-Lee-Yang-Parr (BLYP) functional,[146] and the Perdew-Wang 91 (PW91) functional.[147] However, GGAs struggle to describe systems with strong correlation effects, such as transition metals, and may not always capture long-range dispersion interactions well. To achieve greater accuracy than GGA methods, meta-GGAs and hybrid functionals have been developed.[148] Meta-GGAs use high-order derivatives of the electron density or the orbital kinetic energy density. In contrast, hybrid GGAs combine the GGA and HF methods by combining different ratios of exchange energy from the HF and GGA exchange-correlation. More popular hybrid GGA methods include B3LYP, which uses 20% exact exchange, and PBE0,[149] which uses 25% exact exchange from HF. Recently, ML methods have been used to learn an exact exchange-correlation term, resulting in lower errors compared to hybrid and some double-hybrid methods.[150]

Basis Sets

To perform DFT calculations on a computer, the wave function must be expressed numerically using a linear combination of basis functions. There are two main types of basis sets: local and

plane-wave. Local basis sets are used to express the electron density of individual atoms in a molecule, whereas plane-wave basis sets are used for periodic systems. The Gaussian basis set is a widely used type of local basis set which consists of Gaussian functions centred on each atomic nucleus. Gaussian functions are flexible in shape and size, easily differentiable, and decay with distance, making them ideal for describing the electron density around nuclei. The double- and triple-zeta valence plus polarisation (DZVP and TZVP) are popular Gaussian basis sets that have two and three Gaussian basis functions for each valence electron, respectively. These basis sets include an additional polarisation function to account for the polarisation of the electron density around atoms and a diffusion function to account for the electron density that extends beyond the immediate atom.

The plane-wave basis is a popular alternative basis set, which is naturally suited to describing periodic systems such as crystals or surfaces. In this method, the wavefunction is represented as a combination of plane-waves expressed as periodic functions such as sinusoidal functions. Compared with local basis set functions, plane-wave basis sets are not centred on a single atom, but convergence to properties is often achieved more smoothly than local basis sets. Plane-wave basis sets often use pseudopotentials, which replace electrons in the core with a simplified potential to ease the computational cost of the calculation. Goedecker-Teter-Hutter (GTH) pseudopotentials are an example, which replace the core electrons with a smoothing function. Hybrid basis sets, which combine local Gaussian basis functions with periodic plane-wave basis functions, are also commonly employed. One variation of hybrid functionals is the Gaussian and Plane Wave (GPW) method used in the CP2K computational package.[151] This method uses a combination of Gaussian basis sets to represent wave function terms and plane-waves to represent the density, working towards the goal of linear scaling.

Dispersion Corrections

The accuracy of DFT calculations is often limited in describing long-range dispersion interactions because of their semi-local nature. To account for the long-range dispersion interactions that DFT calculations struggle to describe accurately, a correction method such as DFT-D can be applied.[152] In this method, a dispersion correction is calculated using the sum of the pairwise dispersion energy between atoms, which is then added to the DFT energy of the system, where the equation for the two-body component is:

$$E^{(2)} = -\sum_{i<j} \sum_{n=6,8,10,\ldots} s_{\mathrm{n}} \frac{C_{n,ij}}{r_{ij}^{n}} f_{\mathrm{damp,n}}(r_{ij}) \tag{2.15}$$

where the double sum is taken over all pairs of atoms i and j in the system, summing over even values of n starting from 6, to account for dipole-induced dipole interactions. Here, the $C_{n,ij}$

is the dispersion coefficient for the n-th term in the series, describing the dipole-induced dipole interactions ($n = 6$), and higher order multipole interactions with higher even values of n, s_n is the dispersion scaling factor for the n-th term, r_{ij} is the separation distance between atoms i and j, and $f_{\mathrm{damp,n}}(r_{ij})$ is the damping function that ensures that the contribution to the energy from atoms that are far apart is negligible. This additional dispersion energy term is then added to the Kohn-Sham DFT energy of the system.

The DFT-D3 dispersion correction expands on the original correction function,[152] including a damping function that reduces the contribution of long-range dispersion interactions, given by:

$$f_{\mathrm{damp,n}}^{\mathrm{D3}}\left(r_{ij}\right) = \frac{r_{ij}^{n}}{r_{ij}^{n} + \left(a_1 r_{0,ij} + a_2\right)^{n}} \tag{2.16}$$

where r_{ij} is the separation distance between atoms i and j at multipole order n, a_1 and a_2 are two adjustable parameters used to control the damping of the function, and $r_{0,ij}$ is a reference distance (typically the sum of the van der Waals radii of the atoms). Including dispersion correction terms in the DFT energy can improve the accuracy by accounting for long-range dispersion interactions that are not explicitly considered in the Kohn-Sham DFT method.[153]

The energy calculations in my thesis were carried out using the GPW method implemented within the CP2K Quickstep algorithm.[151] The PBE functional was used for the exchange-correlation term, and the TZVP MOLOPT basis set was employed for all atoms. To account for long-range dispersion interactions, Grimme's D3 correction was applied. Additionally, GTH pseudopotentials were used for core electrons. These parameters were chosen because of their ability to reproduce the experimentally-observed thermodynamic minima of **CC3**, a predominant POC within the literature.[31]

2.2 Machine Learning

In general, the goal of supervised ML is to develop a function that maps the input data to the corresponding output labels. Depending on the type of problem, this function can perform either classification, assigning input to a discrete label, or regression, predicting a continuous value. The function used to describe the relationship can take a variety of different forms, such as linear, non-linear, or a combination of the two. Linear models, such as linear and logistic regression, assume a linear dependency between the input and output. On the other hand, non-linear models, such as decision trees, RFs, and NNs, can create more complex relationships These models have varying levels of complexity, with varying numbers of parameters for each, but their goal remains the same: to accurately predict the output labels for new input data.

2.2.1 Decision Trees

Decision trees are a popular example of supervised ML algorithms, that can be used for both classification and regression tasks, and have found widespread use in the prediction of molecular interactions and biological activity prediction.[154,155] A decision tree is composed of a series of binary tests, where each separation is aimed at maximising the number of correctly classified samples, referred to as sample purity. At the initial root node, a question is asked about the training data, and each sample is classified based on the answer. If the question leads to a pure split, entirely composed of samples of the same class, the node is considered a leaf node and is no longer split further. For impure samples, consisting of samples from multiple classes, questions are repeatedly asked until a completely pure split is achieved.

The purity of the sample is quantified using the Gini impurity, a metric to assess the quality of a binary split :

$$G_i = 1 - \sum_{k=1}^{n} p_{i,k}^2 \tag{2.17}$$

where $p_{i,k}$ is the ratio of samples belonging to class k in node i. The questions asked at each root node are selected based on maximising the Classification and Regression Tree (CART) cost function:

$$J(k, t_k) = \frac{m_{\text{left}}}{m} G_{\text{left}} + \frac{m_{\text{right}}}{m} G_{\text{right}} \tag{2.18}$$

where m is the total number of samples in the node of the left and right subsets, G is the Gini impurity of the left and right subsets, and k and t_k are the feature and threshold used to split the node. A greedy algorithm is used to select the best split at each node, and this process continues until all leaves are pure, or until a specified purity threshold or maximum depth is reached.

The splits in a decision tree can be readily interpreted, allowing an understanding of the factors that influence the model's decision. In addition, the importance of each feature can be computed by investigating the Gini impurity at each split. However, since there are a large number of potential splits, decision trees can easily overfit the training data. To address this, hyperparameters can be tuned to regularise the decision tree, or multiple decision trees can be combined into an ensemble model like a RF.

A RF is an ensemble model that combines multiple decision trees using the bagging method. This approach introduces variation into the decision tree training process by randomly selecting a subset of features and a subset of training data for each decision tree, reducing overfitting and improving the model's ability to generalise to new data. Boosting is another ensemble technique for decision trees, which works by sequentially training models to correct the errors of the previous

model, rather than averaging the predictions of multiple models. One popular boosting method is Extreme Gradient Boosting (XGBoost),[156], which has shown promise in many ML problems. In gradient boosting methods, the models sequentially predict the residual errors of the previous model, iteratively improving the previous model. This process continues until the model reaches a predefined stopping criterion or the maximum number of iterations. In my thesis, a RF model is used for binary classification prediction due to its good performance in a wide range of cheminformatics problems.[154]

2.2.2 Performance Metrics

Evaluating the effectiveness of ML models involves using a performance function that takes true and predicted labels as input. The choice of performance function depends on the specific predictive task and the characteristics of the function. For regression tasks, the root mean square error (RMSE) is a commonly used measure of the difference between predicted and true values. However, in cases where there are many outliers, the mean absolute error (MAE) may be more appropriate, as it is less sensitive to outliers.

In classification tasks, the evaluation of model performance is often performed using a confusion matrix, shown in Figure 2.2 which allows a visual comparison between the predicted and the true labels. For binary classification tasks with two classes, A and B, a true positive occurs when a sample is correctly classified as class A, while a false positive occurs when it is incorrectly classified as class B. A true negative occurs when a sample is correctly classified as class B, while a false negative occurs when it is incorrectly classified as class A. Various performance metrics can be derived by computing these measures, including accuracy (Equation 2.19), as well as precision and recall, both of which are shown in Figure 2.2.

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{2.19}$$

For imbalanced datasets, accuracy, the ratio of the correctly classified samples, is unlikely to be a good measure of performance. For example, a model that always predicts class A would have an accuracy of 90% if 90% of the samples are class A. Instead, precision and recall are commonly used as they provide a more informative evaluation of the model's performance. Precision measures the proportion of correctly classified positive samples among all samples predicted as positive, while recall measures the proportion of correctly classified positive samples among all true positive samples.

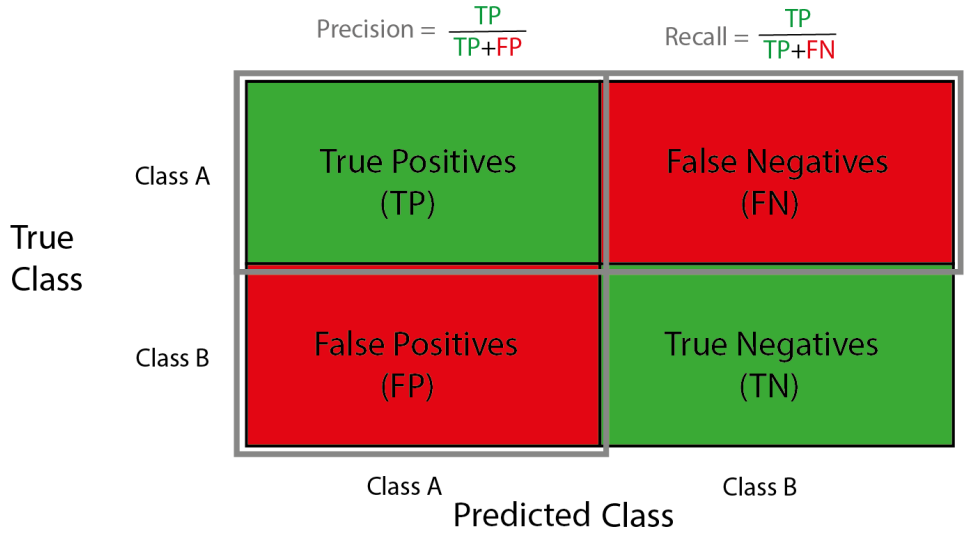


Figure 2.2: An example of a confusion matrix for a binary classification problem, where class A is the positive class and class B is the negative class. Precision and recall can be calculated using the following equations: precision is the fraction of true positives over the sum of true positives and false positives, and recall is the fraction of true positives over the sum of true positives and false negatives.

Precision and Recall Tradeoff

Precision and recall are both important performance metrics, but their relative importance depends on the specific task. In some cases, precision may be significantly more important than recall, such as in the medical domain, where a false negative could lead to a patient being incorrectly diagnosed as healthy. To account for the balance between precision and recall for specific tasks, their harmonic mean is often used as a performance metric, known as the F1 score, where

$$F_1 = 2\frac{Precision \times Recall}{Precision + Recall} \tag{2.20}$$

A weighting term, β, can be introduced into this equation such that precision is weighted β times more than recall, a more informative measure when the costs associated with precision and recall differ, giving rise to

$$F_\beta = (1 + \beta^2)\frac{Precision \times Recall}{\beta^2 Precision + Recall} \tag{2.21}$$

The precision-recall curve provides a visual representation of the trade-off between precision and recall in a binary classification model and is useful for selecting an optimal decision threshold. The decision threshold is applied to the probability that a sample belongs to a specific class, a common output of many ML classification models. Invariably, as the decision threshold is in-

creased, the criteria for belonging to the positive class become stricter, decreasing the number of false positives, increasing precision, but decreasing recall. In contrast, decreasing the threshold relaxes the criteria, increasing the recall, but reducing the precision. The optimal threshold depends on the specific task and can be selected by visualising precision against recall for a range of decision thresholds. The receiver operating characteristic (ROC) curve shows the recall against the false positive rate across a range of decision thresholds, and also illustrates the trade-off between the recall and precision. For imbalanced datasets, where the positive class is rare, PR curves are often better indicators of performance than ROC curves due to the inherently low number of false positives in the dataset.[157]

2.2.3 Chemical Descriptors

Chemical descriptors are numerical representations that encode various molecular properties or structures of chemical compounds. These descriptors are used to quantify the features of molecules, enabling the comparison and analysis of different molecules with mathematical and computational tools. The descriptors can take the form of a vector or matrix, with each element representing a particular aspect of the molecule. There are different types of descriptor, including those based on the tabulated properties of a molecule, such as constitutional descriptors, or those derived from the two-dimensional graph encoding of a molecule. Some descriptors are also derived from three-dimensional coordinate information about a molecule. The choice of descriptor depends on the specific problem or application, and some descriptors may be better suited to predict certain properties over others.

Molecular Fingerprints

Fingerprinting algorithms are widely used in cheminformatics to convert the molecular structure into a vector representation, facilitating efficient comparison between different compounds.[158] These algorithms produce fixed-length bit- or integer-length vectors that are generated using unique combinations of atoms and bonds. There are various methods for generating fingerprints, which can be broadly classified into three types: path-based fingerprints that encode unique molecular paths, atom- or sphere-based fingerprints that capture unique molecular regions, and feature-based fingerprints that encode the presence of specific molecular features.

In this thesis, I use the Extended-Connectivity Fingerprints (ECFP) path-based fingerprints as a molecular representation due to their comparatively high performance in identifying molecules with bioactive properties similar to those of ML and robustness as a molecular descriptor when applied to a range of ML models, including RFs.[159] The fingerprinting process for ECFP is based on circular paths, illustrated in Figure 2.3, and uses the following steps:

1. A unique integer for each atom type is created on the basis of its atomic number and environment.

2. Identifiers of neighbouring atoms are collected, which are then converted to a unique integer. This process is repeated until a maximum path length, called the radius of the fingerprint, is reached.

3. Duplicate structures are identified and removed using the connectivity information collected during the iterative process.

4. All unique identifiers are mapped to a unique bit in a fixed-length vector.

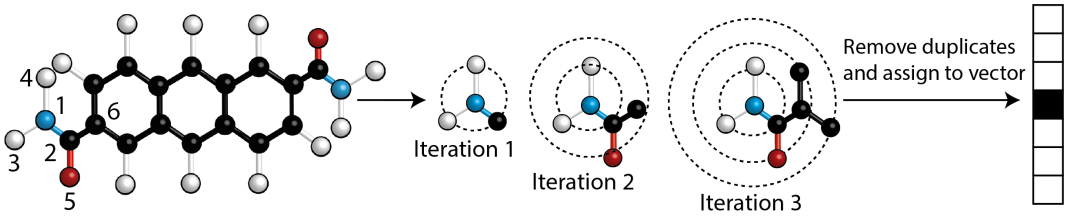


Figure 2.3: An illustration of the Extended-Connectivity Fingerprint (ECFP) algorithm,[159] which generates fingerprints by assigning a unique number to each atom and then performing an iterative updating procedure that identifies atoms within a specified number of bonds. When a threshold radius is reached, the algorithm identifies and maps unique substructures to a unique bit on a vector, discarding duplicate substructures.

The vector length for ECFP fingerprints can be relatively long (more than 512 bits), leading to a sparse vector with many zero values, depending on the number of unique identifiers present in the molecule. Reducing the vector size can reduce sparsity, but it can also lead to bit collisions where the same unique identifier is mapped to multiple bits, resulting in a trade-off between vector sparsity and stored information. Additionally, the counts of each unique identifier can be stored instead of simply indicating their presence to encode additional information.

2.3 Quantifying Synthetic Difficulty

To accelerate the discovery of new materials, it is important to identify candidates that are likely to be synthetically feasible. Quantifying synthetic difficulty can help to achieve this goal by obtaining a function able to map a molecular structure directly to its synthetic difficulty. There are three broad categories of approaches used to quantify synthetic difficulty: those that correlate structural features with increased synthetic difficulty, those that use data-driven techniques to learn a mapping of a molecule to its synthetic difficulty, and those that learn a statistical relationship

using data obtained from chemists. In my thesis, I use a combination of these methods to predict the synthetic accessibility of POC precursors and evaluate their performance.

Early attempts to quantify synthetic difficulty were based on complexity measures applied to the molecular graph. The first general measure of chemical complexity, which considered size, symmetry, bonding, and heteroatoms, was developed using information theory.[101] Later, Boda *et al.* built on this work by proposing a weighted sum of three components: the graph complexity score, a ring complexity score that penalises bridged and spiral ring systems, and the number of stereocenters.[160] However, complexity alone may not be a direct measure of synthetic difficulty, since reactions with readily accessible complex intermediates can increase the complexity of the synthetic route.[161]

In their approach to quantifying synthetic accessibility, Boda *et al.* incorporated a penalty for molecules that contained structural motifs not present in a database of commercially available compounds. Furthermore, they used a retrosynthesis score that favoured molecules that could be fragmented into smaller compounds that were commercially available.[160] The synthetic accessibility score (SAScore) developed by Ertl and Schuffenhauer also followed a similar strategy by combining a fragment score with a complexity penalty to produce a single synthetic difficulty value (Equation 2.22).[103]

$$\text{SAScore} = fragment - complexity \tag{2.22}$$

Here, the fragment component is the logarithmic ratio between the fragments in a molecule and the number of fragments in the top 80% of all fragments in the PubChem database, producing a positive value for fragments that appear frequently in PubChem, and a negative value for those that do not. The complexity penalty is given by:

$$\begin{aligned} \text{complexity} = &\log\left(\sum ring\ bridge\ atoms + 1\right) + \\ &\log\left(\sum spiral\ ring\ atoms + 1\right) + \\ &\log\left(\sum stereocenters + 1\right) + \\ &\log\left(\sum macrocylces + 1\right) + \\ &\sum atoms^{1.005} - \sum atoms \end{aligned} \tag{2.23}$$

The second method to quantify synthetic difficulty explored in my thesis is the synthetic complexity score (SCScore) of Coley *et al.*, which is based on the notion that molecules that require more reaction steps to make are more challenging to synthesise generally. The authors trained a ML model using the 12 million reactions in the Reaxys database, using a NN to perform a

pairwise ranking using a hinge loss function (Equation 2.24). By using a hinge loss, the model ensured that molecules that appeared as products in a reaction were only slightly more complex than their corresponding reactants, rather than simply maximising the difference between reactants and products.

$$\mathrm{L}(\theta) = \sum_{(R,P)} \max(0, f_0 - [f(\mathrm{P};\theta) - f(\mathrm{R};\theta)]) \tag{2.24}$$

Here, R and P are the reactants and products of a reaction, respectively, and f is the NN model with parameters θ. Although both the SAScore and SCScore approaches are mainly benchmarked on molecules used in drug design, their ability to favour easier-to-synthesise molecules can still be valuable for materials discovery despite their different assumptions.

3 Development and Comparison of Synthetic Difficulty Scoring Methods

The work in this chapter led to the publication of:
S. Bennett, F. T. Szczypiński, L. Turcani, M. E. Briggs, R. L. Greenaway and K. E. Jelfs, Materials Precursor Score: Modeling Chemists' Intuition for the Synthetic Accessibility of Porous Organic Cage Precursors, *J. Chem. Inf. Model.*, 2021, 61, 4342–4356.

3.1 Introduction

For an effective transition from computational prediction to the experimental realisation of new materials, it is essential to consider the cost and difficulty involved in synthesising the material. Several factors must be taken into account when attempting to synthesise new materials. These include the synthetic or commercial availability of materials precursors, identifying a viable synthetic method to produce the material from the precursors, and controlling the assembly and solid-state form of the material. For materials composed solely of organic components, it is particularly important to consider whether precursors can be cheaply and affordably synthesised.[100] Without any input from expert chemists, it can be difficult to predict which candidate materials are synthetically viable and will have promising properties prior to any experimental testing.[162]

The increasing popularity of data-driven tools in chemistry is beginning to ease the transition from computational prediction to experimental realisation, especially in the pharmaceutical industry.[82,83] This allows medicinal chemists to allocate their time and resources more efficiently by prioritising candidates with lower synthetic difficulties and greater druglikeness. While there are many methods of quantifying synthetic difficulty, each outlined in Chapter 2, these methods are tailored toward druglike molecules that often do not reflect the synthetic requirements of materials development.[163] Instead, material synthesis is performed on a much larger scale, requiring robust and high-yielding synthetic routes. The presence of impurities is more tolerated in materials synthesis, as pharmaceuticals are subject to stricter regulatory oversight, which limits the type of reactions and solvents that can be used.[164] In drug design, the use of more complex

synthetic routes with costlier starting materials is often tolerated, in part due to the fact that drugs are typically needed in smaller quantities than materials. These contrasting considerations mean that computational methods for predicting synthetic difficulty do not always adequately capture the specific requirements of material synthesis.

In drug discovery, chemist intuition is frequently used to filter undesirable candidates for pharmaceutical discovery.[165] The earliest examples, by Takaoka *et al.*, assigned 3,980 compounds labels of easy-, possible-, and difficult-to-synthesise by five chemists, achieving a classification accuracy of 87.6% on the test set using a support vector machine classifier.[166] Similarly, Sheridan *et al.* used a crowdsourcing exercise to rank molecules between 1 and 5, with 1 being the least complex and 5 being the most complex to synthesise.[167] Altogether, 108,000 votes were obtained from 386 chemists for 2,681 molecules, where the average value for each molecule was strongly correlated with the number of chiral centres present. An average was taken to account for the disagreement between individual chemists, which was found to be greatest around molecules with modest synthetic difficulty within the 2-4 range. Chemists have been shown to disagree when prioritising molecules to carry forward for experimental synthesis,[162,167,168] highlighting the need to establish a consensus in the model, which often requires larger amounts of data and a greater number of chemists.[104] While each of these approaches uses the knowledge of chemists from the drug discovery domain, a model based on chemist intuition has not yet been developed for the identification of synthetically accessible precursors for organic materials.

In this chapter, I describe the development of a novel synthetic difficulty model, aimed at capturing and reproducing the decision-making process of experimental chemists, specifically for the identification of easy-to-synthesise precursors for POC synthesis. The development of a predictive computational model for the synthesis of POC precursors would have a number of benefits. One advantage is the ability to eliminate precursors that are unlikely to be synthesised. This would allow researchers to prioritise computational resources on modelling the properties of materials most likely to be successfully synthesised. A predictive model could also be applied to precursors outside the original training library, allowing unseen precursors to be screened, working towards the goal of accelerated materials discovery. Finally, I compare existing measures of synthetic difficulty (the SAScore and the SCScore) with my model in this chapter,[102,103] evaluating each model's performance at predicting easy-to-synthesise POC precursors.

3.2 Methods

3.2.1 Datasets

For generating a model based on the intuition of chemists, I generated a diverse library of small molecules with varying complexity for analysis by expert chemists. With this library, my goal was

to include approximately equal numbers of easy- and difficult-to-synthesise molecules, ensuring a balanced training dataset for a classification task.

Specifically, this library was a subset of the Reaxys and eMolecules databases,[106,169] primarily those containing functional groups known to undergo common DCC reactions, including di-, tri-, tetra-, penta-, and hexatopic molecules. This decision to concentrate on DCC reactions was made due to the prevalence of this method in the synthesis of POCs in the literature and the desire to encompass a broad range of POC precursor chemical space. The functional groups included not only amines and aldehydes, which are known to undergo imine condensation common in POC syntheses, but also carboxylic acids, alcohols, and halogens. To increase the diversity and size of the library, functional group substitutions were performed using the SMARTS substructure matching language, exchanging every functional group in a molecule with another from a predefined list. This ensured that similar numbers of functional groups were present in the resulting library. I supplemented this library by including precursors that have been used in previous POC syntheses by my collaborators, as well as molecules that have traditionally been avoided due to their challenging synthesis. Altogether, this resulted in a library containing 14,859 different molecules, the full list of which is available at `https://zenodo.org/record/5142390`.

The substructures present in the final library can be seen in Figure 3.1, alongside the counts of the number of functional groups in Table 3.1. In total, there were 1,130 unique ring scaffolds present in the library, according to the definition by Bemis and Murcko.[170] In this definition, a scaffold is defined as a cycle within the graph representation of a molecule that shares an edge with at least one other cycle.[170] While 32% of the library contained a single benzene core and 21% contained a cyclohexane core, the most common scaffolds were still relatively diverse (Figure 3.1), comprising of napthaline, isoquinoline, and pyrrole rings, among others. Indeed, this diversity is also reflected in the number of functional groups present in the library. The molecules in the library were classified into ten different functional group classes (Table 3.1), with each functional group class containing a similar number of molecules. The majority of these molecules were ditopic, but smaller numbers of tri-, tetra-, penta-, and hexatopic molecules were also present. I hoped that by curating a large library of molecules with a range of different functional groups and scaffolds, I would be able to include enough chemical diversity to train a model specifically able to identify easy-to-synthesise POC precursors which is able to generalise to a wide range of diverse potential precursors.

3.2.2 Obtaining Chemist Training Data

To train the RF model, I collected training data from the dataset of 14,859 molecules by asking chemists the question: "Can you make 1 g of this compound in under 5 steps?", using the interface in Figure 3.2. By asking this question, I formulated synthetic accessibility as a binary classification

	Amine					Aldehyde					Carboxylic Acid				
Functionality	2	3	4	5	6	2	3	4	5	6	2	3	4	5	6
Number	1,640	151	20	2	1	1,962	147	21	2	1	1,673	162	22	1	1
	Bromine					Iodine					Fluorine				
Functionality	2	3	4	5	6	2	3	4	5	6	2	3	4	5	6
Number	1,569	178	29	3	4	1,595	143	19	2	1	1,699	222	31	22	14
	Chlorine					Alcohol					Thiol	Methyl			
Functionality	2	3	4	5	6	2	3	4	5	6	2	2	3	4	5
Number	1,682	243	80	6	5	1,049	139	32	3	2	2	59	30	6	2
	Other														
Number	153														

Table 3.1: Breakdown of all molecules in the training dataset by their functional groups. The functionality column indicates the number of times a functional group is present in a molecule, and the row 'Other' refers to molecules that cannot be assigned to any of the other functional group categories.

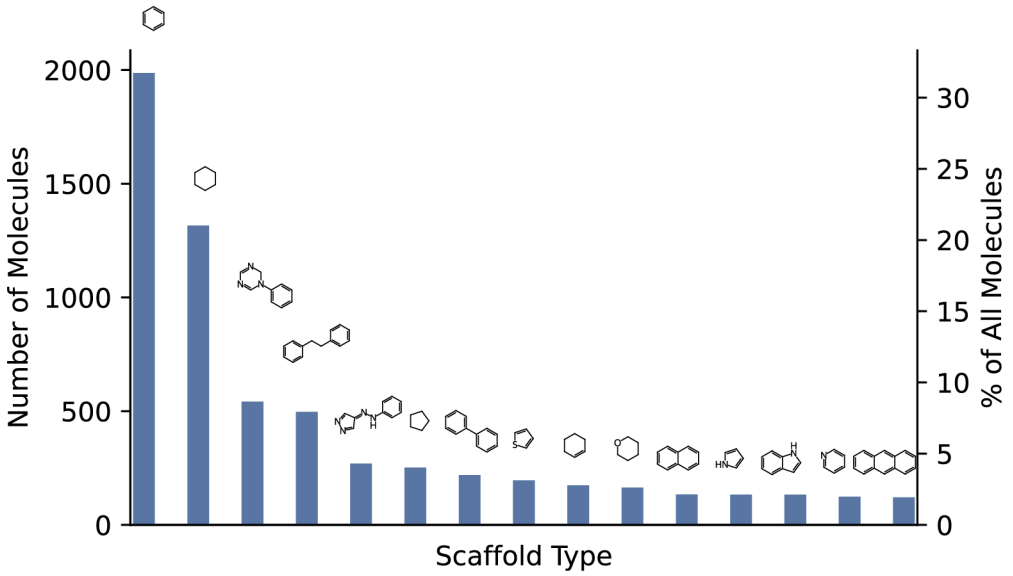


Figure 3.1: Frequency and percentages of the 15 most common ring scaffolds in the training dataset. Ring scaffolds are defined by Bemis and Murcko as cycles within the graph representation of the molecules which share an edge with at least one other cycle.[170]

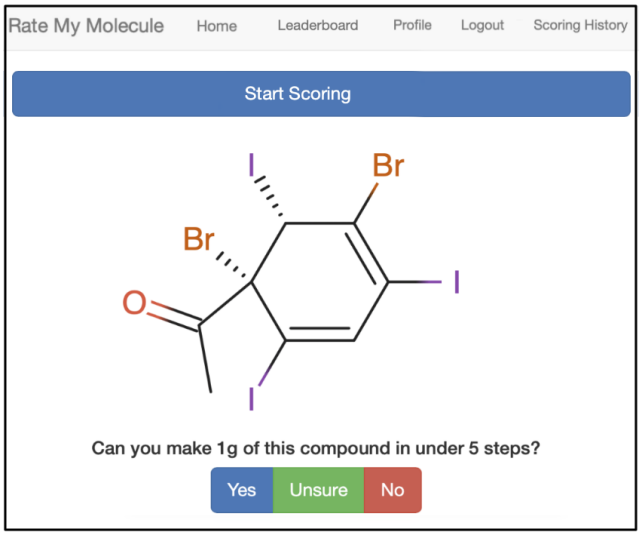


Figure 3.2: The interface used by experimental chemists for labelling molecules and easy- and difficult-to-synthesise.

problem, to create a model that can bias towards more easy-to-synthesise molecules. This specific question was chosen to mirror the precursor requirements of a standard materials screening workflow, where few synthetic steps are required, and synthesis can be readily performed on a larger scale. As collecting training data from chemists is a time-consuming process, I chose a binary classification to reduce the challenge of collecting training data and reduce the number of chemists required to achieve a consensus. Compared with existing synthetic accessibility scores in the literature, the goal was not to identify subtle variations in the ease of synthesis but to rapidly detect and prioritise molecules that are likely to be easier-to-synthesise, favouring the fast judgement of chemists over long retrosynthetic analyses. Overall, I obtained training data from three chemists with at least Ph.D.-level training in synthetic chemistry, with a combined 36 years of experience in the field of synthetic chemistry and 16 years of experience synthesising POC precursors.

For labelling, each chemist was presented with a two-dimensional representation of a molecule and tasked with answering "yes" or "no" to the aforementioned question. The molecules were randomly presented with the same probability of selection to ensure that their opinion was not influenced by previously seen molecules. In addition to the options "yes" and "no", the labelling interface also included an "unsure" option, allowing the chemist to skip to the next molecule in the dataset without appending any label to the database. This option was added to avoid incorrect labelling, allowing chemists to only provide answers for molecules they were confident in evaluating, reducing the chance of obtaining incorrect labels in the training dataset.

3.2.3 Model Selection and Training

By training a binary ML classifier, I aimed to replicate the decision-making process that an experimental chemist may take when selecting precursors for POC synthesis. For this task, I chose a RF model due to its success in predicting properties such as biological activity on small to medium-sized datasets in cheminformatics.[171] Furthermore, RF models are known to work well with sparse datasets and have relatively short training times, allowing extensive hyperparameter testing.[154] Specifically, the *RandomForestClassifier* Python class from *scikit-learn* version 0.24.1 was used for model training.

Hyperparameter Optimisation

Although RF models have relatively few hyperparameters compared to other ML models, they can still have a large impact on model performance. I chose "weighted" as the class-weight option, which weighs each data point inversely to the class frequency, reducing the effect of training on an imbalanced dataset. Applying weights to classes can improve the performance of RF models, even compared to data sampling algorithms such as the Synthetic Minority Oversampling Technique, which counteracts class imbalance by generating synthetic data points.[172] I used a randomised grid search on the parameters listed in Table 3.2 to optimise the RF model, resulting in a potential combination of 302,400 different models. I chose to randomly sample 5% of the total parameter space, resulting in a total of 15,120 models to train and assess.

For encoding molecules, I chose the extended-connectivity fingerprints (ECFP) representation, which is a common choice for RF models, as outlined in Chapter 2. The radius and bit-size parameters were optimised as part of the hyperparameter optimisation procedure, obtaining a best performing bit-size of 1024 and radius of 4.[159] The fingerprint algorithm was implemented using the *RDKit* Python library version 2020.09.4.[173] I used a count-based representation to encode the number of times a substructure appears in a molecule, rather than a single bit to indicate the presence of a substructure, believed to provide additional information in the vector representation of the molecule.[174]

For hyperparameter optimisation and final model training, I used five-fold cross-validation to train and assess models. In K-fold cross-validation, the data set is split randomly into 'K' subsets (folds), where one fold is used as the test set and the remaining folds are used for model training. Performing cross-validation reduces the likelihood that the model will overfit the training data by varying the training set, allowing us to better evaluate how well the model generalises to unseen data. During cross-validation, the total number of trained and evaluated models was 75,600. During hyperparameter optimisation, I specifically aimed to maximise the F_β ($\beta = 0.2$) score, so precision was weighted five times more than recall. The rationalisation for emphasising

Parameters	Options
Number of Decision Trees	20, 50, **80**
Maximum Tree Depth	None, 5, 10, 20, 40, 50, **80**
Minimum Sample Split	**2**, 5, 10, 20
Minimum Samples Leaf	**1**, 2, 5, 10
Maximum Leaf Nodes	**None**, 2, 5, 10, 25
Maximum Features	None, sqrt, **log2**
Maximum Samples	None, **0.1**, 0.2, 0.3, 0.4
Fingerprint Bit Length	512, **1024**, 2046, 4096
Fingerprint Radius	2, 3, **4**

Table 3.2: Hyperparameter options explored in the random forest model using a randomised grid search. For maximum tree depth, the *None* parameter results in each node of the decision trees in the forest being expanded until they are pure or until all leaves contain less than the minimum sample split parameter. Maximum leaf nodes *None* refers to the creation of an unlimited number of leaf nodes in each decision tree, and maximum samples *None* results in the all the training samples being passed to each decision tree. The maximum features parameter indicates the number of features for each training point being used to calculate each split. For this parameter, the argument *sqrt* results in the square route of the number of features being used for each split, *log2* results in $\log_2$ of the number of features used and *None* results in all of the features being used. The best-performing parameters of the model are in bold.

precision over recall will be discussed in the next section. For the final model, I trained on 75% of the dataset and used the remaining 25% of the dataset for calibration, using the metrics from the cross-validation procedure to estimate the performance of the model.

Probability Calibration

For comparison with existing quantitative measures of synthetic difficulty from the literature, I interpret the probability output of the RF model as a continuous measure of synthetic difficulty. Specifically, I interpret the probability that a molecule belongs to a difficult-to-synthesise class as a measure of synthetic difficulty, which I refer to as the materials precursor score (MPScore). This ensures that a molecule with a higher score would be more difficult-to-synthesise, in agreement with the SAScore and SCScore measures.[102,103]

RF models can suffer from poor probability estimates, which reduces the reliability of the probability as a measure of synthetic difficulty.[175] For my purposes, a well-calibrated classifier would correctly classify 40% of the easy-to-synthesise molecules assigned a probability score of 0.4. To highlight the impact of the RF model on final probabilities, I plotted a calibration curve to see the effect of probability scaling using sigmoid (Platt scaling) and isotonic regression. To plot the curve, I split the dataset into a training, calibration, and test set of sizes 56%, 19% and 25%, respectively. I trained an uncalibrated RF model on the largest set, before calculating probabilities

on the smaller test set. For the uncalibrated RF model, I fitted sigmoid and isotonic regressors to perform calibration. Each probability from the uncalibrated, sigmoid- and isotonic-calibrated models was assigned to one of ten equally sized bins, and the average probability of each bin was plotted against the fraction of positive samples expected for that specific bin. For the final MP-Score model, I split the dataset into a training (75%) and a calibration set (25%), using Platt scaling for calibration.

3.3 Results and Discussion

3.3.1 Labelled Training Data

Following the construction of my training library of molecules, I presented the library to three experienced experimental chemists for labelling. Three chemists took part in the labelling exercise, each an author of the manuscript corresponding to this chapter, which will be referred to by their initials throughout: R.L.G, F.S, and M.E.B.[176] Ideally, many chemists would be involved in the labelling process to achieve an overall consensus; however, analysing a large number of molecules is prohibitively time-consuming. In total, it was possible to obtain 12,553 labelled datapoints, with R.L.G labelling 10,000, F.S labelling 1,858, and M.E.B labelling 695. The approximate time to rank molecules was 180 per hour, limiting the amount of data possible to obtain due to time constraints. Each of the three chemists worked in fields related to organic synthesis, and more specifically POC synthesis, making their opinion invaluable when evaluating ideal precursors for POC synthesis. The total number of molecules labelled, in addition to years of experience, can be seen in Table 3.3. Despite the relatively small amount of labelled data from M.E.B and F.S, I believed that even having a small number of labelled data points for the same molecules would improve model performance overall. When evaluating synthetic accessibility, averaging the scores of many chemists can reduce the bias originating from the subjective experience of individuals, providing a more reliable estimate of synthetic accessibility.[104]

Altogether, the training database contained 2,008 easy-to-synthesise and 10,545 difficult-to-synthesise molecules, resulting in 12,553 individual datapoints classified by the three chemists in total. Although a balanced data set of equal numbers of each class was attempted, a relatively small percentage of molecules were labelled as easy-to-synthesise (16%). The smallest percentage of easy-to-synthesise molecules was labelled by R.L.G (11%), followed by F.S (33%) and M.E.B (36%), as shown in Table 3.3. One explanation for the discrepancy in the percentages assigned by each chemist is that they may have different thresholds for what constitutes an easy- versus a difficult-to-synthesise molecule, based on their personal experience and familiarity with various chemical reactions and functional groups. For instance, since R.L.G has more experience in synthesising POC precursors, their threshold for classifying a molecule as easy-to-synthesise might be more

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

Table 3.3: The number of molecules labelled by each chemist, in addition to the number of easy- and difficult-to-synthesise, and their respective percentages. The number of years of synthetic chemistry experience of each chemist is also provided, and the labels at the top refer to the initials of each chemist that rated molecules.

stringent compared to those of the other chemists. The number of molecules scored by multiple chemists was relatively small (11% of the total training library), suggesting that the differences in the percentage of molecules assigned to each class was not due to systematic labelling of one class over the other.

Chemists agreed with each other 73% of the time in total, with 74% agreement on molecules labelled by the three chemists and 73% for molecules labelled by two chemists. This suggests that agreement between chemists is generally low (27% disagreement), highlighting the need to have multiple chemists involved in the labelling process or to average the results of multiple models, each of which is tasked with replicating a single chemist's decision-making process. The 11 molecules on which at least one of the three chemists disagreed on are shown in Figure 3.3 alongside the chemists' respective decisions in Table 3.5. Notably, 9 out of the 11 molecules contained halogen atoms, indicating that these functional groups were a frequent source of disagreement among the chemists. Since halogens are not commonly present in many POC precursors in the literature,[28] it is plausible that some chemists have chosen to avoid these types of functional groups in POC precursors based on their experience. In the final RF model, I accounted for the discrepancies between chemists by providing both positive and negative examples for each molecule, reducing the importance of each sample on the outcome of the model.

In contrast to a thorough retrosynthetic analysis of each molecule, the aim was to maximise the number of datapoints obtained by focussing on the fast intuition of expert chemists, in an attempt to replicate their decision-making process. These decisions may be simply due to the presence of a particular unfamiliar substructure or highly reactive functional group. As my aim was only to bias towards easier-to-synthesise molecules and not to predict the exact synthetic accessibility of a molecule, I believe that this approach and the number of datapoints obtained were sufficient to proceed to model training.

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

Table 3.4: A summary of the molecules labelled by each synthetic materials chemist. Percentages of each label compared to the total number of molecules labelled by each chemist are given in brackets.

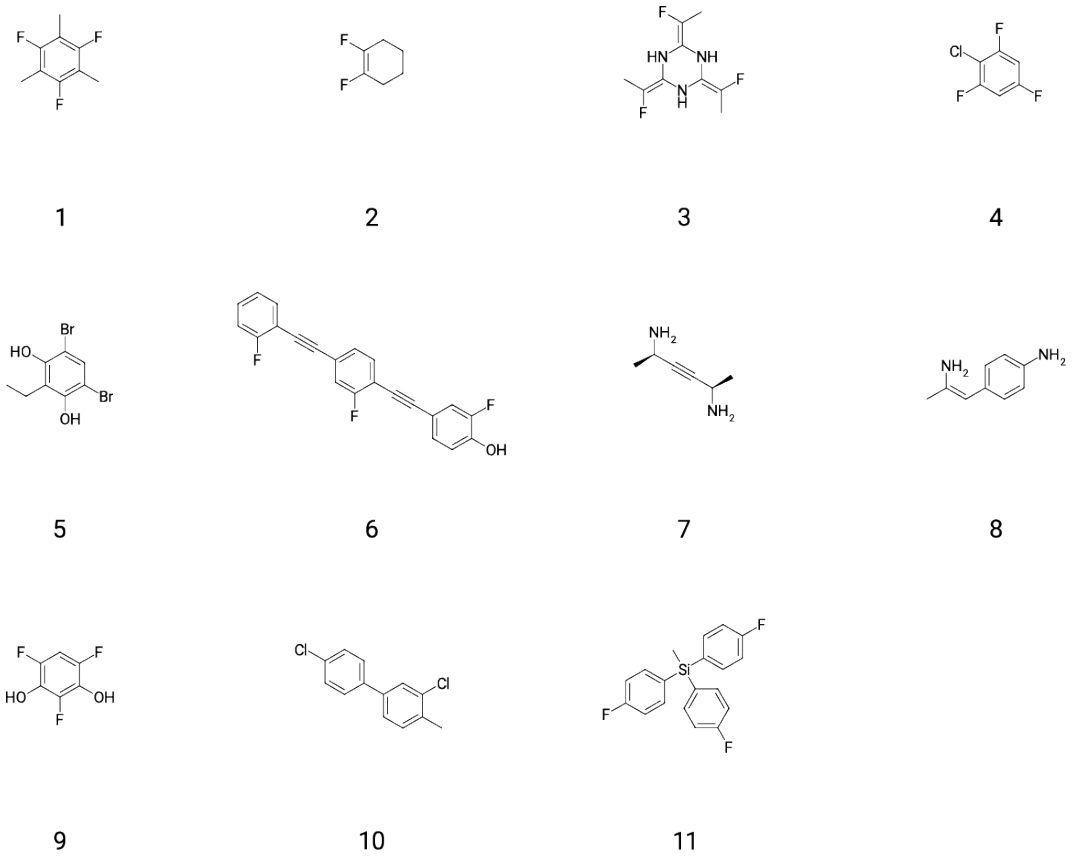


Figure 3.3: The 11 molecules that at least one of the three chemists disagreed on. Individual chemists' decisions corresponding to each of the molecules are shown in Table 3.5.

Molecule Number	R.L.G	F.S	M.E.B
1	No	Yes	Yes
2	No	Yes	No
3	No	No	Yes
4	No	Yes	Yes
5	Yes	Yes	No
6	No	Yes	No
7	No	Yes	Yes
8	No	Yes	Yes
9	No	Yes	Yes
10	No	Yes	Yes
11	No	Yes	Yes

Table 3.5: Chemists' answers to the question, "Can you make 1 g of this compound in under 5 steps?". Each molecule was scored by all three chemists, and at least one chemist disagreed with the others. Molecules corresponding to each of the chemists' decisions are shown in Figure 3.3.

Accuracy	Precision		Recall		F1 Score	F_β (β = 0.2)
	Easy-to-synthesise	Difficult-to-synthesise	Easy-to-synthesise	Difficult-to-synthesise		
0.88 ± 0.004	0.84 ± 0.013	0.88 ± 0.0037	0.32 ± 0.016	0.99 ± 0.00080	0.46 ± 0.017	0.79 ± 0.014

Table 3.6: Evaluation metrics for the MPScore model. Scores are averaged across the five cross-validation folds, and the standard errors are shown

3.3.2 Materials Precursor Score (MPScore) Training and Evaluation

Evaluation Metrics

Using the labelled training dataset, I trained a RF model to predict the probability that a molecule would belong to the easy-to-synthesise or difficult-to-synthesise class, using the outputted probability to bias towards easier-to-synthesise POC precursors. The model was evaluated using 5-fold cross-validation, and I report the average and standard errors across each fold in Table 3.6, in addition to cumulative false positives, false negatives, true positives and true negatives in Table 3.7.

A perfect classifier would maximise both precision and recall, thus achieving an F1 score of 1.0. While both precision and recall are desirable, the impact of false positive or false negative outcomes when predicting the ease of synthesis of a molecule is different. A large number of false positives (resulting in low precision) correspond to the case when a large number of easy-to-synthesise molecules are incorrectly predicted to be difficult-to-synthesise, wasting the resources of experimental chemists if a synthesis is attempted. Conversely, low recall (large number of false negatives) results in easy-to-synthesise molecules being incorrectly classified and missed, which could have formed POCs with promising properties. The latter misclassification has no associated

False Negatives	False Positives	True Positives	True Negatives	Total
1,370	121	638	10,424	12,553

Table 3.7: Cumulative metrics across each of the 5-folds in the cross-validation procedure used to estimate model performance, in addition to the total number of predictions made. False positive outcomes refer to molecules labelled as difficult-to-synthesise by the chemist but easy-to-synthesise by the MPScore, whereas false negative outcomes refer to molecules labelled as easy-to-synthesise by the chemist but difficult-to-synthesise by the MPScore.

experimental cost and is therefore less problematic. I quantified this trade-off in the weighted F_β score, where $\beta = 0.2$ would favour precision five times more than recall. Given that synthetic realisation of precursors is a current bottleneck to POC discovey, I aimed to develop a model that would maximise precision while still obtaining a reasonable number of easy-to-synthesise molecules.

In general, the MPScore achieved an average accuracy of 0.88, as shown in Table 3.6. The mean precision and recall for the difficult-to-synthesise class were 0.88 and 0.99, respectively. These values were lower for the easy-to-synthesise class with 0.84 precision and 0.32 recall, suggesting that the model has a tendency to class a molecule as difficult-to-synthesise. The difference in precision and recall scores is quantified in the F1 score of 0.46, the harmonic mean of precision and recall Unaffected by the low recall score, the weighted F_β score was 0.79, which instead favoured the high precision score (0.84) over the lower recall score (0.32). One explanation for this is due to the imbalanced dataset, where only 16% of labels were easy-to-synthesise. This class imbalance could result in the model being biased towards the majority class, exemplified by the 10,424 false negatives obtained. Despite the imbalance, the high precision suggests that the model mostly identifies easy-to-synthesise molecules correctly (638 true positives). To address the class imbalance, I weighted the splitting of each tree and adjusted the threshold the classifier used to make predictions, further improving the precision, which I will discuss in the next section.

Minimising the Effects of Class Imbalance

Due to the imbalance in the chemist dataset, I aimed to minimise the effects of the imbalance on model performance by adjusting the threshold that the classifier used to assign a molecule to the easy- or difficult-to-synthesise class. Increasing the threshold required for a molecule to be assigned to the minority (easy-to-synthesise class) would reduce the number of false positives obtained, thus increasing precision. In a RF model, the probability of a sample belonging to each class is determined by the proportion of decision trees in the ensemble that assigns the sample to that class. The effect of changing the probability cutoff threshold can be seen in Figure 3.4, where increasing the probability threshold increases precision at the expense of recall. At all probability

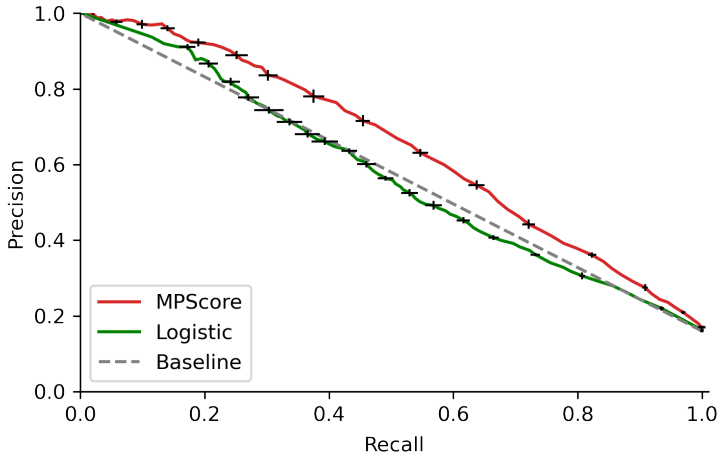


Figure 3.4: Precision-recall curve for the MPScore model, shown in red. Mean precision and recall values over each of the 5-folds are shown at different probability thresholds, alongside the standard error as black lines. The baseline classifier, shown in grey, assigns a probability equal to the fraction of positive samples in each fold, and a logistic regression classifier is shown for comparison in green.

thresholds, the MPScore RF model outperformed the naive baseline, which assigns molecules to each class with probabilities equal to the proportion of molecules in each class, and the logistic regression classifier. Despite potentially missing easy-to-synthesise molecules, selecting a high cut-off threshold ensures that the molecules selected by the MPScore are not false positives and waste synthetic efforts. By adjusting the final probability threshold to 0.21, a decision that was chosen based on the high-throughput screening workflow described in the next section, the MPScore achieved a precision of 0.89 and a recall of 0.25. This was a small improvement over the classifier results without threshold adjustment (Table 3.7).

Model Calibration Curve

As I interpretted the probability of the RF model as a measure of the ease of synthesis of a molecule, I wanted to ensure that the probability estimates returned by the model were reliable. To improve the probability estimates returned by the model, I plotted a calibration curve for the RF model. As shown in Figure 3.5, the base RF classifier produces the worst probability estimates compared to calibration with models calibrated with sigmoid and isotonic regression. Niculescu-Mizil and Caruana suggest the poor probability estimate may be due to "variance in the underlying models".[175] As RFs are ensembles of decision trees, the variance between individual decision trees pushes the probability estimates away from 0 and 1, the ends of the intervals to which the probability estimates are restricted. This is observed in Figure 3.5, where the uncalibrated model

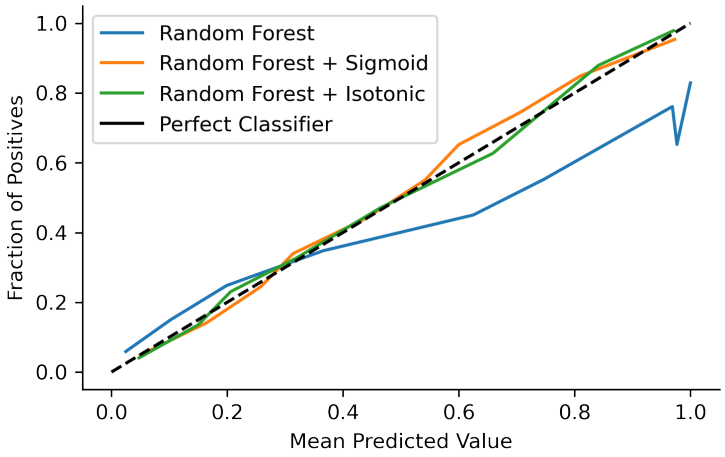


Figure 3.5: Calibration curve for the uncalibrated random forest model, shown in blue. The classifier calibrated with Platt scaling is shown in orange and with isotonic regression in green, compared to a perfectly calibrated model in black. The curve shows the mean predicted probability from each of the ten bins compared to the expected probability, determined by the fraction of positive examples in each of the ten bins.

performed worst at probability values greater than 0.6 and less than 0.2. Scaling with both Platt (sigmoid) scaling and isotonic regression improved the performance of the base classifier around the extremes of the probability interval. For the final MPScore model, I applied Platt scaling, which passes the probabilities outputted by the base classifier through a fitted sigmoid function. To fit this sigmoid function, I used a disjoint 25% of the training dataset for calibration, and the other 75% of the data was used to train the base RF classifier.

Model Explainability

After training and evaluating my model, I sought to understand which molecular substructures were the most important in predicting easy- and difficult-to-synthesise molecules. To accomplish this, I explored the explainability of the RF model, examining the importance of various substructures within the fingerprint encoding. The objective was to identify substructures that could be indicative of molecules that are easy- or difficult-to-synthesise, gaining an insight into the RF model's decision-making process. For RF models, the contribution of each input feature is determined by the mean decrease in the Gini impurity across all the decision trees in the ensemble. The Gini impurity is a measure of how well the input feature separates the datapoints into their respective classes (Equation 2.17). For example, an input feature with a large decrease in Gini impurity is likely to have a large impact on the model outcome.

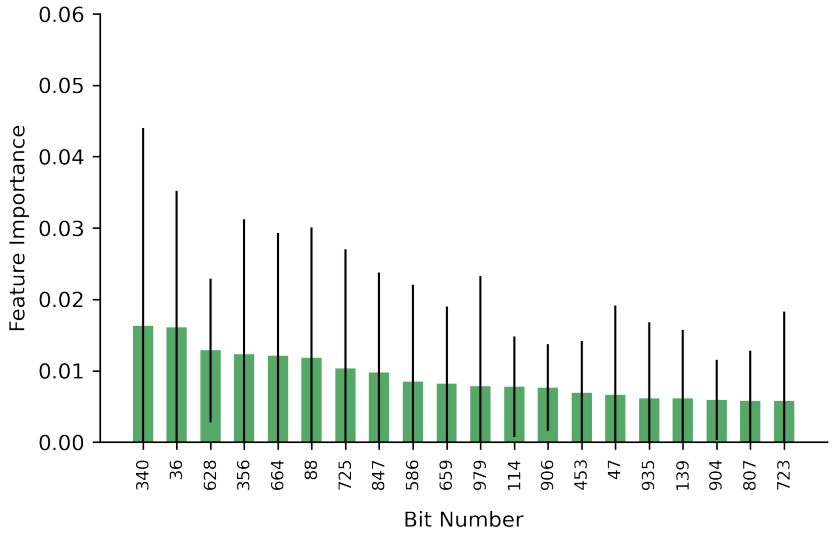


Figure 3.6: Mean importances for the 20 most important fingerprint bits. Feature importances were calculated across all decision trees in the random forest model, and the standard deviation across all decision trees is shown.

Figure 3.6 shows that many different fingerprint bit values contributed a similar amount to the overall prediction of the RF model. This suggests that the chemist's decision-making process was not influenced by the presence of individual substructures but by combinations of multiple substructures. The relatively large confidence intervals in Figure 3.6 suggest that the prediction of each decision tree in the ensemble was affected by a different set of substructures, implying that the model was not overfitting on a single substructure. Overfitting in this context would mean that the model is closely fit to the training data, and therefore, a single molecular substructure in the training data could have a large influence of the model's prediction. Instead, RF models determine their overall prediction based on the collective behaviour of many individual decision trees, each of which uses a different subset of the molecular substructures to make a prediction, avoiding overfitting. This training strategy could also contribute to the large confidence intervals and diverse range of substructures that influence the model's decision.

In addition to analysing feature importances, I aimed to identify molecules that would be the most representative of the easy- and difficult-to-synthesise classes, facilitating the identification of substructures that a chemist may want to avoid or include in a synthesis. To achieve this, I aimed to cluster the training set molecules into ten unique clusters using the K-means clustering algorithm. In K-means, the aim is to partition the input data into k clusters by minimising the variance of the datapoints within each cluster. To improve the performance of the clustering algorithm, I decided to use principal component analysis (PCA) to reduce the dimensionality of the molecular

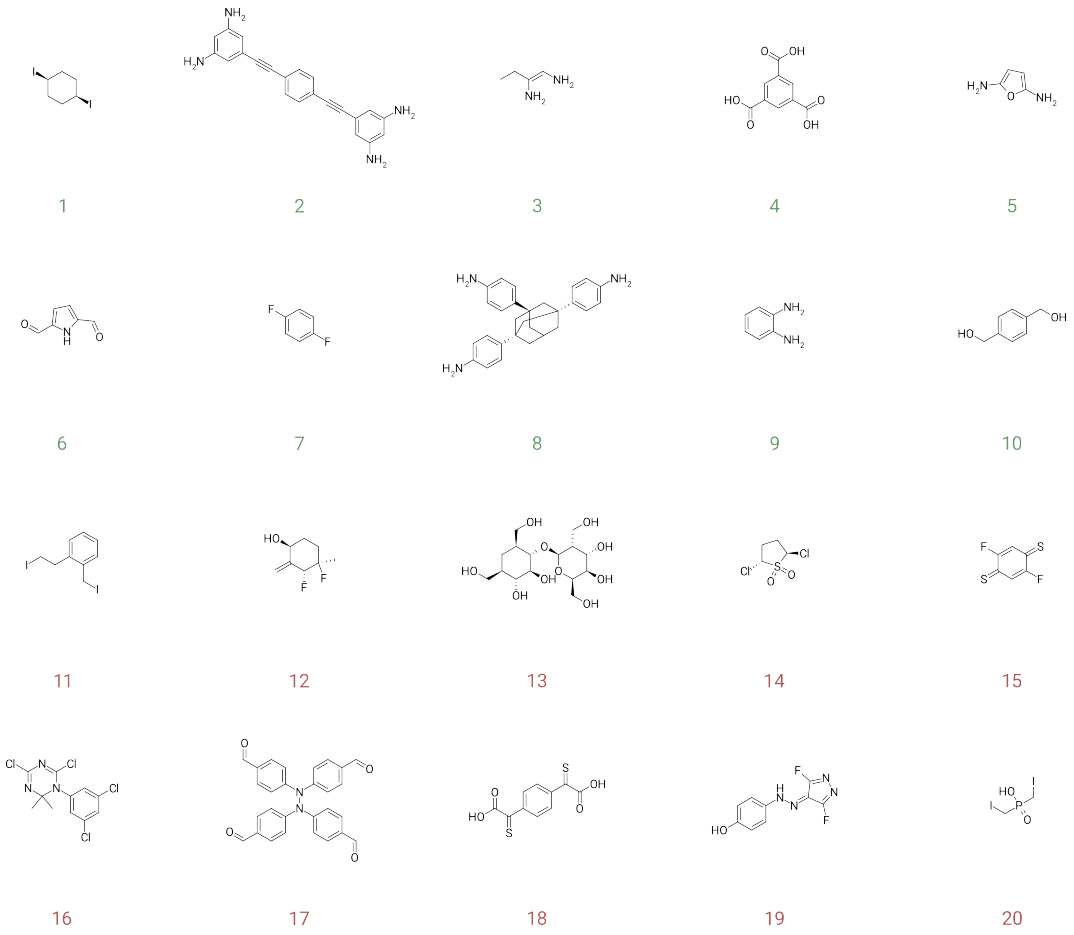


Figure 3.7: Molecules representative of the ten clusters from the easy-to-synthesise class (1-10, green caption) and the difficult-to-synthesise class (1-20, red caption) labelled by synthetic chemists. The molecules selected exhibited the closest Euclidean distance to the centroid of each cluster, where clusters were generated using the K-means algorithm.

fingerprints from 1,024 to 50 dimensions.[177] After performing K-means on the 50 components, I identified the centroid of each of the ten clusters for each class and identified the molecule with the closest Euclidean distance to the centroid as the most representative molecule for that cluster. The ten most representative molecules of each cluster for both the easy- and difficult-to-synthesise classes are shown in Figure 3.7, which could indicate substructures that a chemist may want to avoid in a synthesis.

3.3.3 Comparison with Existing Synthetic Accessibility Scores

To evaluate whether existing synthetic accessibility scores can replicate the decision-making process of expert chemists, I compared chemists' scores with the SAScore and SCScore.[102,103] I found that the correlation between the SAScore and the SCScore was weak, with a Pearson correlation coefficient of 0.27 (Figure 3.8), highlighting the significant qualitative differences between each measure of synthetic difficulty. The overlapping distributions of synthetic difficulty scores for molecules marked as easy- and difficult-to-synthesise suggest that these models cannot readily differentiate between molecules in the two classes. The lack of separation between the two classes highlights the need to develop an alternate heuristic model which can differentiate between molecules in agreement with expert chemists.

The significant differences between the scores could be due to the different approaches that SAScore, SCScore, and MPScore use to quantify synthetic difficulty. The SCScore aims to capture the synthetic complexity, using the number of synthetic steps required to synthesise a molecule as an indicator of synthetic difficulty. Molecules with structurally complex features appearing earlier in a sequence of reactions would have a lower perceived synthetic difficulty than structurally complex molecules that appear more frequently at the end of a reaction pathway.[167] This is exemplified in Figure 3.9a, where the SCScore shows an increasing synthetic difficulty with the number of steps required to synthesise Polmacoxib, chosen as an exemplar synthesis pathway for comparison by Coley *et al.* to evaluate the SCScore against the SAScore.[102,178] In contrast, the SAScore primarily uses a measure of structural complexity to quantify synthetic difficulty that fluctuates over the course of the synthesis of Polmacoxib. However, complex structurally complex features are not always indicative of molecules that are challenging to synthesise. Complex features, such as chiral centres, can often be easily introduced into a molecule from commercially available starting materials and are not always indicative of molecules that are difficult to access. The MPScore, meanwhile, is entirely subjective on the intuition of expert chemists, which may depend on previous experiences and chemical familiarity. In Figure 3.9a, the MPScore is comparatively higher than the other two scores, suggesting that this model does not capture subtle changes in synthetic difficulty over the course of a reaction. For the POC precursors in Figure 3.9b, the MPScore has a score comparatively lower than the molecules featuring in the synthetic route to Polmacoxib,

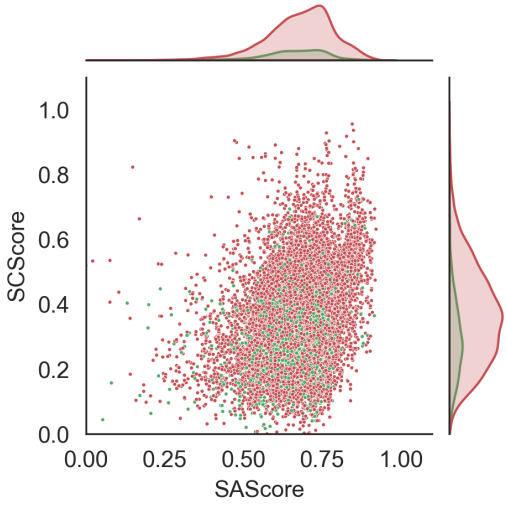


Figure 3.8: The synthetic difficulty scores of molecules from the training data set, which have been calculated using the SCScore and SAScore methods.[102,103]. The scores have been scaled between 0 and 1 to allow for easier comparison, and the colour-coding refers to molecules that are labelled easy-to-synthesise (green) or difficult-to-synthesise (red), based on assessments by synthetic chemists. For visualising the distribution of synthetic difficulty scores, kernel density estimates are presented on the top and the side panels. Here, a lower score indicates a molecule is easier to synthesise, whereas a larger score indicates a molecule is more difficult to synthesise.

suggesting that the model can be used to selectively identify POC precursors but struggles to generalise outside of this domain. The average time to classify a molecule was 20 s by expert chemists, which could limit the model to simply the presence of common structural features or functional groups. The subjective nature of human intuition also invites the possibility of bias and errors, limiting the MPScore to biasing toward easy-to-synthesise precursors and restricting the ability to differentiate the effects of subtle structural changes.

Model Validation by Expert Chemists

Next, I validated the ability of three synthetic difficulty models to identify easy- and difficult-to-synthesise POC precursors, the goal that my synthetic difficulty model aimed to achieve. I presented R.L.G blindly with a set of 120 diamines and trialdehydes with the lowest and highest synthetic difficulty values calculated using the SAScore, SCScore, and MPScore models (20 diamines and 20 trialdehydes per model), presenting them with the same criterion as used to obtain training data: "Can you make 1 g of this compound in under 5 steps?". All molecules are shown in Figure A.1, together with their respective synthetic difficulty scores in Table A.1, comparing the assignment of each the synthetic difficulty scores to the opinion of R.L.G. In a similar evalu-

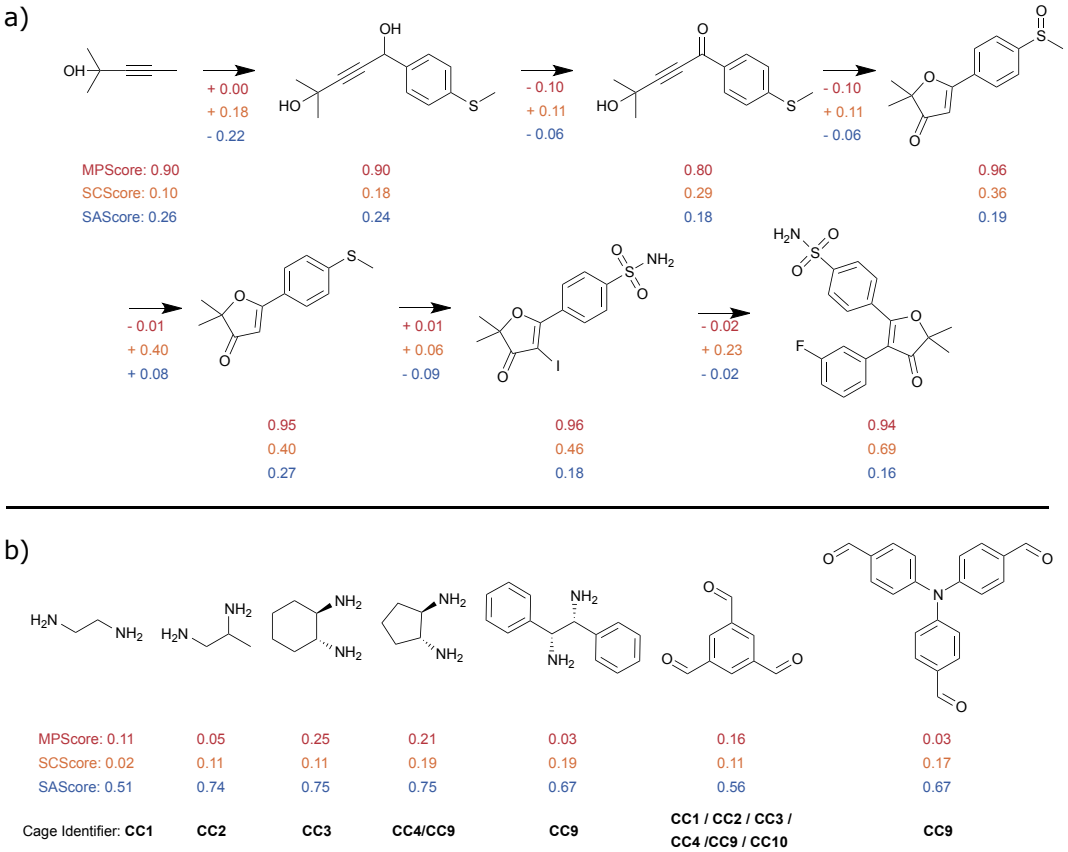


Figure 3.9: Synthesis of Polmacoxib (a),[178] exemplifying the differences between the MPScore (red), SCScore (orange), and SAScore (green) measures of synthetic difficulty. Values underneath each molecule represent the calculated synthetic difficulty score for each molecule using each scoring method, and values underneath each reaction arrow show the difference in synthetic difficulty score between reactants and products at each reaction step. Each molecule in the synthesis is assigned a comparatively high MPScore value. Instead, the MPScore aims to identify synthetically accessible POC precursors, which are lower for precursors that have been used to synthesise POCs in the literature (b). The cage identifier refers to the name of the cage the precursor is used to synthesise.

	Trialdehydes			
	True Positives	**True Negatives**	**False Positives**	**False Negatives**
SAScore	7	9	3	1
SCScore	6	8	4	2
MPScore	9	10	**1**	0
	Diamines			
	True Positives	**True Negatives**	**False Positives**	**False Negatives**
SAScore	3	10	7	0
SCScore	9	10	1	0
MPScore	10	10	**0**	0

Table 3.8: Performance metrics for the blind validation task used to assess the performance of the SAScore, SCScore and MPScore on predicting easy-to-synthesise POC precursors. True positive outcomes are defined as molecules with the lowest calculated synthetic difficulty scores that were also classified as easy-to-synthesise by R.L.G. Values in bold correspond to the score that produced the fewest false positives, corresponding to molecules ranked as difficult-to-synthesise by the expert chemist but which the model assigned a low synthetic difficulty score, implying they are easy-to-synthesise.

ation to the performance of the MPScore model, I present the true positives, true negatives, false negatives, and false positives for each model in Table 3.8.

Generally, models were better at identifying difficult-to-synthesise molecules in agreement with R.L.G, indicated by the higher number of true negatives compared to true positives. Here, a true negative denotes a molecule that was classified as difficult-to-synthesise by both the synthetic difficulty model and the R.L.G, while a true positive indicates a molecule that was classified as easy-to-synthesise by both. On visual inspection of the molecules in Figure A.1, molecules predicted to be the most challenging to synthesise, based on their high synthetic difficulty scores, tended to be larger and more complex than those with lower scores. This suggests that identifying these difficult-to-synthesise molecules could potentially be a simpler task for chemists, resulting in fewer false negatives.

Although minimising false negatives is desirable, this is not a requirement for the task of only identifying easy-to-synthesise molecules as opposed to eliminating precursors that are difficult-to-synthesise. The MPScore was most effective for my intended purpose, resulting in the lowest cumulative false positives across the amine and aldehyde precursors (1 in total). While it was reassuring that the MPScore was able to correctly identify easy-to-synthesise precursors, I note that the sample size of 20 diamines and 20 trialdehydes per model is too small to draw any concrete conclusions about the performance of each model.

3.3.4 Conclusion

In this chapter, I described the development of a new model of synthetic difficulty, the Materials Precursor Score (MPScore), specifically tailored towards the identification of easy-to-synthesise POC precursors. I compared existing models of synthetic difficulty, showing how current methods are unable to replicate the decisions of expert chemists, highlighting the necessity to develop a model specifically able to bias towards easier-to-synthesise molecules for POC synthesis. Altogether, I collected 12,553 labelled training data from three materials chemists in the field of POCs, asking them the question: "Can you make 1 g of this compound in under 5 steps?", and trained a RF binary classifier on the obtained data.

Due to the imbalanced nature of the dataset, primarily skewed towards difficult-to-synthesise molecules, I described an approach to minimise the impact on model performance by reducing the number of difficult-to-synthesise molecules incorrectly classified as easy-to-synthesise (false positives). By altering the probability threshold the model uses to assign a molecule to the easy- or difficult-to-synthesise class, the model can be tuned to maximise precision at the expense of recall, a trade-off that was shown to be beneficial for the task of identifying easy-to-synthesise precursors.

Although the MPScore has not been tested beyond POC precursors, the model can be used to pre-screen precursors for other materials with similar synthetic requirements, such as COFs, and organic precursors for organic-inorganic materials. While the use of a binary classifier was unable to capture the subtle variations in the ease of synthesis, it was able to detect easy-to-synthesise precursors, a significant bottleneck in many computational materials screening workflows. Indeed, the MPScore can predict synthetic difficulty on the order of milliseconds, providing chemists with a means of rapidly filtering potential POC precursor candidates, enabling them to focus on the candidates with the highest probability of being experimentally realised. In addition to its ability to rapidly screen large precursor libraries, the MPScore can also be incorporated into a multi-objective optimisation function to generate experimentally feasible molecules as part of a generative algorithm.[121]

To improve model performance and generalisability, I could extend the library of molecules to include more diversity and greater variation in functional groups. Retraining the MPScore model on a larger dataset could allow the model to be applied to a greater diversity of precursors, facilitating the prediction of synthetic difficulty of precursors in other computational materials discovery workflows. I could also include more chemists in the labelling exercise to obtain multiple labels for individual molecules, to better achieve a consensus on scoring values. In the next chapter, I describe the application of the MPScore in a high-throughput screening workflow to discover novel POCs with promising properties.

4 High-Throughput Screening for Porous Organic Cages

The work in this chapter led to the publication of:

S. Bennett, F. T. Szczypiński, L. Turcani, M. E. Briggs, R. L. Greenaway and K. E. Jelfs, Materials Precursor Score: Modeling Chemists' Intuition for the Synthetic Accessibility of Porous Organic Cage Precursors, *J. Chem. Inf. Model.*, 2021, 61, 4342–4356.

4.1 Introduction

The task of discovering new POCs with desirable properties is a time-consuming and resource-intensive process that has traditionally relied on the expertise of experimental chemists to make incremental modifications to known POCs. However, computational methods have emerged as a powerful tool to expedite this process by enabling the prediction of candidate properties prior to experimental synthesis. In a study by Berardo *et al.*, more than 10,000 combinations of trialdehydes and diamines were computationally screened to identify a promising candidate for cage synthesis.[63] The approach used in this study led to the discovery of a shape persistent cage formed from an unsymmetrical trialdehyde precursor, which would not have been considered by chemists due to its lack of symmetry.

Although computational screening has the potential to predict the properties of many novel POCs, the synthesisability of precursors is often neglected in many high-throughput computational workflows.[100] This oversight can be a major drawback because many simulations are performed on POCs that are unlikely to be synthesised, resulting in computational simulations on candidates that are unlikely to be experimentally realised. For example, Berardo *et al.* employed an evolutionary algorithm to computationally discover shape persistent POCs, but many of the precursors had additional side groups that would not contribute significantly to cage properties and would instead complicate their synthesis.[115] To enhance the efficiency of computational simulations and transition to the experimental realisation of new materials, it is beneficial to incorporate synthetic feasibility as a criterion.

Introducing synthetic feasibility as a criterion in computational screening workflows can reduce the cost and time associated with manually analysing all proposed candidates, which is particularly important for large-scale studies that involve screening thousands of candidates. One example of how computational measures have been applied in high-throughput screening workflows is the workflow by Wen *et al.* to discover new organic dyes for dye-sensitised solar cells.[179] Following their high-throughput screen, they discovered 9,886 dyes with promising percent conversion efficiencies, too many to assess manually. Instead, the SAScore was used to prioritise candidates, resulting in the identification of 8 promising candidates.[103] Similarly, the SAScore was used to reduce the chemical space of high energy density materials from a fragment-based library containing 10^5 different molecules.[180] In both examples, computationally-derived measures of synthetic accessibility were able to bias towards more synthesisable precursors far quicker than manually inspecting all possible candidate molecules, facilitating an easier transition to the experimental realisation of these materials.

To address the challenge of neglecting precursor synthesisability in high-throughput workflows for predicting novel POCs, I incorporated the MPScore model, introduced in Chapter 3, into a computational screening workflow for screening POCs. Specifically, I used the SCScore, SAScore, and MPScore models as filters for easy-to-synthesise precursors in the initial stage of a POC computational screening workflow. By using each model in an initial precursor pre-screening step, I aim to identify shape persistent POCs that can be accessed from the easiest-to-synthesise precursors, thereby increasing the probability of proposing POC candidates that can be synthetically realised.

4.2 Methods

4.2.1 Computational Screening Workflow

The workflow used for computational screening is illustrated in Figure 4.1. To provide a general overview of the entire process, I will initially outline each step of the workflow in broad terms. Later, I will provide details of each step in the subsequent paragraphs. Firstly, a database of trialdehyde and diamine precursors was created, which can be combined *in silico* to form POCs with **Tri**^{4}**Di**^{6} topologies using the *stk* Python package.[56] To bias the dataset towards precursors that are easiest to synthesise, precursors with the lowest 1% synthetic difficulty scores were selected from the initial library using each of three synthetic difficulty scores, namely, the MPScore, SCScore, and SAScore.[102,103,176] Following the precursor selection process, each cage is assembled using *stk*,[56] and undergoes a three-step geometry optimisation procedure to locate the lowest energy conformers. Finally, the internal cavity of the lowest energy conformer is analysed using the *pyWindow* software to characterise each cage as shape persistent or collapsed.[51] The primary goal

of this workflow is to maximise the number of shape persistent cages discovered by identifying promising precursors that are more likely to form a shape persistent cage and are easy to synthesise.

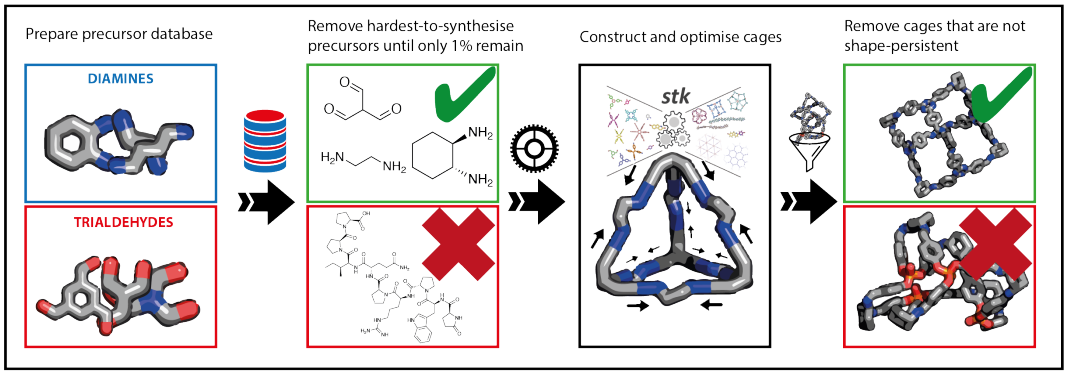


Figure 4.1: The high-throughput computational screening workflow for the discovery of novel porous organic cages from synthetically accessible precursors.

Precursor Database

In this work, my initial goal was to create a database of potential precursors that could form POCs in the **Tri**^{4}**Di**^{6} topology, which are formed through a [4 + 6] imine condensation reaction with a tritopic aldehyde and a ditopic amine. This topology was selected due to the high probability that a **Tri**^{4}**Di**^{6} topology will preferentially form over the other **Tri**^{6}**Di**^{9} and **Tri**^{8}**Di**^{12} topologies known to form in the POC literature. In addition, **Tri**^{4}**Di**^{6} are more likely to be shape persistent than their larger counterparts. This is exemplified in a previous study by Turcani *et al.*, where larger proportions of **Tri**^{4}**Di**^{6} cages were predicted to be shape persistent compared to **Tri**^{8}**Di**^{12} (38% compared to 28%).[89]

To construct the initial precursor database, I used a subset of the eMolecules[169] and Reaxys databases,[106] selecting only molecules that contained exactly three aldehydes or two primary aliphatic amines. An automated detection algorithm, implemented using *RDKit*,[173] was used to automatically identify all functional groups within the molecules.[181] Broadly, this algorithm marks certain atoms (heteroatoms, specific carbon, etc.) present in common functional groups and then merges the marked atoms to define a single unique functional group. Using this algorithm, I was able to differentiate between primary, secondary, tertiary and aromatic amines, but only primary amines were chosen for further investigation due to their prevalence in the POC literature.[32] Aromatic and aliphatic aldehydes were also included in the precursor dataset. Although the precursors selected from the eMolecules and Reaxys databases were restricted to those containing exactly three aldehydes or two primary aliphatic amines, the cost and complexity of synthesising these

molecules can vary widely. Additionally, the dataset was not filtered based on the precursors' ability to form shape persistent cages. The final precursor database consisted of 7,190 ditopic amines and 98 tritopic aldehydes, resulting in a possible 704,620 POC combinations that can be formed when restricted to the **Tri**^{4}**Di**^{6} topology. An initial geometry for these precursors was generated using the ETKDG algorithm implemented within *RDKit*,[54], followed by the MMFF force field to optimise the geometries.[182]

4.2.2 Identifying Synthesisable Precursors

For each combination of trialdehyde and diamine precursor that could be combined to form a POC, I calculated their combined synthetic difficulty score using three scoring methods: the SAScore, SCScore, and MPScore.[102,103,176] For each pair of trialdehyde and diamine precursors, I summed the synthetic difficulty score of each precursor and then scaled the values between 0 and 1 to allow comparison between the three models. In the case of the MPScore, the probability of a molecule being in the "difficult-to-synthesise" class was treated as a continuous score, where higher values were interpreted as greater synthetic difficulty. For the MPScore, I used the calibrated probabilities following the calibration procedure described in Chapter 3. Precursor combinations were selected from the first percentile of synthetic difficulty values (a total of 21,140 pairs, assuming that there were no duplicates) and carried forward for cage construction, optimisation, and property evaluation. Duplicate precursor combinations were identified by concatenating the SMILES strings of the diamine and trialdehyde precursors and finding the overlap between the combinations selected by each score. In addition to the precursor selected from the three synthetic difficulty scores, I randomly selected another 1% of precursor combinations as a control sample to assess whether the synthetic difficulty scores also favoured precursors with a higher likelihood of forming shape persistent cages.

4.2.3 Cage Construction and Conformational Searching

The process of constructing cages using *stk* involves placing and aligning precursors on the vertices of a predefined topology graph, resulting in cages with a **Tri**^{4}**Di**^{6} topology in this case.[56] Subsequently, *in silico* reactions are performed to connect functional groups assigned to the same edge, resulting in the formation of imine bonds from aldehyde and amine functional groups. However, the initial assembly of cages using *stk* results in structures with long unphysical bond lengths between precursors. To obtain more realistic and chemically meaningful structures, an optimisation procedure is typically employed. In this work, I used *stk* to fully automate cage assembly in the **Tri**^{4}**Di**^{6} topology.

To generate a chemically meaningful structure with realistic low energy conformation, I used a three-step conformer search procedure that involved the OPLS3 force field implemented in Schrödinger's MacroModel software.[136,139] In the first step, only the bonds created by *stk* were optimised, while the geometry of the atoms in each precursor placed on the vertices and edges of the cage remained fixed. In the second step, the geometries of all atoms were optimised, and the atoms in the precursors were unconstrained. Finally, MD simulations were performed in the NVE ensemble for 2 ns, sampling 50 conformers along the trajectory using a time step of 1 fs and a temperature of 700 K. Each sampled conformer underwent further optimisation at the force field level, and the lowest energy conformation was selected for further property analysis. The force field optimisation employed a gradient convergence criterion of 0.05 $kJ^{-1}A^{-1}$ with 2,500 as the maximum number of optimisation steps, using the Polak-Ribiere Conjugate Gradient minimisation algorithm.[183] The lowest energy conformation was then selected for subsequent property analysis.

The three-step conformer search procedure using the OPLS3 FF, described in Section 2.1.1, was also used to generate a computational database of over 60,000 organic cages for a ML model to predict shape persistency.[89] While OPLS3 has been validated for various imine organic cages,[31] the FF may not accurately describe the diverse functional groups present in the initial precursor library, which could lead to an unfeasible conformation post-optimisation.[89] Despite this limitation, this optimisation procedure provides a reasonable trade-off between computational cost and the acquisition of low energy conformations for thousands of organic cages. When analysing the most promising POC candidates from the computational screening workflow, I considered the potential inability of the FF to model diverse functional groups, if they are present in POC precursors.

4.2.4 Porosity Evaluation

After the optimisation procedure, an automated procedure was used to identify and filter out cages that lost their shape persistence. To quantify the size of internal pores, I used *pyWindow*, a Python package capable of accurately estimating the pore sizes of cavity-containing discrete molecules.[51] Previously, *pyWindow* has been shown to accurately reproduce the pore sizes of POCs with **Tri**^{4}**Di**^{6} topologies, and can be used to detect and analyse all the windows in a cage.[51] I used an equation developed in a previous study to label the cages as shape persistent or collapsed, with the aim of maximising the number of correctly identified shape persistent cages (Equation 4.1). In Equation 4.1, the average window difference represents the average difference in the window diameter for all possible pairs of windows. A cage was classified as shape persistent if the α parameter was less than 0.035, and the cavity size was greater than 1 Å. Otherwise, the cage

was defined as undetermined. Finally, for organic cages classified as shape persistent, the central cavity diameter was calculated using *pyWindow*.[51]

$$a = \frac{4 \times \text{ average difference in window diameter}}{\max \text{ window diameter } \times \text{ expected number of windows}} \tag{4.1}$$

4.3 Results and Discussion

To evaluate the effectiveness of the MPScore, I applied it in a high-throughput screening with the aim of discovering new promising precursors for POC synthesis. To begin, I will discuss the overall synthetic difficulty scores for all the combinations of precursors from the precursor database, comparing the distributions of synthetic difficulty scores. Next, I will outline the key findings of the computational screen, describing the distributions of cavity sizes of POCs from each of the precursors predicted to be easy-to-synthesise. Finally, I will outline some of the promising precursors that were defined as easy-to-synthesise by each synthetic difficulty score and predicted to form large numbers of shape persistent POCs.

4.3.1 Precursor Database and Synthesisable Precursors

As described in the subsequent section, I interpreted the probability returned by the MPScore model as a continuous score. For each pair of precursors used to form a POC, I assigned an overall synthetic difficulty value, using the sum of individual synthetic difficulty values of each precursor scaled between 0 and 1 to allow for easier comparison between scoring methods. Each overall score was calculated using the MPScore, SAScore, and SCScore models. For our screening workflow, each precursor combination consisted of a tritopic aldehyde and ditopic amine, forming an imine-based cage with a **Tri**^{4}**Di**^{6} topology.

In general, I found the distributions of the scores shown in Figure 4.2 to be relatively broad, with the SAScore systematically assigning higher synthetic difficulty scores than the other two models. While the SAScore and SCScore models exhibit relatively symmetric score distributions, the MPScore model is skewed towards higher values. This suggests that the MPScore model assigns a wider range of scores to precursors, which could be attributed to the imbalanced training set used to train the model. Moreover, the probabilities assigned by the MPScore model appear to be more discontinuous than those of the other two measures, as evidenced by the jagged kernel density estimate plot. This can be explained by the method used to calculate probabilities in RF models, where each probability represents the proportion of decision trees in the ensemble that vote for a sample to belong to a specific class. Therefore, the fewer the number of decision trees, the more discontinuous the probability, resulting in a more jagged distribution.

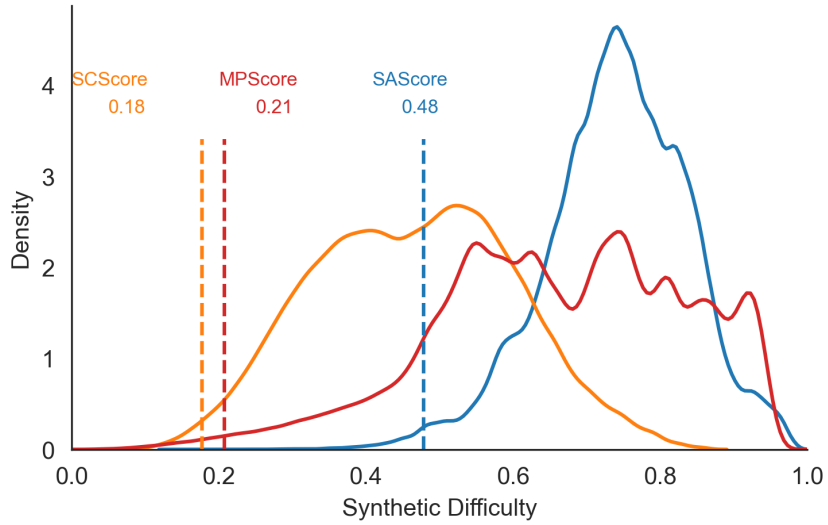


Figure 4.2: Kernel density estimates of combined synthetic difficulty scores for each combination of trialdehyde and diamine precursor, scored using three different models: the SAScore (blue), SCScore (orange), and my MPScore (red). Vertical lines indicate the first percentile for each score, and the corresponding cutoff thresholds are given above the lines and coloured according to the model. Precursor pairs with combined scores equal to or below the respective threshold value were defined as easy-to-synthesise, where a lower score indicates that the precursors in a combination are easier to synthesise and a higher score indicates that they are difficult to synthesise.

The relatively broad distributions observed across the three scores imply that potential precursors have a diverse range of synthetic difficulties, implying that even restricting the initial compound database to molecules previously reported in Reaxys or eMolecules is insufficient to ensure that precursors can be easily synthesised. The synthesis of these molecules could consist of multiple low-yielding reaction steps, toxic or expensive reagents, or challenging purification steps. It is insufficient to assume that the presence of a molecule in a compound database means that it is "synthesisable" without taking into account the aforementioned difficulties. Moreover, even if such molecules are accessible through multi-step syntheses, they are unlikely to be suitable for materials discovery unless there is strong evidence that a particular precursor is the only molecule capable of conferring a desired functionality of high value.[100] Indeed, for POCs, readily accessible precursors are necessary and are essential to create an economically viable and scalable new material.

In the computational screening workflow, I chose cut-off thresholds for each synthetic difficulty scoring method that would remove 99% of all the most difficult-to-synthesise precursor combinations, leaving the 1% of precursors predicted to be easiest to synthesise. This was done to prioritise simulations towards only the easiest-to-synthesise precursors while still exploring a sufficiently diverse chemical space. Given the enormous number of precursor combinations, which

would amount to 704,620 in total, even when considering only the **Tri**^{4}**Di**^{6} topology, I decided to limit the search to only 1% of the library. This filtering process reduced the number of computationally expensive MD simulations that would need to be performed and instead focused on the precursors that were most synthesisable.

The final cut-off thresholds for precursor combinations were: 0.21 for the MPScore, 0.48 for the SAScore, and 0.18 for the SCScore, as shown in Figure 4.2. At this probability threshold for the MPScore, the precision and recall scores for the easy-to-synthesise class were 0.89 and 0.25 respectively. The values suggest that although 75% of easy-to-synthesise molecules may be missed, the selected molecules only have an 11% chance of being a false positive, which is a necessary compromise to ensure that the selected precursors could be synthesised. Precursor combinations that exceeded these thresholds were disregarded, and those that were equal to or below the threshold were retained for further analysis, including POC construction and computational simulations.

In total, I selected a set of 28,185 precursor combinations for cage construction and optimisation. Among them, 7,046 were selected from the MPScore, 7,047 from the SAScore, 7,047 from the SCScore, and 7,045 from a randomly selected control sample. This resulted in 26,830 optimisations, as there were 1,062 overlapping precursor combinations between the four sets of selected precursor combinations. The complete breakdown of overlaps can be seen in Table 4.1. The greatest overlap was observed between the SCScore and MPScore with 621 shared precursor combinations. Fewer overlaps were observed between the SAScore and MPScore with 34 shared precursor combinations, and between the SAScore and SCScore at 205. Despite the 621 overlapping precursor combinations between the SCScore and MPScore, these combinations only account for 8% of the total selected precursor combinations for each score. This indicates that while there may be some similarity between the two scoring methods, the majority of each of the precursor combinations selected were unique to each scoring method.

	SAScore	**SCScore**	**MPScore**	**Control**
SAScore	7,015	205	34	66
SCScore	205	6,877	621	75
MPScore	34	621	6,944	61
Control	66	75	61	7,045

Table 4.1: Number of overlapping precursors scored as easy-to-synthesise by each synthetic difficulty scoring model, in addition to the control sample of randomly selected precursors.

Structural Analysis of Computationally Screened POCs

In this section, I will discuss the next steps of the workflow following the precursor filtering process. After selecting precursor combinations based on their synthetic difficulty scores, I used the

	MPScore	SCScore	SAScore	Control	Previous Work[89]
Cages	6,944	6,877	7,015	7,045	6,018
Shape Persistent	709 (10%)	287 (4%)	76 (1%)	90 (1%)	2,314 (38%)

Table 4.2: The number and percentage of shape persistent cages formed from precursors selected by each synthetic difficulty model, in addition to a control sample of randomly selected precursors. The final column shows the corresponding values from previous work.[89]

stk package to automate cage construction and conformer generation procedure.[56] The resulting structures were analysed using *pyWindow* to detect candidates that are shape persistent by analysing their windows and central internal cavity.[51] For further analysis, I will focus on candidates that were predicted to remain shape persistent with a permanent internal cavity, analysing their property distributions, with the aim of understanding the influence of precursor synthetic difficulty on final material properties.

In previous work, a database of 6,018 **Tri**^{4}**Di**^{6} POCs from trialdehyde and diamine precursors with rigid cores was constructed, out of which 2,314 (38%) remained shape persistent after geometry optimisation, as shown in Table 4.2. As expected, a much lower proportion than 38% remained shape persistent when including a strict consideration of the synthetic difficulty of precursors, without manual assessment of the precursors for their propensity to form a shape persistent POC. As shown in Table 4.2, organic cages that remained shape persistent from each synthetic difficulty score were 10% from the MPScore, 4% from the SCScore, 1% from the SAScore, and 1% from a randomly selected control sample. The percentage of shape persistent cages selected by the MPScore was 9 percentage points greater than the control sample of randomly selected precursor combinations. Both the SCScore and SAScore were much closer to that of the control sample, with percentage point differences of 3 and 0, respectively. The large difference in the number of shape persistent cages suggests that easy-to-synthesise precursors selected by the MPScore are more likely to form a shape persistent POC, compared with precursors selected by other scoring methods. Although my primary aim for the MPScore was to filter for easy-to-synthesise POC precursors, its secondary effect of biasing toward precursors more likely to form a shape persistent POC is nonetheless useful for POC chemists when selecting promising precursors.

In Figure 4.3, the distribution of the cavity diameters of the POCs predicted to be shape persistent is depicted using the three synthetic difficulty models. The results showed that the MPScore-selected precursors formed POCs with the highest cavity diameter, averaging 8.4 Å. This was followed by the control sample of precursors with the lowest synthetic difficulty score, which had a mean of 6.5 Å, the SCScore with 5.5 Å, and finally, the SAScore with 4.7 Å. It is worth noting that a previous study found that cages with cavity sizes ranging from 0-6 are the most common, and there is a lack of cages with diameters greater than 16 Å in the literature.[115] POCs with a larger

cavity are particularly challenging to synthesise due to the likelihood of collapsing after desolvation or catenating with other cages. Therefore, POCs with larger cavities are an interesting target for a computational screening process. The precursors that lead to POCs are often larger, resulting in greater degrees of freedom, competing synthesis pathways for the formation of alternate products, and easier collapse mechanisms.

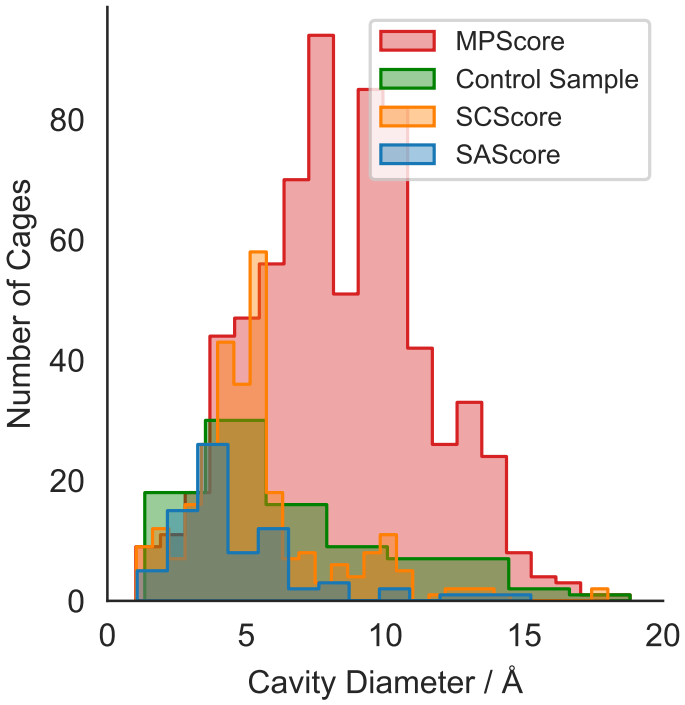


Figure 4.3: The distribution of cavity sizes for POCs predicted to be shape persistent from the high-throughput computational screen. Only POCs with cavity sizes predicted to be greater than 1 Å are included. The POCs were formed from easy-to-synthesise precursor combinations that ranked in the lowest 1% of synthetic difficulty scores, which were calculated using the SAScore (blue), SCScore (orange), MPScore (red), and a control sample of randomly selected precursors (green).

My analysis of the distribution of shape persistent POCs demonstrates that it is possible to uncover POCs from precursors that are predicted to be easy to synthesise, without including specific structural features in precursors aimed at obtaining shape persistent POCs. In particular, my results indicate that the MPScore is the most effective synthetic difficulty score in identifying a higher proportion of shape persistent POCs (709 cages) compared to other scores. In total, I identified 1,072 shape persistent cages that can be formed from easy-to-synthesise precursors, a considerable number compared to the hundreds of POCs discovered to date.[48] Although most

shape persistent POCs have a pore diameter of 5 Å, my approach successfully identified a smaller number of cages with larger pore diameters. This finding is encouraging, as it suggests that precursors capable of forming larger POCs *are* synthetically accessible.

4.3.2 Analysis of Shape Persistent POCs

Structural Features of Selected Precursors

Altogether, a large database of 28,185 hypothetical POCs was generated, of which 1,162 (4% of precursor combinations) were predicted to remain shape persistent, allowing an analysis of the structural features of the precursors that are likely to form a shape persistent POC. Figure 4.4a-d illustrates the differences in the structural characteristics of the precursors selected by each synthetic difficulty score, and the difference between the structural characteristics of the precursor in the cages predicted to be shape persistent and all cages that underwent geometry optimisation. Here, I will begin by contrasting the different structural features of precursors selected by each synthetic difficulty score, and how these relate to the cage cavity sizes shown in Figure 4.3. I will then outline some general design rules for cage precursors that can be derived from the variation in structural characteristics of shape persistent and collapsed cages.

Figure 4.4a shows the variation in the average distance of the bonds between functional groups. This metric refers to the average distance between the nitrogen atoms in the amine functional groups of the diamine and the carbon atoms in the aldehyde functional groups of the trialdehyde, which provides a general proxy for the precursor size. Of all shape persistent cages, the largest precursors were selected by the MPScore, which could explain the trend observed in Figure 4.3. Specifically, shape persistent cages formed from precursor combinations selected by the MPScore exhibited larger cavity sizes compared to those selected by the SAScore, SCScore, or the control sample. When reactive functional groups are connected by rigid structural features, these large precursors would be more likely to form a shape persistent cage with a larger cavity size. A similar trend is observed in Figure 4.4b, where precursor combinations that resulted in shape persistent cages selected by the MPScore contained more atoms on average compared to the SAScore and SCScore, which could again explain the observed tendency to form larger POCs.

The percentage of rotatable bonds and the percentage of double bonds are shown in Figures 4.4c and d, respectively. Each synthetic difficulty score selected precursor combinations with comparable proportions of double bonds and rotatable bonds compared to the control sample, suggesting that there was no systematic preference for precursors with these structural features from each of the synthetic difficulty scores. The large standard deviations for the percentage of double and rotatable bonds further support the absence of selectivity for these features. Indeed, the only structural feature where the synthetic difficulty scores exhibited a marked preference was the aver-

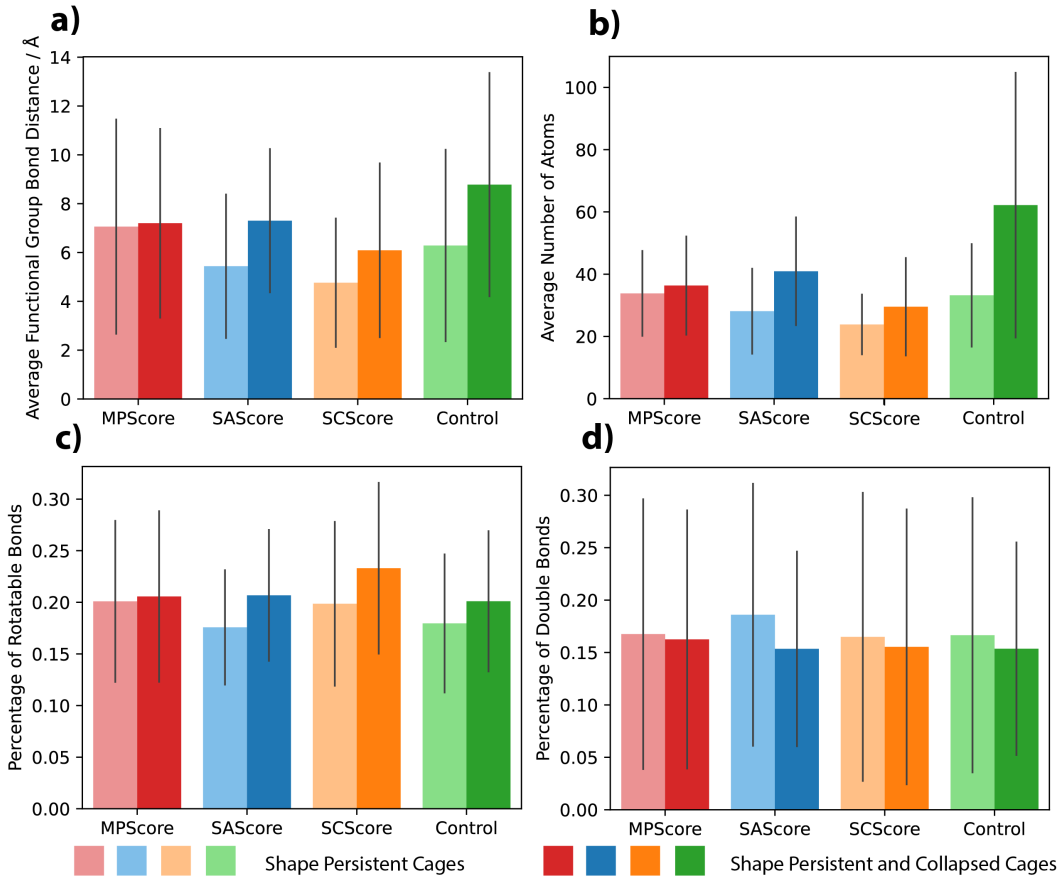


Figure 4.4: Average structural features of the cages formed from precursor combinations selected by the MPScore (red), SAScore (blue), SCScore (orange), and a randomly selected control sample (green). Lighter shades represent the structural features of precursors that resulted in shape persistent cages, while darker shades represent precursors that resulted in both collapsed and shape persistent cages. The black lines indicate the standard deviation of each structural feature. The structural features are the average functional group bond distance (a), which represents the average distance between nitrogen atoms in diamines and the carbon atoms in the aldehyde of trialdehyde precursors, the average number of atoms (b), the percentage of rotatable bonds (c), and the percentage of double bonds (d).

age number of atoms. The three scoring models consistently selected precursors with fewer atoms than the control sample, which can be attributed to the fact that molecules with more atoms are more likely to possess complex structural features that would result in higher synthetic difficulty scores.

Figures 4.4a-d provide some general design principles for cage precursors. Firstly, shape persistent cages were formed from smaller precursors with fewer atoms and shorter distances between the end groups, as indicated by the decrease in these properties for the selected precursors in Figures 4.4a and b. This finding is in contrast to previous studies in which larger precursors were

found to be more likely to form shape persistent cages.[89] However, larger molecules could still be favourable if they contain rigid structural features. Furthermore, shape persistent cages had a lower percentage of rotatable bonds (Figure 4.4c) and a higher percentage of double bonds (Figure 4.4d). These two features indicate a higher degree of rigidity, making the cage less likely to collapse. Generally, these observations suggest that both precursor rigidity and size are important in the POC design process and that, to achieve POC with larger cavity sizes, a larger precursor with more rigid structural features may be preferred.

Figure 4.5 displays some of the most promising precursors selected by the MPScore, each appearing most frequently in shape persistent cages (at least 6 times). As shown in Figure 4.5a, each of the diamines is a 1,2-diamine, which is a common precursor for synthesising **Tri**^{4}**Di**^{6} cages in the literature, however, typically without external functionalities in the cage. Meanwhile, the trialdehydes in Figure 4.5b are relatively symmetrical, generally containing few external functionalities. While external functionalisation can sometimes be beneficial in cage synthesis,[184] these functionalities are unlikely to affect the computationally-predicted cavity size for these cases. Precursors present in the best performing POC candidates in one run of the evolutionary algorithm conducted by Berardo *et al.* are shown in Figure 4.5c, which was targetting POCs with highly symmetrical windows and a cavity size of 5.0 Å.[115] Each of the precursors possessed external functionalities that would not contribute significantly to the size of the internal cavity, but would complicate the synthesis of these molecules. Generally less complex overall, the precursors predicted by the MPScore in Figure 4.5a and b can offer an insight into the types of structures that are easy-to-synthesise and may result in a shape persistent cage experimentally, if the specific **Tri**^{4}**Di**^{6} topology forms experimentally.

POCs with Large Cavity Sizes

During the computational screening workflow, I identified seven distinct shape persistent cages with a cavity size of 16 Å or larger, out of which five were selected using the MPScore, two with the SCSCore, and none with the SAScore. The six largest cages, in descending order of cavity sizes, are shown in Figure 4.6, along with their corresponding precursor pairs. Table 4.3 summarises the properties of these cages, along with their corresponding combined synthetic difficulty scores. In particular, all six cages were predicted to have a cavity size larger than 16 Å, and four were identified using the MPScore model. In contrast, using the SCScore would have only identified cages **2** and **3**, and all 6 cages would have been missed using only the SAScore.

To determine the feasibility of synthesising these cage precursors, I manually analysed the six precursor pairings. The availability of precursors was assessed using ZINC, an extensive commercially available compound database.[185] The diamines required for the cages **1**, **3**, **5**, and **6** were commercially available, with multiple suppliers listed in the ZINC database. Meanwhile, the di-

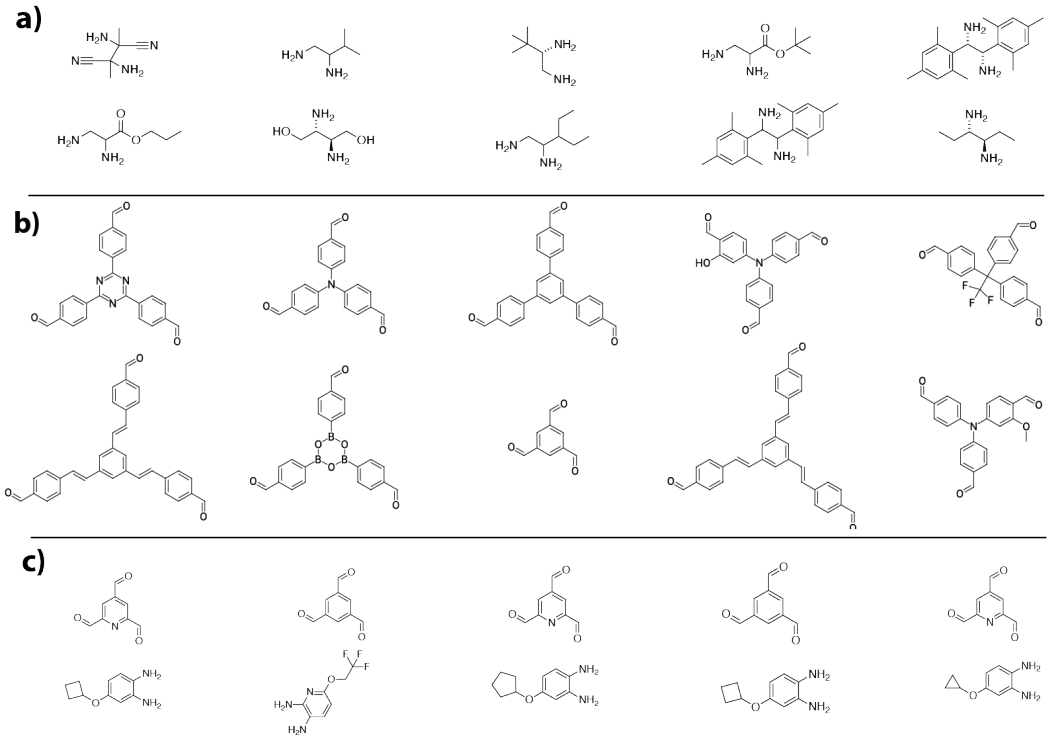


Figure 4.5: The 10 diamines (a) and trialdehydes (b) that appeared most frequently in shape persistent cages. Each precursor in a and b was present in at least 6 shape persistent cages, and each was selected by the MPScore model. In contrast, (c) shows the precursors of the top five cages discovered in a single run of the evolutionary algorithm by Berardo *et al.*[115] Here, the diamine precursors contain external functionalities that would not contribute significantly to the internal cavity size of the cage.

amine for cage **4** can be synthesised in two steps. The enantiopure form of the diamine needed for cage **4**, selected by the SCScore, would be difficult to access, with no commercial suppliers or robust synthetic routes available at the time of investigation. Meanwhile, the precursor 1,3,5-tris(phenylenevinylene)benzene was found in five of the seven remaining shape persistent cages with a cavity diameter larger than 16 Å, suggesting its potential as a favourable choice when designing cages with a large cavity size. The synthesis of this trialdehyde precursor has been reported in two relatively high-yielding reaction steps.[186]

Together, the 1,072 cages predicted to be shape persistent is much larger than the number of POCs that have been experimentally realised.[48] Despite the MPScore potentially missing many easy-to-synthesise precursors due to its low recall, the model is still able to bias towards easy-to-synthesise precursors that can form shape persistent POCs. The results of the high-throughput screen and precursor analysis also imply that easy-to-synthesise cages can produce cavity diameters over 16 Å, if the $\mathbf{Tri}^4\mathbf{Di}^6$ topology can be formed experimentally.

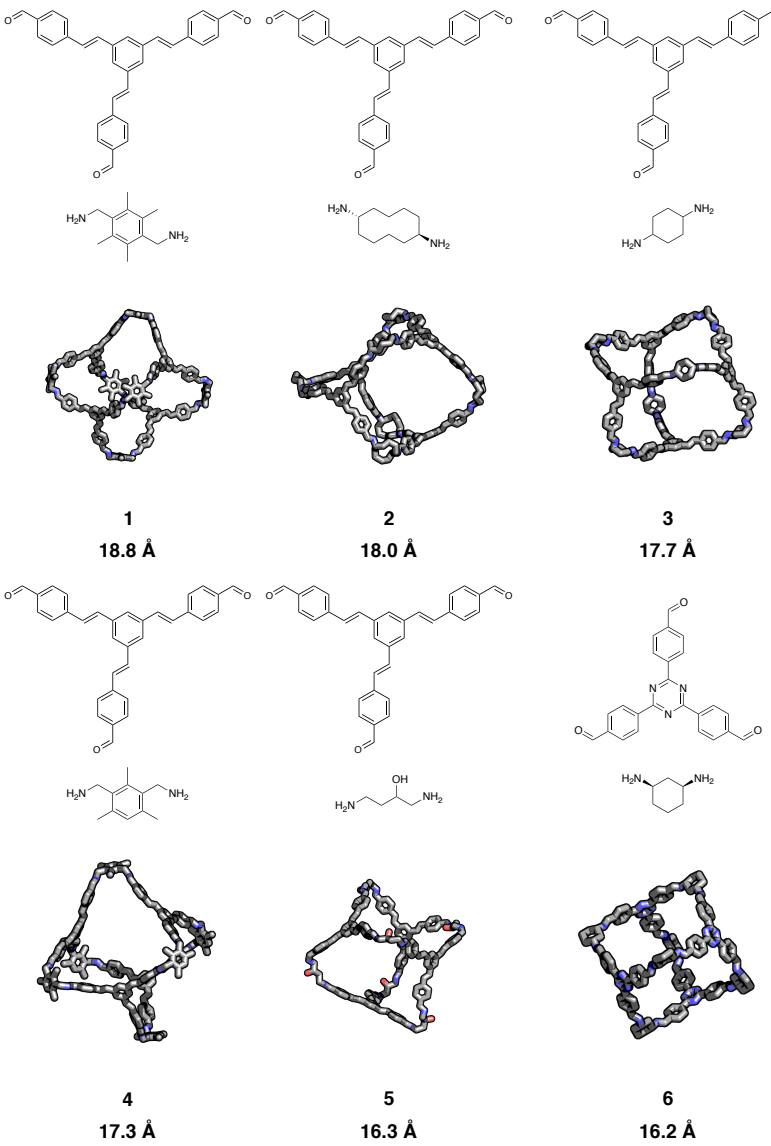


Figure 4.6: The six largest shape persistent POCs, alongside their corresponding precursor pairs, identified through my computational screening workflow. Cages are in descending order of cavity size, with the calculated cavity diameters below each cage. Table 4.3 shows the corresponding synthetic difficulty scores for each cage. The majority of the cages (**1**, **4**, **5**, and **6**) were identified by the MPScore, whereas cages **2** and **3** were identified by the SCScore. The structures of these cages are represented with carbon atoms in grey, nitrogen atoms in blue, and oxygen atoms in red. Hydrogen atoms are omitted for clarity.

		Precursor Synthetic Difficulty		
Number	**Cavity Diameter / Å**	SCScore	SAScore	MPScore
1	18.8	0.23	0.73	**0.17**
2	18.0	**0.17**	0.72	0.48
3	17.7	**0.15**	0.53	0.29
4	17.3	0.29	0.73	**0.19**
5	16.3	0.26	0.73	**0.20**
6	16.2	0.36	0.68	**0.10**

Table 4.3: The six largest shape persistent POCs are presented with their corresponding precursor pair synthetic difficulty scores and calculated cavity diameters. A low score suggests easier synthesis, while a higher score indicates a more challenging synthesis for a given precursor pair. The values highlighted in bold reveal each precursor pair belongs to the lowest 1% of values for each scoring method.

4.4 Conclusions

In this chapter, I used the MPScore model to identify synthesisable precursors as part of a high-throughput computational screening workflow to discover POCs with permanent internal cavities. The MPScore was found to favour precursors that were more likely to form shape persistent POCs than precursors selected by other synthetic difficulty models. Interestingly, even though I limited the precursors to only those predicted to be easiest to synthesise, I was still able to discover POCs with unconventional properties, such as those with cavities sizes in excess of 16 Å, a size that is rarely observed in the literature. In total, I computationally identified 1,072 shape persistent cages, including seven predicted to have a cavity size of greater than 16 Å. I further confirmed that the majority of the six largest cages could be synthesised using commercially available precursors or straightforward synthesis routes described in the literature.

From the large database of 28,185 unique POCs, I derived some general design rules for POC precursors. One key observation was that many of the proposed trialdehydes were highly symmetrical and had increased rigidity due to a high number of double bonds, such as the 1,3,5-tris(phenylenevinylene)benzene trialdehyde that was found in many of the larger cages. Additionally, smaller 1,2-diamines were the most frequent diamines present in shape persistent cages, consistent with previous POC research. Cages with cavity sizes larger than 16 Å require greater separation between amine functional groups, as seen in the 1,3-diamine in cage **6** in Figure 4.6 However, it should be noted that these design rules apply solely to cages that do not collapse following MD simulations, and that there is no guarantee that these cages will form experimentally.

While my high-throughput computational screening workflow focused on identifying easy-to-synthesise precursors for POC synthesis, there are numerous other factors that must be consid-

ered when designing and synthesising POCs. One of the primary challenges is predicting the outcome of imine condensation reactions, which may lead to the formation of different topologies, oligomers, or polymers, rather than the desired cage topology. Moreover, since only the **Tri**^{4}**Di**^{6} topology was modelled in this study, other topologies that could form are likely to have substantially different properties. Additionally, the solubility of the precursor and the possibility of forming insoluble products are critical considerations in POC synthesis.[100]

Unfortunately, incorporating all of these factors into a computational screening workflow is intractable. Thus, human expertise and knowledge still play a critical role in the development of POCs, although computational efforts can provide some direction. In the next chapter, I attempt to address the challenges associated with POC synthesis through a combination of computational and experimental methods. By using some of the top performing precursors from the high-throughput computational screen in this chapter and synthesising them in the lab, I aimed to validate the results of the computational screening and further explore the factors that can influence the formation of the desired topology. The combination of computational predictions and experimental validation can provide a more comprehensive understanding of the complex supramolecular interactions that dictate POC formation.[48]

5 Developing an Automated Experimental Workflow for Cage Discovery

5.1 Introduction

In the previous chapter, I outlined a high-throughput computational screening workflow for the identification of promising precursors to use for POC synthesis. The screening workflow outlined several promising trialdehydes and diamines which were commercially available at a low cost or with a facile synthetic route, each predicted to form at least one **Tri**^{4}**Di**^{6} POC with a shape persistent cavity. However, despite being able to predict whether a precursor combination in the **Tri**^{4}**Di**^{6} topology will remain shape persistent, there is no guarantee that that specific topology will be experimentally observed. Instead, aldehyde and amine precursors can self-assemble into a plethora of different topologies or larger self-assemblies, such as polymers or oligomers. Indeed, the self-assembly process is complex and dictated by not only the internal geometries of the precursors but also external experimental conditions that are difficult to account for computationally, such as solvent choice, temperature, and precursor concentration.

The unpredictability of self-assembly in chemistry poses a challenge in determining the topology that will form from a specific set of precursors. High-throughput experimentation (HTE) has emerged as a powerful tool for addressing this challenge. By automating the process of combining trialdehyde and diamine precursors and analysing the resulting topological results, HTE offers a means of exploring a larger chemical space, increasing the probability of serendipitous discoveries such as novel POC topologies.[48] This approach has gained popularity in various fields of chemistry, such as pharmaceutical development and materials synthesis.[125,187] For example, in the field of porous materials, HTE has been employed to screen crystallisation conditions for gas sorption applications,[68] and size-excluded solvents for porous liquids.[127] Such automated experimentation offers a more efficient way to explore a vast number of combinations and achieve greater throughput compared to conventional batch synthesis.

Combined computational and experimental approaches have also been shown to accelerate the discovery of POCs, combining the predictive ability of computation with the increased number

of reactions that can be carried out in parallel using HTE. Work by Greenaway *et al.* demonstrated accelerated discovery by successfully forming 32 cages with clean conversion from the combination of 3 triamines and 26 aldehydes.[48] Computational simulations were performed alongside the high-throughput experiment to rationalise the observed topology, showing a generally good agreement with topologies that were observed experimentally. In their work, the authors reported a strong energetic preference of 30 kJ mol^{-1} for a specific topology to be experimentally observed, overcoming the additional experimental and entropic factors that complicate the self-assembly process.

Using a similar approach to Greenaway *et al.*,[48] I outline a combined computational and experimental workflow using HTE to combine 9 trialdehydes and 16 diamines. Each precursor was selected based on the design rules and recommendations of the high-throughput screening workflow in the previous chapter, together resulting in 144 combinations of combined precursors. To evaluate the reaction outcome, a fully automated Python workflow was developed here to analyse the NMR and HRMS data, with the aim of identifying whether a cage formed from a reaction, and, if so, what specific topology or topologies were present. Using this workflow, I was able to assign each experimental precursor combination to one of ten possible categories, each corresponding to a different reaction outcome with varying degrees of success in forming a cage. Using these results, I explore the effect of precursor geometries on the self-assembly process, using a diverse range of precursors, predicted to form both shape persistent and collapsed cages, to better understand the internal features of precursors and external reaction conditions that affect cage formation.

Alongside the experimental high-throughput screen, computation was used to predict whether experimentally formed candidates would be shape persistent in a specific topology and to rationalise the experimentally observed topologies using the relative energies of each topology. From the combination of diamines and trialdehydes, the **Tri**^{2}**Di**^{3}, **Tri**^{4}**Di**^{6}, **Tri**^{6}**Di**^{9}, and **Tri**^{8}**Di**^{12} topologies could be formed. To rationalise the experimental outcomes based on energetic preference for specific topologies, I modelled each of the four possible topologies using DFT calculations to understand whether the thermodynamic driving force was able to explain the observed topology.

5.1.1 Precursor Library

The precursor library used in this work is shown in Figure 5.1, and consists of 9 trialdehyde and 16 diamine precursors. This precursor library was primarily selected based on the results of the computational screening workflow in the previous chapter, which predicted that many of these precursors would form shape persistent **Tri**^{4}**Di**^{6} cages of varying sizes. In addition to those precursors predicted to form shape persistent cages, a wide range of flexible diamines with varying

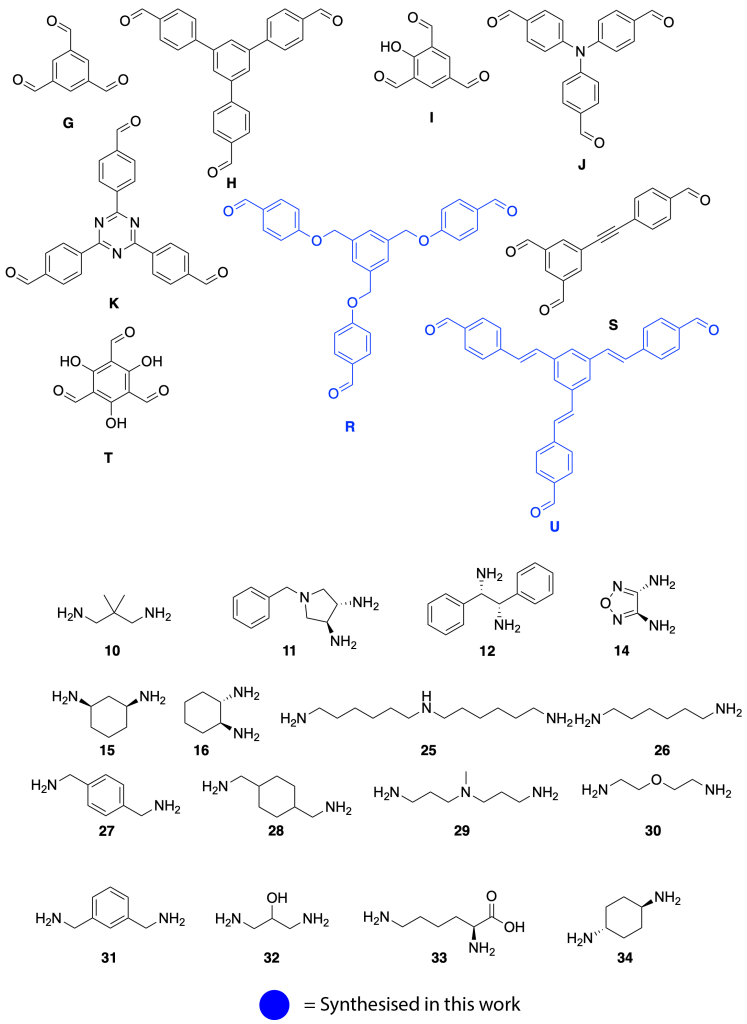


Figure 5.1: The library of precursors used in the high-throughput screening. Each combination of trialdehyde and diamine is represented by the trialdehyde letter followed by the diamine number. The numbering of precursors is not continuous as this library is a subset of a larger precursor library, and the numbering reflects their positions in that library. Precursors in blue were synthesised in this work.

functional groups and geometries were chosen to investigate the effect of precursor geometries on the self-assembly process. Each of the selected aldehydes exhibited relatively rigid aromatic cores of varying sizes, some of which are already known to form shape persistent POCs.[31] Furthermore, trialdehydes with external functionalities, such as hydroxyl groups, were incorporated into the screening process to evaluate their effect on cage formation, given that they have been shown to promote cage formation in previous studies.[188]

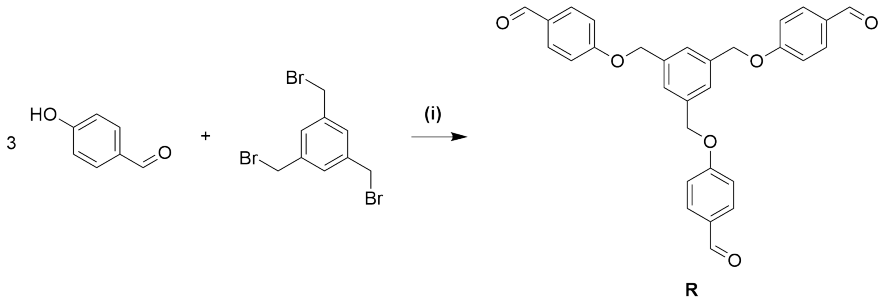
Broadly, the diamines loosely fall under two subcategories: those with large flexible chains and those with more rigid cyclic cores. Diamines **10**, **25**, **26**, **29**, **30**, **32**, **33** fall into the former category, with varying chain lengths connecting the two amine groups. Diamines **11**, **12**, **15**, **16**, **27**, **28**, **31**, and **34** fall into the latter category, each with varying levels of flexibility and different substitution patterns on the cyclic core. The prevalence of vicinal diamines in POCs, such as **11**, **12**, **14**, and **16**, is well established, as they have been reported to form various cage structures, including **CC3**.[38] Meanwhile, 1,3-diamines such as **15** have been shown to form larger cages, including the $\mathbf{Tri}^{8}\mathbf{Di}^{12}$ cage **CC20**.[189] To explore a wide range of precursor geometries, a diverse range of precursors with varying bond angles between reacting functional groups were selected.

The selection of these diamine and trialdehyde precursors was influenced by the design principles and number of shape persistent cages from the previous chapter. Of the 1,072 cages that were predicted to be shape persistent from the previous chapter, 55 contained trialdehyde **U**. The cages that contained trialdehyde **U** possessed a mean cavity size of 12.9 Å, significantly larger than many POCs in the literature.[115] Other notable trialdehydes include **K**, which was present in 131 shape persistent cages, **J** in 108 cages, and **H** in 79 cages. Each of the more rigid diamines in the library (**10**, **25**, **26**, **29**, **30**, **32**, **33**) were present in at least one shape persistent cage, with **16** (CHDA) present in 7 shape persistent cages. Using the number of shape persistent cages as a guide, commercial availability for each precursor was checked using the eMolecules database,[169] prioritising those that were available at a cheap commercial cost or predicted to form a large number of shape persistent cages. Despite not being predicted to form shape persistent cages, the flexible amines in the library were included due to their cheap commercial cost, and to better understand the effect of linker size on cage assembly.

This subset of 144 precursor combinations is part of a larger set of 366 precursor combinations that were investigated in collaboration with Annabel Basford, with the goal of developing a large database of cage precursors, reaction conditions, and the topological outcome of each combination of multi-topic amine and aldehyde precursors. With such a large number of precursor combinations, a data-driven analysis can be conducted to explore the impact of precursor geometries on the self-assembly process.

5.1.2 Synthesis of Trialdehyde Precursors

Of the 25 total precursors, all of the 16 diamines were purchased and 7 of the 9 trialdehydes were either purchased or had been previously synthesised in prior work.[48,63] I synthesised the final two trialdehydes, **U** and **R**, to complete the library of precursors, shown in blue in Figure 5.1.



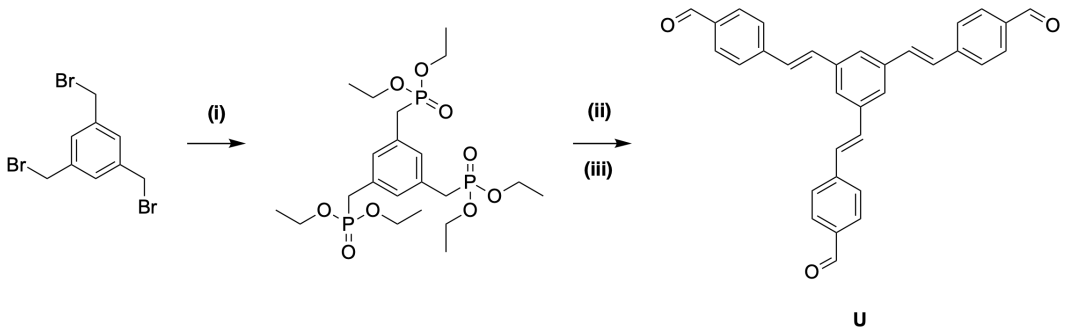
Scheme 5.1: Reaction scheme for the synthesis of 4,4',4"-((benzene-1,3,5- triyltris(methylene)) tris(oxy)) tribenzaldehyde **R**. Reaction conditions: (i) THF, K_2CO_3, 70 °C, 24 h, 81%.

Trialdehdye **R**, which contains three ether moieties, was synthesised in a single step by employing the Williamson ether synthesis, wherein the nucleophilic alkoxide ion displaces an alkyl halide. This reaction was previously reported to form this trialdehyde by Lee *et al.*, where they reported the formation of two-dimensional porous materials.[190] The synthesis of **R** involved the ether synthesis of 1,3,5-tris(bromomethyl)benzene and 4-hydroxybenzaldehyde through a relatively robust reaction, resulting in an overall yield of 81% following purification (Scheme 5.1).

Trialdehyde **U** meanwhile, which was present in many of the larger shape persistent cages described in Chapter 4, was synthesised using a two-step one-pot synthesis (Scheme 5.2). In the first step, the Michaelis-Arbuzov reaction was utilised to form an organophosphate species by the nucleophilic attack of triethyl phosphite on an alkyl halide. The second step involved the use of the Horner-Wadsworth-Emmons reaction to convert the triphosphonate stereoselectively to the *trans* alkene trialdehyde product **U**. This reaction has been used successfully by Garzón *et al.* to produce different molecules with promising optical properties from the 1,3,5-tristyrylbenzene core.[186] The robust nature of this reaction led to a 77% yield following recrystallisation in a 1:1 mixture of chloroform:ethanol.

5.1.3 Experimental Conditions

Following the creation of the precursor library, high-throughput experiments were performed on an Opentrons automated liquid handling platform to identify a single set of reaction conditions that would maximise conversion to cages in as many of the 144 precursor combinations as possible. Key conditions that affect cage conversion are solvent choice, reaction temperature, precursor

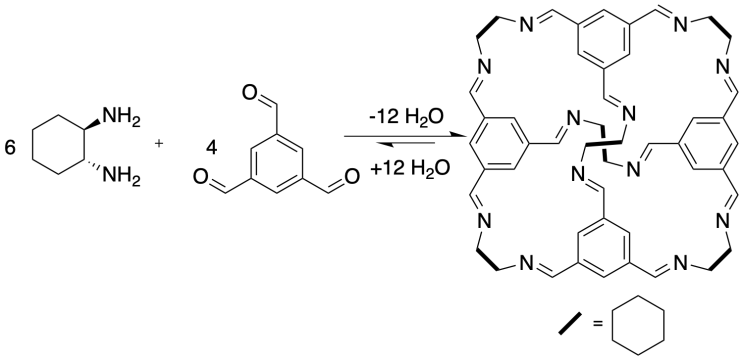


Scheme 5.2: Reaction scheme for the synthesis of 4,4',4" -((1*E*,1'*E*,1"*E*)-benzene - 1,3,5-triyltris(ethene-2,1-diyl)) tribenzaldehyde **U**. Reaction conditions: (i) triethyl phosphite, 150 °C, 12 h, N_2 atm.; (ii) 4-(diethoxymethyl)benzaldehyde, potassium tert-butoxide, THF, 0 °C, 12 h, N_2 atm.; (iii) HCl, RT, 20 min, 77%.

concentrations, and catalyst choice.[191] For the high-throughput screen, an ideal set of conditions would allow all precursors to be soluble, minimise precipitation of kinetic intermediates or cage products, and maximise conversion to soluble cage products. Typically, conditions for cage synthesis are optimised for a single pair of precursors, making the application of these conditions to a set of 144 combinations of precursors challenging.[191] To try to achieve cage formation in the largest number of precursor combinations, a solvent that maximised the solubility of the precursors was chosen and the reaction was performed at a higher dilution than is typically performed for cage synthesis. While reducing the overall rate of reaction, employing a high dilution can reduce the formation of unwanted polymeric products.[192] In addition, using an excess of amine has also been shown to increase cage conversion, minimising oligomer formation by increasing the reversibility of the dynamic equilibrium toward thermodynamic products.[191]

To identify an optimal set of reaction conditions, an initial screen was performed to maximise conversion to a cage prevalent in the literature, **CC3**, formed from the imine condensation of triformylbenzene (TFB) and cyclohexanediamine (CHDA) (Scheme 5.3).[38] The formation of cages from these precursors was assessed to identify the required reaction time, the optimal overall reaction concentration, and the relative amounts of each precursor required for complete conversion to imine cages to occur. To ensure consistency across all 144 precursor combinations, a single set of reaction conditions was used for every precursor combination. The reaction was monitored by ^{1}H NMR spectroscopy to assess consumption of the aldehyde (TFB) starting material and to evaluate conversion to **CC3**. Furthermore, the formation of **CC3** was confirmed using high-resolution mass spectrometry (HRMS), by detecting the mass ion peak corresponding to the expected **Tri**^{4}**Di**^{6} topology.

Initially, the effect of varying the relative ratio of precursors while keeping the overall concentration the same was investigated. Figure 5.2 shows the stacking ^{1}H NMR spectra for precursors



Scheme 5.3: Reaction scheme for the synthesis of **CC3**. This reaction was employed in the preliminary screening process using the Opentrons platform to study the impact of concentration, reaction time, and precursor ratio on the formation of **CC3**.

combined in equivalent ratios of 5:4, 4:4, 4:5, 4:6, 4:7, and 4:8, respectively, using an overall precursor concentration of 0.0092 M of precursors and leaving the reaction for 5 days. Using a larger excess of diamine was shown to lead to a higher conversion of the aldehyde starting material into imine products, as shown by the reduction of the aldehyde peaks in Figure 5.2, but a small amount of starting aldehyde was still present in the final reaction mixture, even after the prolonged reaction time of 5 days. Ideally, complete conversion of the starting aldehyde would have occurred, but the 4:8 equivalents was selected as this led to the highest conversion in this screen.

Next, the effect of varying the overall concentration and reaction time was investigated in a separate experiment. Generally, increasing the concentration from 0.0092 M to 0.0161 M showed little effect on conversion, resulting in the formation of an identical mixture of products as analysed by ^{1}H NMR spectroscopy. The effect of reaction time on the formation of **CC3** is shown in Figure 5.3. Here, it was observed that longer reaction times led to further conversion, with an increase in the intensities of the **CC3** imine and aromatic peaks relative to the other peaks in the aromatic region. Leaving an additional reaction of total concentration 0.0161 M for 3 weeks resulted in full conversion to **CC3**.

Despite the consumption of the aldehyde starting material, a mixture of products was observed by NMR spectroscopy, and not the isolated peaks indicative of the complete formation of **CC3**, each of which is shown in Figure 5.3. These additional peaks could be due to the formation of stable kinetic intermediates formed at high dilution, which have been previously observed in the mechanochemical synthesis of **CC3**,[193] and has been reported as an observable stable intermediate along the reaction pathway of **CC3**.[112] To confirm the formation of **CC3**, HRMS was performed on the reaction mixture at a 0.0161 M concentration after 70 hours. In this reaction, both the mass ions for the [3 + 5] ([M+H]$^{+}$ 895.5885) kinetic intermediate and a [4 + 6] species ([M+H]$^{+}$

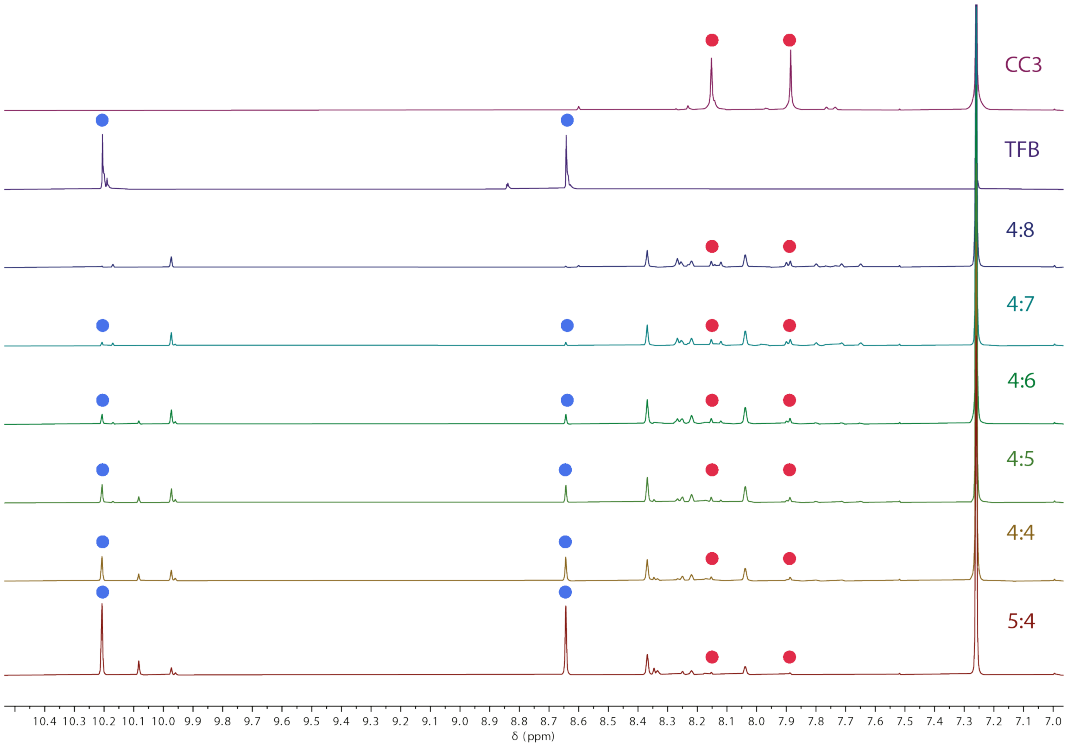


Figure 5.2: Stacked ^{1}H NMR ($CDCl_3$) spectra between 7.0-10.5 ppm of the reaction of trialdehyde **G** with diamine **16** in trialdehyde:diamine equivalent ratios of 5:4, 4:4, 4:5, 4:6, 4:7, and 4:8 using an overall concentration of 0.0092 M for 5 days. ^{1}H NMR ($CDCl_3$) spectra are also shown for **CC3** and TFB, with peaks corresponding to TFB shown below the blue circle and peaks corresponding to **CC3** shown below the red circle. Maximum conversion of the starting aldehyde was observed in the 4:8 ratio, with the largest excess of diamine used.

1117.7052) were present, as shown in Figure 5.4, suggesting that the mixture had not fully equilibrated to the thermodynamic product, although the [4 + 6] species was still able to be detected.

Throughout this chapter, I use the nomenclature [X + Y] to describe the species formed experimentally, where *X* and *Y* are the numbers of trialdehydes and diamines in the product, respectively. For example, a [4 + 6] species would comprise of 4 trialdehydes and 6 diamines, resulting in the loss of 12 water molecules. This notation is used due to the inability of HRMS to differentiate between single discrete cages and catenated cages, which have been observed to form in the literature.[194] In catenated species, two or more cages are interlocked, resulting in a single product comprised of multiple cages. Although **CC3** is well-known in the literature as a cage with a **Tri**^{4}**Di**^{6} topology,[38] HRMS alone would not be able to differentiate between two catenated **Tri**^{2}**Di**^{3} cages or a single **Tri**^{4}**Di**^{6} cage, which would both possess identical molecular masses.

From this initial screen, the lower overall concentration was selected to avoid the formation of insoluble kinetic intermediates during the high-throughput screen (0.0092 M). To maximise

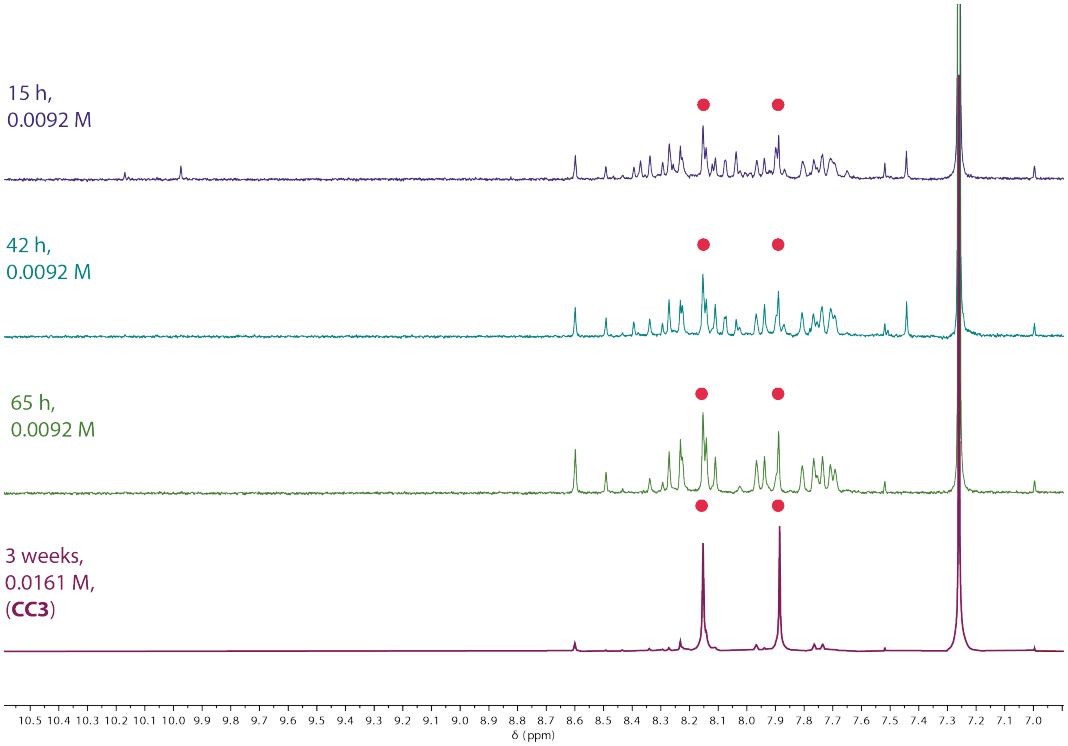


Figure 5.3: Stacked ^{1}H NMR ($CDCl_3$) spectra between 7.0-10.5 ppm of the reaction of trialdehyde **G** with diamine **16** at 15 h, 42 h, 65h, and 3 weeks. Complete conversion to **CC3** was observed after 3 weeks at an overall precursor concentration of 0.0161 M. Peaks corresponding to **CC3** are shown below the red circles.

conversion to imine products, the largest excess of diamine (4:8 trialdehyde:diamine equivalents) was used. After 70 hours of reaction time, a mixture of products was observed by ^{1}H NMR spectroscopy, although HRMS indicated that some formation of the desired thermodynamic product had occurred. Despite observing full conversion to **CC3** after 3 weeks, 5 days was chosen as a compromise to increase the reaction throughput. By selecting 5 days reaction time, only partial conversion to conversion to **CC3** could have occurred, however, it is likely that some of the desired thermodynamic product would be detectable by both ^{1}H NMR spectroscopy and HRMS.

5.2 Computational Methods

5.2.1 Automated Reaction Analysis

Due to the vast number of precursor combinations screened (144 combinations in this thesis, and a larger set totalling 366 different precursor combinations), I developed an automated Python workflow for detecting and analysing the reaction outcome by automating the analysis of ^{1}H

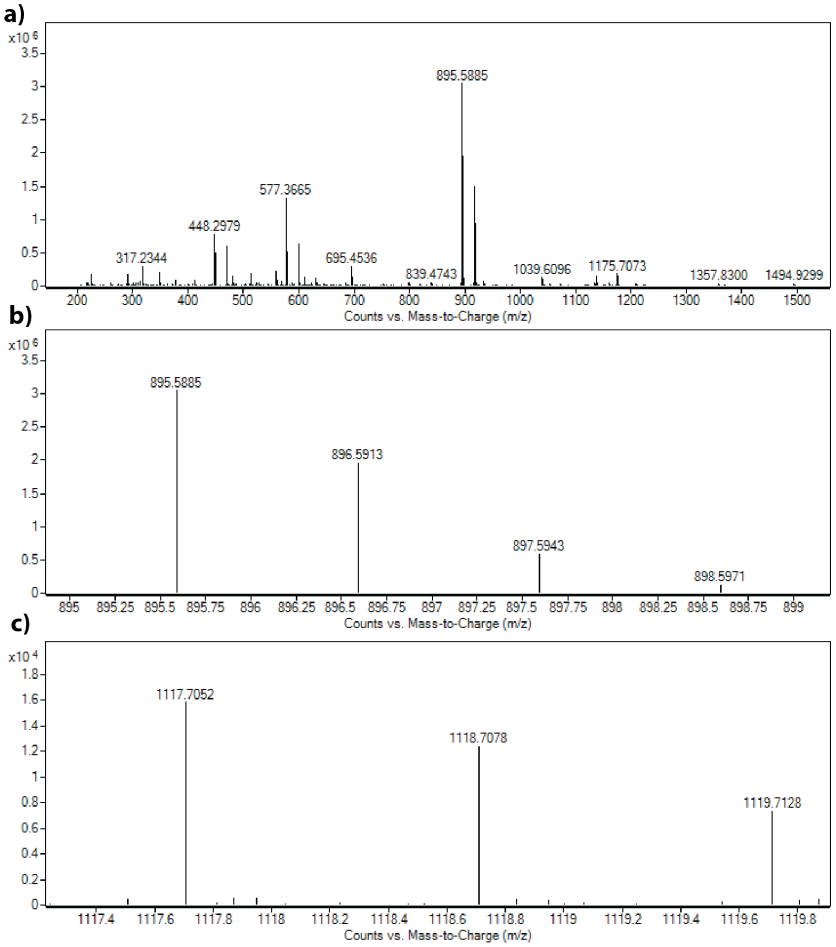


Figure 5.4: The HRMS (a) for the reaction of trialdehyde **G** with diamine **16** using an overall concentration of 0.0161 M after 70 h reaction time, suggesting the formation of a [3 + 5] species ($[M+H]^+$ 895.5885) (b) and a [4 + 6] species ($[M+H]^+$ 1117.7052) (c).

NMR spectra and detecting relevant mass ion peaks in the HRMS. Manual analysis of all NMR and HRMS data would be a time-consuming and error-prone task, which may be influenced by human bias in the identification of relevant mass ion peaks. By developing an automated workflow, most of the human biases in evaluating the reaction outcome could be eliminated and automated analysis could be expanded to include multiple different topologies.

5.2.2 ^{1}H NMR Spectroscopy

In the high-throughput screening workflow, ^{1}H NMR spectroscopy was used to differentiate between complete and incomplete cage formation using an automated analysis workflow. Automation of NMR structure elucidation is becoming increasingly popular, with new workflows such

as DP5 aimed at automating the processing and assignment of ^{13}C and ^{1}H NMR spectra.[195] However, reliable assignments require accurately predicted chemical shifts for a molecule, which can be achieved using either the DFT gauge invariant atomic orbitals (GIAO) method or ML chemical shift prediction.[196,197] In reactions where a mixture of products is likely to form, each with potentially overlapping peaks, it would be extremely challenging to achieve structural elucidation of individual cage molecules through NMR analysis using these workflows. Instead, a simpler NMR analysis was employed to detect the presence of an aldehyde peak, indicating that only partial conversion to imine products had occurred. In addition, the presence of an imine peak was checked to confirm that the reaction had not resulted in complete conversion to insoluble products.

The raw ^{1}H NMR spectra were subjected to an initial standardised pre-processing procedure, which involved baseline correction, phase correction, and Fourier transformation in the proprietary Bruker TopSpin software. Subsequently, the pre-processed spectrum was analysed using the *nmrglue* package within an automated Python workflow.[198] *nmrglue* is an open-source package for processing, manipulating, and analysing NMR spectra within Python, and provides several utility functions for reading Bruker NMR files, peak picking, and integrating NMR data.

Using this software, aldehyde peaks were identified by performing peak picking in the chemical shift region of 9-11 ppm at a threshold of 1×10^4 a.u. using the peak picking function within the *nmrglue* package, using the 'connected' algorithm. The 'connected' algorithm identifies all nodes (peaks) that are above a certain threshold, iteratively determining whether each of those nodes are separated from one another, defining each separate node as an individual peak. Imine peaks and other peaks within the aromatic region were detected within the range of 6.5-9 ppm, using the same peak picking algorithm and threshold. The range of 9-11 ppm was selected for detecting the aldehyde peak to accommodate the additional deshielding effects of the aromatic ring in close proximity to the aldehyde proton, present in many of the trialdehyde precursors. The range of 6.5-9 ppm was selected for the imine peaks due to their known occurrence across a wide range of shifts influenced by external factors that affect the proton environment, such as hydrogen bonding. To avoid detecting false positive peaks from the NMR solvent chloroform, the peak corresponding to this solvent at 7.26 ppm was filtered after peak picking.

The full workflow for peak detection is illustrated in Figure 5.5 using **CC3** from the initial condition screen, in addition to combinations **U14** and **U34** from the high-throughput screen. In this example, the presence of peaks in the imine region of the ^{1}H NMR spectrum of **CC3** indicated that TFB starting material had been fully consumed, whereas aldehyde starting material was present in **U14**. Both **CC3** and **U14** also possessed peaks in the imine region, suggesting the reaction had not resulted in full conversion to an insoluble precipitate. Meanwhile, **U34** did not possess any peaks in the aldehyde or imine region, suggesting that complete conversion to insoluble products had occurred.

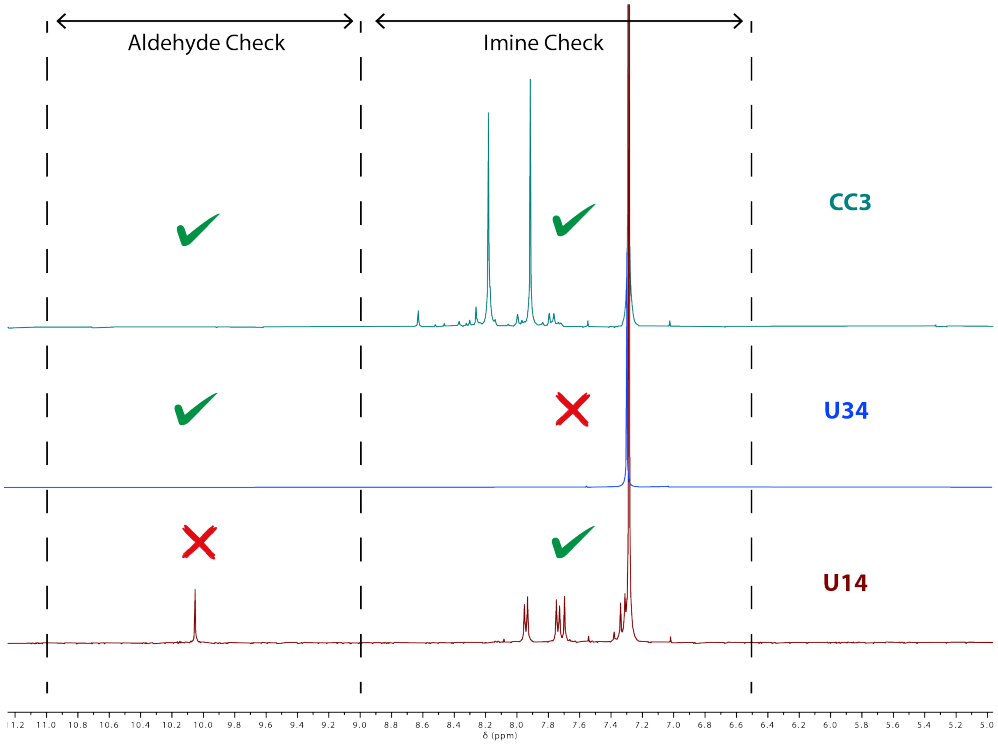


Figure 5.5: Stacked ^{1}H NMR ($CDCl_3$) spectra of **CC3** from the condition screening, the reaction of trialdehyde **U** with diamine **14**, and the reaction of trialdehyde **U** with diamine **34** from the high-throughput screen, demonstrating the automated NMR analysis workflow. The presence of peaks in the imine region (6.5-9 ppm) and the absence of peaks in the aldehyde region (9-11 ppm) indicated that complete conversion to **CC3** had occurred. On the other hand, **U14** showed an aldehyde peak, indicating unreacted trialdehyde starting material, but also had peaks in the imine region, suggesting either partial conversion or no reaction. Meanwhile, **U34** possessed no peaks in both the imine and aldehyde regions, suggesting conversion to insoluble products had occurred. Peaks due to chloroform were excluded from the peak detection process.

5.2.3 HRMS

Next, a high-throughput HRMS workflow was used to detect the presence of specific cage topologies in the reaction mixtures, in addition to the kinetic [3 + 5] species that were observed in the initial condition screen. While the automated HRMS analysis could be extended to detect other kinetic intermediates, my focus was solely on the [3 + 5] species due to its observed formation in the test reactions involving **CC3**. The presence of this particular kinetic intermediate could serve as an indicator that modifying the experimental conditions, such as prolonging the reaction time, might enable the conversion to a larger [4 + 6] species for specific combinations of precursors. The combination of NMR and HRMS analysis provided complementary information, which

was used together to assess the amount of conversion to imine products and the specific imine products that formed.

The proprietary Agilent data format was initially converted to the open-source mzML format using the *MSConvert* tool of the ProteoWizard software suite.[199] The *mzML* is a standardised open data format designed to store raw unprocessed data generated by MS instruments to be read by open-source MS software for further processing. The initial steps of the HRMS workflow involved extracting mass spectra from the total ion count chromatogram. To achieve this, the open-source *mzMine* package was used to automate the extraction of m/z values into a computer-processable format.[200] This process involved the following steps:

1. Mass detection was carried out with the 'centroid' method, which assumes each signal above a certain noise level is a detected ion. This requires centroid data, where each m/z peak is a discrete value with a line width of zero. A noise threshold of 1×10^3 a.u. was used for detecting mass peaks initially.

2. The extracted ion chromatograms for each mass ion peak were constructed using the Automated Data Analysis Pipeline (ADAP).[201] Broadly, the goal of ADAP is to calculate the intensity of each m/z peak over time. The parameters used in this workflow included a minimum group size of 3, which is the number of consecutive scans where the intensity of each m/z peak in the group must be above a pre-defined intensity threshold of 3×10^3 a.u.. The minimum highest intensity of an m/z peak in a group was 1×10^3 a.u., and each m/z ion peak was extracted based on a tolerance of 5 ppm within the scans. Finally, the m/z peaks in each extracted ion chromatogram were written to a text file for comparison with the expected m/z peaks for each cage topology.

3. The extracted m/z peaks were compared with the expected m/z peak for each topology and precursor combination using a Python workflow. The expected m/z of [2 + 3], [3 + 5], [4 + 6], [6 + 9], and [8 + 12] reaction stoichiometries were calculated for each combination, considering the loss of 6, 9, 12, 18, and 24 water molecules, respectively. The m/z peaks corresponding to singly and doubly charged species were also considered, along with the adduct H, 2H, or Na, which form during the interaction with the ionisation source. The *pyOpenMS* library was used to calculate each expected m/z peak,[202] and each peak was compared with the extracted m/z peaks within a tolerance of 5 ppm to detect the presence of each topology.

Differentiating Between Topologies Detecting the topology of the [2 + 3] and [4 + 6] species in mass spectrometry can be challenging, since the singly charged m/z values of the [2 +

3] species will be identical to doubly charged m/z values of the [4 + 6] species. Thus, the identification of the experimentally formed topology requires analysis of the spacing between peaks in the isotopic distribution of the compound, which can complicate the process of identifying mixtures of these species. To overcome this challenge, I calculated the separation between each m/z peak corresponding to a specific topology and the next adjacent peak using Python. A peak separation of 0.5 m/z indicates the presence of a doubly charged species, whereas a separation of 1 m/z indicates a singly charged species.

Similarly, the doubly charged [8 + 12] species would also correspond to the singly charged [4 + 6] species, so a peak separation of 0.5 m/z was required to detect this topology. In the case of mixtures of [8 + 12] and [4 + 6] species that form, detection of the doubly charged [4 + 6] m/z ion was also required to confirm the presence of the [4 + 6] species. A schematic for the detection workflow of different topologies is shown in Figure 5.6, using cage **I15** an example case, which was observed to form a mixture of [4 + 6], [6 + 9], and [8 + 12] species using this workflow. Using this automated procedure, the presence of different topologies can be detected. However, the analysis techniques used are unable to differentiate between the formation of catenated **Tri**^{4}**Di**^{6} topologies and **Tri**^{8}**Di**^{12} topologies, which share the same [8 + 12] reaction stoichiometry, as previously mentioned in the initial condition screening section. Although cage **I15** is known to form a **Tri**^{8}**Di**^{12} topology and not a catenane in the literature,[189] further characterisation techniques, such as diffusion NMR, would be required to differentiate between species that formed **Tri**^{8}**Di**^{12} topologies and catenated species.

Turbidity Analysis Instead of reacting to form discrete molecular cages in solution, precursors can often react to form insoluble polymers or variable-length oligomers.[28] To observe the formation of insoluble species, a manual visual turbidity inspection was performed for each precursor combination to assess whether a solid precipitate was formed in the vial. Typically, POCs are soluble in chloroform, the solvent selected for the high-throughput screen. However, since a diverse range of precursor combinations were investigated in this study, some combinations may have resulted in insoluble kinetic intermediates or even insoluble cages. An ideal POC would be soluble, as solution processability is a requirement for POCs to be processed into membrane materials, and so insoluble species would be undesirable in this context. In my thesis, characterisation of the insoluble species was not pursued, and reactions resulting in the formation of a solid precipitate were categorised independently using the outlined criteria in the next subsection. Indeed, the low concentration chosen for the high-throughput screen was chosen to minimise the formation of insoluble species.

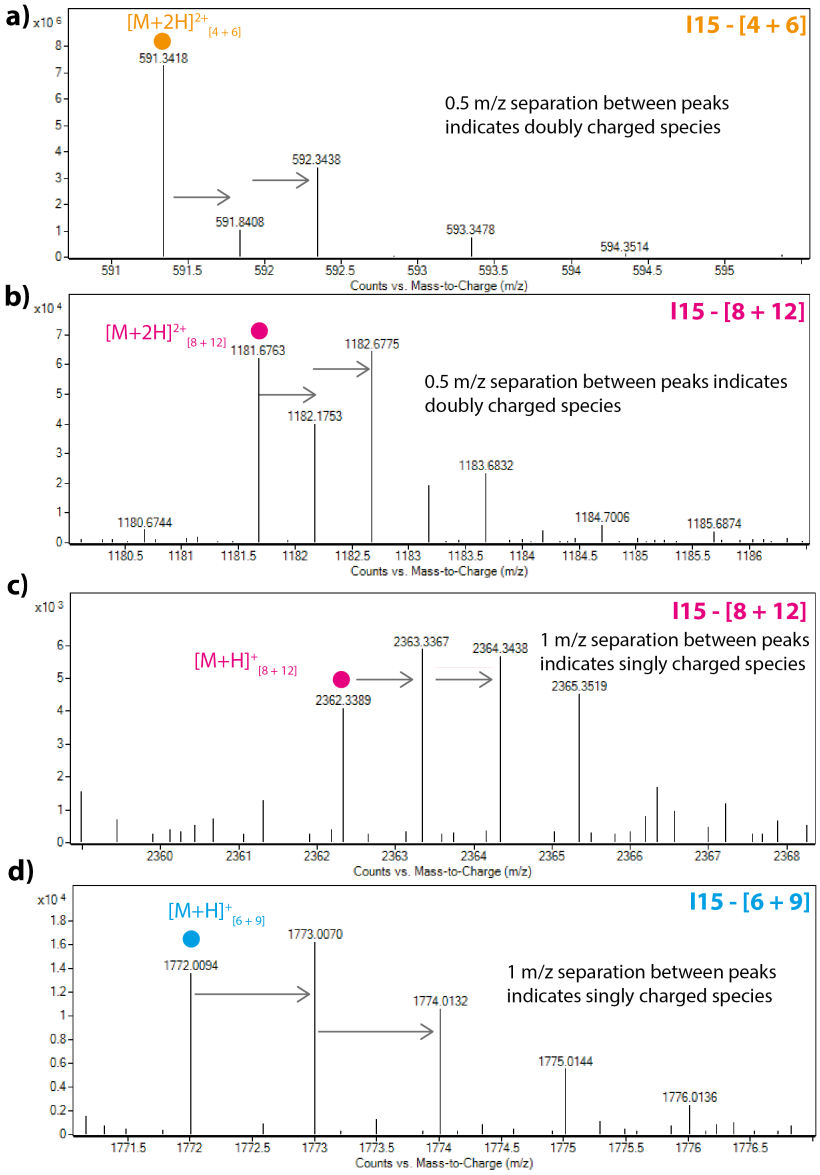


Figure 5.6: Schematic of the HRMS detection workflow for cage **I15**. In this example, **I15** was detected to be a mixture of [4 + 6], [6 + 9], and [8 + 12] species. Both the [4 + 6] (a) and [8 + 12] (b) species were detected by the peak separation of 0.5 present between adjacent peaks, and additional confirmation was obtained by the presence of a singly charged [8 + 12] species (c). In addition, the m/z corresponding to [6 + 9] a species was also detected (d), however, no [2 + 3] species was detected due to the absence of a doubly charged peak corresponding to this species. This analysis procedure was automated in Python for all 144 precursor combinations screened. Dots correspond to mass ion peaks of the [4 + 6] (orange), [6 + 9] (blue), and [8 + 12] (pink) species. Calculated $[M+H]^+_{[4+6]}$ for $C_{72}H_{84}N_{12}O_4$ is 1181.6816; $[M+H]^+_{[6+9]}$ for $C_{108}H_{126}N_{18}O_6$ is 1772.0133; and $[M+H]^+_{[8+12]}$ for $C_{144}H_{168}N_{24}O_8$ is 2362.3555.

5.2.4 Categorisation of High-Throughput Results

Following the automated HRMS, NMR, and manual turbidity analysis, each combination of trialdehyde and diamine precursor was categorised according to the criteria outlined in Table 5.1. In this subsection, I will outline how the outcome of each analysis technique was used to categorise each precursor combination, indicating which categories correspond to clean formation, partial conversion, or no cage formation at all. For materials discovery, the ideal outcome would be the clean formation of a single cage topology.

Category	Category Number	HRMS Check Passed?	NMR Check Passed?	Turbidity Check Passed?
Complete formation of a single cage	1	✓	✓	✓
Complete formation of a single topology, alongside insoluble precipitate	2	✓	✓	✗
Complete formation of a mixture of cages	3	✓	✓	✓
Complete formation of a mixture of cages, alongside insoluble precipitate	4	✓	✓	✗
Incomplete formation of a single cage	5	✓	✗	✓
Incomplete formation of a single cage, alongside insoluble precipitate	6	✓	✗	✗
Incomplete formation of a mixture of cages	7	✓	✗	✓
Incomplete formation of a mixture of cages, alongside insoluble precipitate	8	✓	✗	✗
Reaction observed, but no cage formation	9	✗	✗	✗
No reaction observed	10	✗	✗	✓

Table 5.1: Categorisation criteria of high-throughput screening results. The HRMS check is passed if the mass ion peak corresponding to a [2 + 3], [4 + 6], [6 + 9], or [8 + 12] species is present in the HRMS, indicating formation of a specific topology. The NMR check is passed if there is no aldehyde peak and there is an imine peak in the NMR spectra, indicating complete conversion to a single or mixture of imine products. The turbidity check is passed if the solution is clear, as determined by visually inspecting for the presence of no insoluble polymer or oligomer precipitate.

Each precursor combination underwent three checks, corresponding to the three analysis techniques outlined in the previous subsection. The first check was HRMS analysis, where a precursor combination was considered successful if a mass ion peak corresponding to a [2 + 3], [4 + 6], [6 + 9], or [8 + 12] species was detected in the spectra. If more than one of these species were detected, the precursor combination was classified as a mixture. The second check was NMR analysis, where the presence of an imine peak and the absence of an aldehyde peak indicated complete

conversion to imine products. The third and final check was turbidity analysis, where the absence of a solid precipitate in the vial indicated that there was no conversion to insoluble species.

The first category (category 1), 'complete formation of a single cage' is the ideal outcome. It is assigned only if a single species is observed in the HRMS, and both the NMR and the turbidity analysis indicate complete conversion to a single species. From a materials discovery perspective, this category is the most favourable, requiring the least amount of post-processing to isolate the desired cage. Category 2 is assigned to a combination if a single species is in the HRMS, but an insoluble precipitate is detected, failing the turbidity check. The formation of an insoluble precipitate would necessitate an additional filtration step to isolate the desired cage.

Next, category 3 is assigned if HRMS indicates the formation of two or more cage species, passing both the NMR and the turbidity checks. The formation of a mixture of species is less desirable and would require additional purification steps to isolate the desired cage, such as high-performance liquid chromatography (HPLC). Moreover, this could mean that the reaction conditions are not optimal for this precursor combination, and additional reaction time or increased temperature may be required to maximise conversion to the single most thermodynamically stable cage topology. Although I explored a single set of reaction conditions on the automated platform, if one of the topologies that formed was predicted to be shape persistent, it could be worth exploring different reaction conditions to facilitate complete conversion to this topology.

Category 4 would present the same challenges as category 3, but with the added complication of an insoluble precipitate, also requiring additional purification steps. Categories 5-8 are similar to 1-4, but with the addition of an incomplete conversion to an imine product, as determined by the NMR check failing due to the presence of an aldehyde peak in the NMR spectrum. Incomplete reaction conversion is undesirable and could indicate the formation of kinetic intermediates with unreacted aldehyde functionalities. Similarly to 3, this could mean that the reaction conditions are not optimal for this precursor combination and could require additional optimisation to facilitate complete conversion. Category 9 is assigned if no cage species are detected by HRMS, but NMR or turbidity indicate that a reaction to some imine products has occurred. Finally, category 10 is assigned if no reaction is observed by NMR and no solid precipitate is formed, indicating the two precursors did not undergo imine condensation when combined. Categories 9 and 10 represent the least desirable outcomes and imply that the tested reaction conditions were unsuitable for cage formation between the two precursors. However, it should be noted that altering the reaction conditions, such as temperature, concentration, and reaction time, may facilitate cage formation between the precursors that were not observed to react in my thesis.

5.2.5 Computational Prediction Workflow

Initially, the computational workflow for the construction and geometry optimisation of cages was identical to that outlined in the previous chapter. For each of the 144 precursor combinations, cages were built *in silico* and subjected to a three-step optimisation process. The four topologies (**Tri**^{2}**Di**^{3}, **Tri**^{4}**Di**^{6}, **Tri**^{6}**Di**^{9}, and **Tri**^{8}**Di**^{12}) were assembled using the *stk* Python package,[56] which generated initial structures based on the SMILES strings and stereochemistry of each diamine and trialdehyde precursor, specifying each of the four desired topologies. A constrained FF geometry optimisation was then performed on each of the 576 cages, fixing the precursor geometries and optimising only the bonds created by *stk*. Subsequently, the entire cage structure underwent an unconstrained FF optimisation, optimising the full structure. Finally, a 100 ns MD simulation was performed in the NVE ensemble, following a 100 ps equilibration period, with a time step of 1 fs and a temperature of 700 K. A longer simulation time was used to achieve better sampling of the conformational space, given the smaller number of cages computationally screened compared to previous work.[176] During the simulation, 50 conformers were sampled, each undergoing further FF geometry optimisation to provide a reasonable low energy conformation for each cage. At each stage, the OPLS3 force field was employed,[136] using the Polak-Ribiere Conjugate Gradient minimisation algorithm. The minimisation algorithm was set to converge with a criterion of 0.05 kJ^{-1}A^{-1} and a maximum of 2,500 iterations. The MD and FF optimisation algorithms were implemented using Schrödinger's MacroModel software.[139]

Following the identification of low energy conformers, DFT calculations were employed for further optimisation, and to provide more accurate energetics of each cage system. The lowest energy conformer extracted from the molecular dynamics simulations underwent a DFT optimisation with the PBE functional, employing the triple-zeta basis set (TZVP), and the D3 dispersion correction, using a plane-wave cutoff of 350 Ry. DFT optimisations were performed using the CP2K software package (version 8.1).[151] For each of the optimised cages, the formation energy was calculated using Equation 5.1, where E_{cage} is the total energy of the cage, E_{water} is the total energy of a water molecule, E_{aldehyde} is the total energy of the trialdehyde precursor, and E_{amine} is the total energy of the diamine precursor each used to form the cage. Additionally, m is the number of aldehyde precursors and n is the number of diamine precursors each present in a specific topology, and x is the number of imine bonds present in a specific topology. Using this equation, it is possible to compare between cages formed from different precursors, providing an indication of the thermodynamic preference for each cage topology.

$$E_{\text{formation}} = \frac{\left(E_{\text{cage}} + xE_{\text{water}}\right) - \left(mE_{\text{aldehyde}} + nE_{\text{amine}}\right)}{x} \tag{5.1}$$

The internal cavity size of each cage optimised using DFT was analysed using *pyWindow* to determine whether each cage would be shape persistent or collapsed.[51] Specifically, cages with a calculated cavity diameter in excess of 2.89 Å, corresponding to the kinetic diameter of a hydrogen molecule,[203] would possess a cavity size large enough to host a small guest molecule and could be considered intrinsically porous.

Cage Asymmetry Due to the asymmetry in trialdehydes **I** and **S**, and diamine **33**, it is possible to form a large number of structural isomers experimentally, which would complicate both the computational investigation and experimental characterisation of potential cages formed. Identifying the most stable isomer *a priori* would require enumerating over all structurally unique isomers, computing relative energy values for each, which becomes computationally infeasible when dealing with a large number of unique structural isomers. For instance, when trialdehyde **S** combines with diamine **16**, it could result in 9 unique structural isomers in the **Tri**^{4}**Di**^{6} topology.[63] For larger topologies, the number of possible isomers would be even greater.

Due to the computational challenge of modelling all possible isomers, only a single isomer was generated for each topology using *stk*, which then underwent the MD conformer search and DFT optimisation. However, it is important to note that this single isomer may not form experimentally, and therefore the calculated formation energies for these asymmetric systems (41 potential cages) may not accurately reflect the true experimental structure. The energy differences between isomers can vary significantly. For example, in the case of combination **S16**, energy differences exceeding 60 kJ mol^{-1} have been observed between isomers with the **Tri**^{4}**Di**^{6} topology,[63] whereas for **I16**, the largest energy difference among the 9 possible isomers with the **Tri**^{4}**Di**^{6} topology was 2 kJ mol^{-1}.[188] Nonetheless, computationally modelling a single structural isomer can offer insights into whether a cage is likely to remain shape persistent, and can rationalise the thermodynamic preferences of different topologies when the energetic preference between different isomers is small. For cages that do form from these asymmetric precursors, future work could involve analysis of the potential isomers that could form, to better understand the thermodynamic driving force that results in the formation of specific isomers.

5.3 Results and Discussion

In this section, I will begin by outlining the results of the high-throughput experimental screen that involved the combination of 9 trialdehydes with 16 diamines. Each experimental outcome was categorised into one of the 10 categories listed in Table 5.1, using the automated reaction analysis workflow described earlier. Initially, I will provide a general discussion of the experimental results for each precursor combination, providing context for each outcome within its respective

category. Whenever a cage was observed experimentally and was predicted to be shape persistent using computation, I will use relative energies to rationalise the preference for a particular topology over others or the occurrence of multiple topologies in a mixture. Finally, I will summarise the general trends identified across the precursor library, outlining the features of the most promising precursors for POC formation, and assess the trends in formation energies across precursors to determine whether cage formation can be predicted solely on thermodynamics.

5.3.1 High-Throughput Screening Results

Figure 5.7 shows the categorisation of each of the 144 precursor combinations, and a further visual representation of the breakdowns of each category is shown in Figure 5.8. In total, 68 (47%) of precursor combinations resulted in the formation of at least one cage, and 76 combinations were not observed to form a cage of any topology. Out of the 68 precursor combinations that resulted in cage formation, 44 resulted in complete conversion to a single topology or a mixture of topologies (categories 1-4), and 24 precursor combinations underwent incomplete conversion to a single topology or a mixture of topologies. The complete conversion to a single topology was observed in 27 precursor combinations (categories 1 and 2) and a mixture of topologies in 16 cases (categories 3 and 4). Currently, the number of reported cages is on the order of low hundreds,[48] making the number of cages discovered in this work a significant number.

As shown in Figure 5.8, of the 26 topology mixtures formed, a combination of [2 + 3] and [4 + 6] species was the most prevalent, with 21 combinations resulting in this mixture. In addition, precipitate was observed in 81 precursor combinations, appearing mainly along single topologies (15 precursor combinations) compared with mixtures (10 precursor combinations). Precipitate was observed to form mainly in solutions where complete conversion had occurred. Out of the cages in which clean cage formation was observed, 16 of them resulted in the formation of an insoluble precipitate. In contrast, among the cages where partial conversion was observed, 9 also yielded an insoluble precipitate. Alongside the mixtures of two topologies, 4 cages were observed to form a larger [6 + 9] species, 2 of which were present alongside a species with an [8 + 12] stoichiometry.

5.3.2 Clean Cage Hits

A clean cage hit is the ideal outcome; for a clean reaction outcome, NMR analysis must indicate that complete conversion to imine products has occurred (no residual aldehyde peak), a mass ion peak corresponding to a single topology must be present, and no insoluble precipitate must be formed. In total, 17 (12%) of all precursor combinations fulfilled these criteria, 2 forming [4 + 6] cages predicted to be shape persistent, and 15 forming smaller [2 + 3] cages with a capsule-like

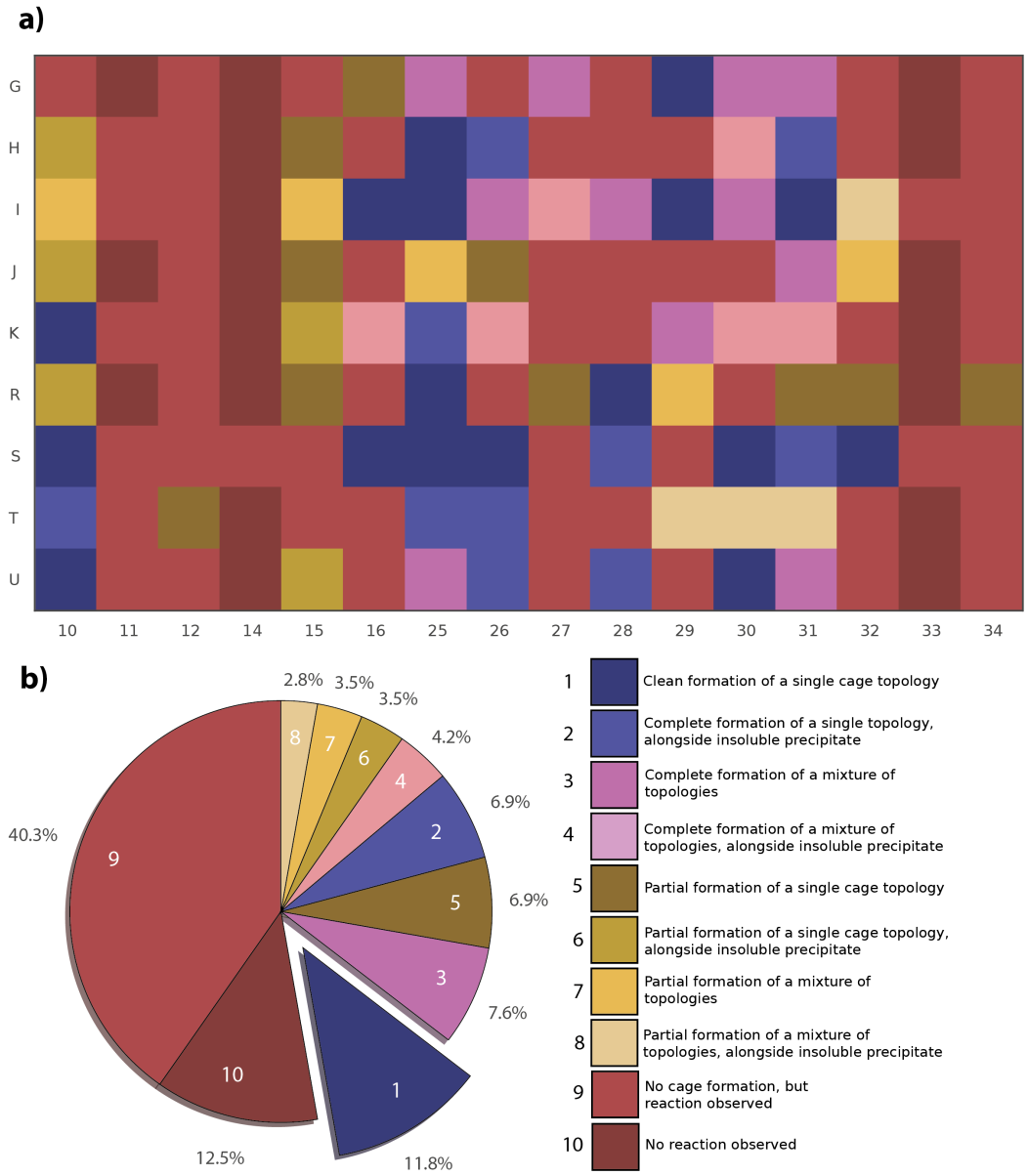


Figure 5.7: Colour map indicating the category assigned to each precursor combination in the high-throughput experimental screen (a), and a pie chart showing the percentage of precursor combinations belonging to each category (b). Assigned categories correspond to those outlined in Table 5.1, and are described in the legend. Letters on the left side of the colour map (a) correspond to trialdehyde precursors, and numbers on the bottom correspond to diamine precursors. Numbers in the pie chart correspond to each category number in Table 5.1.

topology. Figure 5.9 shows the DFT optimised structures of each of the cages that formed cleanly in their respective topologies. Where a cage forms experimentally that is predicted to remain shape

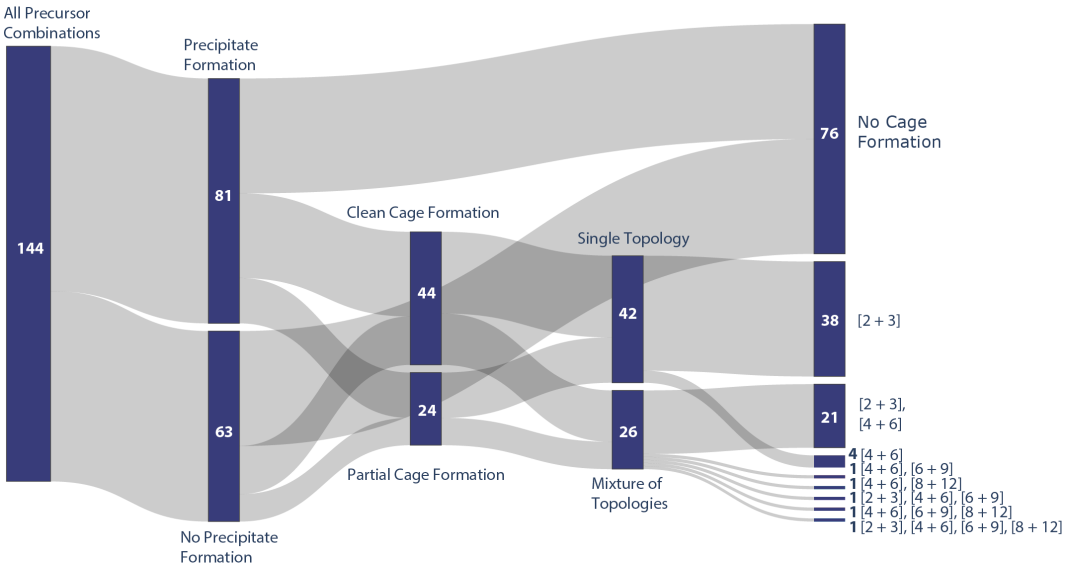


Figure 5.8: Sankey diagram illustrating the results of the high-throughput screen, with the thickness of each connecting line proportional to the number of precursor combinations leading to each outcome. The numbers inside or next to each rectangle represent the number of precursor combinations assigned to that category. When multiple topologies are listed next to a rectangle, it indicates the formation of a mixture of those topologies.

persistent with a permanent internal cavity sufficient to host a small guest molecule, I will use relative energies of each topology to rationalise the experimentally obtained product. Based on the kinetic radius of a hydrogen molecule as a small guest molecule,[203] pore sizes of over 2.89 Å would be considered porous.

Of the cages that were cleanly formed, 10 were formed from asymmetric trialdehyde precursors (6 from trialdehyde **S** and 4 from trialdehdye **I**). Both **S16** and **I16** were observed to form [4 + 6] species, each containing the CHDA used for the synthesis of **CC3**.[38] These two cages have been previously reported in the literature to form **Tri**^{4}**Di**^{6} cages with permanent internal cavities.[63,188] For these hits, assessing the reaction outcome by NMR alone proved difficult due to the complexity introduced by the formation of either a mixture of cage isomers or a single unsymmetric cage isomer. The protons of each distinct isomer could be chemically nonequivalent, resulting in a large number of peaks in the ^{1}H NMR spectrum. To confirm the clean formation of the [4 + 6] species for **S16** and **I16**, a single mass ion peak corresponding to the [4 + 6] was detected by HRMS, and the NMR spectra obtained were identical to that of the crude spectra in the literature.[63,188]

For **S16**, a single highly unsymmetric cage has previously been shown to form using diffusion NMR,[63] while **I16** formed a mixture of different isomers, with different positions of the hydroxy group around each aromatic ring.[188] A comparison of the complex NMR spectra obtained from

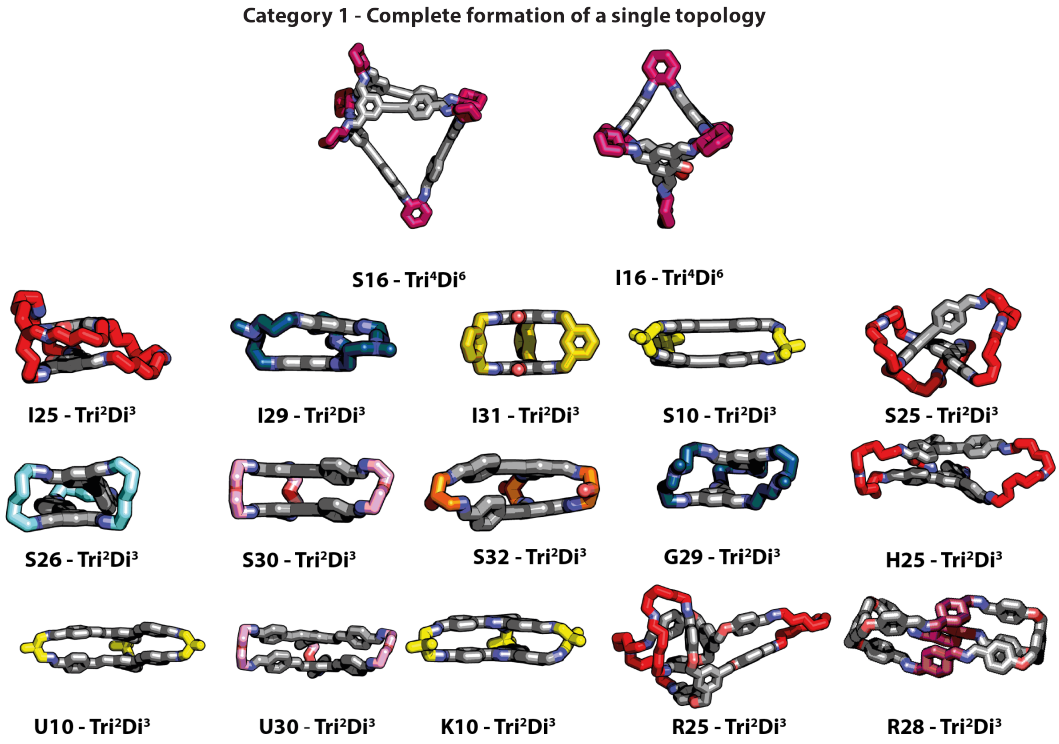


Figure 5.9: Cages optimised at the PBE/TZVP+D3 level which formed a single topology cleanly, corresponding to category 1. Only **S16** and **I16** were predicted to form shape persistent cages. Carbon atoms corresponding to diamine **16** are shown in bright pink, **10** in light yellow, **32** in orange, **30** in pink, **31** in dark yellow, **28** in dark red, **26** in in light blue, **29** in dark blue, and **25** in red. Other carbon atoms are shown in grey, nitrogens in blue and carbons in green.

these cages compared with the highly symmetric **CC3** is shown in Figure 5.10, highlighting how the complexity introduced by cage asymmetry can complicate the identification of single cage species or mixtures. Complex NMR spectra were also observed for the other eight cages that formed [2 + 3] topologies cleanly from asymmetric precursors, namely **I25**, **I29**, **I31**, **S10**, **S25**, **S26**, **S30**, and **S32**.

The other combinations that formed a single topology cleanly were **G29**, **H25**, **K10**, **R25**, and **R28**, **U10**, and **U30**, all forming [2 + 3] species with small internal cavities, each of which is shown in Figure 5.9. Despite only observing a single [2 + 3] species by HRMS, only a clean NMR was observed for **G29** and **K10**, with the other combinations forming complex NMR spectra indicative of multiple imine products. The symmetry of trialdehdyes **H**, **R**, and **U** should result in a clean NMR spectrum, suggesting other imine products could be present in solution amongst the single [2 + 3] topology that was observed by HRMS, although no residual aldehyde group was present.

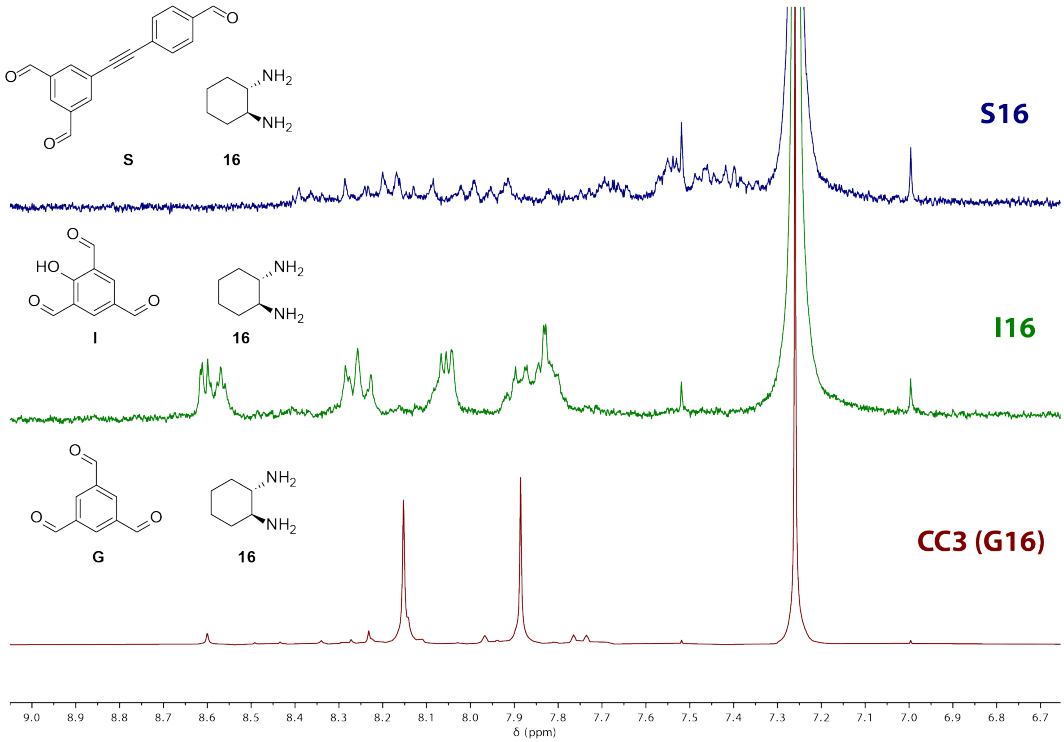


Figure 5.10: Stacked crude ^{1}H NMR ($CDCl_3$) spectra between the region of 6.7-9.2 ppm of precursor combinations **I16**, **S16**, and **CC3** (taken from the experimental condition screening). Asymmetric cages were formed from both **I16** and **S16** precursors, which resulted in more complex NMR spectra in the imine region compared to the highly symmetrical **CC3**. Trialdehyde and diamine precursors corresponding to each combination are shown next to each spectrum.

To rationalise the topology of the two shape persistent cages from this section, **S16** and **I16**, computation was used to compare relative energies per formula unit of each topological outcome, shown in Figures 5.11a and 5.11b. For **S16**, both the **Tri**^{6}**Di**^{9} and **Tri**^{8}**Di**^{12} topologies were found to collapse following the conformer search procedure. Solvent has been shown to provide a scaffolding effect on cage molecules, artificially inflating the cavity by providing a stabilising interaction on open cage conformations.[61] The stabilising interaction of solvent molecules is unable to be accounted for when performing gas-phase MD on the single discrete molecular structure. Of the two open conformations, the **Tri**^{4}**Di**^{6} was found to be 36 kJ mol^{-1} more stable than the **Tri**^{2}**Di**^{3}, and the [4 + 6] species was experimentally observed. For **I16** meanwhile, each of **Tri**^{2}**Di**^{3}, **Tri**^{4}**Di**^{6}, **Tri**^{6}**Di**^{9}, and **Tri**^{8}**Di**^{12} topologies were predicted to form open conformations. The most stable topology, **Tri**^{4}**Di**^{6}, was experimentally formed, and was predicted to be 17 kJ mol^{-1} more stable than the next most stable **Tri**^{6}**Di**^{9} topology.

For both **S16** and **I16**, which can form a range of structural isomers in the **Tri**^{4}**Di**^{6} topology, accurately identifying the most thermodynamically stable topology would first rely on identifying

the isomer (or isomers) that would form experimentally for each topology. Due to the computational expense associated with modelling all possible isomers, I only analysed a single isomer for each precursor combination in each topology. Therefore, it is important to note that the isomer obtained in the simulation may not necessarily correspond to the isomer formed experimentally. For **I16**, only a small energetic preference (2 kJ mol^{-1}) has been observed between isomers, which could explain the complex NMR in Figure 5.10, due to the formation of a mixture of isomers.[188] For **S16**, the energy difference between isomers has been shown to be significant (greater than 60 kJ mol^{-1}), and the isomer I modelled may not be the single isomer that has been shown to experimentally form in the literature.[63]

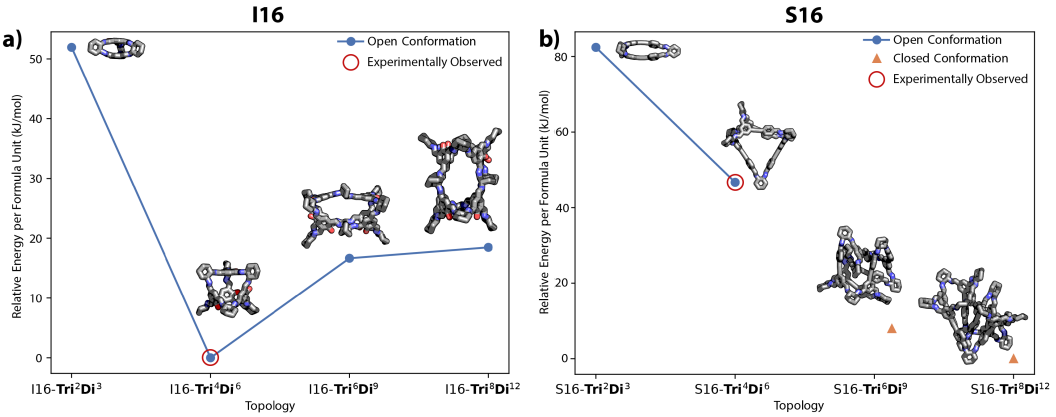


Figure 5.11: Relative energy per formula units of different topologies of combinations **I16** (a) and **S16** (b), each calculated at the PBE/TZVP+D3 level. For **S16**, the **Tri**^{6}**Di**^{9} and **Tri**^{8}**Di**^{12} topologies were predicted to be collapsed, denoted by the orange triangle. Meanwhile, all topologies were predicted to be shape persistent for **I16**. Both precursor combinations formed a single [4 + 6] species cleanly by the presence of a single mass ion corresponding to the [4 + 6] species by HRMS, shown by the red circle.

5.3.3 Formation of a Single Cage Topologies

In the following section, I will discuss the cages that were observed to form a single soluble topology in solution, but were less desirable outcomes due to the formation of an insoluble precipitate (category 2), partial conversion to a single topology (category 5), or both partial conversion and presence of an insoluble precipitate (category 6). In total, 25 combinations were assigned to these 3 categories. The observed insoluble precipitate could be due to the formation of polymers, variable-length oligomers, or even insoluble cage products, each of which would require an additional purification step to isolate the single cage species. Meanwhile, partial conversion indicates that the trialdehyde had not undergone complete conversion to imine products, suggesting the presence of kinetic intermediates and incomplete equilibration to the thermodynamically most

stable product. For these cases, altering the reaction conditions by increasing the temperature or reaction time could facilitate a complete conversion to this single topology.

Clean Formation of Single Topologies

In total, 10 (7%) combinations of precursors were observed to form a single topology alongside an insoluble precipitate, belonging to category 5. The DFT optimised structures of each of the topologies that were observed to form are shown in Figure 5.12. Of these 10 cages, 2 were formed from trialdehyde **U**, the *trans* alkene trialdehyde precursor predicted to form shape persistent cages in Chapter 4. The large size of this trialdehyde could result in a greater propensity to form longer insoluble chains, and trialdehyde **U** has also recently been shown to form a COF with diamine linkers, which could explain the formation of some of the insoluble precipitates formed with this precursor.[204] Each of the cages formed from this precursor were smaller [2 + 3] species in contrast to the larger [4 + 6] species that were predicted to be shape persistent for this trialdehyde in the previous chapter. Of the 2 cages formed with trialdehyde **U**, both **U26** and **U28** were observed to form more complex NMR spectra with many peaks in the imine region, which could be due to additional imine species being present or the imine protons in the [2 + 3] species being in different environments. While **Tri**^{2}**Di**^{3} topologies are typically too small to accommodate even a small guest molecule, **U28** was predicted to possess a larger cavity size of 6.8 Å due to the larger size of the trialdehyde precursor, making this a potentially porous **Tri**^{2}**Di**^{3} cage. Each of the **Tri**^{4}**Di**^{6}, **Tri**^{6}**Di**^{9}, and **Tri**^{8}**Di**^{12} topologies were predicted to collapse for this precursor combination.

In addition, combinations **T10**, **T25**, **T26**, **S31**, **H26**, **H31**, and **K25** were also observed to form a smaller [2 + 3] topology, and **S28** was observed to form a larger [4 + 6] topology, each alongside insoluble precipitate (category 2). Combination **H31** formed the [2 + 3] topology cleanly by ^{1}H NMR, as indicated by the single imine peak formed from the reaction of these symmetric precursors. In contrast, **K25** possessed a more complex ^{1}H NMR spectra, despite only a mass ion peak corresponding to the [2 + 3] topology being observed by HRMS.

Similarly to **S16**, both **S28** and **S31** each displayed complex NMR spectra due to the asymmetry of trialdehyde **S** resulting in cages with multiple distinct imine environments, or a mixture of cage isomers. Although **S31** formed a smaller [2 + 3] species, **S28** formed a larger [4 + 6] species which was predicted to possess a relatively large permanent internal cavity (11.6 Å). This topology was the most stable open topology, 38 kJ mol^{-1} more stable than the smaller [2 + 3] topology, as shown in Figure 5.13. This predicted large cavity size and formation of only a single topology suggest that this combination could be a promising candidate for optimisation and scale-up, to identify reaction conditions that would reduce the amount of insoluble precipitate produced.

Meanwhile, combinations **T10**, **T25**, and **T26** exhibited a more complex ^{1}H NMR spectra, which would not be anticipated based on the high symmetry of trialdehyde **T**, indicative of the

formation of a mixture of imine products. One possible cause of the complex NMR spectra could be the formation of tautomers of cages **T10**, **T25**, and **T26**, which have been reported to form from trialdehyde **T**, and would be in equilibrium with the imine cage.[205]

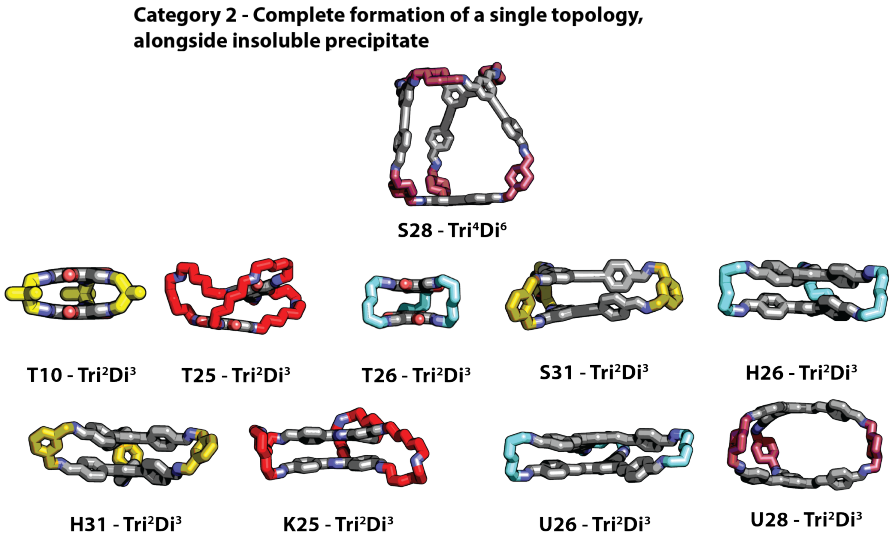


Figure 5.12: Cages optimised at the PBE/TZVP+D3 level which formed a single topology alongside an insoluble precipitate, corresponding to category 2. Only precursor combination **S28** was predicted to be shape persistent. Carbon atoms corresponding to diamine **10** are shown in light yellow, **25** in red, **26** in light blue, **28** in dark red, and **31** in dark yellow. Other carbon atoms are shown in grey, nitrogens in blue and carbons in green.

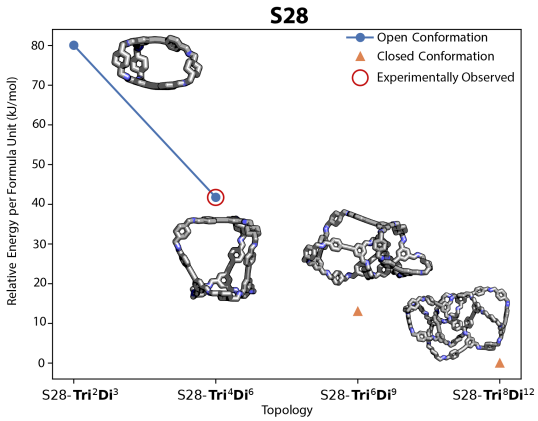


Figure 5.13: Relative energies per formula unit of different topologies of combination **S28**, each calculated at the PBE/TZVP+D3 level. All energies are calculated with respect to the most stable open **Tri**^{8}**Di**^{12} topology. For **S28**, both the **Tri**^{6}**Di**^{9} and **Tri**^{8}**Di**^{12} topologies were predicted to be collapsed. This precursor combination resulted in the formation of a single [4 + 6] species, shown by the red circle, alongside an insoluble precipitate.

Partial Formation of Single Cage Topologies

Next, I will explore the combinations that resulted in the formation of a single topology, but resulted in only partial conversion to imine products, characterised by an aldehyde peak present in the ^{1}H NMR spectrum (categories 5 and 6). The DFT optimised structures of each of the precursor combinations that were assigned to these categories are shown in Figure 5.14. The presence of residual aldehyde, especially when excess diamine is used, may be due to the formation of kinetic intermediates, where the variable number of imine condensations leads to unreacted aldehyde groups. Stable kinetic traps have been observed for **CC3**, **CC21**, and **CC13**, being produced along the formation pathway to the thermodynamically stable [4 + 6] species.[112,206]

Out of all combinations, 15 partially formed a single topology, 10 of which did not form a solid precipitate (category 5), and 5 of which formed alongside an insoluble precipitate (category 6), each of which is shown in Figure 5.14. Of the 15 combinations which formed cages in these two categories, only one formed a larger [4 + 6] species (**G16**), which is the **Tri**^{4}**Di**^{6} cage **CC3** used as a control study in the screen. In the initial condition screening for this study, **CC3** was observed to only partially form the **Tri**^{4}**Di**^{6} topology, forming in a mixture with the [3 + 5] kinetic intermediate. For these partially formed cages, kinetic intermediates largely form complex NMR spectra due to the number of distinct imine peaks that form, making identification of single cage topologies by NMR alone challenging.

Of the 10 cases that did not form a solid precipitate, 5 were formed with trialdehyde **R** (**R15**, **R27**, **R31**, **R32**, and **R34**). The other combinations were **H15**, **J15**, **J26**, **T12**, and **G16** (**CC3**). The 5 combinations that resulted in partial conversion alongside an insoluble precipitate were **H10**, **J10**, **K15**, **R10**, and **U15**, each forming a smaller [2 + 3] species. For all of these combinations, multiple imine peaks were observed in the ^{1}H NMR spectra, and each resulted in the formation of complex NMR spectra with varying amounts of conversion to the single cage species, complicating analysis. Due to the absence of shape persistency in all of the smaller [2 + 3] cages, none were predicted to form porous cages.

Figure 5.15 shows the relative energies of each of the topologies of **G16** (**CC3**). For **G16**, the **Tri**^{4}**Di**^{6} topology was predicted to be the most stable, 59 kJ mol^{-1} more stable than the **Tri**^{2}**Di**^{3} topology and 6 kJ mol $^{-1}$ more stable than the partially collapsed **Tri**^{6}**Di**^{9} topology. The **Tri**^{4}**Di**^{6} topology of **CC3** being the most stable topology is in agreement with the findings of Santolini *et al.*, where the **Tri**^{4}**Di**^{6} topology was reported to be more stable than the larger collapsed **Tri**^{6}**Di**^{9} and **Tri**^{8}**Di**^{12} topologies,[31] and is the topology I experimentally observed.

In this section, I have outlined the 42 precursor combinations that resulted in the formation of a single cage topology, encompassing 29% of all the precursor combinations explored. Clean formation of a single topology is the most desirable property, requiring the fewest additional steps to purify and isolate the desired material. These combinations were categorised by their ability

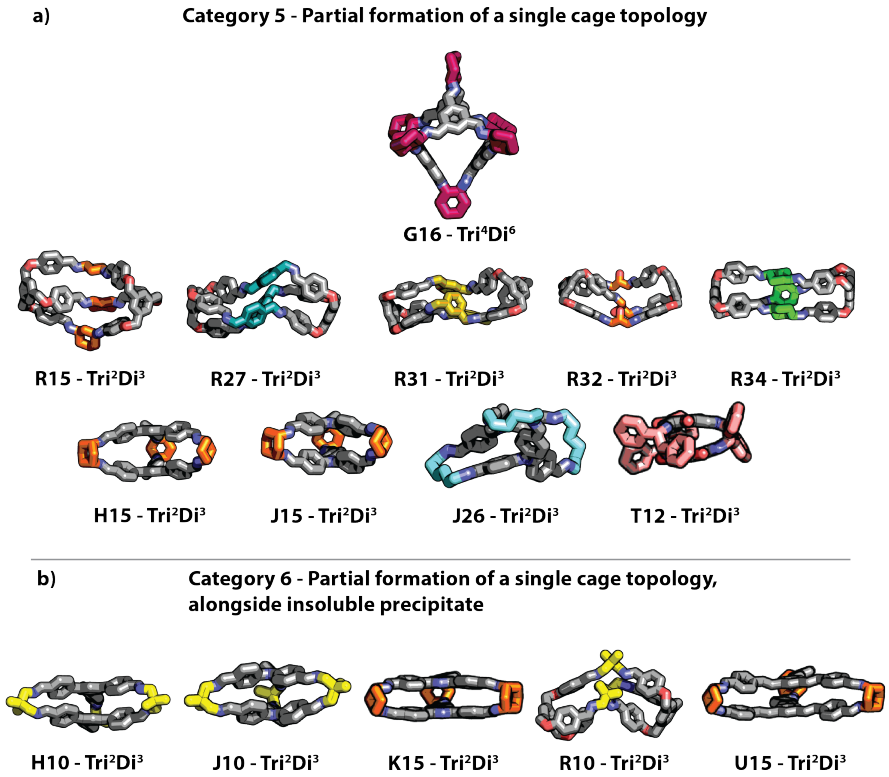


Figure 5.14: Cages optimised at the PBE/TZVP+D3 level which partially converted to a single topology, forming without an insoluble precipitate present (a), and alongside an insoluble precipitate (b), corresponding to categories 5 and 6 respectively. Only precursor combination **G16** was predicted to be shape persistent. Carbon atoms corresponding to diamine **10** are shown in light yellow, **12** in pink, **15** and **32** in orange, **16** in bright pink, **26** in light blue, **27** in teal, **31** in dark yellow, and **34** in green. Other carbon atoms are shown in grey, nitrogens in blue and carbons in green.

to form the topology either cleanly or partially, and with or without a solid precipitate. From the clean hits, both **I16** and **S16** were observed to form shape persistent **Tri^{4}Di^6** cages, in agreement with their respective reports in the literature.[63,188] Of the other combinations that formed cages, **S28** was a promising candidate due to the formation of a larger [4 + 6] species which was predicted to be shape persistent. For this candidate, an additional purification step would be required to filter the insoluble precipitate, however, alternate reaction conditions could facilitate complete conversion to the [4 + 6] species. Of the remaining 39 cages discussed in this section, 38 formed smaller [2 + 3] species too small to accommodate a guest molecule. Combination **U28** also formed a [2 + 3] species cleanly, but was predicted to form a larger cavity of 6.8 Å due to the large size of the trialdehdye **U**.

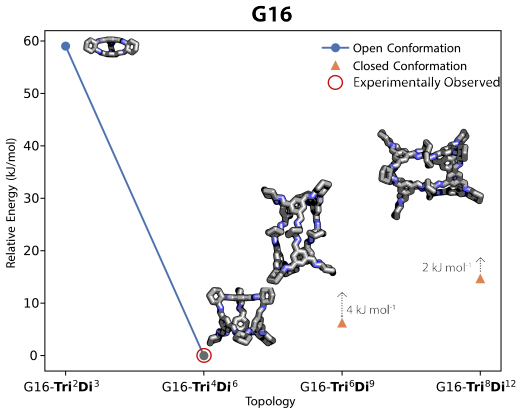


Figure 5.15: Relative energies per formula unit of different topologies of combination **G16** (**CC3**), each calculated at the PBE/TZVP+D3 level. All energies are calculated with respect to the most stable topology, **Tri**^{4}**Di**^{6}. For **G16**, the **Tri**^{6}**Di**^{9} and **Tri**^{8}**Di**^{12} topologies were predicted to be collapsed, denoted by the orange triangle. A [4 + 6] species was observed to partially form experimentally, shown by the red circle. The grey arrows represent the energy difference between collapsed and metastable open conformations of the **Tri**^{6}**Di**^{9} and **Tri**^{8}**Di**^{12} topologies observed by Santolini *et al.*[31]

5.3.4 Formation of a Mixture of Topologies

In this next section, I will discuss precursor combinations that resulted in a mixture of topologies forming, either cleanly or partially. Similarly to the previous section, I will split the discussion into whether each combination resulted in the formation of an insoluble precipitate or dissolved completely. To detect mixtures of topologies, mass ions corresponding to multiple cage topologies must be present in the HRMS. Complex NMR spectra may also suggest a mixture of products, but the proton environments between different topologies may not always produce separate, distinct peaks.[207] Relative proportions of each topology were not quantified due to the inability of the quadrupole time-of-flight (QTOF) mass spectrometer used in my work to provide quantitative amounts of each product. Instead, I use computation to rationalise the formation of mixtures of products that are predicted to be shape persistent by comparing the relative energies of different topologies to provide an insight into the relative stabilities of each topology.

Overall, mixtures are less desirable than the formation of a single cage topology, requiring isolation of a single cage species or alternative reaction conditions to promote conversion to a single topology. Despite being less attractive overall, some precursor combinations resulted in larger topologies formed in this section, which were predicted to be shape persistent *in silico*, and would make them ideal candidates to further explore alternate reaction conditions to facilitate complete conversion to the larger topology.

In total, 26 (18%) precursor combinations resulted in either clean or partial formation of a mixture of cage species, with and without precipitate formation. Of these 26 combinations, 21 were mixtures of [2 + 3] and [4 + 6] species, 2 formed larger [6 + 9] species in a mixture with other species, and 3 were found to form [8 + 12] species in a mixture with other species. Out of the 26 combinations that formed a mixture, 11 formed a clean mixture with no solid precipitate (category 3), 6 formed alongside a solid precipitate (category 4), 5 partially formed a mixture of imine products with no solid precipitate (category 7), and, finally, 4 formed a partial mixture alongside solid precipitate (category 8).

Clean Formation of Mixtures of Topologies

The first category I will discuss in this section, clean formation with no solid precipitate, is the most desirable of the four, and suggests that no insoluble side products are formed. In this category are 11 combinations, namely, **G25**, **G27**, **G30**, **G31**, **I26**, **I28**, **I30**, **J31**, **K29**, **U25**, and **U31**. All the aforementioned combinations formed a mixture of [2 + 3] and [4 + 6] species, with the exception of **I28**, which formed a mixture of a larger [6 + 9] species and a [4 + 6] species, and **J31** which formed a mixture of [2 + 3], [4 + 6], and [6 + 9] species. Each of the cages formed in this section is shown in Figure 5.16.

Unsurprisingly, all combinations comprised of diamines with a more flexible acyclic backbone (diamines **25**, **26**, **29**, and **30**) were predicted to collapse following the high temperature MD simulations, and each of the [2 + 3] species formed possessed only a small internal cavity. Despite possessing less flexibility, combinations containing diamine **31** (**G31**, **J31**, and **U31**) were also predicted to collapse upon desolvation, including the [6 + 9] species formed in combination **J31**.

Combinations **G27** and **I28** were more interesting, with the **Tri**^{4}**Di**^{6} topology predicted to be shape persistent for both combinations, and **I28** forming an additional [6 + 9] species, also predicted to form a shape persistent **Tri**^{6}**Di**^{9} topology. Calculated cavity sizes were relatively large, with **G27** possessing a calculated cavity diameter of 8.2 Å in the **Tri**^{4}**Di**^{6} topology, and **I28** possessing cavity sizes of 10.2 Å for the **Tri**^{4}**Di**^{6} topology and 4.5 Å for the **Tri**^{6}**Di**^{9} topology. Despite containing more atoms, the calculated cavity diameter was due to the tubular structure predicted to form in the **Tri**^{6}**Di**^{9} topology, resulting in a smaller spherical volume in the central cavity that is calculated using *pyWindow*.[51] For these two combinations, the ^{1}H NMR spectra was complex in the aromatic region for both **G27** and **I28**, suggesting that other imine products could be present in addition to the cage topologies that were identified. For **I28** meanwhile, the asymmetry of trialdehyde **I** introduced additional complexity into the NMR spectrum due to the many distinct imine environments.

To further rationalise the formation of a mixture of topologies for these two cases, I used the calculated relative energies of the **Tri**^{2}**Di**^{3}, **Tri**^{4}**Di**^{6}, **Tri**^{6}**Di**^{9}, and **Tri**^{8}**Di**^{12} topologies to identify

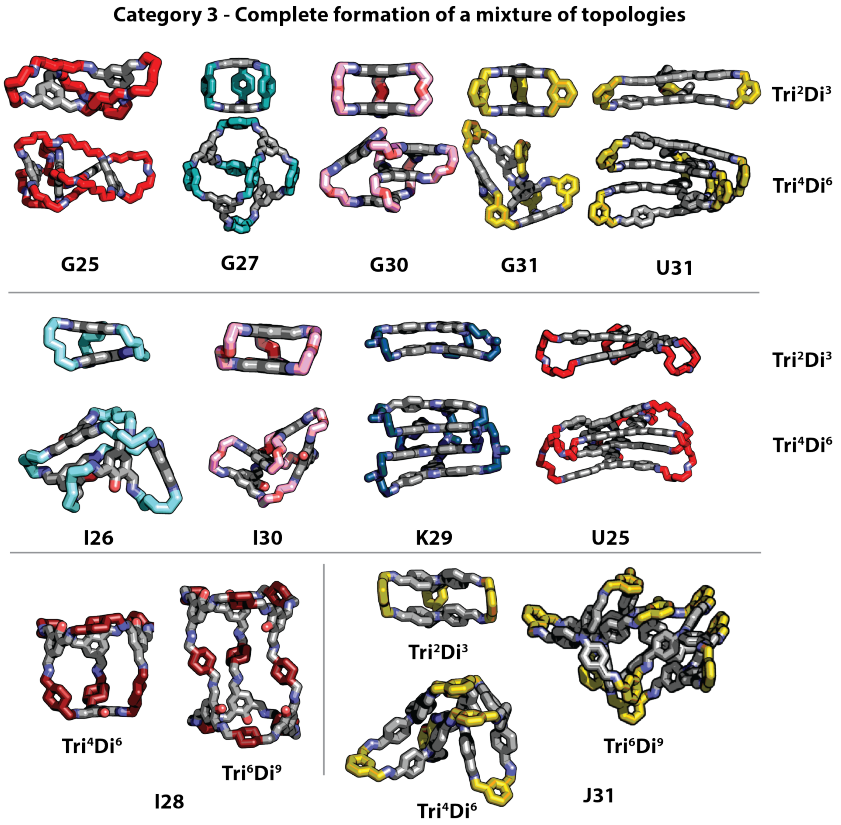


Figure 5.16: Cages optimised at the PBE/TZVP+D3 level which cleanly formed a mixture of cages with no solid precipitate, corresponding to category 3. Out of all the precursor combinations, the **Tri**^{4}**Di**^{6} species of both **G27** and **I28** were predicted to be shape persistent, in addition to the **Tri**^{6}**Di**^{9} topology of **I28**. For each precursor combination, all the species where a mass ion was observed by HRMS are shown, with **Tri**^{2}**Di**^{3} topologies above the **Tri**^{4}**Di**^{6} topologies in rows 1 and 2. Carbon atoms corresponding to diamine **25** are shown in red, **26** in light blue, **27** in teal, **28** in dark red, **29** in dark blue, **30** in pink, **31** in dark yellow, and **34** in green. Other carbon atoms are shown in grey, nitrogens in blue and carbons in green.

the most stable topology. The relative energies for **I28** are shown in Figure 5.18b, indicating the **Tri**^{4}**Di**^{6} and **Tri**^{6}**Di**^{9} topologies are similar in energy, in agreement with the observed experimental outcome of both these topologies forming. The energetically most stable topology, **Tri**^{8}**Di**^{12}, was predicted to collapse upon desolvation, which may not resemble the structure of the solution phase molecule due to the templating effects of solvent molecules.[61] Meanwhile, for **G27** the energy difference between the **Tri**^{2}**Di**^{3} and **Tri**^{4}**Di**^{6} cages was smaller than **I28**, with the **Tri**^{4}**Di**^{6} cage being 25 kJ mol^{-1} more stable, as shown in Figure 5.18a. Both [2 + 3] and [4 + 6] species were observed for **G27**.

Next, 6 combinations formed a mixture of topologies cleanly, but alongside an insoluble precipitate, which were combinations **H30**, **I27**, **K16**, **K26**, **K30**, and **K31**, and the DFT optimised structures are shown in Figure 5.17. Combinations **H30**, **I27**, **K26**, and **K30** each formed a mixture of [2 + 3] and [4 + 6] species, whereas **K31** formed mixtures of [4 + 6], [6 + 9], and [8 + 12] species, and **K16** formed a combination of [4 + 6] and [8 + 12] species. Out of the combinations that formed, only the **Tri^{4}Di^6** topologies of **I27** and **K16** were predicted to be shape persistent, with calculated cavity sizes of 10.7 Å and 11.5 Å, respectively, with all others predicted to collapse due to the flexibility of the diamine linkers.

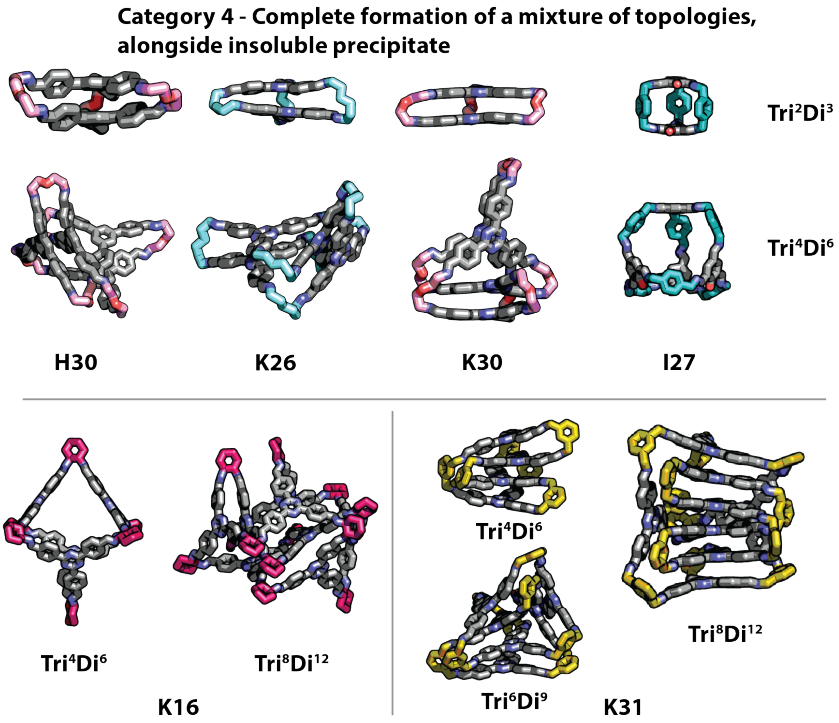


Figure 5.17: Cages optimised at the PBE/TZVP+D3 level which cleanly formed a mixture of cages alongside a solid precipitate, corresponding to category 4. From these six precursor combinations, the **Tri^{4}Di^6** topologies of **K16** and **I27** were predicted to be shape persistent. For each precursor combination, all the species where a mass ion was observed by HRMS are shown, with **Tri^{2}Di^3** topologies above the **Tri^{4}Di^6** topologies in row 1. Carbon atoms corresponding to diamine **26** are shown in light blue, **27** in teal, **29** in dark blue, **30** in pink, and **31** in dark yellow. Other carbon atoms are shown in grey, nitrogens in blue and carbons in green.

Promisingly, **K16** has already been reported in the literature to form a **Tri^{4}Di^6** topology with a high BET surface area of 1,181 m^2 g^{-1}.[208] The maximum diameter obtained using *pyWindow* (32.8 Å) is similar to the experimentally obtained diameter from the reported crystal structure (30 Å), although the calculated minimum internal cavity diameter of 11.5 Å is slightly lower than

the 16 Å experimentally observed, .[208] This disagreement in calculated minimum internal cavity diameter could be due to differences in how *pyWindow* calculated the value compared with how it is experimentally calculated. The ^{1}H NMR spectrum for **K16** was relatively simple, consistent with the literature and the high symmetry of the two precursors used. In addition to the [4 + 6] species observed, a species with an [8 + 12] reaction stoichiometry was also observed by HRMS. In the crystal structure of **K16**, Ding *et al.* report the formation of 1D polymeric chains formed by the interlinking of [4 + 6] cage units. One plausible explanation for the observed [8 + 12] species could be due to the formation of two interlocked [4 + 6] species in solution, a possible intermediate to the polymeric self-assembly.[208] Relative energy calculations for **K16**, shown in Figure 5.18d, indicated the **Tri**^{4}**Di**^{6} would be 15 kJ mol^{-1} more stable than the **Tri**^{2}**Di**^{3} topology, and is experimentally observed. Both the **Tri**^{6}**Di**^{9} and **Tri**^{8}**Di**^{12} topologies were predicted to collapse. In this example, the interlocked species could be more stable than the isolated **Tri**^{4}**Di**^{6}, driving the formation of the catenated **Tri**^{4}**Di**^{6} species, which is not considered when modelling the isolated molecular unit.

Meanwhile, precursor combination **I27** was observed to form a [2 + 3] and [4 + 6] species in a mixture alongside an insoluble precipitate, which has not been previously reported. For this promising candidate, the energy of the **Tri**^{2}**Di**^{3} was 25 kJ mol^{-1} higher than the **Tri**^{4}**Di**^{6} topology (Figure 5.18c), suggesting altering reaction conditions could facilitate conversion to the more stable product. Both the **Tri**^{6}**Di**^{9} and **Tri**^{8}**Di**^{12} topologies were predicted to collapse for this precursor combination.

Partial Formation of Mixtures of Topologies

The final two outcomes corresponded to the partial formation of a mixture of topologies, without solid precipitate formation (5 combinations) or with insoluble precipitate formation (4 combinations). The combinations that formed mixtures of topologies partially with no precipitate were **I10**, **I15**, **J25**, **J32**, and **R29**, with **I15** forming a mixture of [4 + 6], [6 + 9], and [8 + 12] species, and the other 4 forming mixtures of [2 + 3] and [4 + 6] species. Each of the DFT optimised cages observed in this section are shown in Figure 5.19a. Of the combinations that formed cages, the **Tri**^{4}**Di**^{6} topology of **J32** was predicted to be shape persistent, in addition to the **Tri**^{4}**Di**^{6}, **Tri**^{6}**Di**^{9}, **Tri**^{8}**Di**^{12} topologies of **I15**. All other combinations were predicted to collapse.

Combination **J32**, which was observed to form a mixture of [2 + 3] and [4 + 6] species, has not been previously reported. While the [2 + 3] species was predicted to be non-porous due to its small size, the [4 + 6] species was predicted to possess a modest internal cavity diameter of 5.8 Å. The small size of diamine **32** combined with the larger distance between aldehyde functional groups in trialdehyde **J** resulted in the formation of an octahedron with asymmetric pores in the **Tri**^{4}**Di**^{6} topology, shown in Figure 5.20b.

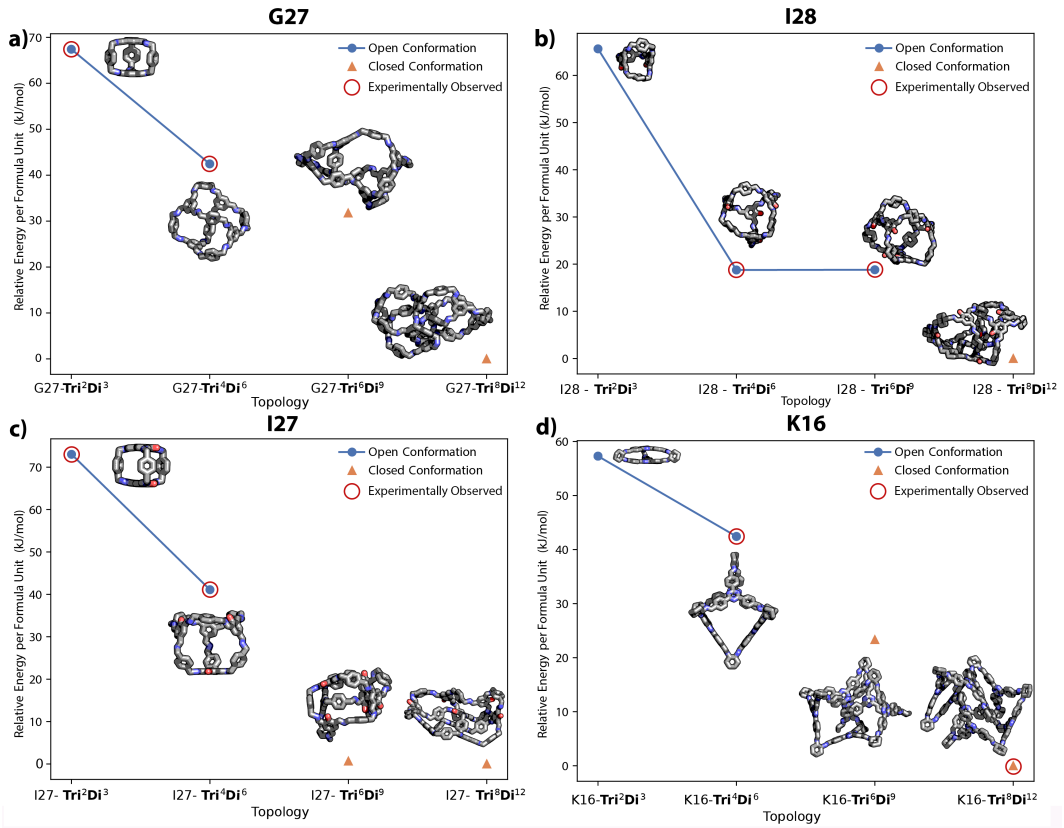


Figure 5.18: Relative energies per formula unit of different topologies of combinations **G27** (a), **I28** (b), **I27** (c), and **K16** (d) each calculated at the PBE/TZVP+D3 level. All energies are calculated with respect to the most stable collapsed **Tri**^{8}**Di**^{12} topology for all four precursor combinations. For **G27**, **I27**, and **K16** the **Tri**^{2}**Di**^{3} and **Tri**^{4}**Di**^{6} topologies were predicted to be shape persistent, whereas for **I28** the **Tri**^{6}**Di**^{9} was also predicted to be shape persistent. **G27** and **I27** formed clean mixtures of the [2 + 3] and [4 + 6] species, whereas **I28** formed a mixture of [4 + 6] and [6 + 9] species, and **K16** formed a mixture of [4 + 6] and [8 + 12] species. All experimentally formed species are shown by a red circle.

In the literature, cage **I15** has previously been isolated in the **Tri**^{8}**Di**^{12} topology (**CC20**).[189] This **Tri**^{8}**Di**^{12} is one of the largest discovered cages to date, with an SA_{BET} of 1752 m^2g^{-1}. Here, I also observed the formation of an [8 + 12] species by HRMS, corresponding to the large **Tri**^{8}**Di**^{12} topology, albeit with incomplete conversion. The conditions used by Teng *et al.* were selected to optimise the formation of the **Tri**^{8}**Di**^{12} topology, employing a reaction time of 2 weeks, DCM/MeOH as solvent, and layering a dilute solution of diamine **15** onto solid trialdehyde **I** to maximise conversion.[189] My conditions of 5 days may have been insufficient to allow full conversion to the **Tri**^{8}**Di**^{12} topology, instead forming less energetic stable topologies alongside additional kinetic intermediates. Relative energies of each topology also supported the presence of a mixture of topologies form-

ing, with the **Tri**^{4}**Di**^{6}, **Tri**^{6}**Di**^{9}, and **Tri**^{8}**Di**^{12} each within a range of 30 kJ mol^{-1} of each other, as shown in Figure 5.20a. The open **Tri**^{8}**Di**^{12} topology, being lowest in energy, is predicted to be the most thermodynamically stable topology, which is what is observed experimentally. In contrast, the **Tri**^{2}**Di**^{3} topology is 138 kJ mol^{-1} greater in energy than the most stable topology, rationalising the absence of this topology experimentally. Each of the **Tri**^{4}**Di**^{6}, **Tri**^{6}**Di**^{9}, and **Tri**^{8}**Di**^{12} topologies was also predicted to be porous, with cavity sizes of 8.4, 11.8, and 15.2 Å, respectively.

The final category, partial formation of a mixture of topologies alongside insoluble precipitate, is the least desirable topological outcome, requiring the separation of insoluble products and also individual cage species. Combinations **I32**, **T29**, **T30**, and **T31**, were categorised here, and each was observed to form a mixture of [2 + 3] and [4 + 6] species. Each of the DFT optimised cages are shown in Figure 5.19b. Despite forming larger [4 + 6] species, all of the species were predicted to collapse.

In this section, I have outlined the 26 experimentally formed cages which formed in a mixture of topologies. Interesting novel candidates in this section include **G27**, which formed a mixture of [2 + 3] and [4 + 6] species, **I28**, which formed a mixture of [2 + 3], [4 + 6], and [6 + 9] species, and, finally, **I27**, which formed a mixture of [2 + 3] and [4 + 6] species. Each of these candidates were predicted to form at least one shape persistent topology which has not been previously reported in the literature. Generally, relative energies between shape persistent topologies of cages that formed in a mixture were lower than those that only formed a single topology. This is exemplified by the **Tri**^{4}**Di**^{6} topology being 25 kJ mol^{-1} lower in energy than the **Tri**^{2}**Di**^{3} for both **G27** and **I27**, both of which formed a mixture of [2 + 3] and [4 + 6] species. In comparison, the **Tri**^{4}**Di**^{6} topology was 59 kJ mol^{-1} more stable than the **Tri**^{2}**Di**^{3} for **G16** (**CC3**), which was only observed to form a [4 + 6] species. In addition to these novel shape persistent cages, **I15** was observed to form a mixture [4 + 6], [6 + 9], and [8 + 12] species, each predicted to be shape persistent, in agreement with the cage **CC20** from the literature.[189] Combination **K16** was also observed to form a [4 + 6] species, in agreement with the reported literature.[208]

5.3.5 Unsuccessful Combinations

Altogether, 76 (53%) precursor combinations did not result in any cage formation. Of these 76 unsuccessful outcomes, 43 resulted from combinations involving diamines **11**, **12**, **14**, **33**, and **34**, suggesting these are generally poor candidates for cage synthesis. Diamine **14** was only observed to react with a single trialdehyde, and diamine **33** with 2 trialdehydes, indicating these precursors are unlikely to undergo the required imine condensation reactions required for cage synthesis. Generally, diamines **11**, **12**, and **34** were more reactive than diamines **14** and **33**, reacting with the majority of the trialdehdye precursors. However, reactions involving these diamines mostly resulted in no cage formation, with two exceptions (**T12** and **R34**). The bulky phenyl groups on

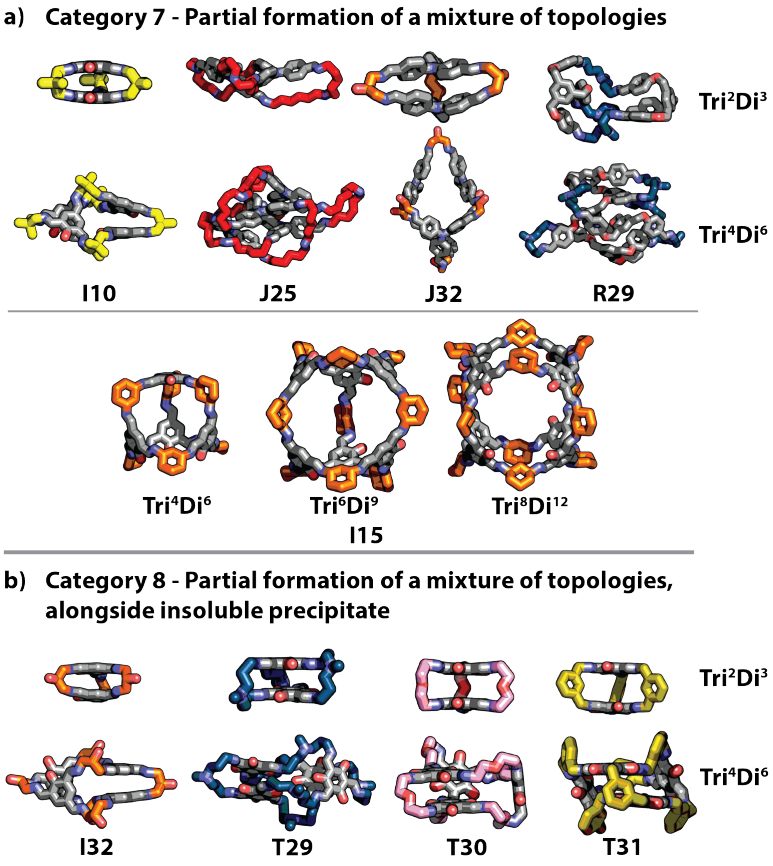


Figure 5.19: Cages optimised at the PBE/TZVP+D3 level which partially converted to a mixture of topologies without insoluble precipitate (a), and with insoluble precipitate (b), belonging to categories 7 and 8 respectively. Here, the **Tri**^{4}**Di**^{6}, **Tri**^{6}**Di**^{9}, and **Tri**^{8}**Di**^{12} topologies of **I15** were predicted to be shape persistent, in addition to the **Tri**^{4}**Di**^{6} topology of **J32**. For each precursor combination, all the species where a mass ion was observed by HRMS are shown, with **Tri**^{2}**Di**^{3} topologies above the **Tri**^{4}**Di**^{6} topologies in rows 1 and 3. Carbon atoms corresponding to diamine **10** are shown in yellow, **25** in red, **27** in teal, **32** and **15** in orange, **29** in dark blue, **30** in pink, and **31** in dark yellow. Other carbon atoms are shown in grey, nitrogens in blue and carbons in green.

diamines **11** and **12** could be inhibiting cage formation, but it is difficult to predict the effects of external groups on the cage formation mechanism using computation alone.

Meanwhile, diamine **34** reacted with all 9 trialdehydes, but only formed a cage with trialdehyde **R**. The absence of cage formation could be attributed to the large distance between the two amine groups on the cyclohexane core of the diamine, preventing self-assembly into specific cage topologies. In the literature, diamine **34** has been reported to form COFs with both trialdehdyes **G** and **T** to form **PPN-30** and **PPN-31** respectively.[209] As both **PPN-30** and **PPN-31** were reported to be insoluble in many organic solvents, the solid precipitate observed in the reaction with **G** and **T**

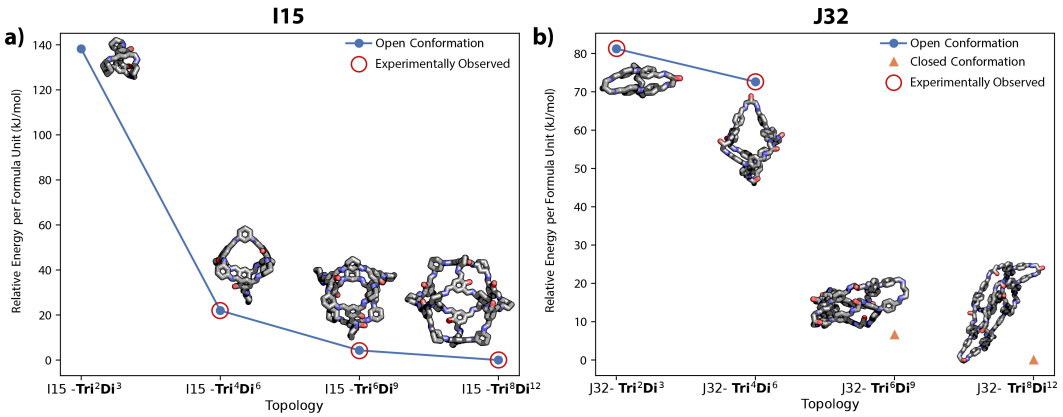


Figure 5.20: Relative energies per formula unit of different topologies of combination **I15** (a), and **J32** (b), each calculated at the PBE/TZVP+D3 level. All energies are calculated with respect to the most stable topology, **Tri**^{8}**Di**^{12} for both combinations. For **I16**, all topologies were predicted to be shape persistent, but for **J32** only the **Tri**^{2}**Di**^{3} and **Tri**^{4}**Di**^{6} topologies were predicted to be shape persistent. **I15** partially formed a mixture of [4 + 6], [6 + 9], and [8 + 12] species, whereas **J32** formed a mixture of [2 + 3] and [4 + 6] species. An orange triangle indicates a closed conformation, whereas a blue circle indicates an open conformation.

could have been the formation of insoluble COFs. Indeed, a solid precipitate was observed to form from the reaction of diamine **34** with 7 trialdehdyes, with the exception of trialdehydes **R** and **J**. While no conclusions can be drawn about the nature of the solids in my thesis, it would be interesting to further analyse these solids to differentiate between the formation of COFs, and other oligomeric or polymeric products. To further analyse the insoluble precipitate, high-throughput powder X-ray diffraction (PXRD) could be performed, which would be able to elucidate their structures and differentiate between insoluble cages and COFs.[210]

5.3.6 Considering Kinetic Intermediates

While thermodynamic products typically dominate on longer timescales, kinetic intermediates have been observed to form during the formation of **CC3**-*S* and **CC3**-*R*.[112,193] Zhu *et al.* suggest that the larger [4 + 6] species are a result of the conformational rearrangement of smaller intermediates, such as [1 + 1], [1 + 2], and [1 + 3].[112] Using the automated HRMS analysis, 20 precursor combinations were observed to form a [3 + 5] kinetic intermediate, the full breakdown of which is shown in Table 5.2. Despite the potential to form a number of different kinetic intermediates, only the [3 + 5] intermediate was investigated due to its presence in the initial condition screening with **CC3** and because it is a commonly reported kinetic intermediate in the literature.[112,206] Here, the detection of the [3 + 5] species is used as an indicator that the formation of larger cage species might be possible for specific precursor combinations by altering the reaction conditions.

Of these intermediates, 15 (75%) were observed alongside either a [2 + 3] species, and 17 (85%) were present in a mixture of other topologies. In my thesis, I did not consider the energetic preference for the formation of kinetic intermediates, hypothesising that the conditions selected for high-throughput screening would maximise conversion to a thermodynamic product for the vast majority of cases. For cages where a stable kinetic intermediate was observed, a further computational investigation could be performed to further probe the stability of these intermediates. Further analysis of these intermediates could be crucial to fully understanding POC formation mechanisms, facilitating more accurate structural predictions for these precursor combinations.

Species Formed	Precursor Combination
[3 + 5]	**J34**, **T27**, **T28**
[2 + 3], [3 + 5]	**J10**, **J15**, **J26**, **G29**, **S26**, **T12**, **T25**, **T26**
[2 + 3], [4 + 6], [3 + 5]	**I10**, **I26**, **J25**, **I27**, **J32**, **T29**, **T30**
[4 + 6], [6 + 9], [3 + 5]	**I28**
[4 + 6], [6 + 9], [8 + 12], [3 + 5]	**I15**

Table 5.2: The breakdown of precursor combinations where a [3 + 5] kinetic intermediate was observed to form by HRMS. Multiple topologies indicate a mixture of these topologies formed in solution, in addition to the [3 + 5] kinetic intermediate.

5.3.7 Precursor Trends

Following the analysis of the reactions for each precursor combination, some general trends can be obtained about the potential of each precursor to form a cage. The number of times a precursor appeared in a precursor combination that formed a cage at all (categories 1-8) is shown in Figure 5.21. Generally, the more frequently a precursor is involved in successful cage formations, the more promising it could be as a candidate for cage synthesis. From an examination of Figure 5.21, trialdehyde **I** would be the most promising, appearing in 11 precursor combinations where some evidence of cage formation was observed. Diamines **11**, **14**, and **33** would be the least promising, forming no cages. In this section, I will further discuss the structural features of precursors that could inhibit or promote cage formation, and rationalise these observations using calculated formation energies.

Upon analysing the reactions of each trialdehyde, the asymmetric trialdehyde **S** displayed promising results as a precursor for cages, yielding 6 clean cages and 2 with an insoluble precipitate. Other notable trialdehydes include **I**, which yielded 4 clean cages and 7 mixtures of topologies, and **U**, which produced 2 clean cages, 2 with insoluble precipitate, 1 partial formation, and 2 mixtures of topologies. In total, 7 out of the 9 trialdehydes resulted in at least one clean cage formation, with only **J** and **T** failing to produce any cages cleanly. Trialdehyde **G**, a commonly used precur-

sor in the POC literature,[32] was less successful than **S** or **I**, generating 1 clean topology, 1 partial formation, and 4 mixtures of topologies.

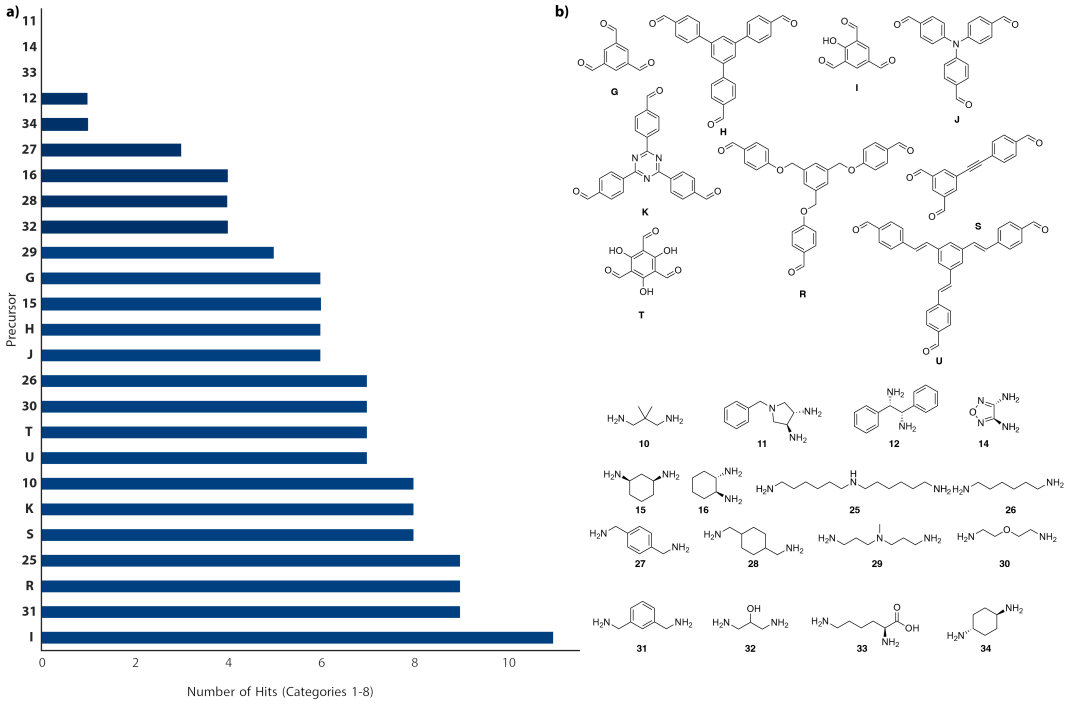


Figure 5.21: The number of times each precursor appeared in precursor combinations that successfully formed a cage (a), corresponding to categories 1-8 using the high-throughput reaction analysis workflow. Precursor letters and numbers on the bar plot (a) correspond to the molecules on the right (b).

The success of trialdehydes **S** and **I** is unsurprising given their previous use as POC precursors.[63,188,189] Trialdehyde **S** has been used in a high-throughput experimental workflow resulting in the formation of [4 + 6] species with 7 different diamines, many of which were predicted to be shape persistent *in silico*.[63] Meanwhile, trialdehyde **I**, which differs from **G** by the presence of a hydroxy group around the aromatic ring, has been shown to facilitate cage formation due to stabilising intramolecular hydrogen bonding interactions.[189] Trialdehyde **U** meanwhile has not been reported for cage synthesis previously and was discovered as a promising precursor from the high-throughput computational screen in the previous chapter. Indeed, this precursor showed promise as a POC precursor due to the number of clean cages observed.

Cages were found to form most readily with long flexible amines, with diamine **25**, having the longest acyclic backbone, producing clean cages with 4 different trialdehydes, a single topology with insoluble precipitate with 2, and mixtures of topologies with 3. Diamine **26** formed 2 clean cages, 1 with insoluble precipitate, and 4 mixtures of topologies. However, despite their

cage-forming potential, none of these candidates were predicted to be shape persistent computationally, as illustrated in Figure 5.24. Examples of cages predicted to be collapsed, but formed experimentally are shown in Figure 5.22. The additional flexibility of the long diamine linker may account for its ability to accommodate cage formation across a wide range of trialdehyde sizes. This feature likely contributes to the high cage-forming probability of these long flexible amines across a wide range of trialdehyde geometries. In previous studies, the use of larger chain diamines with trialdehyde **G** has resulted in the formation of cages, albeit predicted to collapse. For instance, 1,6-hexane-diamine yielded a [4 + 6] species and 1,7-heptane-diamine resulted in a [2 + 3] species using the same trialdehyde precursor.[110]

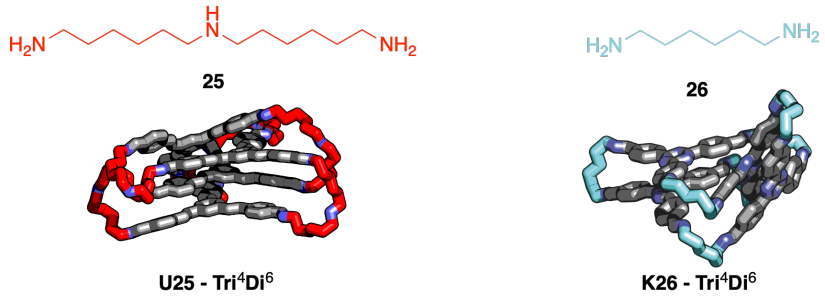


Figure 5.22: The two diamines (**25** and **26**) with the longest flexible chain between reacting amine groups used in the high-throughput screen. The **Tri**^{4}**Di**^{6} topologies of **U25** and **K26** are shown, each optimised at the PBE/TZVP+D3 level. Both **K26** and **U25** formed [2 + 3] and [4 + 6] species in a mixture, but were predicted to collapse. Carbon atoms in diamine **25** are shown in red and **26** in blue, nitrogen in dark blue, and other carbons in grey.

Smaller, more rigid diamine linkers, which are more likely to form a shape persistent cage, were found to be more selective with regard to the geometry of the trialdehyde, forming fewer cages overall. This category includes diamines with fewer degrees of rotation in their backbone, such as **10**, **15**, **16**, **27**, **28**, **31**, and **34**, and examples of each of these diamines in experimentally-formed cages are shown in Figure 5.23. Diamine **16**, a commonly used precursor in POC literature,[32] was able to form 2 clean cages and 2 mixtures of different topologies, although it formed a cage with fewer trialdehydes than the longer and more flexible diamines. Diamine **28** was similarly selective, resulting in 1 clean cage, 2 with precipitate, and 1 mixture of different topologies. As expected, these rigid diamines were more likely to result in a shape persistent cage computationally, as shown by the greater numbers of shape persistent cages for these diamines in Figure 5.24.

Diamine **10** is a notable success among the shorter amine linkers, as it formed single topologies with 4 trialdehydes and mixtures of topologies with 5 trialdehydes. On the other hand, diamine **32** had the same 3-atom backbone chain length but only formed a cage with at least one topology in 6 cases. The structural differences between diamine **10** and **32** were that **10** possessed two methyl

groups on the central carbon in the diamine backbone, whereas **32** possessed a single hydroxy group on the central carbon. This observation can be attributed to the two methyl substituents on the backbone of diamine **10**, which exhibit the *gem*-dimethyl effect that has been shown to promote cage formation.[184] In this effect, macrocycle cyclisation is favoured via two mechanisms: the Thorpe-Ingold theory and the reactive rotamer theory.[211] The former mechanism reduces the bond angle between reactive end groups, facilitating cyclisation, while the latter suggests the larger energy barrier to rotation of the amine functional groups results in a larger population of the diamine precursors adopting the gauche conformation required for cage formation.

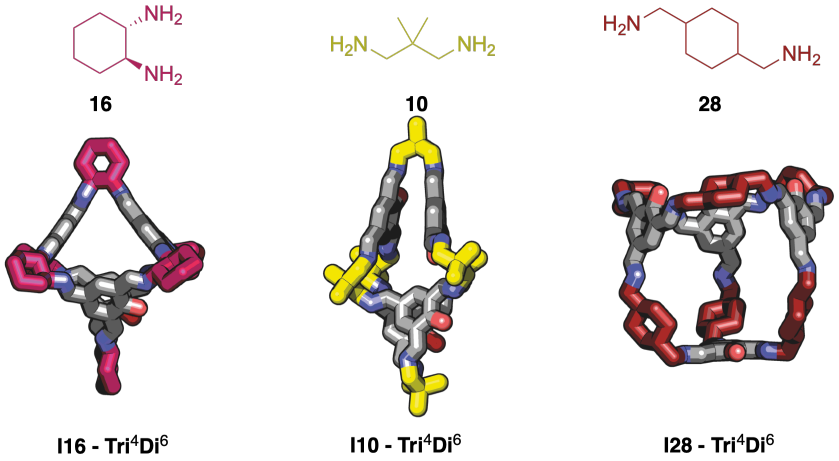


Figure 5.23: Three of the more rigid diamines (**16**, **10**, and **28**) used in the high-throughput screen. The **Tri**^{4}**Di**^{6} topologies of **I16**, **I10**, and **I28** are shown, each optimised at the PBE/TZVP+D3 level, and each of these cages was observed experimentally. Carbon atoms of diamine **16** are shown in purple, **10** in yellow, and **28** in dark red. Nitrogen atoms are shown in blue, oxygen in red, and other carbon atoms in grey.

In this section, I explored some of the general trends of diamine and trialdehyde precursors from the high-throughput experimental screen, focusing on key features that promote the formation of organic cages. In the next section, I will use the results of the DFT calculations to rationalise the observed topologies of each precursor, analysing the formation energies of each cage and topology. This analysis will provide insight into the thermodynamic preferences for each reaction, and how different precursor features could lead to different topological outcomes.

Formation Energy Calculations Figure 5.24 shows the formation energy per imine bond for each precursor combination in each of the four topologies (**Tri**^{2}**Di**^{3},**Tri**^{4}**Di**^{6}, **Tri**^{6}**Di**^{9}, and **Tri**^{8}**Di**^{12}), where cages with a cavity diameter of greater than 1 Å are shown. For comparing formation energies and cavity sizes, I use a smaller value of 1 Å instead of the kinetic diameter of the hydrogen atom (2.89 Å). While cages with cavity diameters below 2.89 Å are unlikely to be in-

trinsically porous, many of the smaller cages with a **Tri**^{2}**Di**^{3} topology were predicted to possess a smaller cavity size than this. For cages with larger **Tri**^{4}**Di**^{6}, **Tri**^{6}**Di**^{9}, and **Tri**^{8}**Di**^{12} topologies, a calculated cavity diameter of smaller than 1 Å would indicate that the cage is collapsed. Only shape persistent cages are shown as in solution solvent molecules are likely to create a templating effect on cage formation, influencing the topological outcome.[31] A computational analysis of the isolated molecular structure cannot account for templating effects, and many of the collapsed structures possessed stabilising dispersion or π-π interactions that would result in low energy structures. Unsurprisingly, many of the larger **Tri**^{4}**Di**^{6} cages were predicted to be shape persistent compared

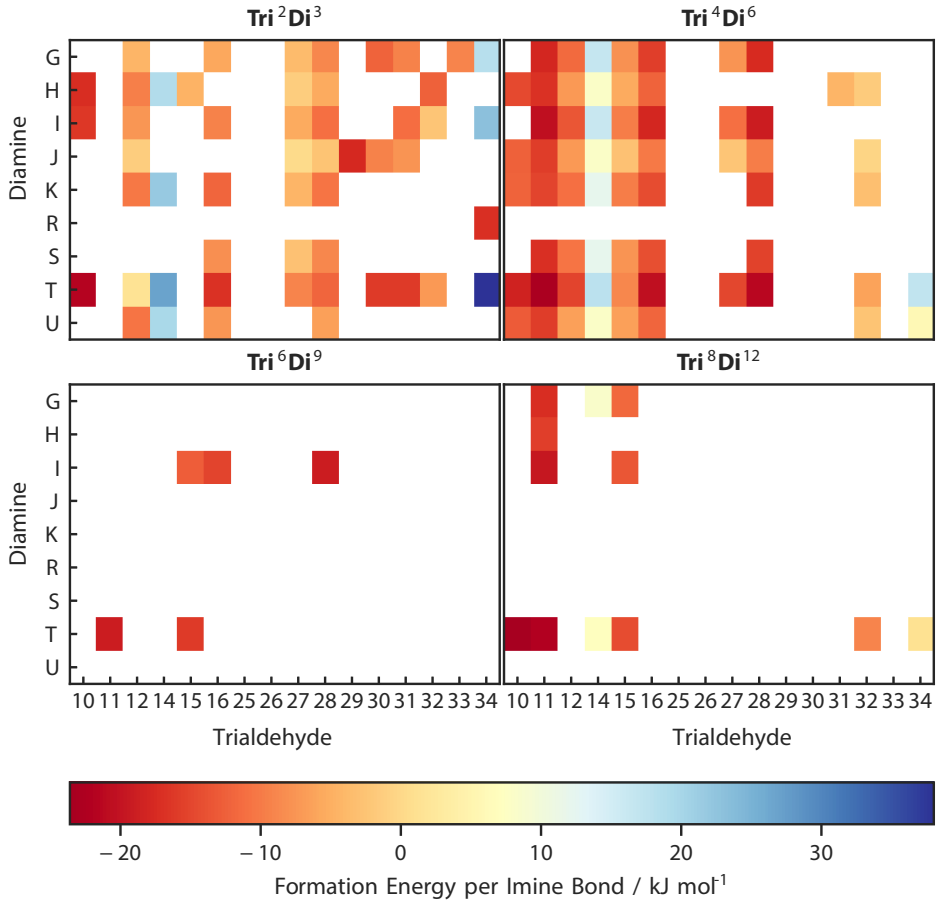


Figure 5.24: Colour map showing the formation energy per imine bond for each precursor combination in four different topologies, each calculated at the PBE/TZVP+D3 level. Cooler colours correspond to higher, and therefore, less stable formation energies, whereas warmer colours correspond to more negative, favourable formation energies. White squares represent cages that were collapsed following the DFT optimisation process, calculated as those cages with a cavity size of less than 1 Å. Larger **Tri**^{6}**Di**^{9} and **Tri**^{8}**Di**^{12} were also manually inspected for shape persistency.

to **Tri**^{2}**Di**^{3}, **Tri**^{6}**Di**^{9}, and **Tri**^{8}**Di**^{12} cages, with 63 cages predicted to have cavity diameters greater than 1 Å for **Tri**^{4}**Di**^{6} cages, compared with 52, 5, and 12 cages for the other topologies, respectively. For the smaller **Tri**^{2}**Di**^{3} the absence of porosity can be attributed to the small capsule-like structure many of these cages form, resulting in a small pore size. For larger topologies, the increased flexibility afforded by the greater number of atoms would be more likely to result in a collapsed conformation. This increased propensity for larger topologies to collapse was also observed by Turcani *et al.*, as discussed in Chapter 4.[89]

Starting with the **Tri**^{2}**Di**^{3} topology, diamine **10** was found to have the most favourable formation energy, with values ranging from -16 to -22 kJ mol^{-1} with trialdehydes **H**, **I**, and **T**. This prediction was experimentally confirmed as a [2 + 3] species was observed with 8 of the 9 trialdehydes with diamine **10**, supporting the *gem*-dimethyl effect described in the previous section. For **H10** and **T10**, the **Tri**^{2}**Di**^{3} topology had lower formation energy than the **Tri**^{4}**Di**^{6} topology (-17 compared with -15 kJ mol^{-1} and -22 compared with -18 kJ mol^{-1}). On the other hand, for **I10**, the **Tri**^{4}**Di**^{6} topology had a lower formation energy (-20 compared with -16 kJ mol^{-1}) and was observed to form in a mixture of the two topologies. Other highly favourable combinations for this topology include **J29**, **T16**, **T30**, **T31**, and **R34**. With the exception of **J29** and **T16**, all of these combinations were experimentally observed to form a [2 + 3] species.

Looking next at the **Tri**^{4}**Di**^{6} topologies, diamines **11**, **12**, **15**, **16**, and **28** would appear best for synthesising a shape persistent POC, based on the number of shape persistent cages predicted and their relatively favourable formation energies. Simulations indicated that each of these diamines was predicted to form a shape persistent **Tri**^{4}**Di**^{6} cage with at least 6 of 9 of the trialdehydes. While **11** and **12** did not form many cages experimentally (which will be discussed later), 4 [4 + 6] species formed with diamine **16**, 1 with diamine **15**, and 2 with diamine **28**, confirming the usefulness of these diamines in forming a shape persistent **Tri**^{4}**Di**^{6} cage. Particularly favourable combinations in this topology include **T28**, **T16**, **I28**, **I16**, **G28**, **K28**, **G16** (**CC3**), and **S28**, each with formation energies below -15 kJ mol^{-1}. Among these combinations, **I28**, **I16**, **G16**, and **S28** were observed to form [4 + 6] species in solution, and each was predicted to form a more stable **Tri**^{4}**Di**^{6} topology when compared to the **Tri**^{2}**Di**^{3} topology. Combination **G16** exhibited the largest difference in formation energy at -16 kJ mol^{-1} for the **Tri**^{4}**Di**^{6} topology compared to -6 kJ mol^{-1} for the **Tri**^{2}**Di**^{3} topology. This difference was -19 kJ mol^{-1} compared to -11 kJ mol^{-1} for **I28**, -18 kJ mol^{-1} compared to -9 kJ mol^{-1} for **I16**, and -15 kJ mol^{-1} compared to -9 kJ mol^{-1} for **S28**.

Out of the larger **Tri**^{6}**Di**^{9} cages, only five were predicted to be shape persistent. However, each of them had a favourable formation energy lower than -12 kJ mol^{-1}. Among these five, two cages (**I15** and **I28**) were observed experimentally, but they formed as a mixture with other topologies. On the other hand, for the **Tri**^{8}**Di**^{12} topology, 12 cages were predicted to be shape persistent after

high temperature MD, but only 9 had a favourable formation energy of lower than -9 kJ mol^{-1}. Of these 9 cages, only **I15** was observed experimentally forming as a mixture with other topologies.

In general, the formation energies were able to predict that cage formation with diamines **14** and **34** was unlikely. Diamine **14** did not form any cages, while diamine **34** formed an impure cage with trialdehdye **R**. The positive formation energies for both diamines, shown in Figure 5.24, indicate highly unfavourable reactions. However, the formation energy for the reaction between diamine **34** and trialdehyde **R** for the [2 + 3] species was an exception, with a value of -17 kJ mol^{-1}, indicating a favourable reaction. Indeed, this topology was observed experimentally.

In contrast, the formation energies were unable to predict the absence of reactivity for diamines **11** and **12**, which were the two most sterically hindered diamines. Although the formation energies for these two diamines were mainly favourable, with an average of -17 kJ mol^{-1} for diamine **11** across all the shape persistent cages, no cage formed experimentally with diamine **11** and only one formed with diamine **12**. Each of these diamines has a sterically hindered phenyl substituent attached to a 1,2-diamine. The effect of these sterically hindered functionalities on cage formation cannot be captured using a computational analysis of the single molecule, and it is challenging to understand the precise effects without investigating the intermediates to cage formation.

Cavity Sizes of Experimentally Formed Cages Figure 5.25 shows the cavity sizes of cages that were observed experimentally, including cages that formed mixtures of topologies, and each of the shape persistent cages that formed are shown in Figure 5.26. As expected many of the $\mathbf{Tri}^2\mathbf{Di}^3$ cages were predicted to have smaller cavity diameters of less than 2.89 Å. However, **G27** and **I27** possessed slightly larger cavity sizes of 4.2 Å due to the increased size of the diamine linker, and each was observed in a mixture with a [4 + 6] species. Meanwhile, **U28** formed a larger [2 + 3] with a predicted cavity diameter of 6.8 Å due to the large size of the trialdehyde and diamine linker.

Many of the $\mathbf{Tri}^4\mathbf{Di}^6$ cages were larger, each having cavity diameters in excess of 5 Å. **G16** (**CC3**), **I16**, **J32**, and **S16** possessed cavity sizes between 5-6.1 Å, a value POCs frequently exhibit in the literature.[115] In contrast, **G27** and **I15** were predicted to be slightly larger with cavity diameters of 8.1 and 8.4 Å, respectively. The other four cages predicted to have cavity diameters in excess of 10 Å were **I27**, **I28**, **S28**, and **K16**, which are shown in Figure 5.26 alongside their respective cavity sizes. Of the larger $\mathbf{Tri}^6\mathbf{Di}^9$ topology, only **I15** had a large cavity diameter of 11.8 Å, but **I28** had a smaller spherical cavity diameter of 4.5 Å due to the tubular structure. Finally, only **I15** was observed to form a shape persistent cage in the largest $\mathbf{Tri}^8\mathbf{Di}^{12}$ topology with the largest cavity size of 15.3 Å.

From examining the results of shape persistent cages and their cavity diameters, some general design rules can be determined. Larger cavity sizes are formed from diamine linkers with a larger

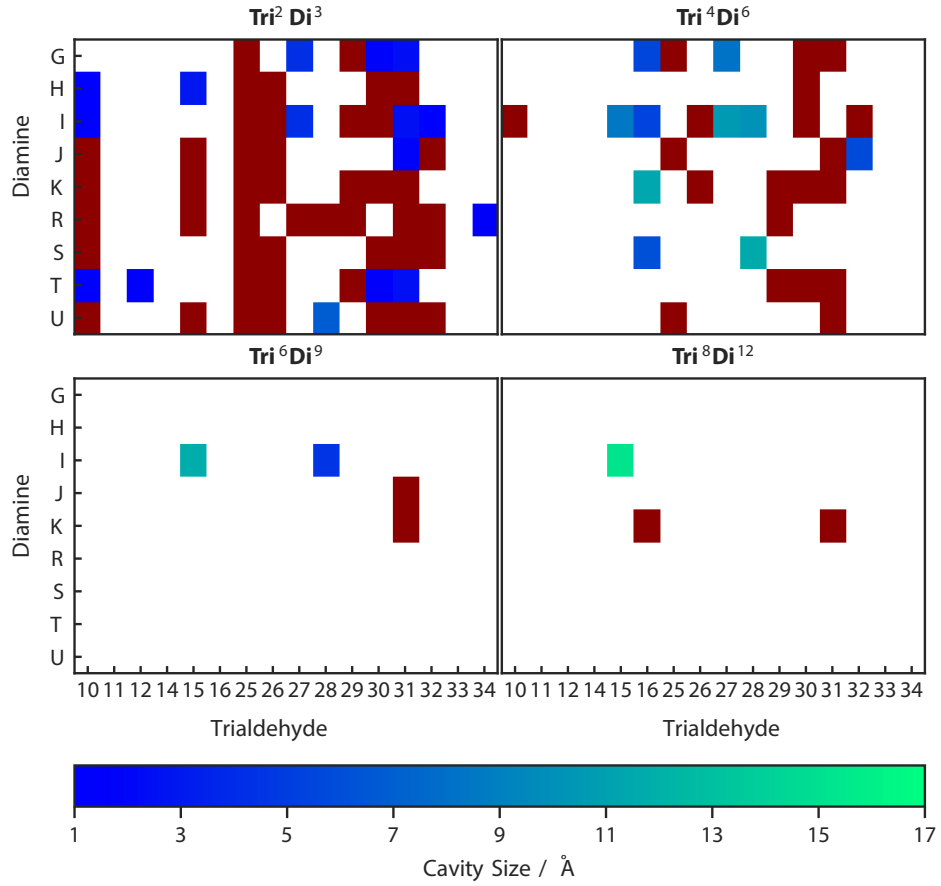


Figure 5.25: Colour map showing the cavity diameter of experimentally formed cages optimised at the PBE/TZVP+D3 level, calculated using *pyWindow*.[10] Only cages with cavity diameter above 1 Å are shown. Green represents larger cavity sizes, and blue represents smaller cavity sizes. White squares represent cages that did not form experimentally, and red squares represent cages that formed experimentally but *pyWindow* calculates as having no internal cavity.

distance between amine functional groups, but they must be connected by a relatively rigid core, such as diamines **27** and **28** which possess a cyclohexane and phenyl core, respectively. Using a smaller bond angle between amine functional groups, such as in 1,2-diamines, is more likely to produce a shape persistent cage, albeit with a smaller cavity size. Instead of using a 1,2-diamine such as diamine **16**, a 1,3-diamine can be used to increase the cage's cavity size, as demonstrated in **I15**. While increasing the bond angle may favour larger topologies, as suggested by Teng *et al.*,[189] this precursor resulted in fewer shape persistent cages overall.

Looking at the results as a whole, the most promising cages from the high-throughput screen are shown in 5.26, each with cavity diameters above 2.89 Å and predicted to be porous. While **S16** and **I16** were the only larger [4 + 6] species formed cleanly using the reaction conditions selected

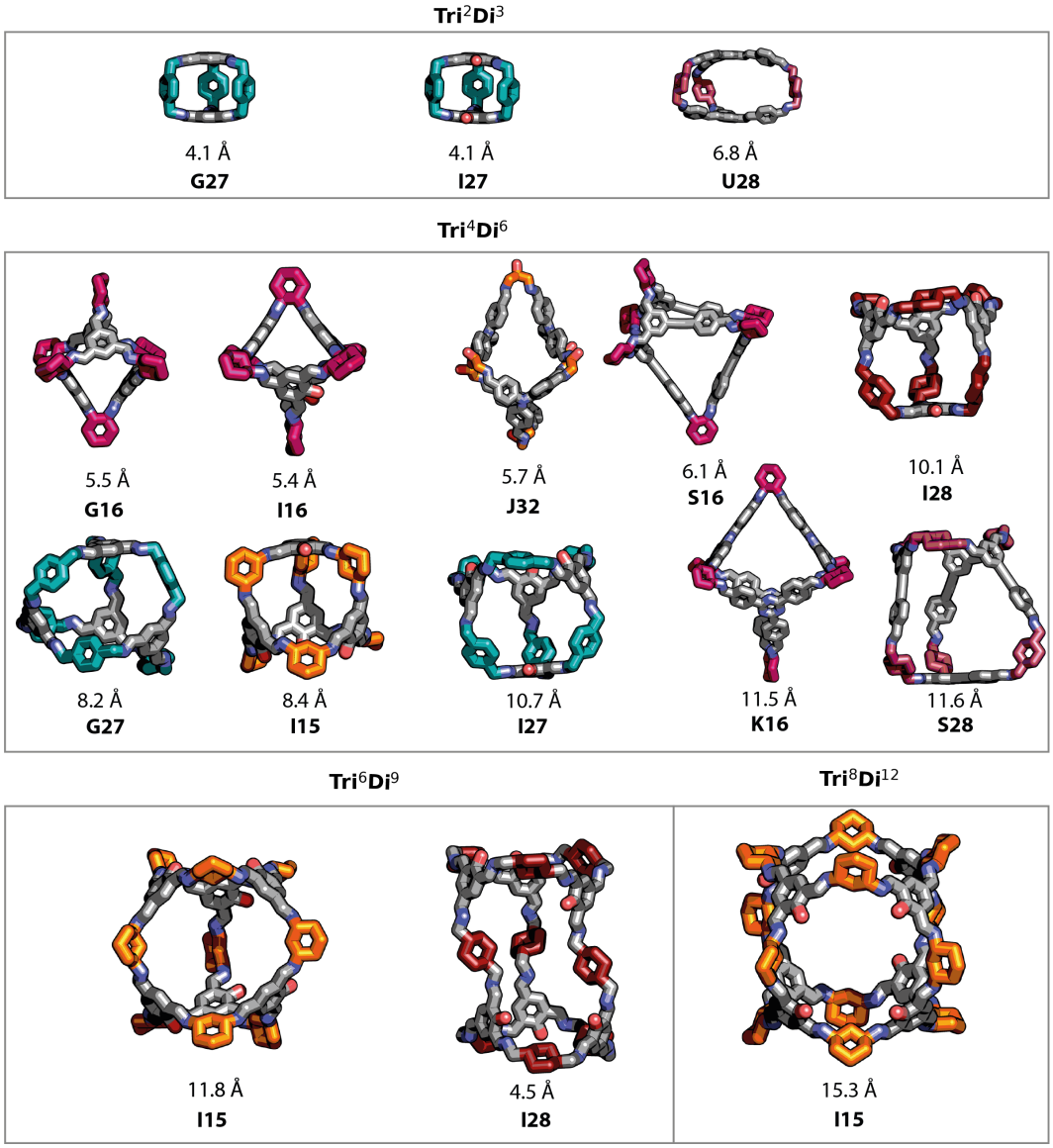


Figure 5.26: Experimentally observed cages predicted to shape persistent following optimisation at the PBE/TZVP+D3 level, with cavity diameters in excess of 2.89 Å. Cavity diameters are shown below each cage label. Combinations **I28**, **I27**, **G27**, and **I15** formed at least two topologies predicted to be shape persistent, each of which are shown.

for this high-throughput screen, optimising the reaction conditions for each of the candidates that formed a mixture of topologies could result in complete conversion to these candidates predicted to be shape persistent.

5.4 Conclusions

In this work, I demonstrate the potential of combining HTE and automated reaction evaluation to accelerate cage discovery. By using HTE to combine 9 trialdehdyes and 16 diamines, I was able to identify 17 combinations that showed clear formation of a single cage topology, and a further 51 that showed evidence of forming at least one topology. The development of an automated reaction workflow allowed for the rapid assessment of reaction outcomes, allowing me to screen cage hits much faster than would be possible through manual inspection. For HTE, reaction analysis can be a significant bottleneck, which becomes significant when performing hundreds of reactions in parallel. The automated reaction analysis workflow also provided insights into the kinetic intermediates that formed during the self-assembly process.

Out of the 68 combinations that resulted in cage formation, 2 [4 + 6] species formed cleanly with predicted cavity diameters of around 5 Å, 15 formed [2 + 3] species cleanly, while another 51 formed cages in either a mixture of topologies or single topologies with varying amounts of conversion. The relatively low number of precursor combinations (10) that resulted in a shape persistent cage across all of the categories highlights the challenges of the POC discovery process. Typically, the majority of candidates would either collapse or not result in cage formation at all.

For the shape persistent cages that did form, relative energies were able to rationalise the single observed topological outcome. In cases where a shape persistent [4 + 6] species was formed, a large energy difference of over 30 kJ mol^{-1} was typically observed, as seen in combinations **S16**, **I16**, and **S28**. For mixtures of topologies forming, this energy difference was lower, such as in combinations **I15** and **J32**, rationalising the formation of a mixture of topologies. In reality, entropic factors and reaction conditions, such as solvent, are likely to play a large role in driving the formation of different topologies, which are unable to be accounted for by comparing the relative internal energies between topologies.[31] Despite these limitations of the simulations employed, the trends across formation energies were somewhat predictive. For example, cages containing diamines **14** and **34** were predicted to be unfavourable to form, and, indeed, resulted in the lowest number of cages formed.

Combining computational simulations with reaction outcome prediction was able to not only provide information on the relative stability of each topology, but enabled prioritisation of candidates predicted to be shape persistent. For the six combinations where at least one shape persistent topology was observed to form in a mixture with other topologies (**G27**, **I28**, **I27**, **K16**, **I15**, and **J32**), altering the reaction conditions could favour the formation of a single shape persistent topology, and would make these interesting candidates for further scale-up and optimisation.

The next steps for the automated high-throughput analysis workflow would be to expand the number of reactions performed and analysed, either by systematically exploring reaction condi-

tions that were not tested in this work or by incorporating different precursors into the reaction screening. Furthermore, the use of ML algorithms could be incorporated into the workflow, by training on a large dataset of known cage structures and reaction outcomes, facilitating a data-driven prediction about the likelihood of a given reaction leading to a successful cage formation.

6 Conclusions and Future Work

The broad goal of my PhD was to combine computational and experimental efforts to accelerate the discovery of POCs. Towards this aim, I developed a binary classification model in Chapter 3, called the Materials Precursor Score (MPScore), which was able to categorise a POC precursor as easy- or difficult-to-synthesise. The objective of this model was to replicate the decision-making process of expert chemists when selecting cage precursors, which met the synthesis requirements for a typical POC discovery programme. To train the model, I generated a dataset by posing the question, "Can you synthesise 1 g of this molecule in fewer than 5 steps?", to three expert synthetic chemists, and investigated approaches to train and evaluate the ML model using imbalanced datasets. In this chapter, I compared the performance of the developed MPScore model with other quantitative computational measures of predicting the synthetic difficulty of small organic molecules, showing how other measures do not capture the specific synthetic requirements of precursors for materials discovery. For an economically viable, scalable new material, precursors are required in high yields from few robust reaction steps, and for a cheap commercial cost.

With the MPScore model in hand, I then applied it to a POC computational screening workflow in Chapter 4. Here, I used the SAScore, SCScore, and MPScore synthetic difficulty models to filter a library of 7,190 diamines and 98 trialdehydes, using these easy-to-synthesise precursors to construct 21,138 hypothetical cages in the $\mathbf{Tri}^4\mathbf{Di}^6$ topology. Following the computational screening workflow, I was able to identify 1,072 shape persistent cages from precursor combinations that were predicted to be easy-to-synthesise by each of the synthetic difficulty scores. Of these shape persistent cages, 7 were predicted to possess a central cavity diameter larger than 16 Å, a cavity size rarely exhibited in the literature. By analysing the structural features of precursors that formed shape persistent cages, I was able to identify trialdehyde **U** as a promising precursor, a *trans* alkene trialdehyde present in many of the cages with a large internal cavity diameter.

In Chapter 5, I synthesised two trialdehydes (**U** and **R**) as part of a library of 16 diamines and 9 trialdehydes that would be combined using high-throughput experimentation. I developed an automated reaction analysis workflow to quickly determine the outcome of each reaction, identifying whether each pair of precursors cleanly formed a single cage topology or a mixture of topologies, partially converted to a single topology or mixture of topologies, or whether no cage formation occurred. Altogether, I observed the clean formation of 17 cages with a single topology, 11

cages that formed a mixture of topologies cleanly, and a further 40 precursor combinations that showed evidence of cage formation. Of the 68 total combinations that showed evidence of cage formation, 10 were predicted to form a POC in a topology predicted to be shape persistent. Future work could involve the scale-up and property analysis of these promising candidates.

The relatively low percentage of the 144 precursor combinations that formed a shape persistent POC (7%) is representative of many POC discovery workflows, which typically have a low success rate.[48] Despite this low percentage, increasing the throughput of reactions that can be performed in parallel using high-throughput experimentation, and analysing the reaction outcome autonomously, has the potential to rapidly accelerate the number of precursors that are tested for POC formation.

Computation also proved useful to rationalise the observed topologies based on the thermodynamic driving force between different topologies, and to identify cages that formed experimentally that would remain shape persistent. Using DFT calculations, I was able to rationalise the formation of a mixture of topologies for precursor combination **I15**, where the difference in energy per formula unit of the experimentally observed **Tri**^{4}**Di**^{6}, **Tri**^{6}**Di**^{9}, and **Tri**^{8}**Di**^{12} topologies was less than 30 kJ mol^{-1}. Similarly, I was able to rationalise the formation of the single **Tri**^{4}**Di**^{6} topology formed with precursor combinations **G16** and **I16**, which both were predicted to be lowest in energy. However, there were limitations of the computational calculations, most notably the inability to account for many experimental factors that would influence cage formation. Factors such as the formation of stable kinetic intermediates, solvent, and the formation of insoluble species can either inhibit or promote the formation of specific cage topologies. These factors pose computational challenges when attempting to account for them accurately. In addition, I did not account for the number of different structural isomers that could form from asymmetric precursors. For these asymmetric precursors where a shape persistent cage was observed to form, a full analysis of the different structural isomers could help elucidate the isomer or statistical distribution of isomers that are experimentally observed.

To help address some of the challenges of computationally predicting the outcome of imine condensation reactions, work is underway in collaboration with Annabel Basford to curate a large database of precursor combinations and their reaction outcomes, to train a data-driven model for predicting the formation of cages. By combining my library of 144 precursor combinations with an additional 222 precursor combinations, a library of 366 precursor combinations will be assembled. Using this larger precursor library, it would be easier to identify structural features of precursors that are more likely to form a shape persistent cage, working further towards the goal of being able to predict whether a shape persistent POC will form from a set of precursors. Using the HTE workflow described in Chapter 5, the number of reactions performed in parallel can also be scaled up to explore the effect of altering reaction conditions, such as changing solvents, reaction

temperatures, and concentrations, to better understand how external factors drive cage formation. Indeed, precursor combinations that were found to be unsuccessful in my high-throughput screen might result in cage formation under a different set of experimental conditions.

In this thesis, I have shown that combining computational screening with high-throughput automation can assist in not only proposing POC new precursors that are easy-to-synthesise, but can be used to predict whether those precursors will form a shape persistent POC *in silico*. While a large number of hypothetical POCs were predicted to be shape persistent computationally, the current bottleneck remains in being able to predict the reaction outcome of a set of precursors. Nonetheless, my thesis contributes to accelerating POC discovery, establishing a foundation to rapidly perform and evaluate imine condensation reactions, in addition to the computational prediction of promising precursors that could form a shape persistent POC. As computationalists and experimentalists continue to refine and expand these methods, the gap between computational predictions and synthetic realisation will continue to diminish, paving the way for accelerated materials discovery.

6.1 Future Work

There are multiple avenues of research that could be investigated following the work in my thesis. Firstly, the MPScore described in Chapter 3 could be extended to precursors of other materials with similar synthetic requirements, including COFs, polymers, or even organic linkers for MOFs. For this to be achieved, a larger, more diverse training library of molecules would need to be constructed, canvassing a wider region of chemical space. Obtaining more training data from more chemists would also allow a consensus to be obtained, facilitating a more accurate prediction of synthetic difficulty. The high-throughput computational screen in Chapter 4 proved fruitful in proposing POC precursors that were predicted to be easy-to-synthesise and form a shape persistent cage. More precursors could be included in the initial library before screening for synthetic difficulty to increase the diversity of hypothetical POCs screened. Alternatively, the MPScore model can be used to bias towards more synthesisable precursors in generative computational workflows. By using the MPScore as part of a multi-objective optimisation function, a greater diversity of synthesisable POC precursors could be generated using evolutionary algorithms, reinforcement learning, or generative adversarial networks.

Finally, the workflow in the final chapter allows for the systematic generation and analysis of large quantities of reaction data. By expanding the number of precursor combinations and reaction conditions explored, and uploading this data to a central database, a ML model could be trained to predict what the reaction outcome will be when combining a specific set of precursors in specific experimental conditions. In the future, it is possible to envisage a similar reaction anal-

ysis technique being applied to the closed-loop optimisation of POCs. Here, an algorithm would determine what changes to make to this reaction to maximise the yield of shape persistent POCs.

7 Experimental Data

7.1 General Synthetic and Analytical Methods

7.1.1 Synthesis

All reactions were heated, when required, using a stirrer hotplate with a heating block, using an external probe attached to the heating block. Stirring was conducted using Teflon-coated magnetic stirrer bars, and residual solvent was removed using a rotary evaporator.

7.1.2 HRMS

High resolution mass spectrometry (HRMS) was carried out using Agilent Technologies 6545XT AdvanceBio LC/Q-TOF spectrometer. For each reaction in the high-throughput screen, a 20 μL portion of the reaction mixture was transferred into 980 μL of 50:50 MeOH:DCM. The mobile solvent was MeOH + 0.1% formic acid at a flow rate of 0.25 mL/min, capillary voltage 4000 V, fragmentor 225 V, in positive ion mode.

7.1.3 NMR

^{1}H and ^{13}C NMR spectra for precursor synthesis were recorded on a Bruker Avance-400 (400 MHz/ 101 MHz) at ambient probe temperature, using an internal deuterium lock ($CDCl_3$ - δ = 7.26 ppm or $(CD_3)_2SO$ δ = 2.52ppm for ^{1}H NMR; and $CDCl_3$ - δ = 77.16 ppm or $(CD_3)_2SO$ - δ = 39.52 ppm for ^{13}C NMR). For conducting high-throughput NMR experiments, the machine was equipped with a Bruker SampleJet autosampler, which allows for queuing of up to 5 blocks of 96 samples each. NMR tubes for the SampleJet had a height of 103.5 mm and a diameter of 3.0 mm.

Data are presented in the form: chemical shift, integration, peak multiplicity (s = singlet, d = doublet, t = triplet, q = quartet, qt = quintet, m = multiplet, br = broad), and coupling constants, followed by assignment. Chemical shifts are presented in ppm, relative to δ_{CDCl_3}, δ_{TMS}, or $\delta_{(CD_3)_2SO}$, as appropriate. Coupling constants are given in Hz and are referenced to the proton with the highest chemical shift.

7.1.4 IR

Infrared spectra were recorded using an Agilent Technologies Cary 630 FTIR spectrometer, across a 4000-400 cm^{-1} range, in ATR mode, using a solid sample with compression, 32 background scans, a resolution of 16 and a Happ-Genzel Apodization.

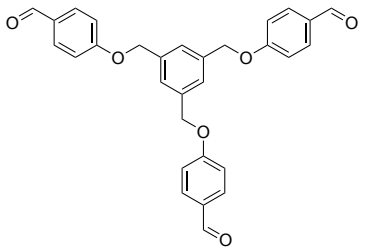
7.1.5 High-Throughput Experimental Details

The high-throughput experiment was conducted using the *Opentrons OT-2* liquid handling platform, with a gantry speed of 100 ms^{-1}, plunger flow rates (for aspirating and dispensing) of 40 μLs^{-1} using the 300 µL *Opentrons* proprietary pipette. Before transferring a new substance, 100 µL of the substance was aspirated into the pipette and lifted to the top of the vial, followed by a 10 s delay to swell the tip, which was found to reduce the dripping during transfer across the deck. A maximum amount of 230 µL of a substance was transferred at a time, using a 15 µL air gap. Where more than 230 µL of a substance was required, multiple transfers were made. The code required to replicate the protocol is available at `https://github.com/GreenawayLab/Opentrons`.

7.2 Chapter 6 Experimental

7.2.1 Synthesis of Trialdehydes

Synthesis of Trialdehyde R
4,4′,4′′-((Benzene-1,3,5-triyltris(methylene))tris(oxy))tribenzaldehyde



This synthesis employed the procedure by Lee *et al.*[190] To a stirred solution of 1,3,5-tris (bromomethyl) benzene (500 mg, 1.4 mmol, 1.0 eq.) in THF (25 mL) was added 4-hydroxy benzaldehyde (546 mg, 4.5 mmol, 3.2 eq.) and anhydrous potassium carbonate (1.36 g, 9.8 mmol, 7 eq.). The mixture was refluxed at 70 °C for 24 h before being allowed to cool to room temperature. The mixture was concentrated *in vacuo* and partitioned between saturated aqueous $NaHCO_3$ solution and chloroform (3 x 25 mL). The resulting combined chloroform layers were dried using $MgSO_4$ and filtered. Chloroform was removed by rotary evaporation from the resulting mixture to afford the desired product as a colourless solid (547 mg, 1.1 mmol, 81%).

IR: (υ_{max}/cm^{-1}) 1685 (**C=O**), 1583 (Ar**C**-Ar**C**), 1250 (Ar**C**-**O**), 1213 (Ar**C**-**O**), 1156 (**C**-**O**); **^{1}H NMR:** (400 MHz, $CDCl_3$) δ_H 9.89 (3H, s, O=C**H**), 7.84 (6H, d, Ar**H**), 7.50 (3H, s, Ar**H**) 7.07 (6H, d, O=C**H**), 5.19 (6H, s, O-C**H**); **^{13}C NMR:** (101 MHz, $CDCl_3$) δ_C 190.78 (O=**C**), 163.43 (Ar**C**), 137.31 (Ar**C**), 132.06 (Ar**C**), 130.33 (Ar**C**), 126.24 (Ar**C**), 115.13 (Ar**C**), 69.74 (O-**C**); **HRMS:** (ES-) calculated for $C_{30}H_{24}O_6$ 480.1573; found $[M+Cl]^-$ 515.1275.
Data in accordance with literature values.[190]

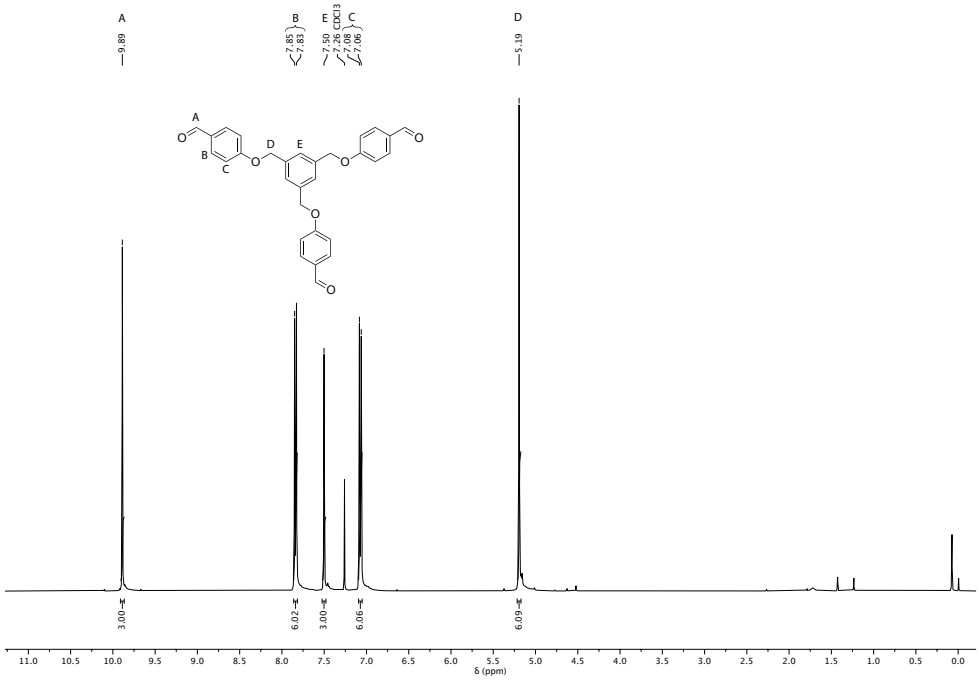


Figure 7.1: ^{1}H NMR spectrum ($CDCl_3$) of trialdehyde **R**.

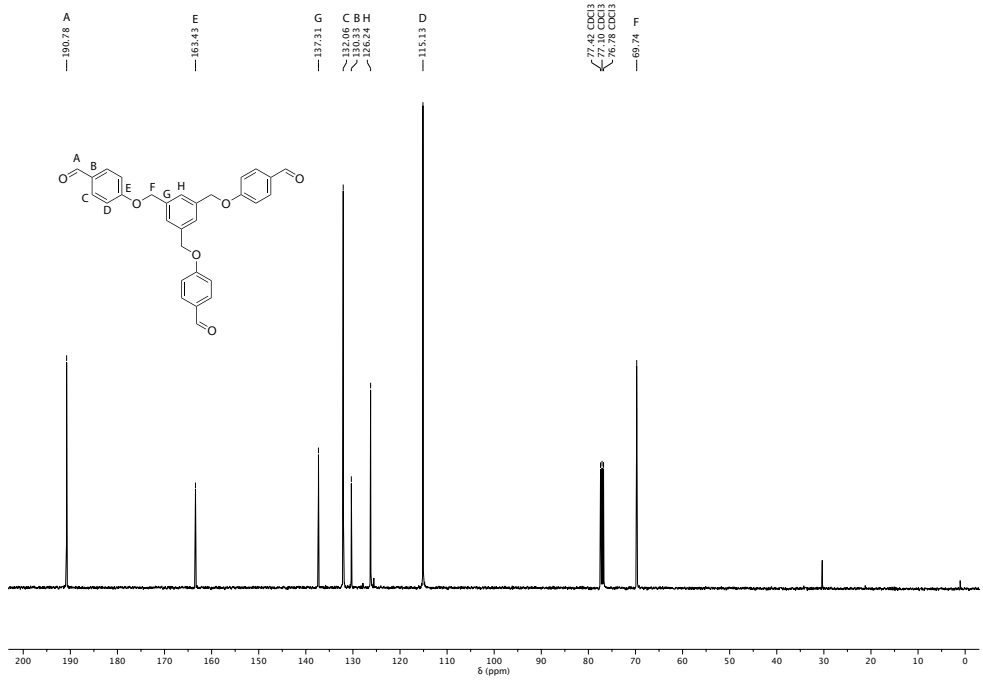
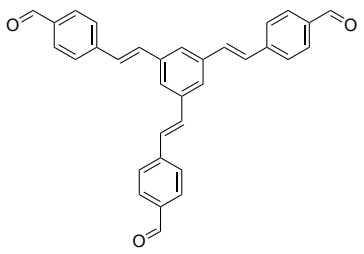


Figure 7.2: ^{13}C NMR spectrum ($CDCl_3$) of trialdehyde **R**.

Synthesis of Trialdehyde U

4,4′,4′′-((1*E*,1′1*E*,1′′*E*)-Benzene-1,3,5-triyltris(ethene-2,1-diyl))tribenzaldehyde



This synthesis employed the procedure by Díez-Barra *et al.*[212] To an oven-dried and degassed round-bottomed flask equipped with stirrer bar was added 1,3,5-tris(bromomethyl) benzene (2.00 g, 5.6 mmol, 1.0 eq.) and triethyl phosphite (12 mL, 70.0 mmol, 12.5 eq.). The resulting solution was stirred at 150 °C under N_2 atmosphere. After cooling to room temperature, the excess triethyl phosphite was removed under reduced pressure to yield a viscous orange oil.

The mixture was cooled to 0 °C using an ice bath followed by the addition of 4-(diethoxymethyl) benzaldehyde (3.50 g, 16.8 mmol, 3.0 eq.). To the resulting solution was slowly added potassium tert-butoxide (5.66 g, 50.4 mmol, 9.0 eq.) under an inert atmosphere at 0 °C. The mixture was stirred for 12 h at room temperature. To the resulting mixture was added 1M HCl (40 mL) and stirred for additional 20 minutes at room temperature. The mixture was extracted with DCM (3 x 30 mL) and washed with brine (2 x 30 mL). The resulting solution was dried *in vacuo* to afford a yellow solid. The crude product was recrystallised using 1:1 chloroform:ethanol (140 mL), before being dried in a vacuum oven at 50 °C to yield the pure product as a yellow solid (2.02 g, 4.3 mmol, 77%).

IR: (υ_{max}/cm^{-1}) 1687 (**C**=**O**), 1590 (**C**=**C**), 1160 (Ar**C**-Ar**C**); **^{1}H NMR:** (400 MHz, $(CD_3)_2SO$) δ_H 10.02 (3H, s, O=C**H**), 7.98-7.87 (15H, m, Ar**H**), 7.84 (6H, s, C=C**H**); **^{13}C NMR:** (101 MHz,$(CD_3)_2SO$) δ_C 197.71 (O=**C**), 148.16 (Ar**C**), 142.79 (Ar**C**), 140.45 (Ar**C**), 136.54 (C=**C**), 135.36 (Ar**C**), 133.56 (C=**C**), 132.32 (Ar**C**), 130.57 (Ar**C**); **HRMS:** (ES-) calculated for $C_{33}H_{24}O_3$ 468.1730; found $[M+Cl]^-$ 503.1422.

Data in accordance with literature values.[212]

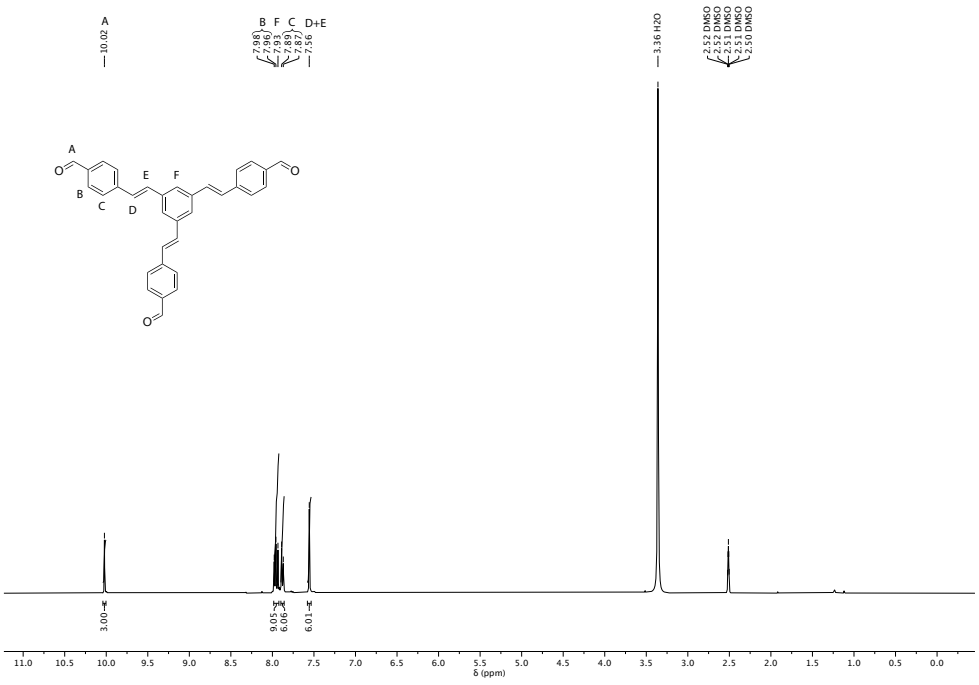


Figure 7.3: ^{1}H NMR spectrum ($(CD_3)_2SO$) of trialdehyde **U**.

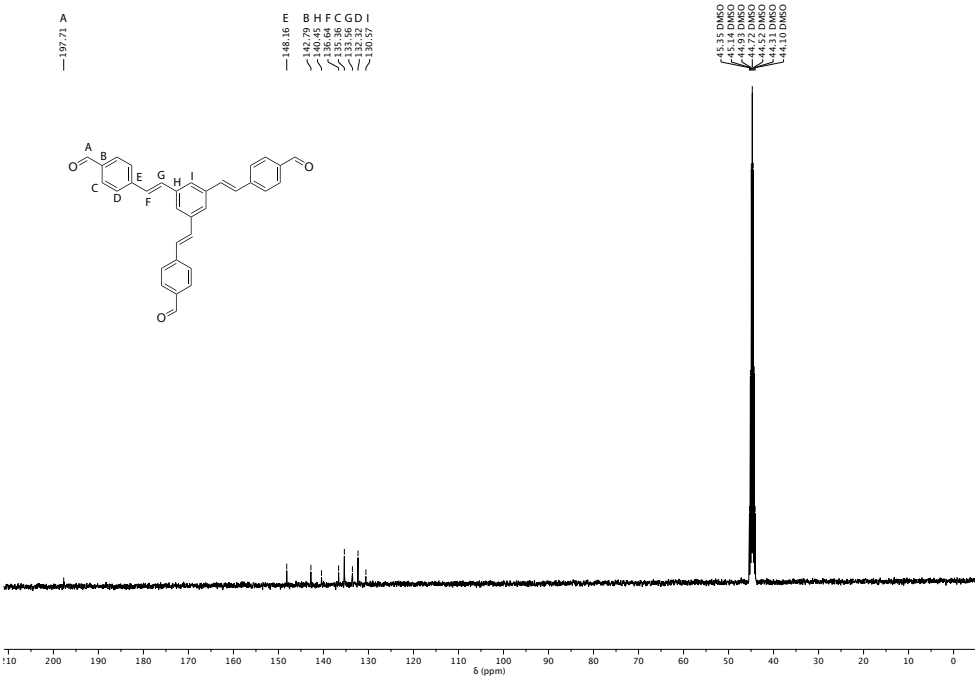


Figure 7.4: ^{13}C NMR spectrum ($(CD_3)_2SO$) of trialdehyde **U**.

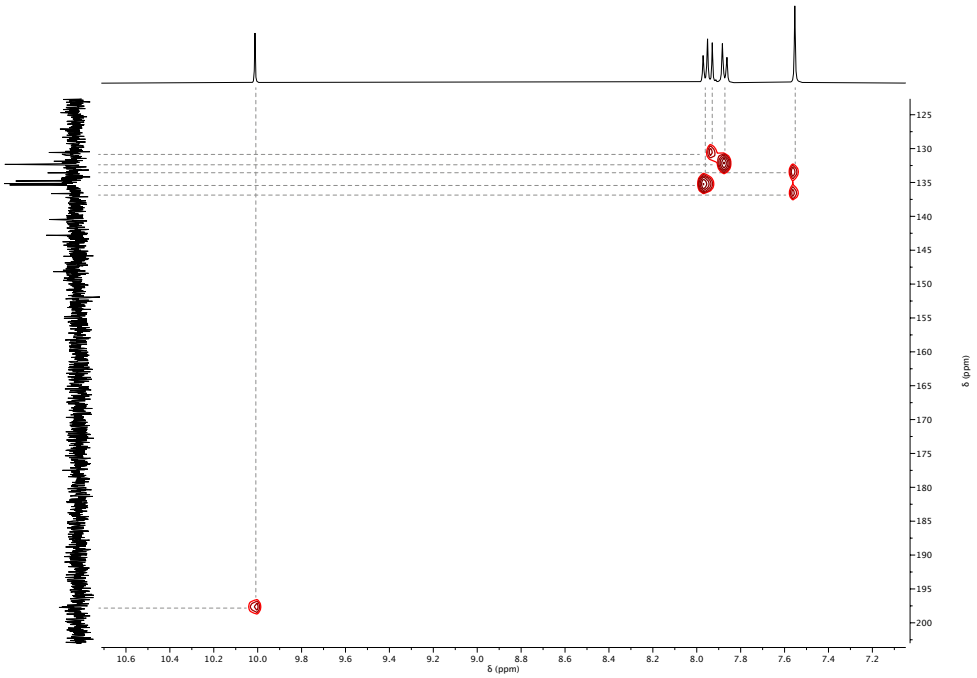


Figure 7.5: ^{1}H-^{13}C HSQC spectrum ($(CD_3)_2SO$) of trialdehyde **U**.

7.2.2 High-Throughput Cage Screen

General Procedure:

This high-throughput screen employed an adapted procedure to that of Greenaway *et al.*[48] Stock solutions of all trialdehyde and diamine precursors were prepared in chloroform (5 mg/mL, 5mL) (Table 7.1). To the stock solution containing diamine **33**, triethylamine (TEA) (94 µL, 0.68 mmol, 4.0 eq.) was added. Using the *Opentrons OT-2* reaction platform, the required volume of each trialdehyde (0.0031 mmol, 4.0 eq.) and diamine (0.0063 mmol, 8.0 eq.) was added to vials using liquid dispensing. To each vial, chloroform was added to make the total volume 1 mL (Table 7.2). The vials were then sealed in the reaction plate, transferred to a stirring deck, and stirred at room temperature for 5 days. Following the reaction, the reaction plate was unsealed and each vial underwent a manual turbidity check. The remaining solvent was removed under a constant flow of nitrogen using the EquaVAP 48-well evaporator, before being redissolved in $CDCl_3$ (0.7 mL) for subsequent analysis.

For HRMS analysis, samples were prepared using the *Opentrons* platform by transferring 20 µL from each of the vials into mass spectrometry vials. An additional 980 µL of 50:50 MeOH:DCM solution was transferred to each of the mass spectrometry vials. NMR samples were prepared

manually by pipetting the remaining reaction solution from each reaction vial into NMR tubes compatible with the Bruker SampleJet autosampler.

Table 7.1: Precursor stock solutions used in the high-throughput screen, each dissolved in chloroform.

Reactant	MW / g mol^{-1}	Stock Solution Concentration / mg mL^{-1}	Stock Solution Concentration / mmol mL^{-1}
Trialdehyde G	162.14	5	0.031
Trialdehyde H	390.44	5	0.013
Trialdehyde I	178.14	5	0.028
Trialdehyde J	329.36	5	0.015
Trialdehyde K	393.40	5	0.013
Trialdehyde R	480.52	5	0.010
Trialdehyde S	262.26	5	0.019
Trialdehyde T	210.14	5	0.024
Trialdehyde U	468.55	5	0.011
Diamine 10	102.18	5	0.049
Diamine 11	191.28	5	0.026
Diamine 12	212.30	5	0.024
Diamine 14	100.08	5	0.050
Diamine 15	114.19	5	0.044
Diamine 16	60.10	5	0.083
Diamine 25	215.38	5	0.023
Diamine 26	116.21	5	0.043
Diamine 27	136.20	5	0.037
Diamine 28	142.25	5	0.035
Diamine 29	145.25	5	0.034
Diamine 30	104.15	5	0.048
Diamine 31	136.20	5	0.037
Diamine 32	90.13	5	0.055
Diamine 33[a]	146.19	5	0.034
Diamine 34	114.19	5	0.044

(a) Diamine 33 was purchased in trihydrochloride salt form of MW 182.65 g mol^{-1}. To this solution, triethylamine (TEA) (95 µL, 0.68 mmol, 4 eq.) was added.

Table 7.2: Reaction volumes used for each combination of trialdehyde and diamine in the high-throughput screen.

Cage	Trialdehyde Reactant	Amount of Trialdehyde / mmol	Volume of Trialdehyde / µL	Diamine Reactant	Amount of Diamine / mmol	Volume of Diamine / µL	Volume of Chloroform / µL
G10	G	0.0031	101	10	0.0062	127	772
G11	G	0.0031	101	11	0.0062	238	661
G12	G	0.0031	101	12	0.0062	264	635
G14	G	0.0031	101	14	0.0062	125	774
G15	G	0.0031	101	15	0.0062	142	757
G16	G	0.0031	101	16	0.0062	74	824
G25	G	0.0031	101	25	0.0062	268	631
G26	G	0.0031	101	26	0.0062	145	754
H10	H	0.0031	240	10	0.0061	126	634
H11	H	0.0031	240	11	0.0061	235	525
H12	H	0.0031	240	12	0.0061	261	499
H14	H	0.0031	240	14	0.0061	123	637
H15	H	0.0031	240	15	0.0061	140	620
H16	H	0.0031	240	16	0.0061	73	686
H25	H	0.0031	240	25	0.0061	265	495
H26	H	0.0031	240	26	0.0061	143	617
I10	I	0.0031	109	10	0.0061	125	766
I11	I	0.0031	109	11	0.0061	234	657
I12	I	0.0031	109	12	0.0061	260	631
I14	I	0.0031	109	14	0.0061	122	769
I15	I	0.0031	109	15	0.0061	140	751
I16	I	0.0031	109	16	0.0061	73	817
I25	I	0.0031	109	25	0.0061	264	627
I26	I	0.0031	109	26	0.0061	142	749
J10	J	0.0031	207	10	0.0063	128	665
J11	J	0.0031	207	11	0.0063	240	553
J12	J	0.0031	207	12	0.0063	267	526
J14	J	0.0031	207	14	0.0063	126	667
J15	J	0.0031	207	15	0.0063	144	649
J16	J	0.0031	207	16	0.0063	75	717
J25	J	0.0031	207	25	0.0063	271	522
J26	J	0.0031	207	26	0.0063	146	647
K10	K	0.0031	247	10	0.0063	128	625
K11	K	0.0031	247	11	0.0063	240	513
K12	K	0.0031	247	12	0.0063	267	486
K14	K	0.0031	247	14	0.0063	126	627
K15	K	0.0031	247	15	0.0063	143	610

K16	K	0.0031	247	16	0.0063	75	678
K25	K	0.0031	247	25	0.0063	270	483
K26	K	0.0031	247	26	0.0063	146	607
R10	R	0.0031	302	10	0.0063	128	570
R11	R	0.0031	302	11	0.0063	240	458
R12	R	0.0031	302	12	0.0063	267	431
R14	R	0.0031	302	14	0.0063	126	572
R15	R	0.0031	302	15	0.0063	144	554
R16	R	0.0031	302	16	0.0063	75	622
R25	R	0.0031	302	25	0.0063	271	427
R26	R	0.0031	302	26	0.0063	146	552
G27	G	0.0031	101	27	0.0063	171	727
G28	G	0.0031	101	28	0.0063	179	719
G29	G	0.0031	101	29	0.0063	183	715
G30	G	0.0031	101	30	0.0063	131	767
G31	G	0.0031	101	31	0.0063	171	727
G32	G	0.0031	101	32	0.0063	113	785
G33	G	0.0031	101	33	0.0063	184	714
G34	G	0.0031	101	34	0.0063	144	754
H27	H	0.0031	240	27	0.0063	171	584
H28	H	0.0031	240	28	0.0063	179	576
H29	H	0.0031	240	29	0.0063	182	573
H30	H	0.0031	240	30	0.0063	131	624
H31	H	0.0031	240	31	0.0063	171	584
H32	H	0.0031	240	32	0.0063	113	642
H33	H	0.0031	240	33	0.0063	183	572
H34	H	0.0031	240	34	0.0063	143	612
I27	I	0.0031	109	27	0.0063	171	717
I28	I	0.0031	109	28	0.0063	179	709
I29	I	0.0031	109	29	0.0063	183	705
I30	I	0.0031	109	30	0.0063	131	757
I31	I	0.0031	109	31	0.0063	171	717
I32	I	0.0031	109	32	0.0063	113	775
I33	I	0.0031	109	33	0.0063	184	704
I34	I	0.0031	109	34	0.0063	144	744
J27	J	0.0031	207	27	0.0063	171	622
J28	J	0.0031	207	28	0.0063	179	614
J29	J	0.0031	207	29	0.0063	183	610
J30	J	0.0031	207	30	0.0063	131	662
J31	J	0.0031	207	31	0.0063	171	622
J32	J	0.0031	207	32	0.0063	113	680
J33	J	0.0031	207	33	0.0063	184	609
J34	J	0.0031	207	34	0.0063	144	649
K27	K	0.0031	247	27	0.0063	171	582

K28	K	0.0031	247	28	0.0063	179	574
K29	K	0.0031	247	29	0.0063	182	571
K30	K	0.0031	247	30	0.0063	131	622
K31	K	0.0031	247	31	0.0063	171	582
K32	K	0.0031	247	32	0.0063	113	640
K33	K	0.0031	247	33	0.0063	184	569
K34	K	0.0031	247	34	0.0063	143	610
R27	R	0.0031	302	27	0.0062	170	530
R28	R	0.0031	302	28	0.0062	178	522
R29	R	0.0031	302	29	0.0062	181	519
R30	R	0.0031	302	30	0.0062	130	570
R31	R	0.0031	302	31	0.0062	170	530
R32	R	0.0031	302	32	0.0062	113	587
R33	R	0.0031	302	33	0.0062	183	517
R34	R	0.0031	302	34	0.0062	143	557
S10	S	0.0031	165	10	0.0063	129	706
S11	S	0.0031	165	11	0.0063	241	594
S12	S	0.0031	165	12	0.0063	267	568
S14	S	0.0031	165	14	0.0063	126	709
S15	S	0.0031	165	15	0.0063	144	691
S16	S	0.0031	165	16	0.0063	75	759
S25	S	0.0031	165	25	0.0063	271	564
S26	S	0.0031	165	26	0.0063	146	689
T 10	T	0.0031	132	10	0.0063	128	740
T 11	T	0.0031	132	11	0.0063	240	628
T 12	T	0.0031	132	12	0.0063	267	601
T 14	T	0.0031	132	14	0.0063	126	742
T 15	T	0.0031	132	15	0.0063	143	725
T 16	T	0.0031	132	16	0.0063	75	792
T 25	T	0.0031	132	25	0.0063	271	597
T 26	T	0.0031	132	26	0.0063	146	722
S27	S	0.0031	165	27	0.0063	171	664
S28	S	0.0031	165	28	0.0063	179	656
S29	S	0.0031	165	29	0.0063	183	652
S30	S	0.0031	165	30	0.0063	131	704
S31	S	0.0031	165	31	0.0063	171	664
S32	S	0.0031	165	32	0.0063	113	722
S33	S	0.0031	165	33	0.0063	184	651
S34	S	0.0031	165	34	0.0063	144	691
T27	T	0.0031	132	27	0.0063	171	697
T28	T	0.0031	132	28	0.0063	179	689
T29	T	0.0031	132	29	0.0063	182	686
T30	T	0.0031	132	30	0.0063	131	737
T31	T	0.0031	132	31	0.0063	171	697

T32	T	0.0031	132	32	0.0063	113	755
T33	T	0.0031	132	33	0.0063	184	684
T34	T	0.0031	132	34	0.0063	143	725
U10	U	0.0031	295	10	0.0063	129	576
U11	U	0.0031	295	11	0.0063	241	464
U12	U	0.0031	295	12	0.0063	267	438
U14	U	0.0031	295	14	0.0063	126	579
U15	U	0.0031	295	15	0.0063	144	561
U16	U	0.0031	295	16	0.0063	75	629
U25	U	0.0031	295	25	0.0063	271	434
U26	U	0.0031	295	26	0.0063	146	559
U27	U	0.0031	295	27	0.0063	172	533
U28	U	0.0031	295	28	0.0063	179	526
U29	U	0.0031	295	29	0.0063	183	522
U30	U	0.0031	295	30	0.0063	131	574
U31	U	0.0031	295	31	0.0063	172	533
U32	U	0.0031	295	32	0.0063	113	592
U33	U	0.0031	295	33	0.0063	184	521
U34	U	0.0031	295	34	0.0063	144	561

Table 7.3: Categorisation of all cage hits from the high-throughput screening workflow, alongside their respective predicted cavity sizes following DFT optimisation. Cavity sizes reported with a '-' values indicate that topology was predicted to be collapsed using the *pyWindow* software, and categories assigned correspond to those listed in Table 5.1. For categories 3, 4, 7, and 8, which correspond to the identification of a mixture of topologies, each topology observed is reported, alongside the calculated cavity size for that topology. Hits are ordered in descending order of cavity sizes.

Key	Species Observed	Cavity Size / Å	Category Number
S16	[4 + 6]	6.1	1
I16	[4 + 6]	5.4	1
I31	[2 + 3]	2.5	1
I29	[2 + 3]	0.6	1
S25	[2 + 3]	0.5	1
K10	[2 + 3]	0.4	1
G29	[2 + 3]	-	1
H25	[2 + 3]	-	1
I25	[2 + 3]	-	1
R25	[2 + 3]	-	1
R28	[2 + 3]	-	1
S10	[2 + 3]	-	1
S26	[2 + 3]	-	1
S30	[2 + 3]	-	1
S32	[2 + 3]	-	1
U10	[2 + 3]	0.3	2
U30	[2 + 3]	-	2
S28	[4 + 6]	11.6	2
U28	[2 + 3]	6.8	2
T10	[2 + 3]	1.2	2
T26	[2 + 3]	0.8	2
S31	[2 + 3]	0.6	2
H26	[2 + 3]	-	2
H31	[2 + 3]	-	2
K25	[2 + 3]	-	2
T25	[2 + 3]	-	2
U26	[2 + 3]	-	2

I28	[2 + 3]	4.1	3
	[4 + 6]	10.2	3
	[6 + 9]	4.5	3
G27	[2 + 3]	4.2	3
	[4 + 6]	8.2	3
G31	[2 + 3]	2.4	3
	[4 + 6]	-	3
J31	[2 + 3]	1.7	3
	[4 + 6]	-	3
	[6 + 9]	-	3
G30	[2 + 3]	1.7	3
	[4 + 6]	-	3
I30	[2 + 3]	0.9	3
	[4 + 6]	-	3
G25	[2 + 3]	-	3
	[4 + 6]	-	3
U31	[2 + 3]	-	3
	[4 + 6]	-	3
I26	[2 + 3]	-	3
	[4 + 6]	-	3
K29	[2 + 3]	-	3
	[4 + 6]	-	3
U25	[2 + 3]	-	3
	[4 + 6]	-	3
K16	[2 + 3]	1.5	4
	[4 + 6]	11.5	4
	[8 + 12]	0.7	4
I27	[2 + 3]	4.2	4
	[4 + 6]	10.7	4
K31	[2 + 3]	0.9	4
	[4 + 6]	-	4
	[6 + 9]	-	4
	[8 + 12]	-	4
K30	[2 + 3]	0.3	4
	[4 + 6]	-	4
K26	[2 + 3]	-	4
	[4 + 6]	-	4
H30	[2 + 3]	-	4
	[4 + 6]	-	4

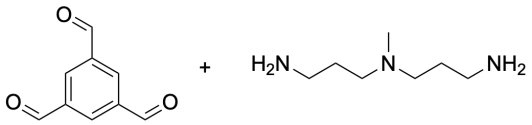
G16	[4 + 6]	5.5	5
H15	[2 + 3]	2.9	5
R34	[2 + 3]	1.6	5
T12	[2 + 3]	1.3	5
J15	[2 + 3]	-	5
R15	[2 + 3]	-	5
R27	[2 + 3]	-	5
R31	[2 + 3]	-	5
R32	[2 + 3]	-	5
J26	[2 + 3]	0.8	5
H10	[2 + 3]	1.1	6
K15	[2 + 3]	1.0	6
U15	[2 + 3]	0.6	6
J10	[2 + 3]	-	6
R10	[2 + 3]	-	6
I15	[4 + 6]	8.4	7
	[6 + 9]	11.8	7
	[8 + 12]	15.3	7
J32	[2 + 3]	-	7
	[4 + 6]	5.8	7
I10	[2 + 3]	1.1	7
	[4 + 6]	1.0	7
J25	[2 + 3]	-	7
	[4 + 6]	-	7
R29	[2 + 3]	-	7
	[4 + 6]	-	7
T31	[2 + 3]	2.5	8
	[4 + 6]	-	8
T30	[2 + 3]	1.5	8
	[4 + 6]	-	8
T29	[2 + 3]	-	8
	[4 + 6]	-	8
I32	[2 + 3]	1.1	8
	[4 + 6]	-	8

7.2.3 High-Throughput Results

Clean Cage Hits - Category 1

In this section, I will characterise the precursor combinations that were observed to completely convert to a single topology, corresponding to category 1. For each precursor combination, the crude ^{1}H NMR spectra and HRMS spectra are presented. The HRMS spectra for the other precursor combinations where a cage was observed to form, corresponding to categories 2-8, will be shown in the appendix, in addition to the stacked ^{1}H NMR spectra for all 144 precursor combinations.

G29$_{[2+3]}$



^{1}H NMR: (400 MHz, $CDCl_3$) δ_H 8.13 (6H, s), 7.86 (6H, s), 3.61 (12H, t, J = 6.6 Hz), 2.35 (12H, t, J = 6.8 Hz), 2.24 (9H, s), 1.82 (12H, qt, J = 6.7 Hz); **HRMS:** (ES+) calc. for [2 + 3] cage $C_{39}H_{57}N_9$ 651.4737, found $[M+H]^+$ 652.4836 and $[M+2H]^{2+}$ 326.7441.

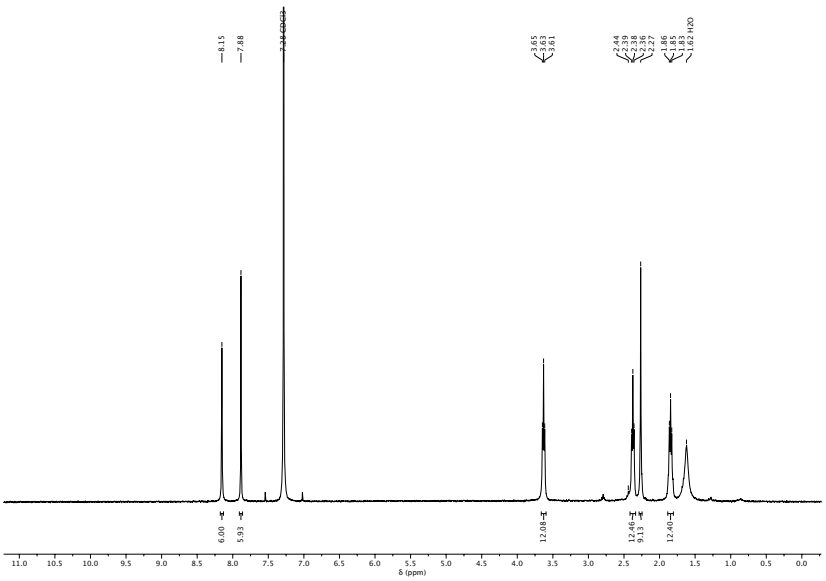


Figure 7.6: Crude ^{1}H NMR spectrum ($CDCl_3$) for **G29$_{[2+3]}$** taken directly from the high-throughput screen.

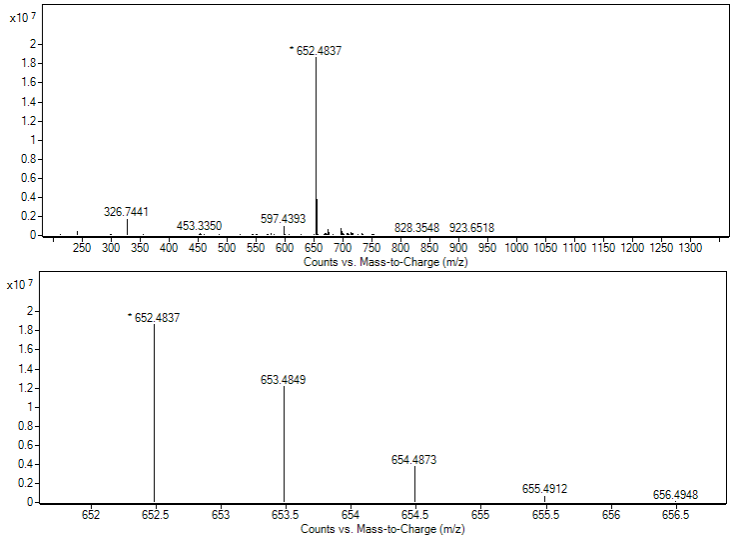
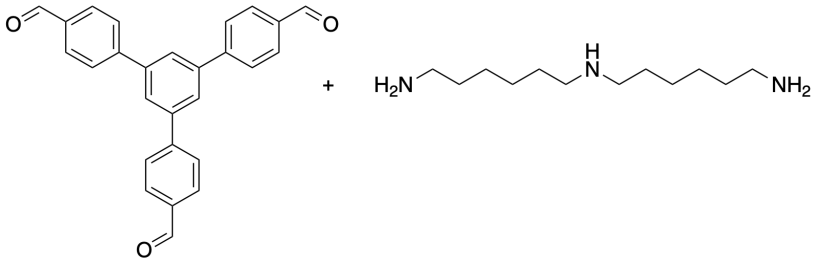


Figure 7.7: HRMS spectra for crude reactions taken from the high-throughput screen for cage **G29**$_{[2+3]}$, indicating the formation of a [2 + 3] cage — calc. for $C_{39}H_{57}N_9$ 651.4737, found $[M+H]^+$ 652.4836 and $[M+2H]^{2+}$ 326.7441.

H25$_{[2+3]}$



^{1}H NMR: (400 MHz, $CDCl_3$) δ_H 8.38–8.26 (m, 6H), 7.95–7.56 (30H, m), 3.70–3.57 (11H, m), 2.80–2.53 (25H, m), 1.80–1.68 (15H, m), 1.59–1.28 (78H, m); **HRMS:** (ES+) calc. for [2 + 3] cage $C_{90}H_{111}N_9$ 1317.8962, found $[M+H]^+$ 1318.9006 and $[M+2H]^{2+}$ 659.9522.

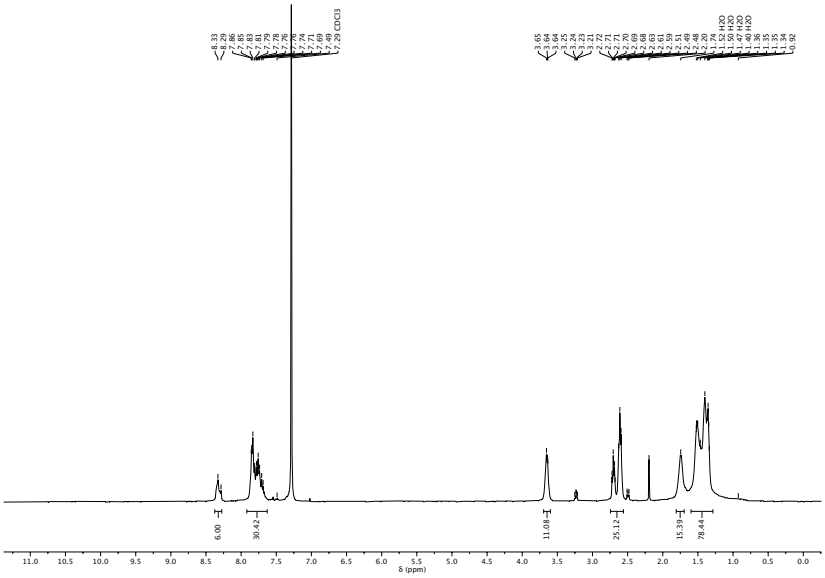


Figure 7.8: Crude ^{1}H NMR spectrum ($CDCl_3$) for **H25$_{[2+3]}$** taken directly from the high-throughput screen.

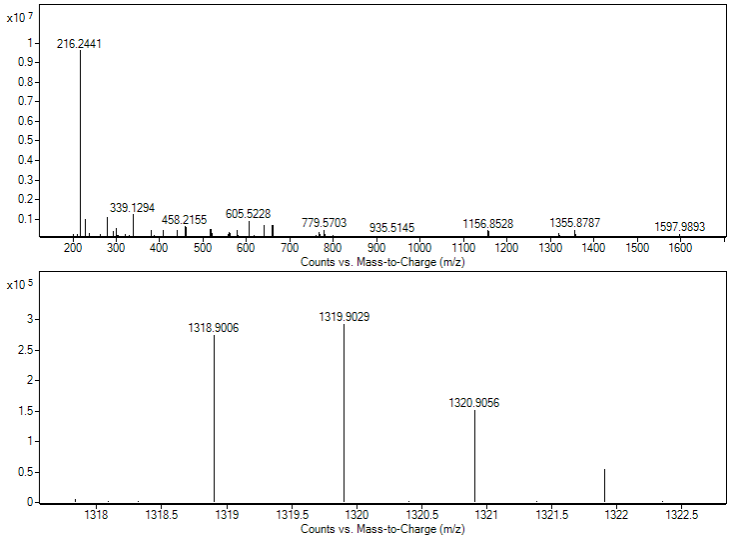
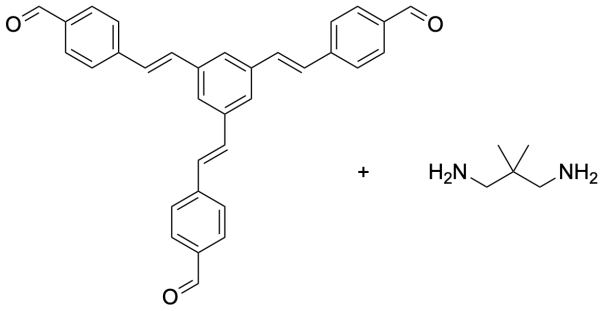


Figure 7.9: HRMS spectra for crude reactions taken from the high-throughout screen for cage **H25**$_{[2+3]}$, indicating the formation of a [2 + 3] cage — calc. for $C_{90}H_{111}N_9$ 1317.8962, found $[M+H]^+$ 1318.9006 and $[M+2H]^{2+}$ 659.9522.

U10$_{[2+3]}$



^{1}H NMR: (400 MHz, $CDCl_3$) δ_H 7.57 (5H, d, J = 5.0 Hz), 7.19 (2H, d, J = 8.2 Hz), 4.55 (1H, s), 2.85 (4H, s), 2.69 (1H, s), 2.58 (1H, s), 2.20 (2H, s), 1.33 (1H, s), 1.17 (3H, s), 0.92 (2H, s), 0.88 (5H, d, J = 3.4 Hz); **HRMS:** (ES+) calc. for [2 + 3] cage $C_{81}H_{78}N_6$ 1135.6366, found $[M+H]^+$ 1136.6325 and $[M+2H]^{2+}$ 568.3182.

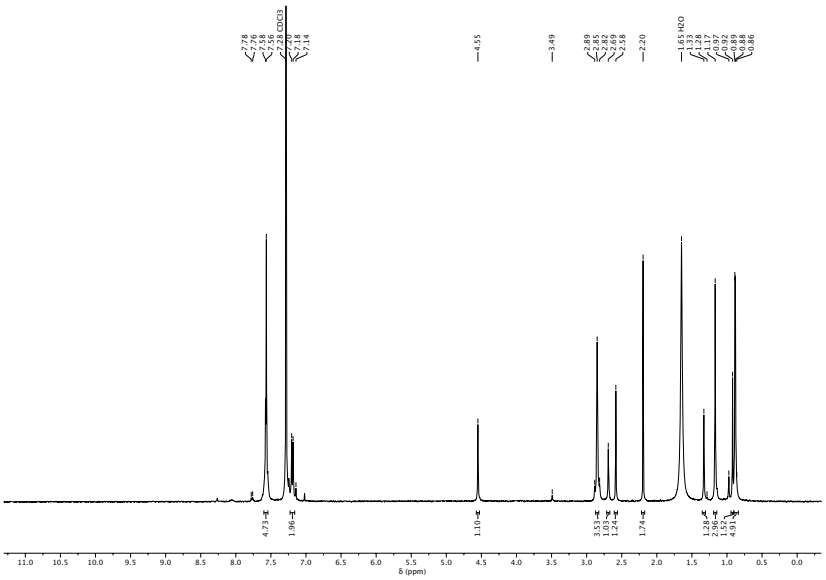


Figure 7.10: Crude ^{1}H NMR spectrum ($CDCl_3$) for **U10$_{[2+3]}$** taken directly from the high-throughput screen.

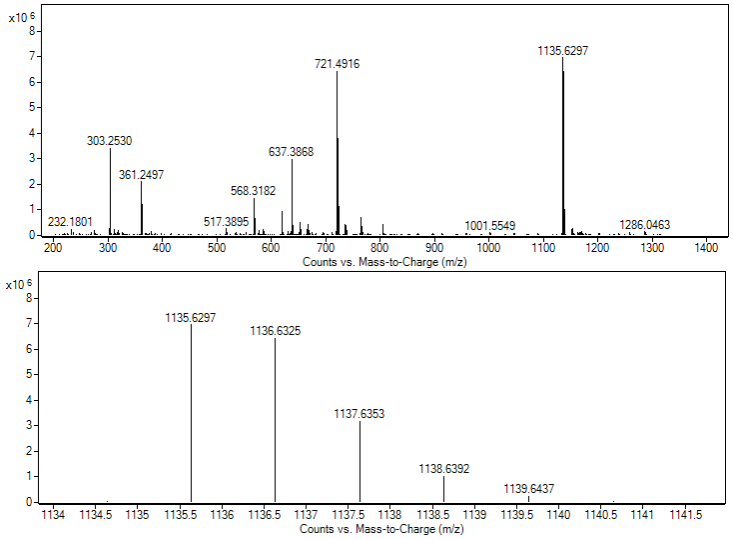
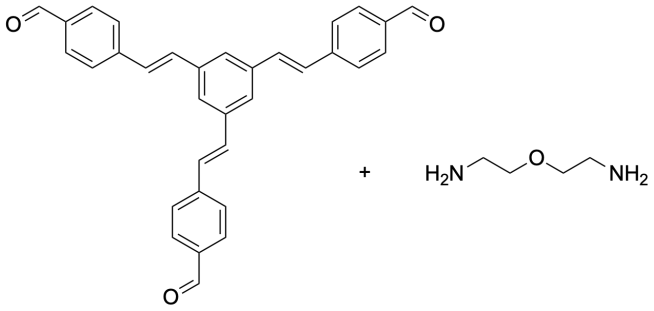


Figure 7.11: HRMS spectra for crude reactions taken from the high-throughput screen for cage **U10**$_{[2+3]}$, indicating the formation of a [2 + 3] cage — calc. for $C_{81}H_{78}N_6$ 1135.6366, found $[M+H]^+$ 1136.6325 and $[M+2H]^{2+}$ 568.3182.

U30$_{[2+3]}$



^{1}H NMR: (400 MHz, $CDCl_3$) δ_H 8.14 (1H, s), 7.37 (2H, d, J = 8.0 Hz), 7.29–7.21 (13H, m), 7.06–6.90 (3H, m), 3.81 (4H, s), 3.52 (3H, t, J = 5.2 Hz), 2.90 (3H, t, J = 5.2 Hz), 2.20 (3H, s); **HRMS:** (ES+) calc. for [2 + 3] cage $C_{78}H_{72}N_6O_3$ 1140.5666, found $[M+H]^+$ 1141.5661 and $[M+2H]^{2+}$ 571.2862.

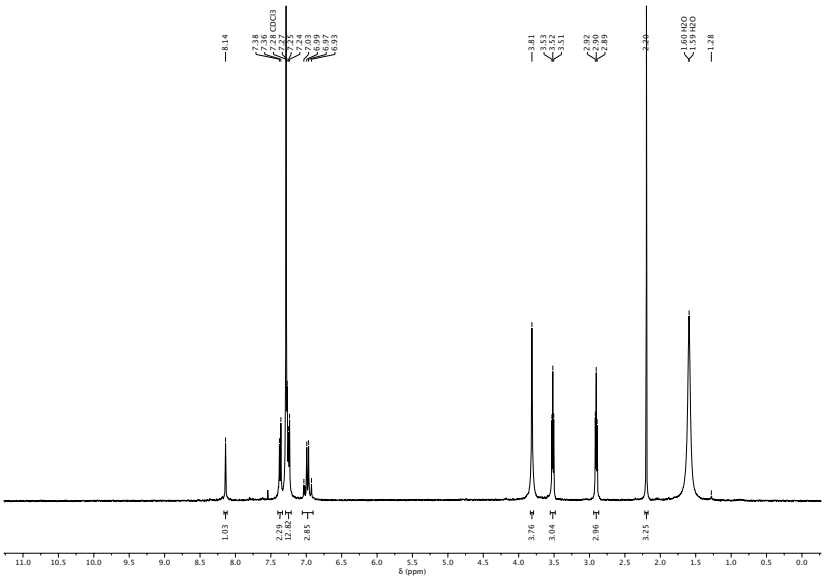


Figure 7.12: Crude ^{1}H NMR spectrum ($CDCl_3$) for **U30$_{[2+3]}$** taken directly from the high-throughput screen.

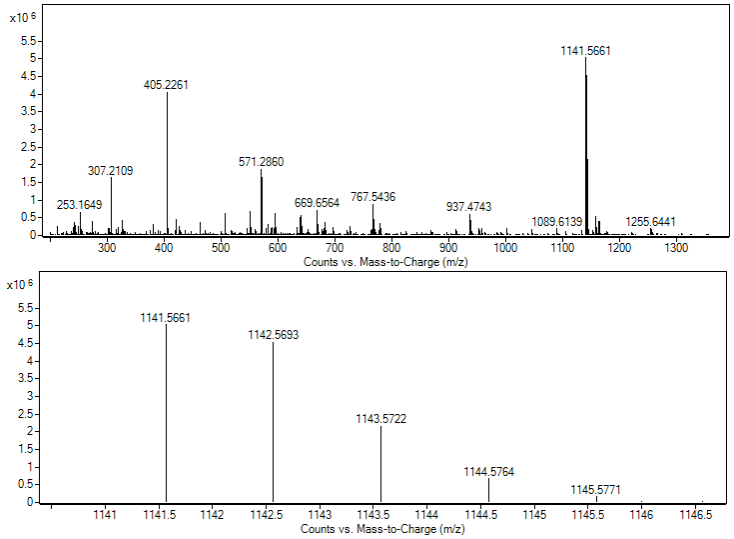
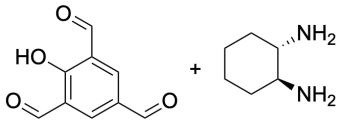


Figure 7.13: HRMS spectra for crude reactions taken from the high-throughput screen for cage **U30**$_{[2+3]}$, indicating the formation of a [2 + 3] cage — calc. for $C_{78}H_{72}N_6O_3$ 1140.5666, found $[M+H]^+$ 1141.5661 and $[M+2H]^{2+}$ 571.2862.

I16$_{[4+6]}$



^{1}H NMR: (400 MHz, $CDCl_3$) δ_H 14.56-14.44 (4H, m), 8.64 – 8.53 (4H, m), 8.30 – 8.21 (4H, m), 8.09 – 8.00 (4H, m), 7.92 – 7.77 (8H, m), 3.34 – 3.30 (14H, m), 2.07 – 1.20 (100H, m); **HRMS:** (ES+) calc. for [4 + 6] cage $C_{72}H_{84}N_{12}O_4$ 1180.6739, found $[M+H]^+$ 1181.6796 and $[M+2H]^{2+}$ 591.3408.

Data in agreement with literature values.[188]

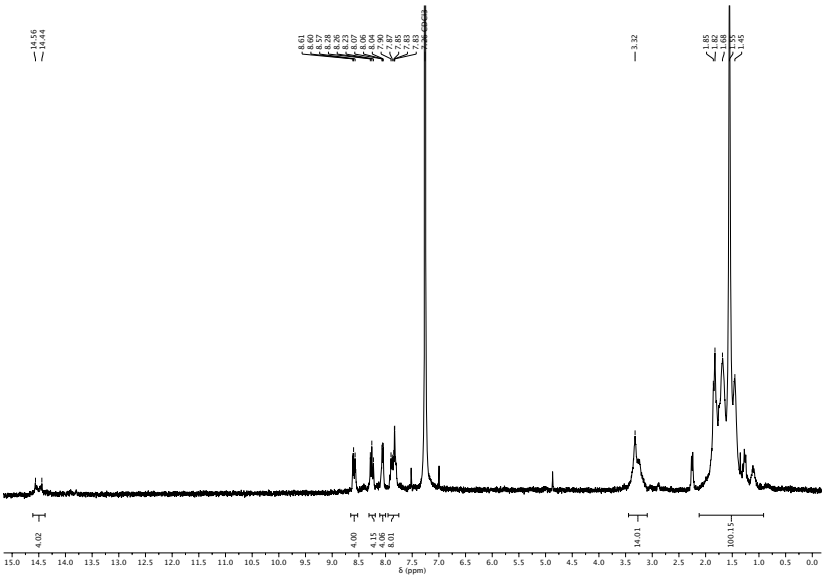


Figure 7.14: Crude ^{1}H NMR spectrum ($CDCl_3$) for **I16$_{[4+6]}$** taken directly from the high-throughput screen.

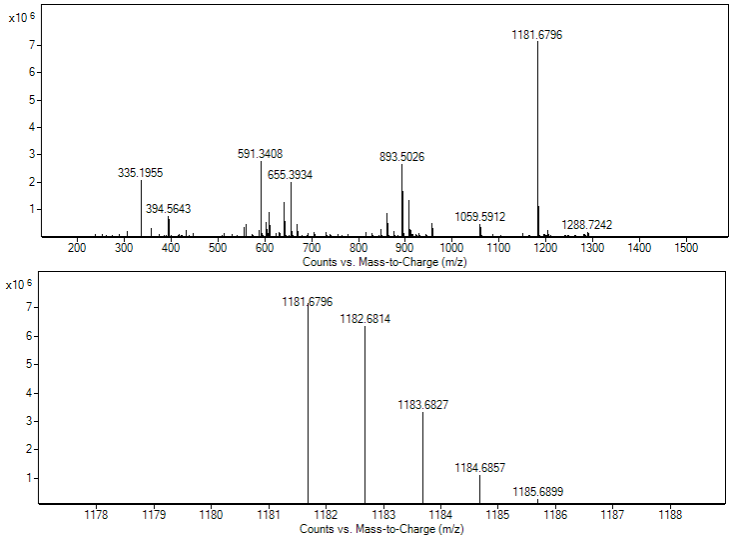
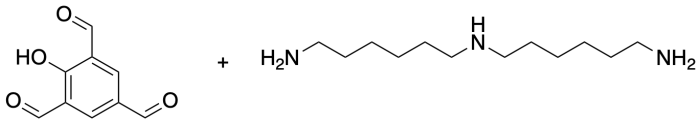


Figure 7.15: HRMS spectra for crude reactions taken from the high-throughout screen for cage **I16**$_{[4+6]}$, indicating the formation of a [4 + 6] cage — calc. for $C_{72}H_{84}N_{12}O_4$ 1180.6739, found $[M+H]^+$ 1181.6796 and $[M+2H]^{2+}$ 591.3408.

I25$_{[2+3]}$



^{1}H NMR: (400 MHz, $CDCl_3$) δ_H 8.87–7.79 (m, 8H), 4.83 (1H, s), 3.74-3.69 (1H, m), 3.67-3.62 (2H, m), 3.28-3.16 (1H, m), 2.70 (9H, t,J = 6.9 Hz), 2.61 (12H, t,J = 7.2 Hz), 1.73 (3H, s), 1.53-1.50 (9H, m), 1.50-1.43 (16H, m), 1.42 – 1.31 (29H, m); **HRMS:** (ES+) calc. for [2 + 3] cage $C_{54}H_{87}N_9O_2$ 893.692, found $[M+H]^+$ 894.7037 and $[M+2H]^{2+}$ 447.8546.

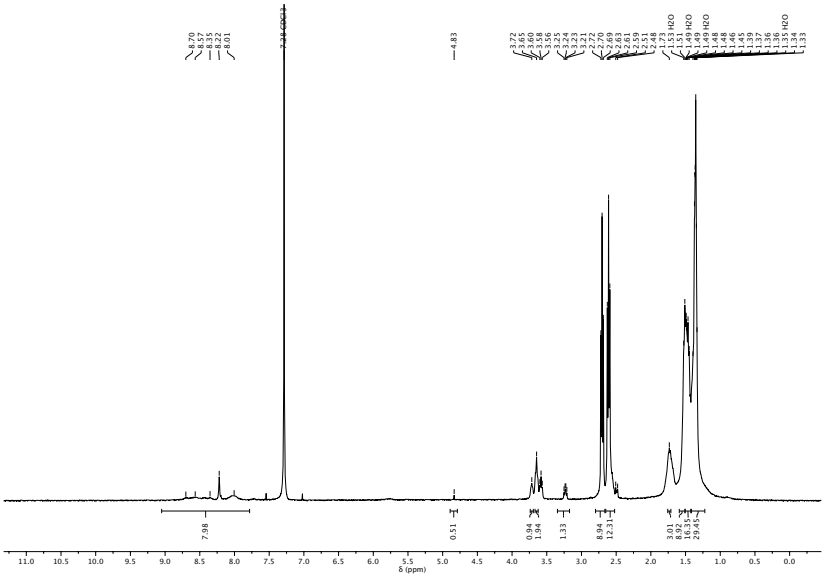


Figure 7.16: Crude ^{1}H NMR spectrum ($CDCl_3$) for **I25$_{[2+3]}$** taken directly from the high-throughput screen.

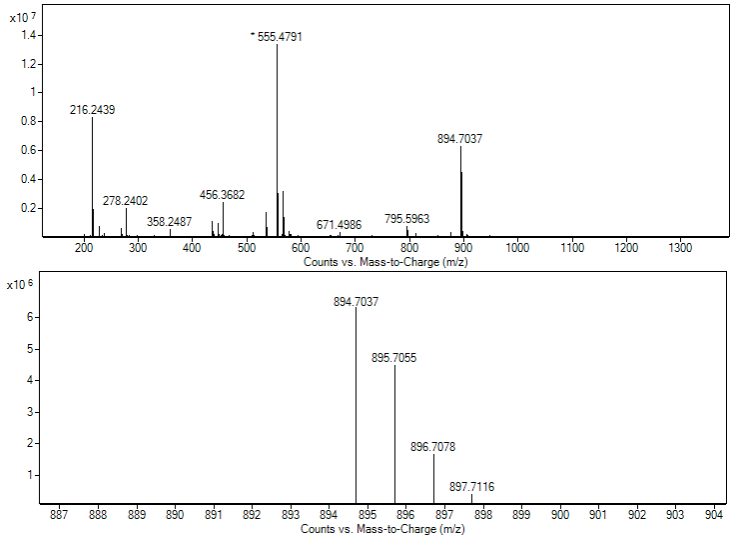
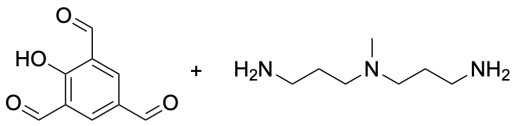


Figure 7.17: HRMS spectra for crude reactions taken from the high-throughput screen for cage **I25**$_{[2+3]}$, indicating the formation of a [2 + 3] cage — calc. for $C_{54}H_{87}N_9O_2$ 893.6983, found $[M+H]^+$ 894.7037 and $[M+2H]^{2+}$ 447.8546.

I29$_{[2+3]}$



^{1}H NMR: (400 MHz, $CDCl_3$) δ_H 8.62-8.58 (1H, m), 8.21-8.17 (1H, m), 8.09-7.98 (3H, m), 7.73-7.68 (1H, m), 3.78-3.45 (15H, m), 2.77 (10H, t, J = 6.8 Hz), 2.45–2.38 (13H, m), 2.24 (15H, d, J = 4.3 Hz), 1.83 (13H, m), 1.72–1.53 (30H, m); **HRMS:** (ES+) calc. for [2 + 3] cage $C_{39}H_{57}N_9O_2$ 683.4635, found $[M+H]^+$ 684.4731 and $[M+2H]^{2+}$ 342.7396.

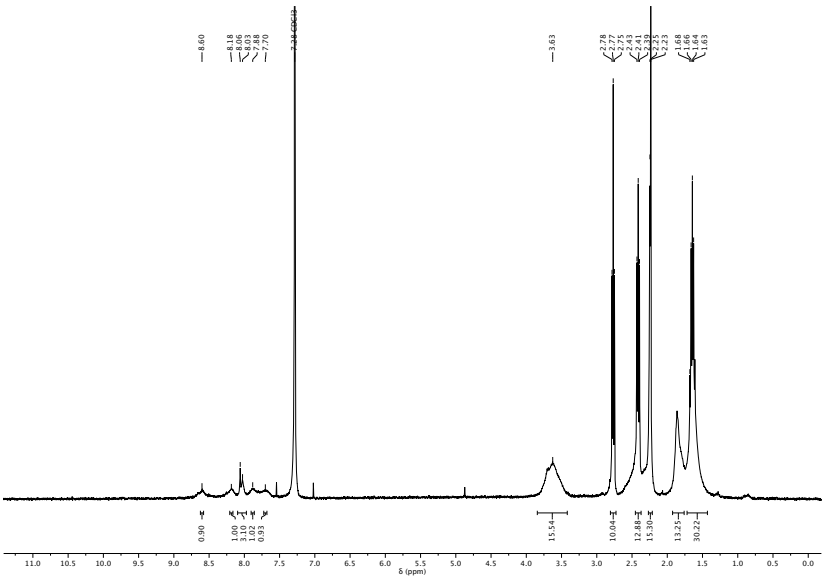


Figure 7.18: Crude ^{1}H NMR spectrum ($CDCl_3$) for **I29**$_{[2+3]}$ taken directly from the high-throughput screen.

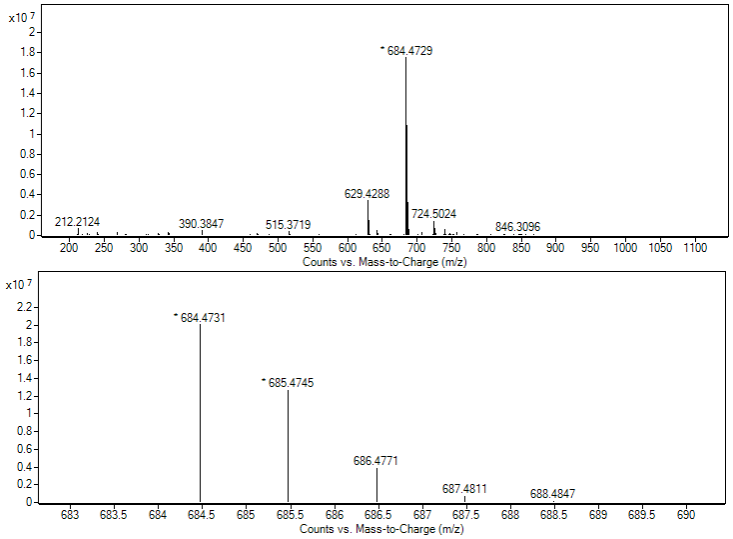
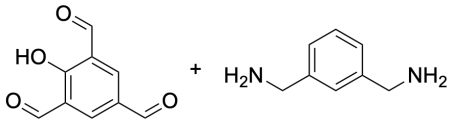


Figure 7.19: HRMS spectra for crude reactions taken from the high-throughput screen for cage **I29**$_{[2+3]}$, indicating the formation of a [2 + 3] cage — calc. for $C_{39}H_{57}N_9O_2$ 683.4635, found $[M+H]^+$ 684.4731 and $[M+2H]^{2+}$ 342.7396.

I31$_{[2+3]}$



^{1}H NMR: (400 MHz, $CDCl_3$) δ_H 8.83-8.77 (1H, m), 8.49-8.35 (3H, m), 8.26-8.18 (2H, m), 8.17-8.03 (3H, m), 7.84-7.77 (1H, m), 7.69-7.60 (1H, m), 7.42-7.30 (11H, m), 7.26-7.19 (8H, m), 6.64-6.59 (1H, m), 5.29-5.02 (4H, m), 4.98-4.72 (7H, m), 4.67-4.49 (4H m), 3.93 (9H, s), **HRMS:** (ES+) calc. for [2 + 3] cage $C_{42}H_{36}N_6O_2$ 656.2900, found $[M+H]^+$ 657.2990 and $[M+2H]^{2+}$ 329.1524.

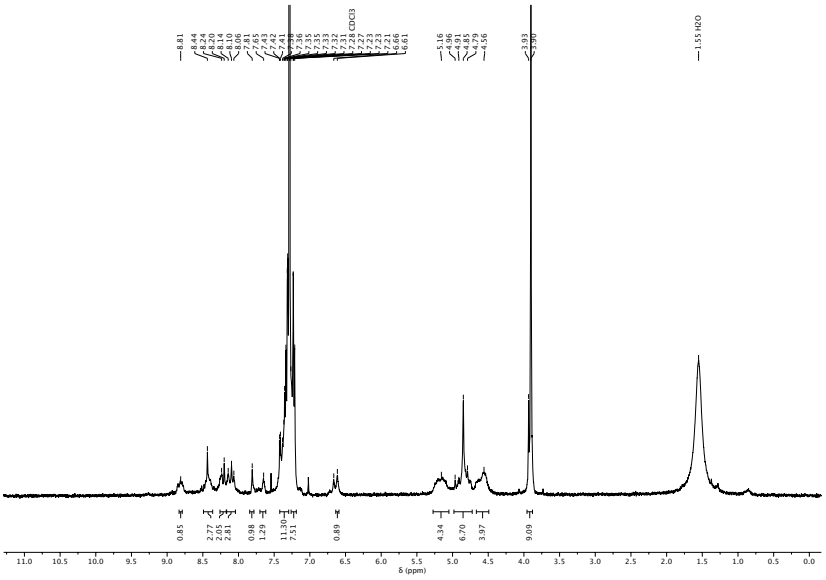


Figure 7.20: Crude ^{1}H NMR spectrum ($CDCl_3$) for **I31$_{[2+3]}$** taken directly from the high-throughput screen.

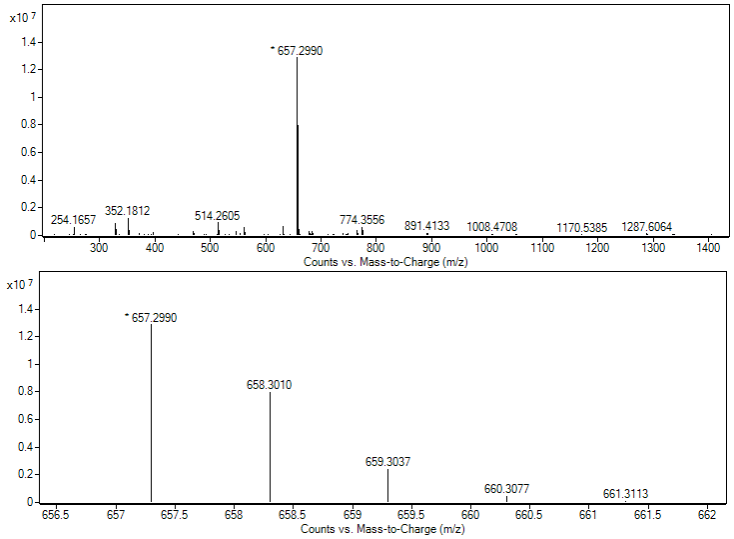
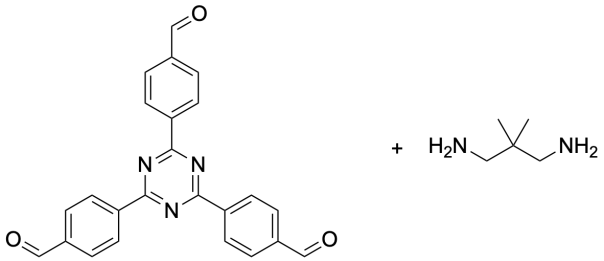


Figure 7.21: HRMS spectra for crude reactions taken from the high-throughput screen for cage **I31**$_{[2+3]}$, indicating the formation of a [2 + 3] cage — calc. for $C_{42}H_{36}N_6O_2$ 656.2900, found $[M+H]^+$ 657.2990 and $[M+2H]^{2+}$ 329.1524.

K10$_{[2+3]}$



^{1}H NMR: (400 MHz, $CDCl_3$) δ_H 8.75 (12H, d, J = 8.3 Hz), 7.75 (12H, d, J = 8.2 Hz), 4.64 (6H, s), 2.94-2.82 (27H, m), 1.18 (18H, s), 0.88 (22H, s); **HRMS:** (ES+) calc. for [2 + 3] cage $C_{63}H_{60}N_{12}$ 984.5064, found $[M+H]^+$ 985.5088 and $[M+2H]^{2+}$ 493.2598.

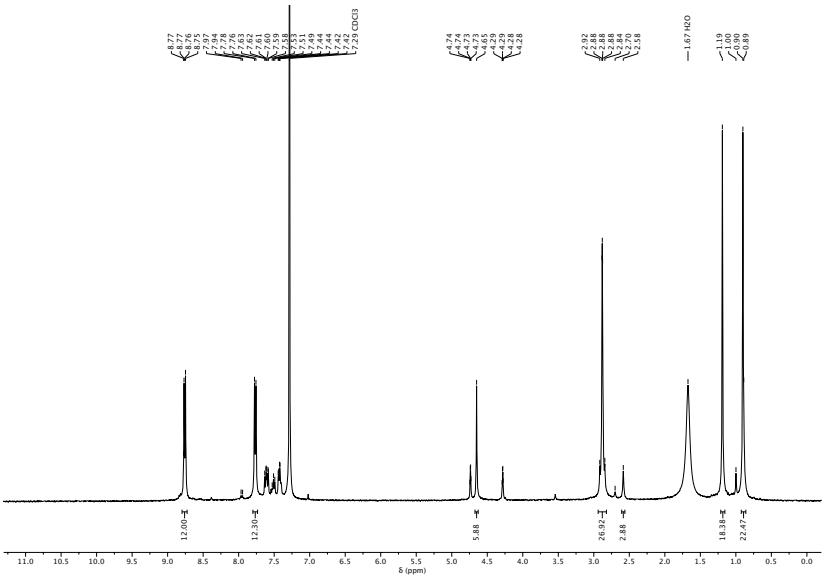


Figure 7.22: Crude ^{1}H NMR spectrum ($CDCl_3$) for **K10$_{[2+3]}$** taken directly from the high-throughput screen.

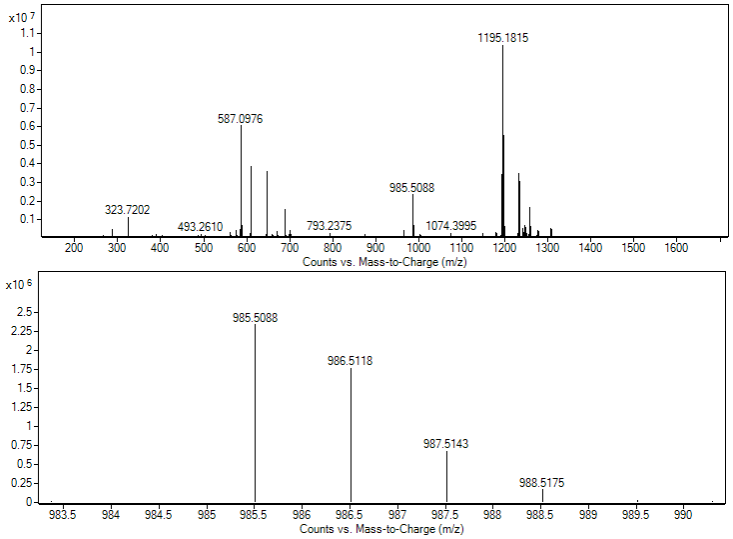
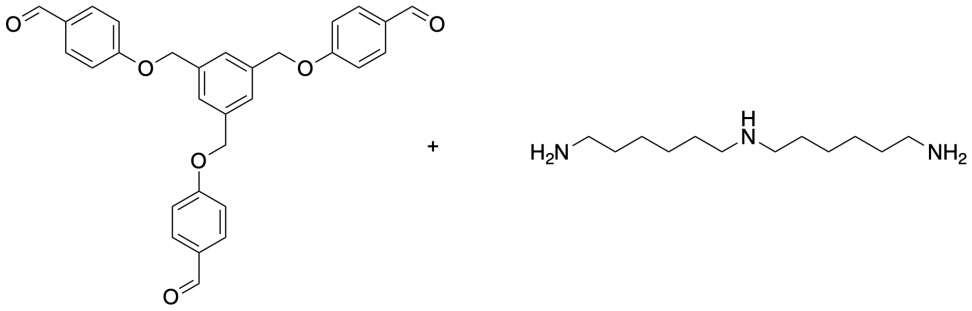


Figure 7.23: HRMS spectra for crude reactions taken from the high-throughout screen for cage **K10**$_{[2+3]}$, indicating the formation of a [2 + 3] cage — calc. for $C_{63}H_{60}N_{12}$ 984.5064, found $[M+H]^+$ 985.5088 and $[M+2H]^{2+}$ 493.2598.

R25$_{[2+3]}$



^{1}H NMR: (400 MHz, $CDCl_3$) δ_H 8.24-8.14 (10H, m), 7.69-7.61 (12H, m), 7.60-7.53 (7H, m), 7.45 (8H, d, J = 6.9 Hz), 5.22 (6H, s), 5.15-5.10 (9H, m), 3.61-3.56 (15H, m), 3.29-3.17 (5H, m), 2.74-2.67 (14H, m), 2.65-2.46 (36H, m), 1.62-1.13 (168H, m); **HRMS:** (ES+) calc. for $C_{96}H_{123}N_9O_6$ 1498.9585, found $[M+H]^+$ 1499.9620 and $[M+2H]^{2+}$ 750.9741.

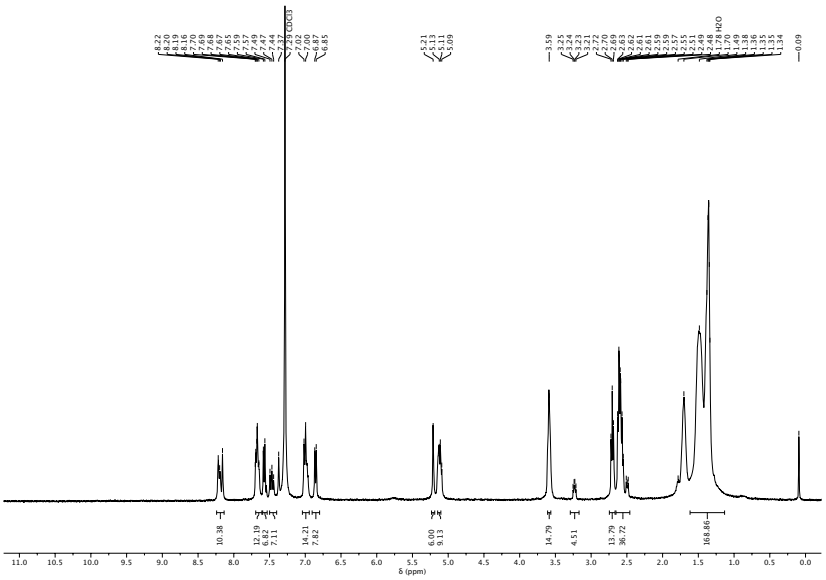


Figure 7.24: Crude ^{1}H NMR spectrum ($CDCl_3$) for **R25$_{[2+3]}$** taken directly from the high-throughput screen.

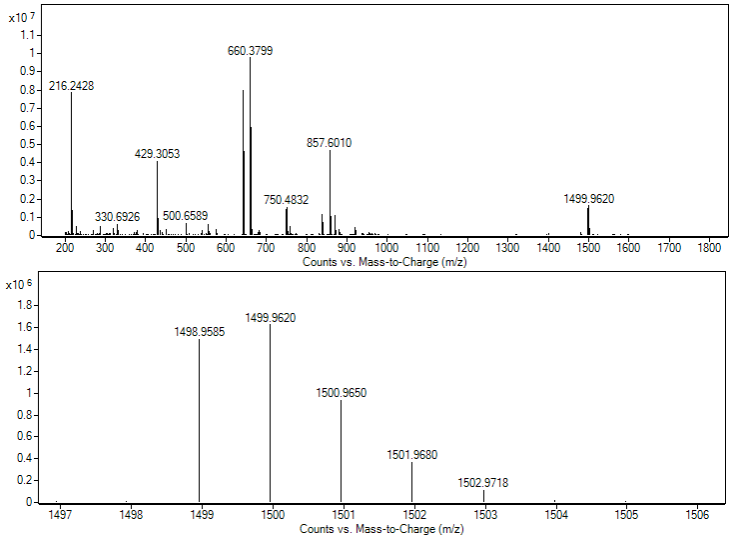
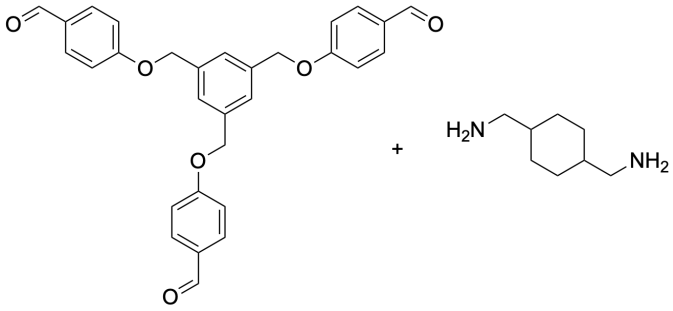


Figure 7.25: HRMS spectra for crude reactions taken from the high-throughput screen for cage **R25**$_{[2+3]}$, indicating the formation of [2 + 3] cages — calc. for $C_{96}H_{123}N_9O_6$ 1498.9585, found $[M+H]^+$ 1499.9620 and $[M+2H]^{2+}$ 750.9741.

R28$_{[2+3]}$



^{1}H NMR: (400 MHz, $CDCl_3$) δ_H 8.20-8.13 (4H, m), 7.72-7.61 (7H, m), 7.52-7.44 (4H, m), 7.02-6.98 (4H, m), 6.98-6.94 (3H, m), 5.17-5.12 (6H, m), 3.57-3.53 (4H, m), 3.48-3.43 (4H, m), 2.69-2.60 (7H, m), 2.56 (5H, d, J = 6.4 Hz), 1.98-1.94 (2H, m), 1.86-1.82 (5H, m), 1.55-1.50 (15H, m), 1.30-1.25 (4H, m), 1.06-1.02 (3H, m), 1.01-0.89 (9H, m), 0.88-0.84 (3H, m); **HRMS:** (ES+) calc. for [2 + 3] cage $C_{84}H_{90}N_6O_6$ 1278.6922, found $[M+H]^+$ 1279.6996 and $[M+2H]^{2+}$ 640.3535.

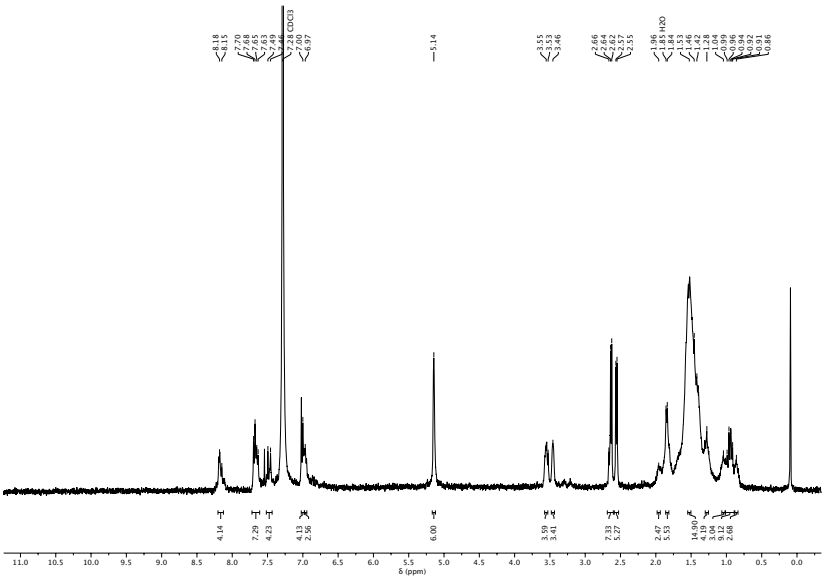


Figure 7.26: Crude ^{1}H NMR spectrum ($CDCl_3$) for **R28** taken directly from the high-throughput screen.

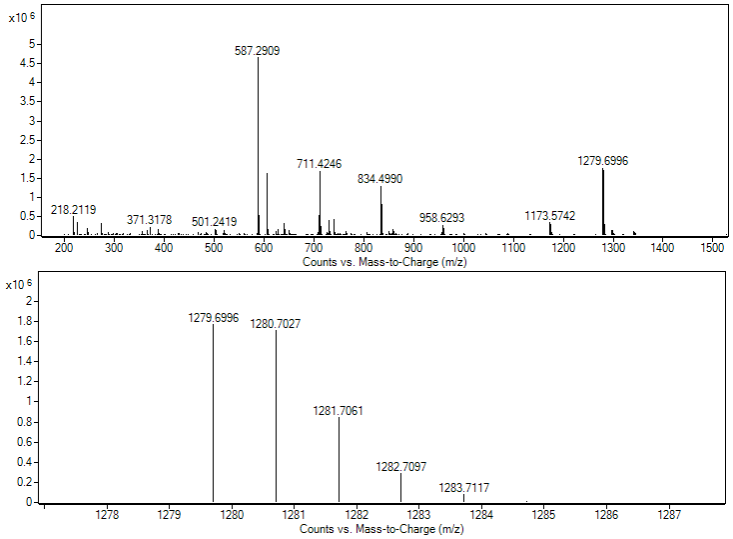
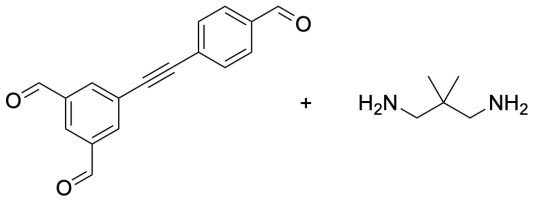


Figure 7.27: HRMS spectra for crude reactions taken from the high-throughput screen for cage **R28**$_{[2+3]}$, indicating the formation of a [2 + 3] cage — calc. for $C_{84}H_{90}N_6O_6$ 1278.6922, found $[M+H]^+$ 1279.6996 and $[M+2H]^{2+}$ 640.3535.

S10$_{[2+3]}$



^{1}H NMR: (400 MHz, $CDCl_3$) δ_H 7.68 (8H, d, J = 10.7 Hz), 7.55–7.49 (12H, m), 5.71 (11H, s), 3.65 (38H, q, J = 7.2 Hz), 2.82 (38H, d, J = 5.7 Hz), 2.62 (17H, s), 2.19 (12H, s), 1.50 (52H, t, J = 7.3 Hz), 1.34 (8H, s), 1.15 (14H, d, J = 2.5 Hz), 0.92 (32H, d, J = 7.6 Hz), 0.87 (18H, d, J = 2.8 Hz); **HRMS:** (ES+) calc. for [2 + 3] cage $C_{49}H_{50}N_6$ 722.4097, found $[M+H]^+$ 723.4141.

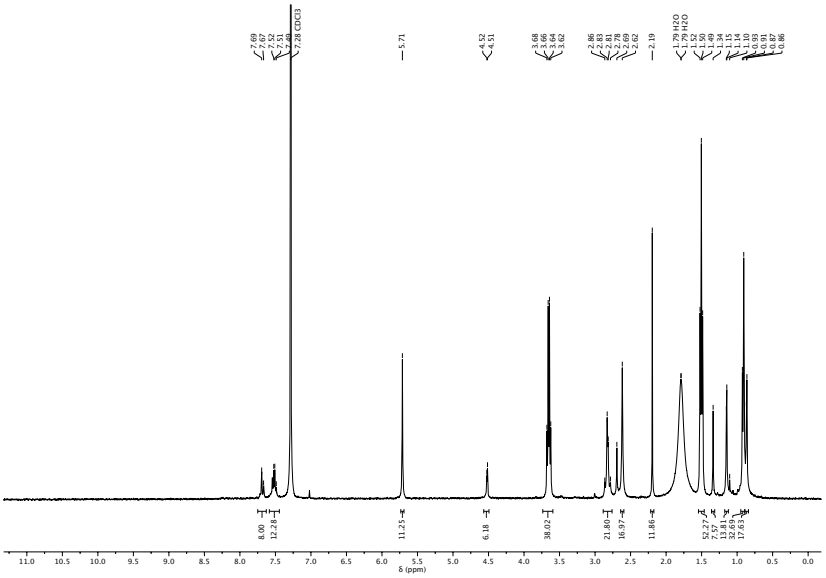


Figure 7.28: Crude ^{1}H NMR spectrum ($CDCl_3$) for **S10$_{[2+3]}$** taken directly from the high-throughput screen.

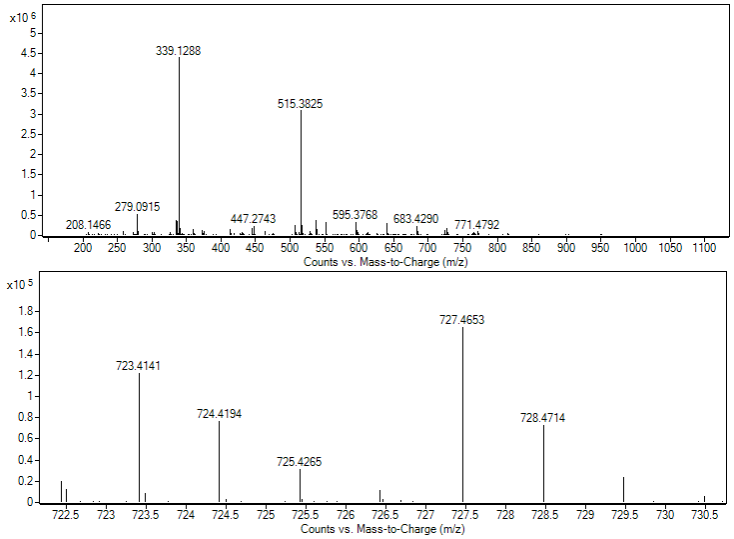
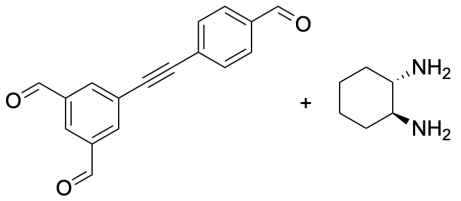


Figure 7.29: HRMS spectra for crude reactions taken from the high-throughput screen for cage **S10**$_{[2+3]}$, indicating the formation of a [2 + 3] cage — calc. for $C_{49}H_{50}N_6$ 722.4097, found $[M+H]^+$ 723.4141.

S16$_{[4+6]}$



^{1}H NMR: (400 MHz, $CDCl_3$) δ_H 8.70-7.08 (18H, m), 5.72 (4H, s), 3.65 (14H, q, J = 7.2 Hz), 3.43 (1H, m), 2.19 (45H, s), 1.51 (22H, t, J = 7.3 Hz), 1.27 (2H, s); **HRMS:** (ES+) calc. for [4 + 6] cage $C_{104}H_{100}N_{12}$ 1516.8194, found $[M+2H]^{2+}$ 759.4134.
Data in agreement with literature values.[63]

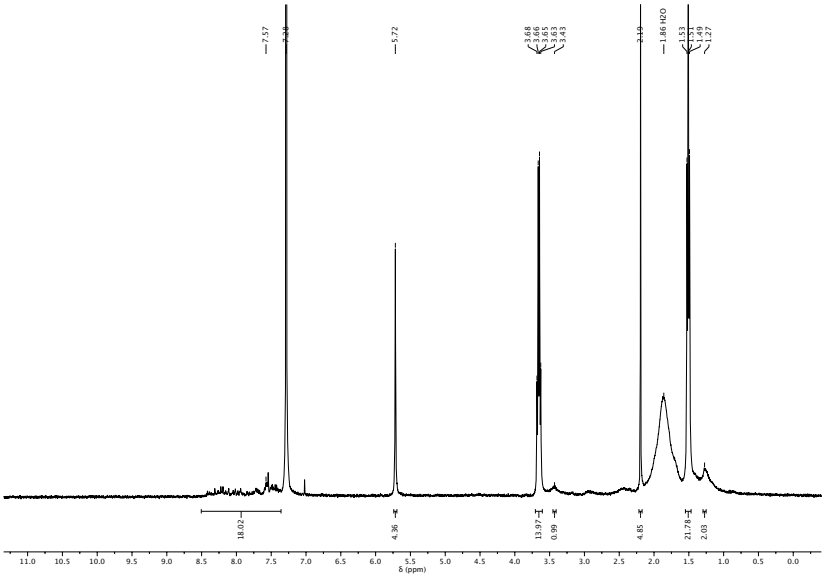


Figure 7.30: Crude ^{1}H NMR spectrum ($CDCl_3$) for **S16$_{[4+6]}$** taken directly from the high-throughput screen.

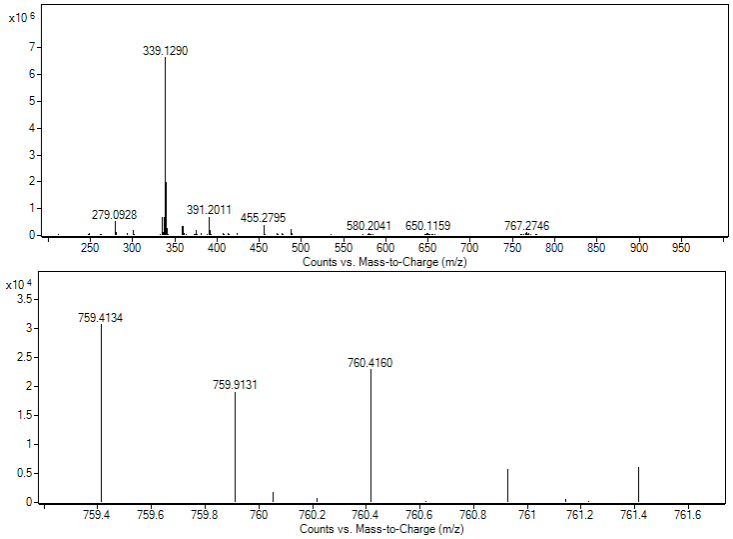
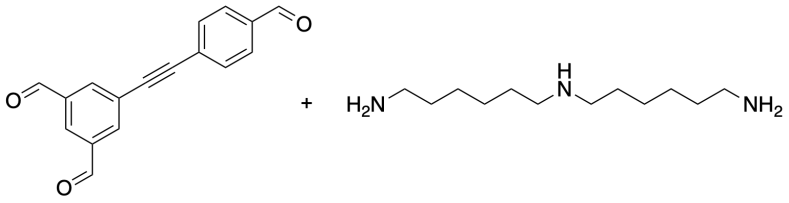


Figure 7.31: HRMS spectra for crude reactions taken from the high-throughout screen for cage **S16**$_{[4+6]}$, indicating the formation of a [4 + 6] cage — calc. for $C_{104}H_{100}N_{12}$ 1516.8194, found $[M+2H]^{2+}$ 759.4134.

S25$_{[2+3]}$



^{1}H NMR: (400 MHz, $CDCl_3$) δ_H 8.39–8.25 (1H, m), 8.09–7.96 (1H, m), 7.83–7.68 (1H, m), 7.64–7.44 (1H, m), 5.71 (2H, s), 3.65 (6H, q, J = 7.3 Hz), 2.70 (6H, t, J = 6.8 Hz), 2.62 (5H, t, J = 7.3 Hz), 2.19 (2H, s), 1.73 (2H, m), 1.56–1.43 (29H, m), 1.35 (12H, m); **HRMS:** (ES+) calc. for [2 + 3] cage $C_{70}H_{95}N_9$ 1061.7710, found $[M+H]^+$ 1062.7700 and $[M+2H]^{2+}$ 684.4731.

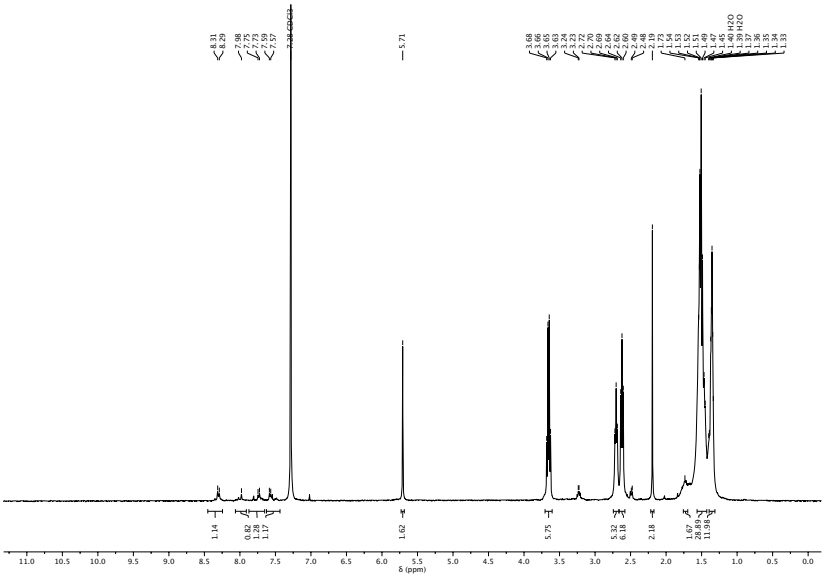


Figure 7.32: Crude ^{1}H NMR spectrum ($CDCl_3$) for **S25$_{[2+3]}$** taken directly from the high-throughput screen.

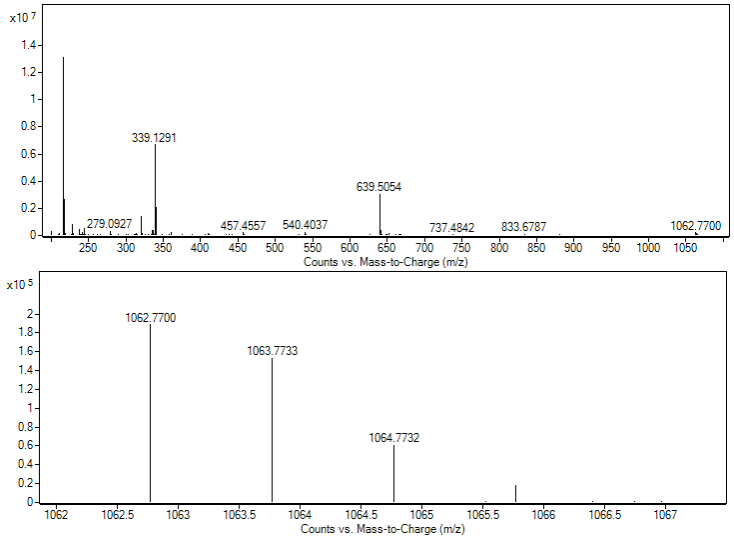
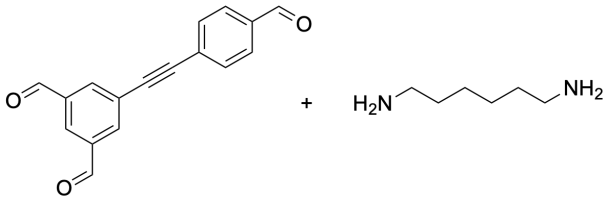


Figure 7.33: HRMS spectra for crude reactions taken from the high-throughout screen for cage **S25**$_{[2+3]}$, indicating the formation of a [2 + 3] cage — calc. for $C_{70}H_{95}N_9$ 1061.7710, found $[M+H]^+$ 1062.7700 and $[M+2H]^{2+}$ 684.4731.

S26$_{[2+3]}$



^{1}H NMR: (400 MHz, $CDCl_3$) δ_H 8.41-8.13 (1H, m), 8.05-7.90 (1H, m), 7.84-7.63 (1H, m), 7.63-7.43 (1H, m), 5.70 (2H, s), 3.66 (5H, t, J = 7.3 Hz), 2.71 (1H, s), 2.19 (1H, s), 1.72-1.08 (22H, s); **HRMS:** (ES+) calc. for [2 + 3] cage $C_{52}H_{56}N_6$ 764.4566, found $[M+H]^+$ 765.4581 and $[M+2H]^{2+}$ 383.2350.

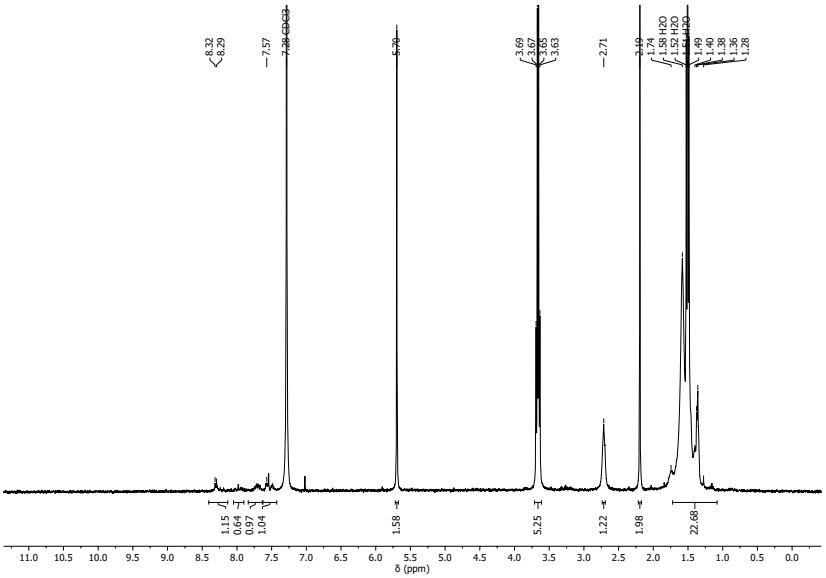


Figure 7.34: Crude ^{1}H NMR spectrum ($CDCl_3$) for **S26$_{[2+3]}$** taken directly from the high-throughput screen.

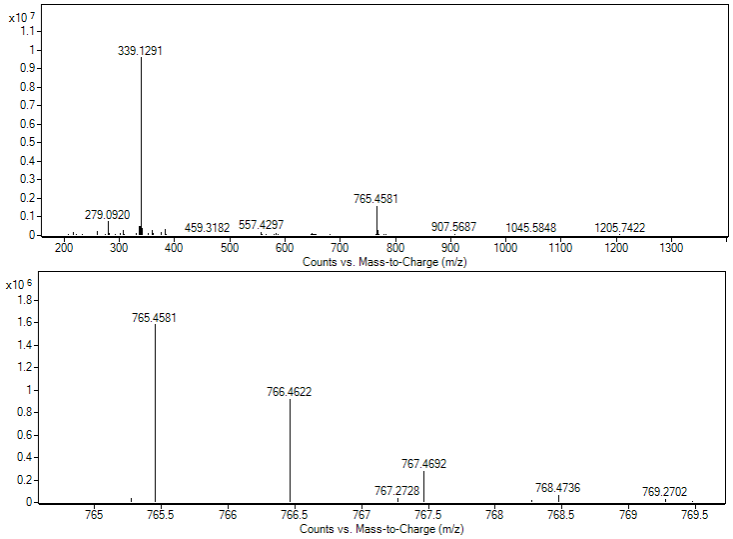
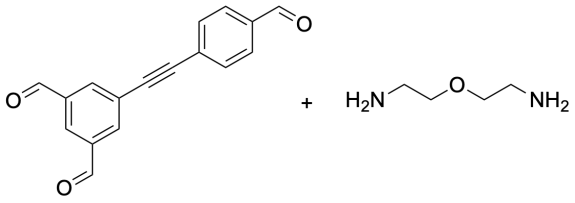


Figure 7.35: HRMS spectra for crude reactions taken from the high-throughput screen for cage **S26**$_{[2+3]}$, indicating the formation of a [2 + 3] cage — calc. for $C_{52}H_{56}N_6$ 764.4566, found $[M+H]^+$ 765.4581 and $[M+2H]^{2+}$ 383.2350.

S30$_{[2+3]}$



^{1}H NMR: (400 MHz, $CDCl_3$) δ_H 8.08 (2H, d, J = 3.8 Hz), 7.68 (2H, s), 7.60 (1H, s), 7.25-7.20 (2H, m), 5.70 (4H, s), 3.82-3.72 (5H, m), 3.66 (13H, q, J = 7.3 Hz), 3.61-3.51 (4H, m); 3.54 (4H, s), 2.93 (1H, s) 2.19 (4H, s), 1.50 (15H, d, J = 7.2 Hz); **HRMS:** (ES+) calc. for [2 + 3] cage $C_{46}H_{44}N_6O_3$ 728.3475, found $[M+H]^+$ 729.3494 and $[M+2H]^{2+}$ 365.1803.

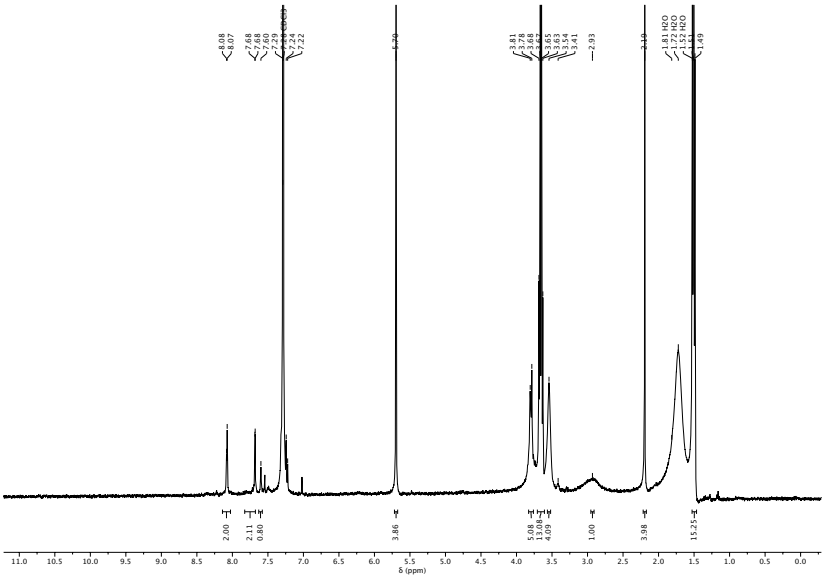


Figure 7.36: Crude ^{1}H NMR spectrum ($CDCl_3$) for **S30$_{[2+3]}$** taken directly from the high-throughput screen.

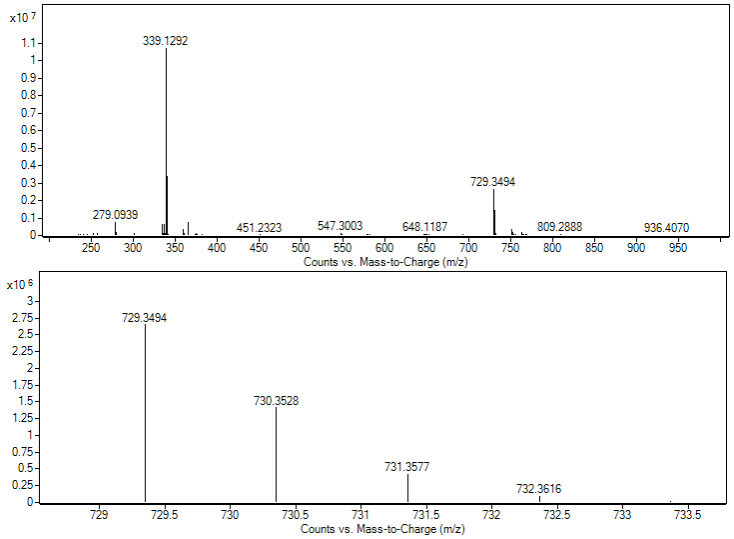
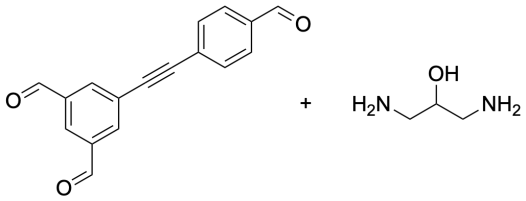


Figure 7.37: HRMS spectra for crude reactions taken from the high-throughput screen for cage **S30**$_{[2+3]}$, indicating the formation of a [2 + 3] cage — calc. for $C_{46}H_{44}N_6O_3$ 728.3475, found $[M+H]^+$ 729.3494 and $[M+2H]^{2+}$ 365.1803.

S32$_{[2+3]}$



^{1}H NMR: (400 MHz, $CDCl_3$) δ_H 7.87-7.40 (3H, m), 5.69 (1H, s), 3.66 (4H, q, J = 7.3 Hz), 3.56-3.07 (4H, m), 2.85 (1H, dd, J = 12.7, 3.8 Hz), 2.65 (1H, dd, J = 12.7, 7.7 Hz), 1.51 (6H, t, J = 7.3 Hz); **HRMS:** (ES+) calc. for [2 + 3] cage $C_{43}H_{38}N_6O_3$ 686.3005, found $[M+H]^+$ 687.3048 and $[M+2H]^{2+}$ 344.1634.

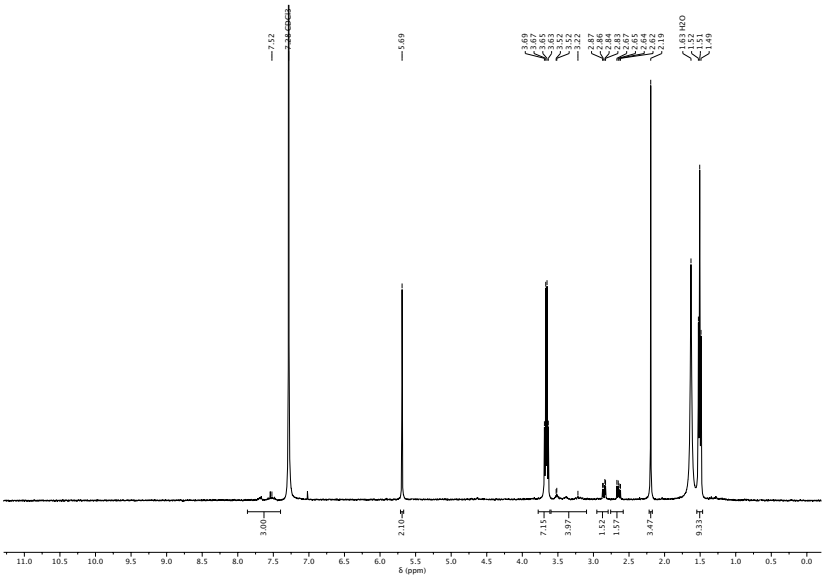


Figure 7.38: Crude ^{1}H NMR spectrum ($CDCl_3$) for **S32$_{[2+3]}$** taken directly from the high-throughput screen.

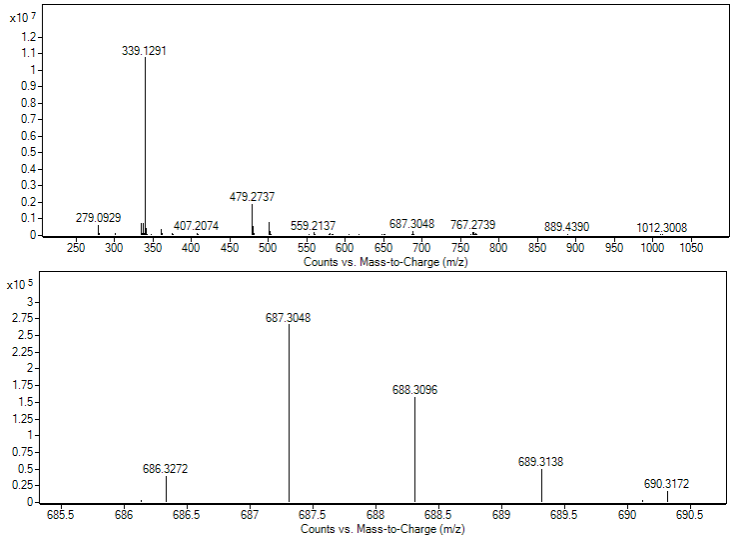


Figure 7.39: HRMS spectra for crude reactions taken from the high-throughput screen for cage **S32**$_{[2+3]}$, indicating the formation of a [2 + 3] cage — calc. for $C_{43}H_{38}N_6O_3$ 686.3005, found $[M+H]^+$ 687.3048 and $[M+2H]^{2+}$ 344.1634.

Bibliography

1. G. S. Day, H. F. Drake, H.-C. Zhou, and M. R. Ryder, *Commun. Chem.*, 2021, **4**, 1–4.
2. E. Koohsaryan, M. Anbia, and M. Maghsoodlu, *J. Environ. Chem. Eng.*, 2020, **8**, 104287.
3. R. L. Siegelman, E. J. Kim, and J. R. Long, *Nat. Mater.*, 2021, **20**, 1060–1072.
4. K. B. Idrees, Z. Li, H. Xie, K. O. Kirlikovali, M. Kazem-Rostami, X. Wang, X. Wang, T.-Y. Tai, T. Islamoglu, J. F. Stoddart, R. Q. Snurr, and O. K. Farha, *J. Am. Chem. Soc.*, 2022, **144**, 12212–12218.
5. BCS Inc., Laurel, MD (United States), *Materials for Separation Technologies. Energy and Emission Reduction Opportunities*, 1218755, 2005, p. 1218755.
6. D. S. Sholl and R. P. Lively, *Nature*, 2016, **532**, 435–437.
7. T. D. Bennett, F.-X. Coudert, S. L. James, and A. I. Cooper, *Nat. Mater.*, 2021, **20**, 1179–1187.
8. R. M. Barrer, *J. Chem. Soc.*, 1948, 2158–2163.
9. J. Naber, K. De Jong, W. Stork, H. Kuipers, and M. Post, in *Studies in Surface Science and Catalysis*, Elsevier, 1994, pp. 2197–2219.
10. D. W. Breck, W. G. Eversole, R. M. Milton, T. B. Reed, and T. L. Thomas, *J. Am. Chem. Soc.*, 1956, **78**, 5963–5972.
11. E. L. First, C. E. Gounaris, J. Wei, and C. A. Floudas, *Phys. Chem. Chem. Phys.*, 2011, **13**, 17339–17358.
12. I. M. Hönicke, I. Senkovska, V. Bon, I. A. Baburin, N. Bönisch, S. Raschke, J. D. Evans, and S. Kaskel, *Angew. Chem. Int. Ed.*, 2018, **57**, 13780–13783.
13. J. K. Schnobrich, K. Koh, K. N. Sura, and A. J. Matzger, *Langmuir*, 2010, **26**, 5808–5814.
14. J. R. Holst, A. Trewin, and A. I. Cooper, *Nat. Chem.*, 2010, **2**, 915–920.
15. A. I. Cooper, *ACS Cent. Sci.*, 2017, **3**, 544–553.
16. J. D. Evans, D. M. Huang, M. Haranczyk, A. W. Thornton, C. J. Sumby, and C. J. Doonan, *CrystEngComm*, 2016, **18**, 4133–4141.

17. R. M. Barrer and V. H. Shanson, *J. Chem. Soc., Chem. Commun.*, 1976, 333–334.
18. A. P. Dianin, *J. Russ. Phys. Chem. Soc*, 1914, **46**, 1310–1319.
19. G. O. Lloyd, M. W. Bredenkamp, and L. J. Barbour, *Chem. Commun.*, 2005, **0**, 4053–4055.
20. E. Eikeland, M. A. Spackman, and B. B. Iversen, *Cryst. Growth Des.*, 2016, **16**, 6858–6866.
21. H. R. Allcock, *Acc. Chem. Res.*, 1978, **11**, 81–87.
22. K. J. Msayib, D. Book, P. M. Budd, N. Chaukura, K. D. M. Harris, M. Helliwell, S. Tedds, A. Walton, J. E. Warren, M. Xu, and N. B. McKeown, *Angew. Chem. Int. Ed.*, 2009, **48**, 3273–3277.
23. W. Yang, A. Greenaway, X. Lin, R. Matsuda, A. J. Blake, C. Wilson, W. Lewis, P. Hubberstey, S. Kitagawa, N. R. Champness, and M. Schröder, *J. Am. Chem. Soc.*, 2010, **132**, 14457–14469.
24. Y. He, S. Xiang, and B. Chen, *J. Am. Chem. Soc.*, 2011, **133**, 14570–14573.
25. D. V. Soldatov, I. L. Moudrakovski, and J. A. Ripmeester, *Angew. Chem. Int. Ed.*, 2004, **43**, 6308–6311.
26. G. Zhang and M. Mastalerz, *Chem. Soc. Rev.*, 2014, **43**, 1934–1947.
27. G. Zhang, O. Presly, F. White, I. M. Oppel, and M. Mastalerz, *Angew. Chem. Int. Ed.*, 2014, **53**, 1516–1520.
28. M. A. Little and A. I. Cooper, *Adv. Funct. Mater.*, 2020, **30**, 1909842.
29. S. Lee, A. Yang, T. P. I. Moneypenny, and J. S. Moore, *J. Am. Chem. Soc.*, 2016, **138**, 2182–2185.
30. K. R. West, K. D. Bake, and S. Otto, *Org. Lett.*, 2005, **7**, 2615–2618.
31. V. Santolini, M. Miklitz, E. Berardo, and K. E. Jelfs, *Nanoscale*, 2017, **9**, 5280–5298.
32. T. Tozawa, J. T. A. Jones, S. I. Swamy, S. Jiang, D. J. Adams, S. Shakespeare, R. Clowes, 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. I. Cooper, *Nat. Mater.*, 2009, **8**, 973–978.
33. P. Skowronek and J. Gawronski, *Org. Lett.*, 2008, **10**, 4755–4758.
34. A. Galán, E. C. Escudero-Adán, and P. Ballester, *Chem. Sci.*, 2017, **8**, 7746–7750.
35. A. Giri, A. Sahoo, T. K. Dutta, and A. Patra, *ACS Omega*, 2020, **5**, 28413–28424.
36. D. Hu, J. Zhang, and M. Liu, *Chem. Commun.*, 2022, **58**, 11333–11346.

37. R. L. Greenaway, V. Santolini, A. Pulido, M. A. Little, B. M. Alston, M. E. Briggs, G. M. Day, A. I. Cooper, and K. E. Jelfs, *Angew. Chem. Int. Ed.*, 2019, **58**, 16275–16281.

38. 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. I. Cooper, *J. Am. Chem. Soc.*, 2016, **138**, 1653–1659.

39. J. Morales-Sanfrutos, M. Ortega-Muñoz, J. Lopez-Jaramillo, F. Hernandez-Mateo, and F. Santoyo-Gonzalez, *J. Org. Chem.*, 2008, **73**, 7772–7774.

40. W. Liu, S. Bobbala, C. L. Stern, J. E. Hornick, Y. Liu, A. E. Enciso, E. A. Scott, and J. F. Stoddart, *J. Am. Chem. Soc.*, 2020, **142**, 3165–3173.

41. T. J. Ryan, G. Lecollinet, T. Velasco, and A. P. Davis, *Proceedings of the National Academy of Sciences*, 2002, **99**, 4863–4866.

42. M. Delecluse, C. Colomban, B. Chatelet, S. Chevallier-Michaud, D. Moraleda, J.-P. Dutasta, and A. Martinez, *J. Org. Chem.*, 2020, **85**, 4706–4711.

43. Z. Wang, N. Sikdar, S.-Q. Wang, X. Li, M. Yu, X.-H. Bu, Z. Chang, X. Zou, Y. Chen, P. Cheng, K. Yu, M. J. Zaworotko, and Z. Zhang, *J. Am. Chem. Soc.*, 2019, **141**, 9408–9414.

44. H. Wang, Y. Jin, N. Sun, W. Zhang, and J. Jiang, *Chem. Soc. Rev.*, 2021, **50**, 8874–8886.

45. S. Tothadi, M. A. Little, T. Hasell, M. E. Briggs, S. Y. Chong, M. Liu, and A. I. Cooper, *CrystEngComm*, 2017, **19**, 4933–4941.

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

47. A. G. Slater and A. I. Cooper, *Science*, 2015, **348**, aaa8075.

48. R. L. Greenaway, V. Santolini, M. J. Bennison, B. M. Alston, C. J. Pugh, M. A. Little, M. Miklitz, E. G. B. Eden-Rump, R. Clowes, A. Shakil, H. J. Cuthbertson, H. Armstrong, M. E. Briggs, K. E. Jelfs, and A. I. Cooper, *Nat. Commun.*, 2018, **9**, 2849.

49. R.-B. Lin, Y. He, P. Li, H. Wang, W. Zhou, and B. Chen, *Chem. Soc. Rev.*, 2019, **48**, 1362–1389.

50. D. Holden, S. Y. Chong, L. Chen, K. E. Jelfs, T. Hasell, and A. I. Cooper, *Chem. Sci.*, 2016, **7**, 4875–4879.

51. M. Miklitz and K. E. Jelfs, *J. Chem. Inf. Model.*, 2018, **58**, 2387–2391.

52. I. Gómez García and M. Haranczyk, *CrystEngComm*, 2020, **22**, 7242–7251.

53. E. Berardo, R. L. Greenaway, M. Miklitz, A. I. Cooper, and K. E. Jelfs, *Mol. Syst. Des. Eng.*, 2020, **5**, 186–196.

54. S. Riniker and G. A. Landrum, *J. Chem. Inf. Model.*, 2015, **55**, 2562–2574.
55. L. Turcani, E. Berardo, and K. E. Jelfs, *J. Comput. Chem.*, 2018, **39**, 1931–1942.
56. L. Turcani, A. Tarzia, F. T. Szczypiński, and K. E. Jelfs, *J. Chem. Phys.*, 2021, **154**, 214102.
57. T. A. Young, R. Gheorghe, and F. Duarte, *J. Chem. Inf. Model.*, 2020, **60**, 3546–3557.
58. P. C. D. Hawkins, *J. Chem. Inf. Model.*, 2017, **57**, 1747–1756.
59. A. K. Rappe, C. J. Casewit, K. S. Colwell, W. A. Goddard, and W. M. Skiff, *J. Am. Chem. Soc.*, 1992, **114**, 10024–10035.
60. H. Sun, *J. Phys. Chem. B*, 1998, **102**, 7338–7364.
61. V. Santolini, G. A. Tribello, and K. E. Jelfs, *Chem. Commun.*, 2015, **51**, 15542–15545.
62. W. L. Jorgensen, D. S. Maxwell, and J. Tirado-Rives, *J. Am. Chem. Soc.*, 1996, **118**, 11225–11236.
63. E. Berardo, R. L. Greenaway, L. Turcani, B. M. Alston, M. J. Bennison, M. Miklitz, R. Clowes, M. E. Briggs, A. I. Cooper, and K. E. Jelfs, *Nanoscale*, 2018, **10**, 22381–22388.
64. A. R. Oganov, C. J. Pickard, Q. Zhu, and R. J. Needs, *Nat. Rev. Mater.*, 2019, **4**, 331–348.
65. P. Zhang, G. P. F. Wood, J. Ma, M. Yang, Y. Liu, G. Sun, Y. A. Jiang, B. C. Hancock, and S. Wen, *Cryst. Growth Des.*, 2018, **18**, 6891–6900.
66. C. Zhao, L. Chen, Y. Che, Z. Pang, X. Wu, Y. Lu, H. Liu, G. M. Day, and A. I. Cooper, *Nat. Commun.*, 2021, **12**, 817.
67. E. O. Pyzer-Knapp, H. P. G. Thompson, F. Schiffmann, K. E. Jelfs, S. Y. Chong, M. A. Little, A. I. Cooper, and G. M. Day, *Chem. Sci.*, 2014, **5**, 2235–2245.
68. P. Cui, D. P. McMahon, P. R. Spackman, B. M. Alston, M. A. Little, G. M. Day, and A. I. Cooper, *Chem. Sci.*, 2019, **10**, 9988–9997.
69. G. Kupgan, L. J. Abbott, K. E. Hart, and C. M. Colina, *Chem. Rev.*, 2018, **118**, 5488–5538.
70. L. J. Abbott, N. B. McKeown, and C. M. Colina, *J. Mater. Chem. A*, 2013, **1**, 11950–11960.
71. 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.
72. I. Rose, C. G. Bezzu, M. Carta, B. Comesaña-Gándara, E. Lasseuguette, M. C. Ferrari, P. Bernardo, G. Clarizia, A. Fuoco, J. C. Jansen, K. E. Hart, T. P. Liyana-Arachchi, C. M. Colina, and N. B. McKeown, *Nat. Mater.*, 2017, **16**, 932–937.

73. B. D. Egleston, A. Mroz, K. E. Jelfs, and R. L. Greenaway, *Chem. Sci.*, 2022, **13**, 5042–5054.

74. G. Melaugh, N. Giri, C. E. Davidson, S. L. James, and M. G. D. Pópolo, *Phys. Chem. Chem. Phys.*, 2014, **16**, 9422–9431.

75. O. Kravchenko, A. Varava, F. T. Pokorny, D. Devaurs, L. E. Kavraki, and D. Kragic, *J. Chem. Inf. Model.*, 2020, **60**, 1302–1316.

76. J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, and D. Hassabis, *Nature*, 2021, **596**, 583–589.

77. T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, *Language Models Are Few-Shot Learners*, 2020, `http://arxiv.org/abs/2005.14165` (visited on 11/14/2022), preprint.

78. K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, and A. Walsh, *Nat. Mater.*, 2018, **559**, 547–555.

79. Y. LeCun, Y. Bengio, and G. Hinton, *Nature*, 2015, **521**, 436–444.

80. W. S. McCulloch and W. Pitts, *Bulletin of Mathematical Biophysics*, 1943, **5**, 115–133.

81. A. Korotcov, V. Tkachenko, D. P. Russo, and S. Ekins, *Mol. Pharmaceutics*, 2017, **14**, 4462–4475.

82. S. Bennett, A. Tarzia, M. A. Zwijnenburg, and K. E. Jelfs, in *Machine Learning in Chemistry: The Impact of Artificial Intelligence*, ed. H. M. Cartwright, Royal Society of Chemistry, Cambridge, 2020, pp. 280–310.

83. J. Vamathevan, D. Clark, P. Czodrowski, I. Dunham, E. Ferran, G. Lee, B. Li, A. Madabhushi, P. Shah, M. Spitzer, and S. Zhao, *Nat. Rev. Drug. Discov.*, 2019, **18**, 463–477.

84. K. M. Jablonka, D. Ongari, S. M. Moosavi, and B. Smit, *Chem. Rev.*, 2020, **120**, 8066–8129.

85. B. Kim, S. Lee, and J. Kim, *Sci. Adv.*, 2020, **6**, eaax9324.

86. S. Bennett and K. E. Jelfs, in *AI-Guided Design and Property Prediction for Zeolites and Nanoporous Materials*, John Wiley & Sons, Ltd, 2023, pp. 251–282.
87. S. M. Moosavi, H. Xu, L. Chen, A. I. Cooper, and B. Smit, *Chem. Sci.*, 2020, **11**, 5423–5433.
88. C. Altintas, O. F. Altundal, S. Keskin, and R. Yildirim, *J. Chem. Inf. Model.*, 2021, **61**, 2131–2146.
89. L. Turcani, R. L. Greenaway, and K. E. Jelfs, *Chem. Mater.*, 2019, **31**, 714–727.
90. A. Bender, H. Y. Mussa, R. C. Glen, and S. Reiling, *J. Chem. Inf. Comput. Sci.*, 2004, **44**, 1708–1718.
91. P. Nguyen, D. Loveland, J. T. Kim, P. Karande, A. M. Hiszpanski, and T. Y.-J. Han, *J. Chem. Inf. Model.*, 2021, **61**, 2147–2158.
92. A. P. Bartók, R. Kondor, and G. Csányi, *Phys. Rev. B*, 2013, **87**, 184115.
93. S. Li, Y. Liu, D. Chen, Y. Jiang, Z. Nie, and F. Pan, *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, 2022, **12**, e1558.
94. Y. Lee, S. D. Barthel, P. Dłotko, S. M. Moosavi, K. Hess, and B. Smit, *Nat. Commun.*, 2017, **8**, 15396.
95. A. Li, R. Bueno-Perez, and D. Fairen-Jimenez, *Chem. Sci.*, 2022, 10.1039.D2SC03171J.
96. J. Yang, S. De, J. E. Campbell, S. Li, M. Ceriotti, and G. M. Day, *Chem. Mater.*, 2018, **30**, 4361–4371.
97. A. Sturluson, M. T. Huynh, A. H. P. York, and C. M. Simon, *ACS Cent. Sci.*, 2018, **4**, 1663–1676.
98. A. C. Mater and M. L. Coote, *J. Chem. Inf. Model.*, 2019, **59**, 2545–2559.
99. Q. Yuan, F. T. Szczypiński, and K. E. Jelfs, *Digital Discovery*, 2022, **1**, 127–138.
100. F. T. Szczypiński, S. Bennett, and K. E. Jelfs, *Chem. Sci.*, 2021, **12**, 830–840.
101. S. H. Bertz, *J. Am. Chem. Soc.*, 1981, **103**, 3599–3601.
102. C. W. Coley, L. Rogers, W. H. Green, and K. F. Jensen, *J. Chem. Inf. Model.*, 2018, **58**, 252–261.
103. P. Ertl and A. Schuffenhauer, *J. Cheminform.*, 2009, **1**, 8.
104. Y. Baba, T. Isomura, and H. Kashima, *J. Mol. Graph.*, 2018, **80**, 217–223.
105. S. M. Kearnes, M. R. Maser, M. Wleklinski, A. Kast, A. G. Doyle, S. D. Dreher, J. M. Hawkins, K. F. Jensen, and C. W. Coley, *J. Am. Chem. Soc.*, 2021, **143**, 18820–18826.

106. A. J. Lawson, J. Swienty-Busch, T. Géoui, and D. Evans, in *The Future of the History of Chemical Information*, ACS Publications, 2014, pp. 127–148.

107. M. H. S. Segler, M. Preuss, and M. P. Waller, *Nature*, 2018, **555**, 604–610.

108. P. Schwaller, T. Laino, T. Gaudin, P. Bolgar, C. A. Hunter, C. Bekas, and A. A. Lee, *ACS Cent. Sci.*, 2019, **5**, 1572–1583.

109. T. Klucznik, B. Mikulak-Klucznik, M. P. McCormack, H. Lima, S. Szymkuć, M. Bhowmick, K. Molga, Y. Zhou, L. Rickershauser, E. P. Gajewska, A. Toutchkine, P. Dittwald, M. P. Startek, G. J. Kirkovits, R. Roszak, A. Adamski, B. Sieredzińska, M. Mrksich, S. L. J. Trice, and B. A. Grzybowski, *Chem*, 2018, **4**, 522–532.

110. K. E. Jelfs, E. G. B. Eden, J. L. Culshaw, S. Shakespeare, E. O. Pyzer-Knapp, H. P. G. Thompson, J. Bacsa, G. M. Day, D. J. Adams, and A. I. Cooper, *J. Am. Chem. Soc.*, 2013, **135**, 9307–9310.

111. S. Bera, A. Basu, S. Tothadi, B. Garai, S. Banerjee, K. Vanka, and R. Banerjee, *Angew. Chem. Int. Ed.*, 2017, **56**, 2123–2126.

112. 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.

113. S. P. Collins, T. D. Daff, S. S. Piotrkowski, and T. K. Woo, *Sci. Adv.*, 2016, **2**, e1600954.

114. Y. Kwon, S. Kang, Y.-S. Choi, and I. Kim, *Sci. Rep.*, 2021, **11**, 17304.

115. E. Berardo, L. Turcani, M. Miklitz, and K. E. Jelfs, *Chem. Sci.*, 2018, **9**, 8513–8527.

116. M. Miklitz, L. Turcani, R. L. Greenaway, and K. E. Jelfs, *Commun. Chem.*, 2020, **3**, 10.

117. E. S. Henault, M. H. Rasmussen, and J. H. Jensen, *PeerJ Phys. Chem.*, 2020, **2**, e11.

118. A. Deshwal, C. M. Simon, and J. R. Doppa, *Mol. Syst. Des. Eng.*, 2021, **6**, 1066–1086.

119. P. I. Frazier, *A Tutorial on Bayesian Optimization*, 2018, `http://arxiv.org/abs/1807.02811` (visited on 11/25/2022), preprint.

120. E. O. Pyzer-Knapp, L. Chen, G. M. Day, and A. I. Cooper, *Sci. Adv.*, 2021, **7**, eabi4763.

121. W. Gao and C. W. Coley, *J. Chem. Inf. Model.*, 2020, **60**, 5714–5723.

122. M. Shevlin, *ACS Med. Chem. Lett.*, 2017, **8**, 601–607.

123. T. Cernak, N. J. Gesmundo, K. Dykstra, Y. Yu, Z. Wu, Z.-C. Shi, P. Vachal, D. Sperbeck, S. He, B. A. Murphy, L. Sonatore, S. Williams, M. Madeira, A. Verras, M. Reiter, C. H. Lee, J. Cuff, E. C. Sherer, J. Kuethe, S. Goble, N. Perrotto, S. Pinto, D.-M. Shen, R. Nargund, J. Balkovec, R. J. DeVita, and S. D. Dreher, *J. Med. Chem.*, 2017, **60**, 3594–3605.

124. M. L. Green, I. Takeuchi, and J. R. Hattrick-Simpers, *J. Appl. Phys.*, 2013, **113**, 231101.

125. D. W. Robbins and J. F. Hartwig, *Science*, 2011, **333**, 1423–1427.

126. A. Buitrago Santanilla, E. L. Regalado, T. Pereira, M. Shevlin, K. Bateman, L.-C. Campeau, J. Schneeweis, S. Berritt, Z.-C. Shi, P. Nantermet, Y. Liu, R. Helmy, C. J. Welch, P. Vachal, I. W. Davies, T. Cernak, and S. D. Dreher, *Science*, 2015, **347**, 49–53.

127. R. J. Kearsey, B. M. Alston, M. E. Briggs, R. L. Greenaway, and A. I. Cooper, *Chem. Sci.*, 2019, **10**, 9454–9465.

128. H. S. Stein and J. M. Gregoire, *Chem. Sci.*, 2019, **10**, 9640–9649.

129. M. Christensen, L. P. E. Yunker, F. Adedeji, F. Häse, L. M. Roch, T. Gensch, G. dos Passos Gomes, T. Zepel, M. S. Sigman, A. Aspuru-Guzik, and J. E. Hein, *Commun. Chem.*, 2021, **4**, 1–12.

130. F. Häse, M. Aldeghi, R. J. Hickman, L. M. Roch, and A. Aspuru-Guzik, *Appl. Phys. Rev.*, 2021, **8**, 031406.

131. N. H. Angello, V. Rathore, W. Beker, A. Wołos, E. R. Jira, R. Roszak, T. C. Wu, C. M. Schroeder, A. Aspuru-Guzik, B. A. Grzybowski, and M. D. Burke, *Science*, 2022, **378**, 399–405.

132. L. Porwol, D. J. Kowalski, A. Henson, D.-L. Long, N. L. Bell, and L. Cronin, *Angew. Chem. Int. Ed.*, 2020, **59**, 11256–11261.

133. S. Langner, F. Häse, J. D. Perea, T. Stubhan, J. Hauch, L. M. Roch, T. Heumueller, A. Aspuru-Guzik, and C. J. Brabec, *Adv. Mater.*, 2020, **32**, 1907801.

134. B. P. MacLeod, F. G. L. Parlane, T. D. Morrissey, F. Häse, L. M. Roch, K. E. Dettelbach, R. Moreira, L. P. E. Yunker, M. B. Rooney, J. R. Deeth, V. Lai, G. J. Ng, H. Situ, R. H. Zhang, M. S. Elliott, T. H. Haley, D. J. Dvorak, A. Aspuru-Guzik, J. E. Hein, and C. P. Berlinguette, *Sci. Adv.*, 2020, **6**, eaaz8867.

135. X. Jia, A. Lynch, Y. Huang, M. Danielson, I. Lang'at, A. Milder, A. E. Ruby, H. Wang, S. A. Friedler, A. J. Norquist, and J. Schrier, *Nature*, 2019, **573**, 251–255.

136. 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.

137. S. A. Hollingsworth and R. O. Dror, *Neuron*, 2018, **99**, 1129–1143.

138. L. Verlet, *Phys. Rev.*, 1967, **159**, 98–103.

139. L. L. C. Schrödinger, *Schrödinger Release 2018-1: MacroModel*, New York, NY, 2018.

140. L. H. Thomas, *Math. Proc. Camb. Phil. Soc.*, 1927, **23**, 542–548.

141. P. Hohenberg and W. Kohn, *Phys. Rev.*, 1964, **136**, B864–B871.

142. W. Kohn and L. J. Sham, *Phys. Rev.*, 1965, **140**, A1133–A1138.

143. R. O. Jones, *Rev. Mod. Phys.*, 2015, **87**, 897–923.

144. D. M. Ceperley and B. J. Alder, *Phys. Rev. Lett.*, 1980, **45**, 566–569.

145. J. P. Perdew, M. Ernzerhof, and K. Burke, *J. Chem. Phys.*, 1996, **105**, 9982–9985.

146. C. Lee, W. Yang, and R. G. Parr, *Phys. Rev. B*, 1988, **37**, 785–789.

147. J. P. Perdew and Y. Wang, *Phys. Rev. B*, 1992, **45**, 13244–13249.

148. S. F. Sousa, P. A. Fernandes, and M. J. Ramos, *J. Phys. Chem. A*, 2007, **111**, 10439–10452.

149. C. Adamo and V. Barone, *J. Chem. Phys.*, 1999, **110**, 6158–6170.

150. J. Kirkpatrick, B. McMorrow, D. H. P. Turban, A. L. Gaunt, J. S. Spencer, A. G. D. G. Matthews, A. Obika, L. Thiry, M. Fortunato, D. Pfau, L. R. Castellanos, S. Petersen, A. W. R. Nelson, P. Kohli, P. Mori-Sánchez, D. Hassabis, and A. J. Cohen, *Science*, 2021, **374**, 1385–1389.

151. T. D. Kühne, M. Iannuzzi, M. Del Ben, V. V. Rybkin, P. Seewald, F. Stein, T. Laino, R. Z. Khaliullin, O. Schütt, F. Schiffmann, D. Golze, J. Wilhelm, S. Chulkov, M. H. Bani-Hashemian, V. Weber, U. Borštnik, M. Taillefumier, A. S. Jakobovits, A. Lazzaro, H. Pabst, T. Müller, R. Schade, M. Guidon, S. Andermatt, N. Holmberg, G. K. Schenter, A. Hehn, A. Bussy, F. Belleflamme, G. Tabacchi, A. Glöß, M. Lass, I. Bethune, C. J. Mundy, C. Plessl, M. Watkins, J. VandeVondele, M. Krack, and J. Hutter, *J. Chem. Phys.*, 2020, **152**, 194103.

152. S. Grimme, J. Antony, S. Ehrlich, and H. Krieg, *J. Chem. Phys.*, 2010, **132**, 154104.

153. J. Witte, N. Mardirossian, J. B. Neaton, and M. Head-Gordon, *J. Chem. Theory Comput.*, 2017, **13**, 2043–2052.

154. A. Kensert, J. Alvarsson, U. Norinder, and O. Spjuth, *J. Cheminform.*, 2018, **10**, 49.

155. J. B. O. Mitchell, *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, 2014, **4**, 468–481.

156. T. Chen and C. Guestrin, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794.

157. G.-H. Fu, L.-Z. Yi, and J. Pan, *Biom. J.*, 2019, **61**, 652–664.

158. A. Cereto-Massagué, M. J. Ojeda, C. Valls, M. Mulero, S. Garcia-Vallvé, and G. Pujadas, *Methods*, 2015, **71**, 58–63.

159. D. Rogers and M. Hahn, *J. Chem. Inf. Model.*, 2010, **50**, 742–754.

160. K. Boda, T. Seidel, and J. Gasteiger, *J. Comput. Aided. Mol. Des.*, 2007, **21**, 311–325.

161. V. J. Gillet, G. Myatt, Z. Zsoldos, and A. P. Johnson, *Perspect. Drug. Discov. Des.*, 1995, **3**, 34–50.

162. P. S. Kutchukian, N. Y. Vasilyeva, J. Xu, M. K. Lindvall, M. P. Dillon, M. Glick, J. D. Coley, and N. Brooijmans, *PLoS ONE*, 2012, **7**, e48476.

163. P. Bonnet, *Eur. J. Med. Chem.*, 2012, **54**, 679–689.

164. K. Grodowska and A. Parczewski, *Acta. Pol. Pharm.*, 2010, **67**, 3–12.

165. D. G. Brown and J. Boström, *J. Med. Chem.*, 2018, **61**, 9442–9468.

166. Y. Takaoka, Y. Endo, S. Yamanobe, H. Kakinuma, T. Okubo, Y. Shimazaki, T. Ota, S. Sumiya, and K. Yoshikawa, *J. Chem. Inf. Comput. Sci.*, 2003, **43**, 1269–1275.

167. R. P. Sheridan, N. Zorn, E. C. Sherer, L.-C. Campeau, C. Chang, J. Cumming, M. L. Maddess, P. G. Nantermet, C. J. Sinz, and P. D. O'Shea, *J. Chem. Inf. Model.*, 2014, **54**, 1604–1616.

168. M. S. Lajiness, G. M. Maggiora, and V. Shanmugasundaram, *J. Med. Chem.*, 2004, **47**, 4891–4896.

169. *eMolecules*, `https://www.emolecules.com/` (visited on 01/09/2023).

170. G. W. Bemis and M. A. Murcko, *J. Med. Chem.*, 1996, **39**, 2887–2893.

171. V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan, and B. P. Feuston, *J. Chem. Inf. Comput. Sci.*, 2003, **43**, 1947–1958.

172. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, *JAIR*, 2002, **16**, 321–357.

173. G. Landrum, *RDKit: Open-source Cheminformatics*, `http://www.rdkit.org`.

174. N. M. O'Boyle and R. A. Sayle, *J. Cheminform.*, 2016, **8**, 36.

175. A. Niculescu-Mizil and R. Caruana, Proceedings of the 22nd International Conference on Machine Learning - ICML '05, The 22nd International Conference, ACM Press, Bonn, Germany, 2005, pp. 625–632.

176. S. Bennett, F. T. Szczypiński, L. Turcani, M. E. Briggs, R. L. Greenaway, and K. E. Jelfs, *J. Chem. Inf. Model.*, 2021, **61**, 4342–4356.

177. A. Zimek, E. Schubert, and H.-P. Kriegel, *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 2012, **5**, 363–387.

178. A. C. Flick, H. X. Ding, C. A. Leverett, R. E. Kyne, K. Liu, S. J. Fink, and C. J. O'Donnell, *J. Med. Chem.*, 2017, **60**, 6480–6515.

179. Y. Wen, L. Fu, G. Li, J. Ma, and H. Ma, *Sol. RRL*, 2020, **4**, 2000110.

180. L. Wen, B. Wang, T. Yu, W. Lai, J. Shi, M. Liu, and Y. Liu, *Fuel*, 2022, **310**, 122241.

181. P. Ertl, *J. Cheminform.*, 2017, **9**, 36.

182. P. Tosco, N. Stiefl, and G. Landrum, *J. Cheminform.*, 2014, **6**, 37.

183. E. Polak and G. Ribiere, *R.I.R.O.*, 1969, **3**, 35–43.

184. S. Madhu, E. V. Rashmi, R. G. Gonnade, and G. J. Sanjayan, *New J. Chem.*, 2017, **41**, 8721–8724.

185. J. J. Irwin and B. K. Shoichet, *J. Chem. Inf. Model.*, 2005, **45**, 177–182.

186. A. Garzón, M. P. Fernández-Liencres, M. Moral, T. Peña-Ruiz, A. Navarro, J. Tolosa, J. Canales-Vázquez, D. Hermida-Merino, I. Bravo, J. Albaladejo, and J. C. García-Martínez, *J. Phys. Chem. C*, 2017, **121**, 4720–4733.

187. A. Ludwig, *npj Comput. Mater.*, 2019, **5**, 1–7.

188. M. Petryk, J. Szymkowiak, B. Gierczyk, G. Spólnik, Ł. Popenda, A. Janiak, and M. Kwit, *Org. Biomol. Chem.*, 2016, **14**, 7495–7499.

189. B. Teng, M. A. Little, T. Hasell, S. Y. Chong, K. E. Jelfs, R. Clowes, M. E. Briggs, and A. I. Cooper, *Cryst. Growth Des.*, 2019, **19**, 3647–3651.

190. S. Lee, D. Kim, J. Kim, and O.-S. Jung, *Cryst. Growth Des.*, 2020, **20**, 3601–3604.

191. M. E. Briggs and A. I. Cooper, *Chem. Mater.*, 2017, **29**, 149–157.

192. Y. Jin, Q. Wang, P. Taynton, and W. Zhang, *Acc. Chem. Res.*, 2014, **47**, 1575–1586.

193. B. D. Egleston, M. C. Brand, F. Greenwell, M. E. Briggs, S. L. James, A. I. Cooper, D. E. Crawford, and R. L. Greenaway, *Chem. Sci.*, 2020, **11**, 6582–6589.

194. Z. Sun, P. Li, S. Xu, Z.-Y. Li, Y. Nomura, Z. Li, X. Liu, and S. Zhang, *J. Am. Chem. Soc.*, 2020, **142**, 10833–10840.

195. A. Howarth and J. M. Goodman, *Chem. Sci.*, 2022, **13**, 3507–3518.

196. K. Wolinski, J. F. Hinton, and P. Pulay, *J. Am. Chem. Soc.*, 1990, **112**, 8251–8260.

197. Y. Guan, S. V. S. Sowndarya, L. C. Gallegos, P. C. S. John, and R. S. Paton, *Chem. Sci.*, 2021, **12**, 12012–12026.

198. J. J. Helmus and C. P. Jaroniec, *J Biomol. NMR*, 2013, **55**, 355–367.

199. R. Adusumilli and P. Mallick, *Methods Mol. Biol.*, 2017, **1550**, 339–368.

200. T. Pluskal, S. Castillo, A. Villar-Briones, and M. Orešič, *BMC Bioinformatics*, 2010, **11**, 395.

201. O. D. Myers, S. J. Sumner, S. Li, S. Barnes, and X. Du, *Anal. Chem.*, 2017, **89**, 8696–8703.

202. H. L. Röst, U. Schmitt, R. Aebersold, and L. Malmström, *Proteomics*, 2014, **14**, 74–77.

203. 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.

204. W. Dong, Z. Qin, K. Wang, Y. Xiao, X. Liu, S. Ren, and L. Li, *Angew. Chem. Int. Ed.*, 2023, **62**, e202216073.

205. P. Kieryk, J. Janczak, J. Panek, M. Miklitz, and J. Lisowski, *Org. Lett.*, 2016, **18**, 12–15.

206. R. J. Kearsey, A. Tarzia, M. A. Little, M. C. Brand, R. Clowes, K. E. Jelfs, A. I. Cooper, and R. L. Greenaway, *Chem. Commun.*, 2023, **59**, 3731–3734.

207. C. J. Pugh, V. Santolini, R. L. Greenaway, M. A. Little, M. E. Briggs, K. E. Jelfs, and A. I. Cooper, *Cryst. Growth Des.*, 2018, **18**, 2759–2764.

208. H. Ding, Y. Yang, B. Li, F. Pan, G. Zhu, M. Zeller, D. Yuan, and C. Wang, *Chem. Commun.*, 2015, **51**, 1976–1979.

209. L. Zou, X. Yang, S. Yuan, and H.-C. Zhou, *CrystEngComm*, 2017, **19**, 4868–4871.

210. W. Zhao, P. Yan, B. Li, M. Bahri, L. Liu, X. Zhou, R. Clowes, N. D. Browning, Y. Wu, J. W. Ward, and A. I. Cooper, *J. Am. Chem. Soc.*, 2022, **144**, 9902–9909.

211. M. E. Jung and G. Piizzi, *Chem. Rev.*, 2005, **105**, 1735–1766.

212. E. Díez-Barra, J. C. García-Martínez, S. Merino, R. del Rey, J. Rodríguez-López, P. Sánchez-Verdú, and J. Tejeda, *J. Org. Chem.*, 2001, **66**, 5664–5670.

A Appendix

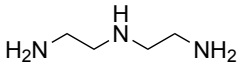
In the appendix, I will present the 120 diamines and trialdehydes re-ranked by R.L.G, alongside their respective synthetic difficulty scores, calculated by the SAScore, SCScore, and MPScore, corresponding to the results of Chapter 3.

Next, I will present additional data related to the high-throughput screen in Chapter 5. I will present the stacked ^{1}H NMR spectra for all 144 precursor combinations from the crude reaction mixture, ordered by diamine. In addition, I will present the mass spectra for the 51 precursor combinations where cage formation occurred but formed a mixture of topologies, insoluble precipitate, or only partial conversion to a cage was observed. In the stacked ^{1}H NMR spectra, the presence of an aldehyde peak (9-11 ppm) would indicate only partial conversion to imine products has occurred. Meanwhile, the presence of a mass ion peak in the HRMS corresponding to a specific cage species would indicate that that species has formed experimentally.

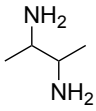
A.1 Chapter 3 Supplementary Data

Figure A.1: Molecules provided to R.L.G for the blind classification task are shown across the next 21 pages. The molecules selected were diamines and trialdehydes with the highest and lowest synthetic difficulty scores calculated by SAScore, SCScore and MPScore, and were presented to R.L.G, who was then asked: "Can you make 1 g of this compound in under 5 steps?" Synthetic difficulty scores for each molecule are shown in Table A.1.

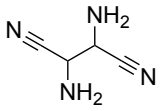
Lowest SCScore Amines



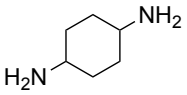
1



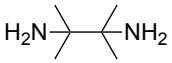
2



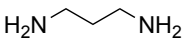
3



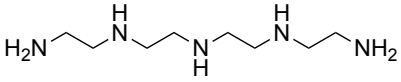
4



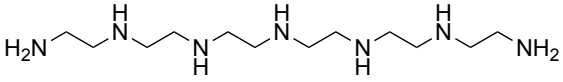
5



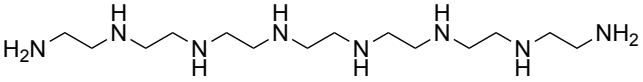
6



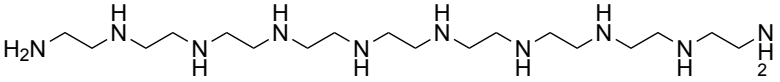
7



8

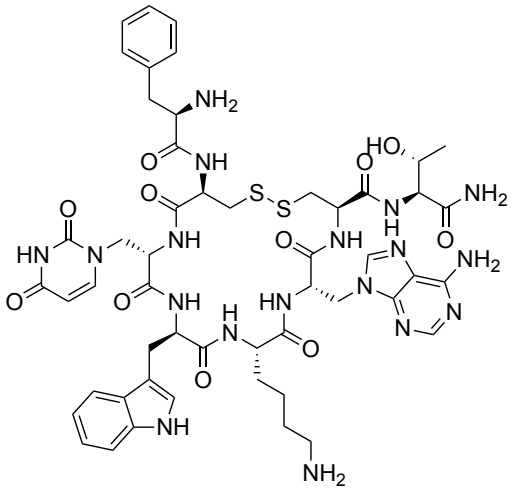


9

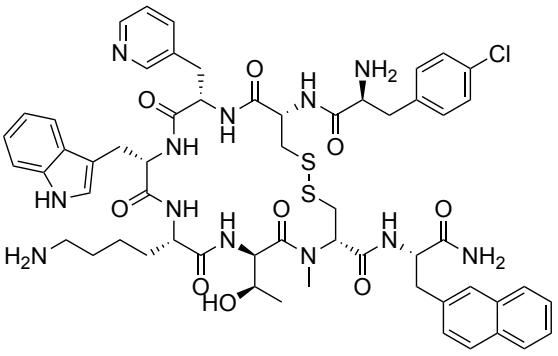


10

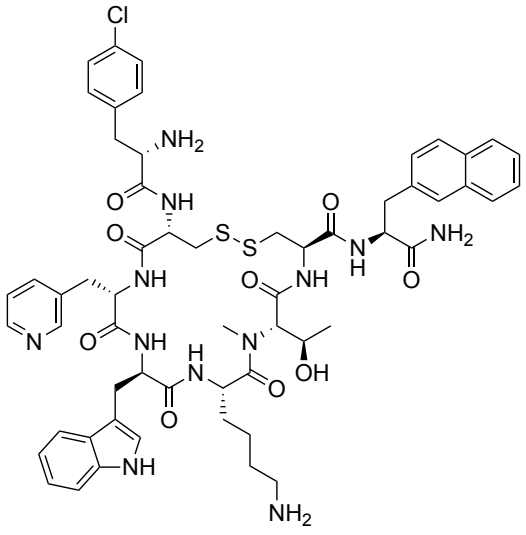
Highest SCScore Amines



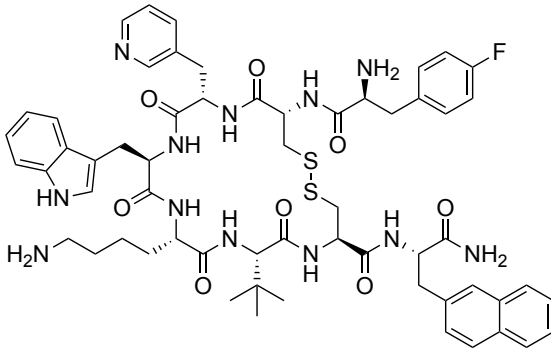
11



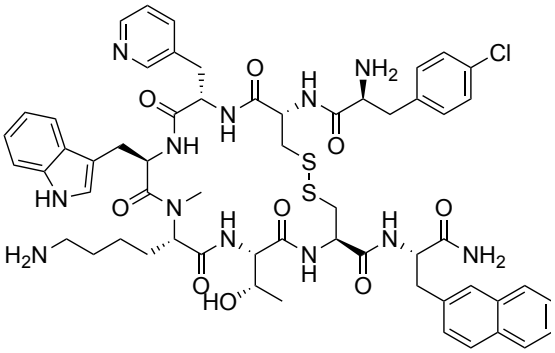
12



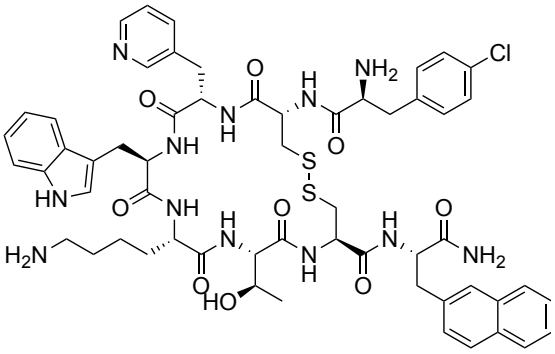
13



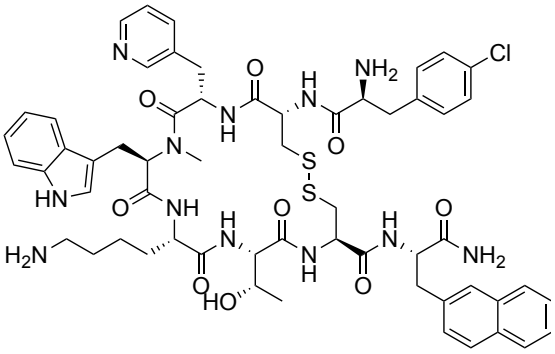
14



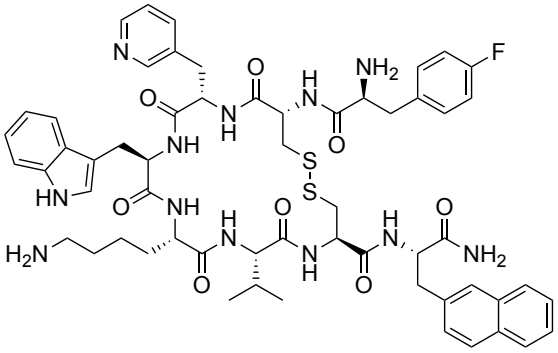
15



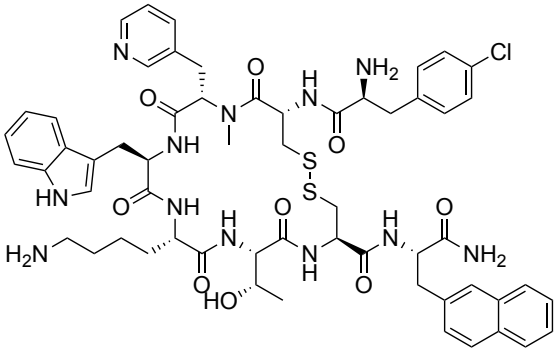
16



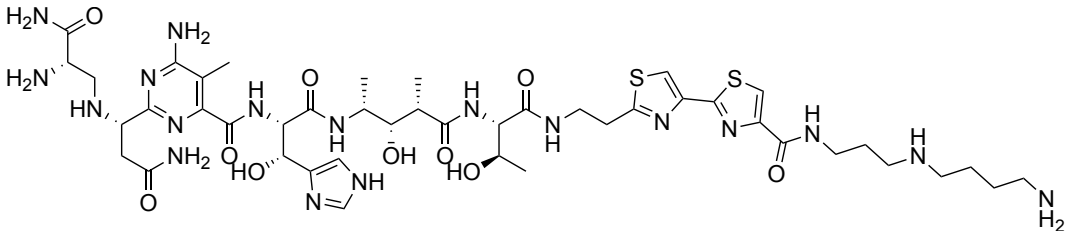
17



18

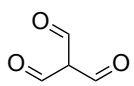


19

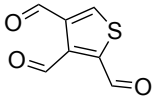


20

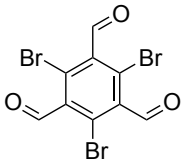
Lowest SCScore Aldehydes



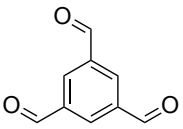
21



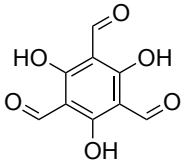
22



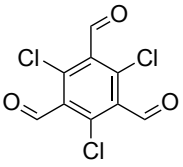
23



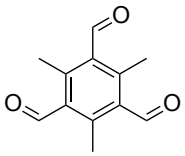
24



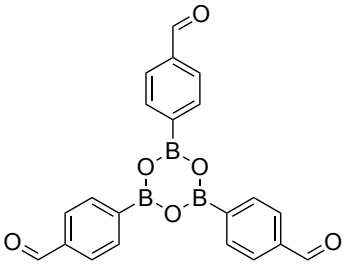
25



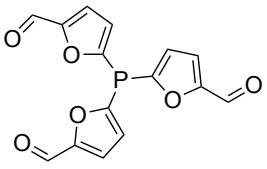
26



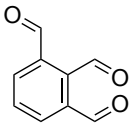
27



28

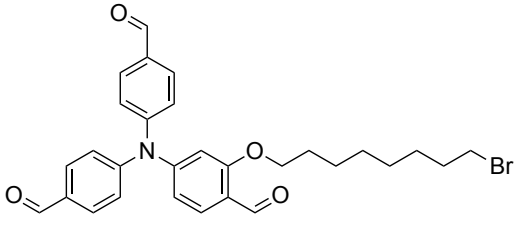


29

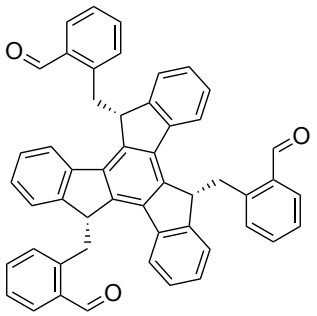


30

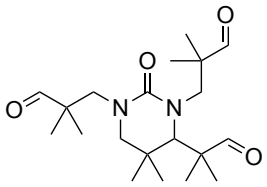
Highest SCScore Aldehydes



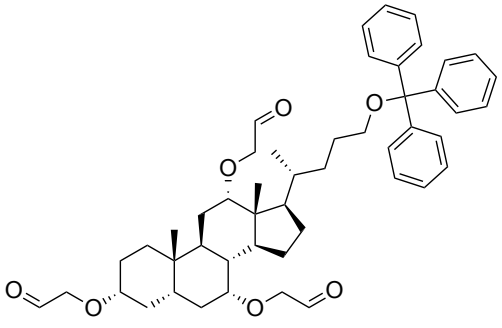
31



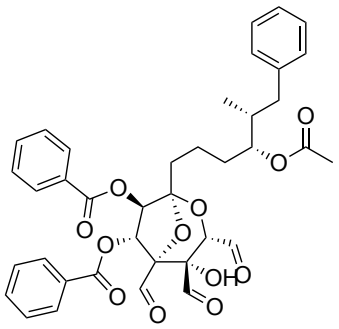
32



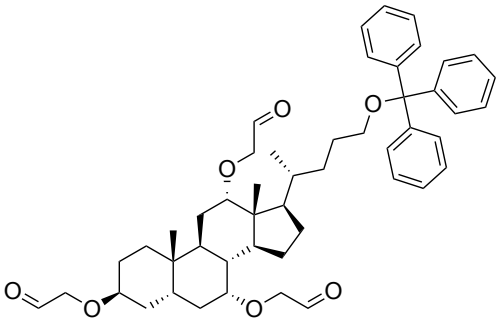
33



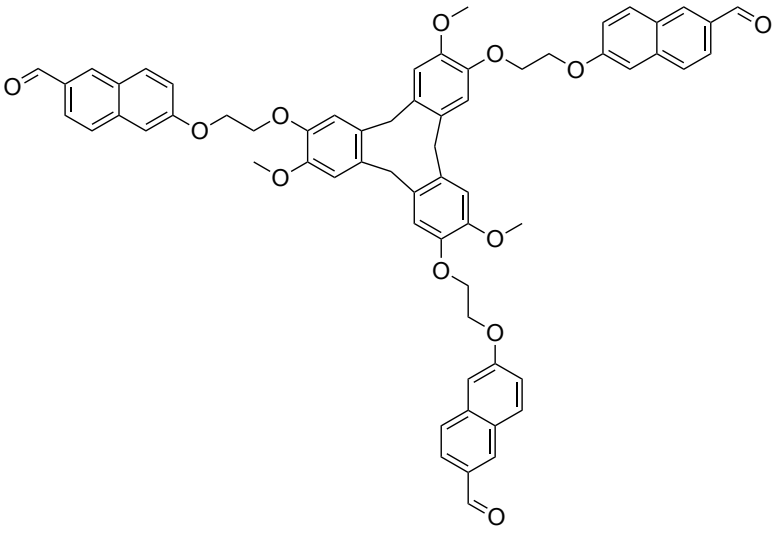
34



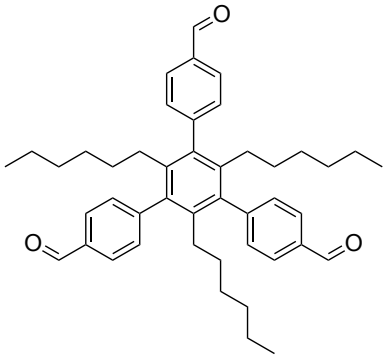
35



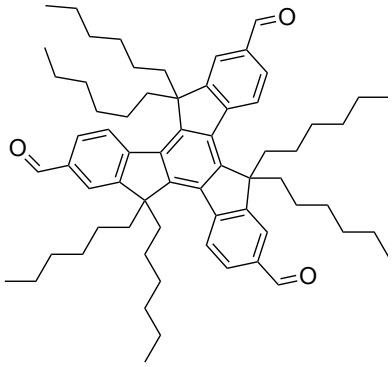
36



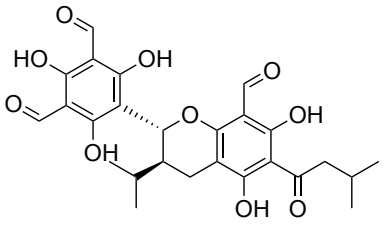
37



38

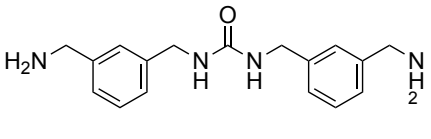


39

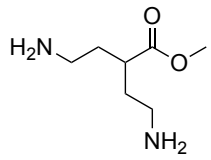


40

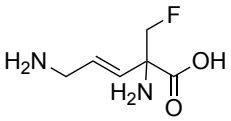
Lowest SAScore Amines



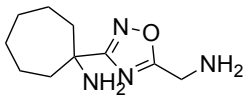
41



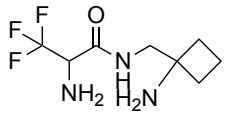
42



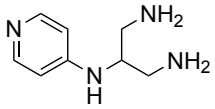
43



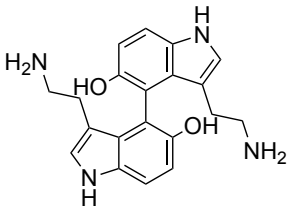
44



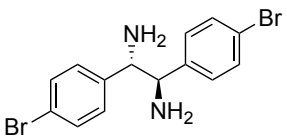
45



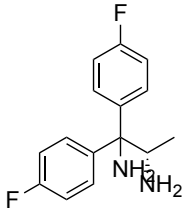
46



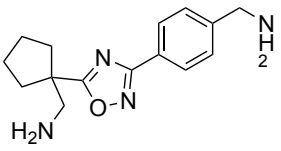
47



48

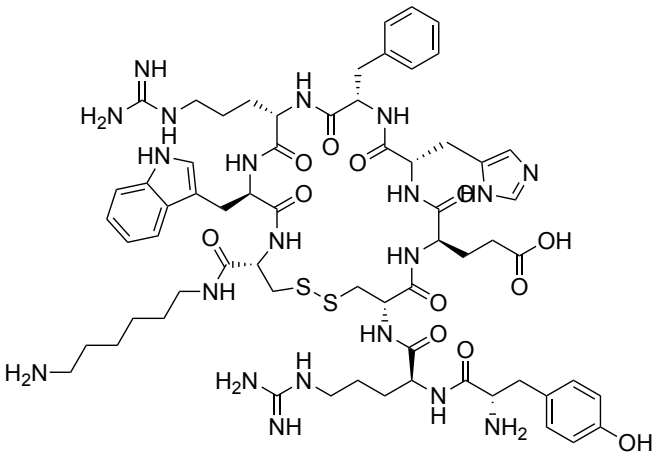


49

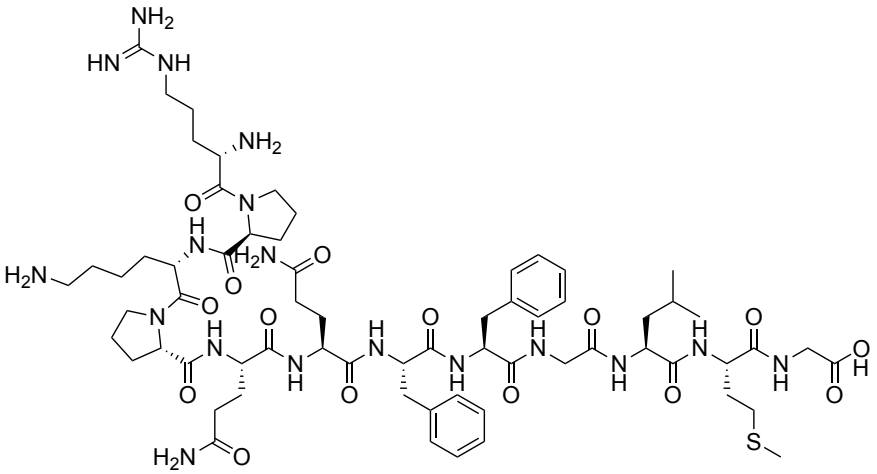


50

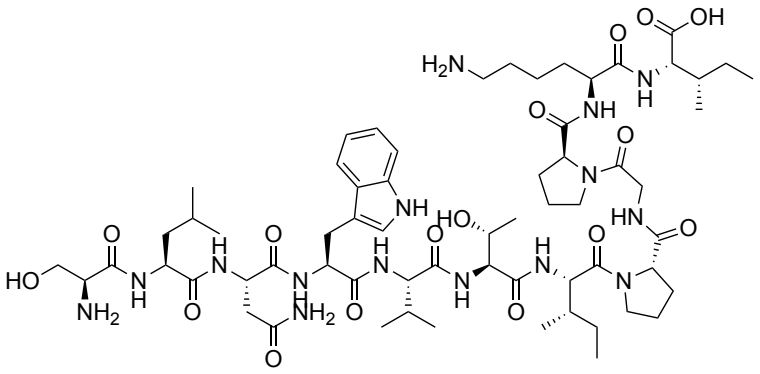
Highest SAScore Amines



51

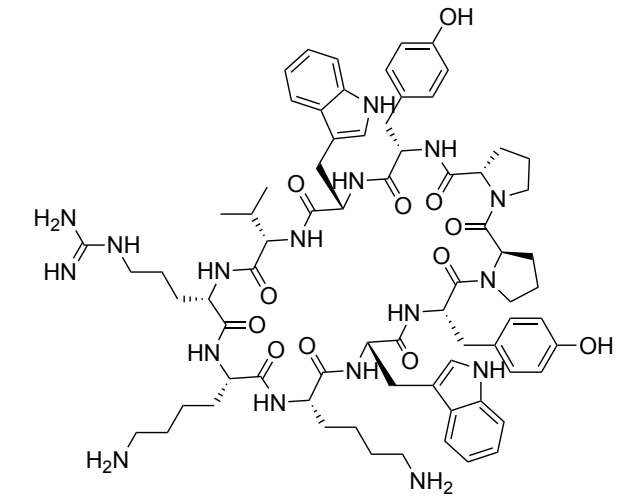


52

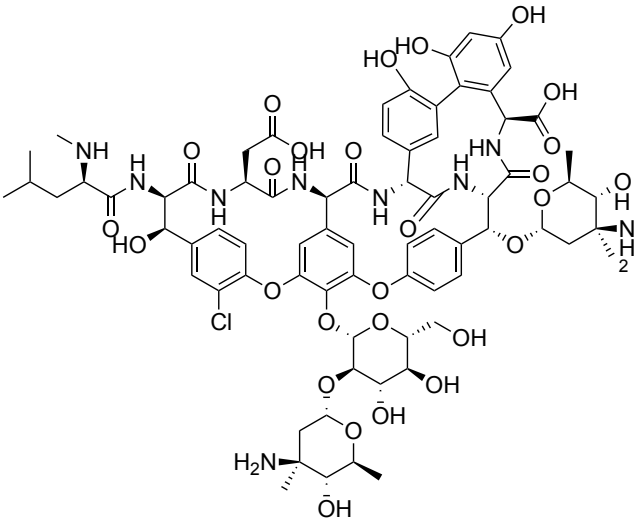


53

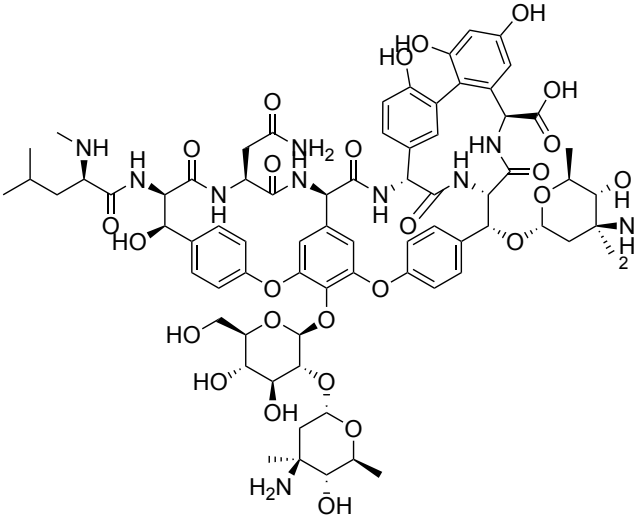
Highest SAScore Amines



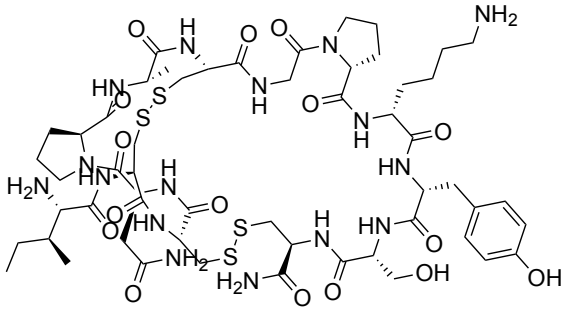
54



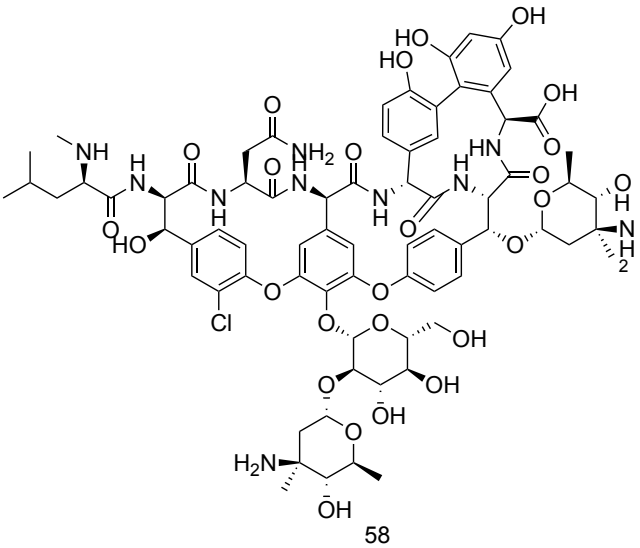
55



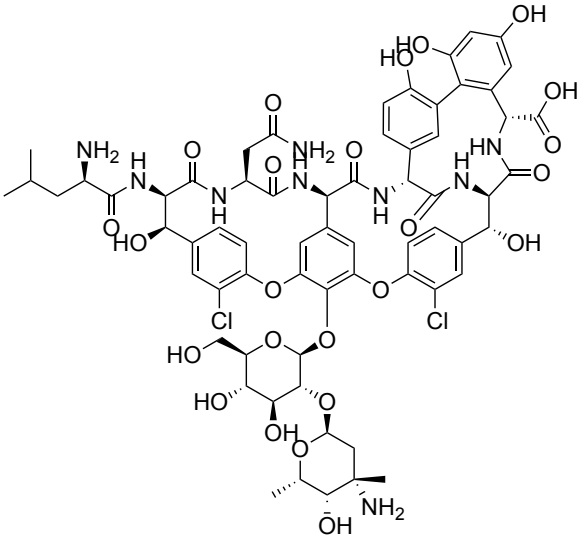
56



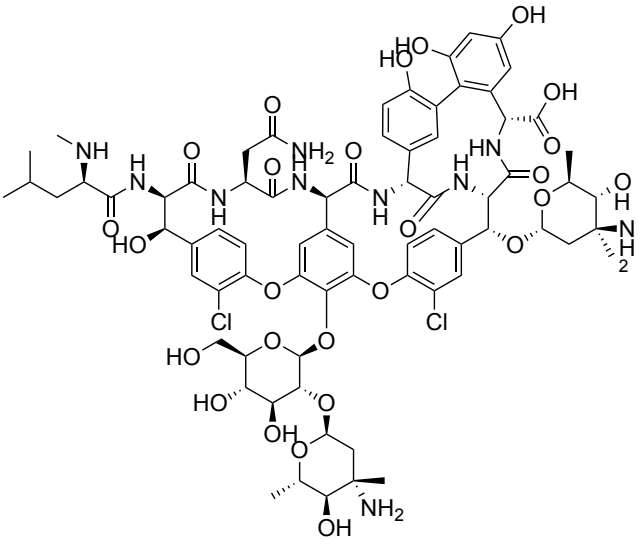
57



58

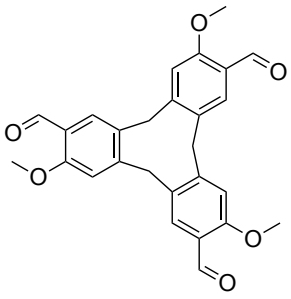


59

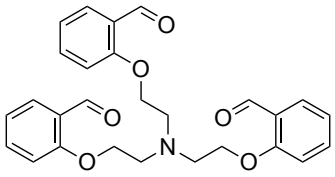


60

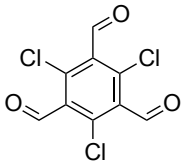
Lowest SAScore Aldehydes



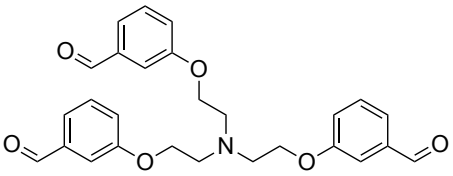
61



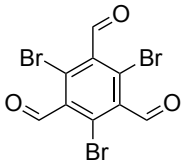
62



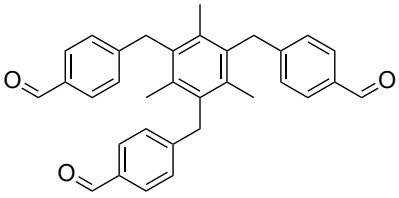
63



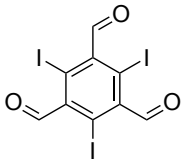
64



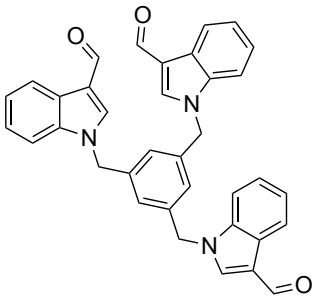
65



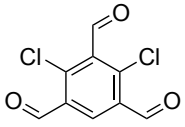
66



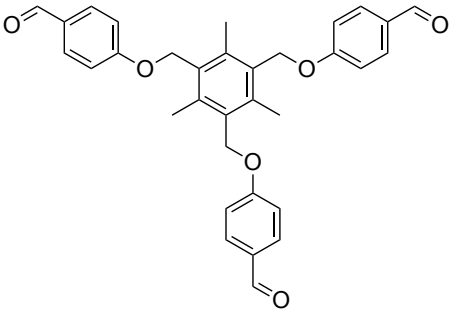
67



68

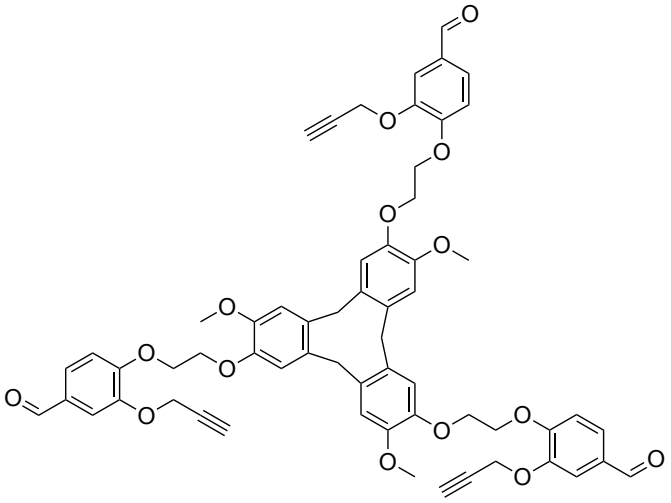


69

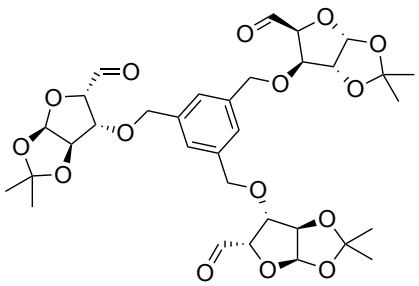


70

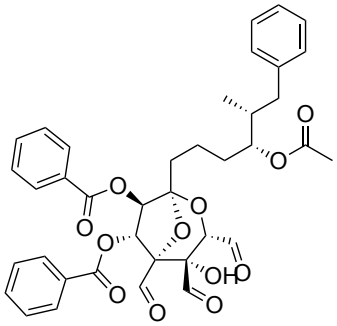
Highest SAScore Aldehydes



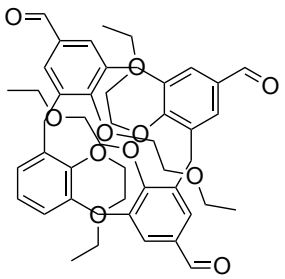
71



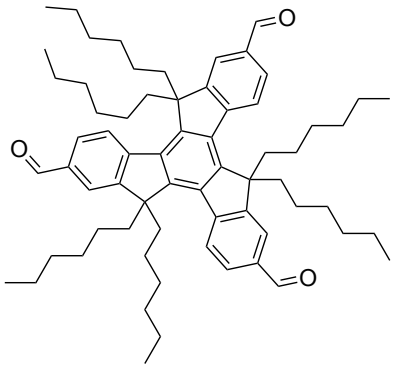
72



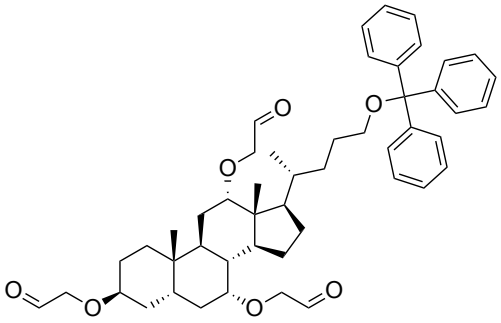
73



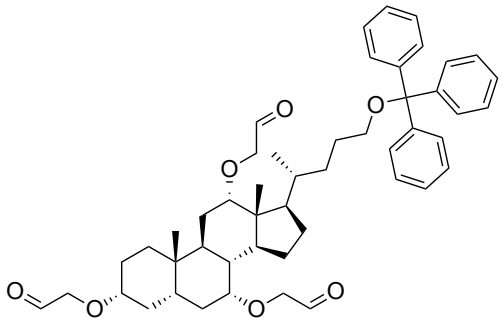
74



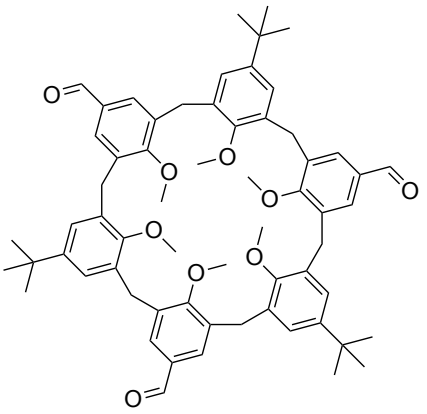
75



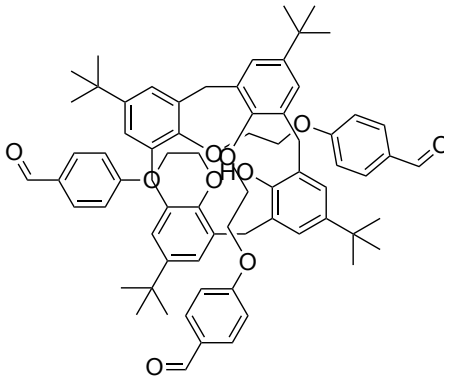
76



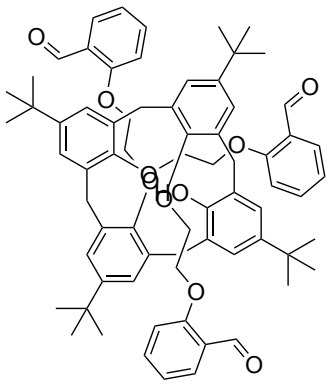
77



78

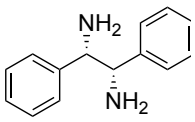


79

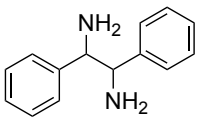


80

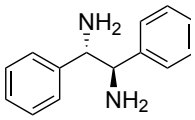
Lowest MPScore Amines



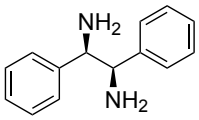
81



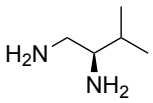
82



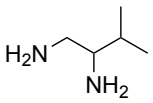
83



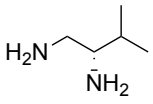
84



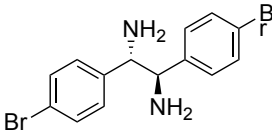
85



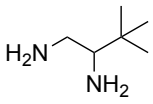
86



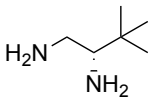
87



88

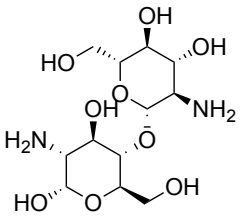


89

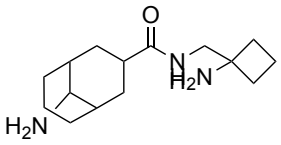


90

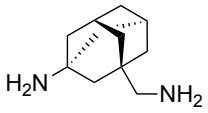
Highest MPScore Amines



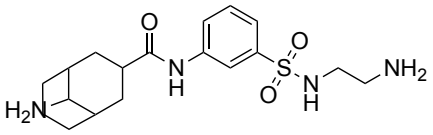
91



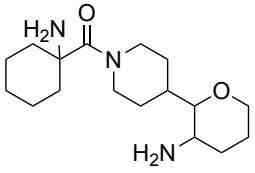
92



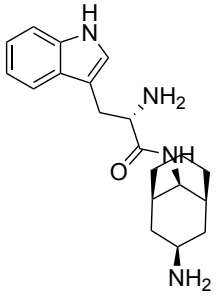
93



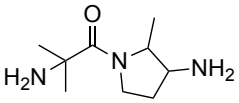
94



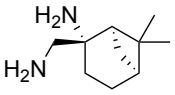
95



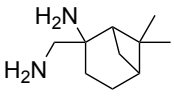
96



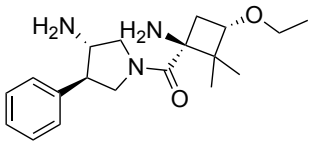
97



98

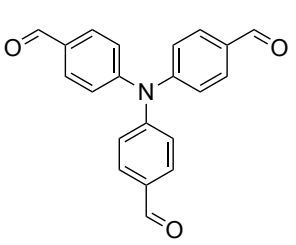


99

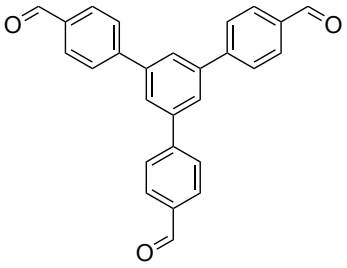


100

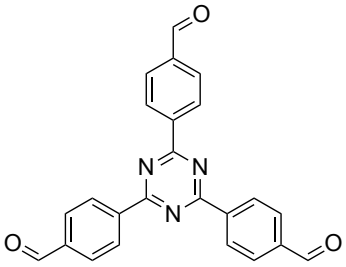
Lowest MPScore Aldehydes



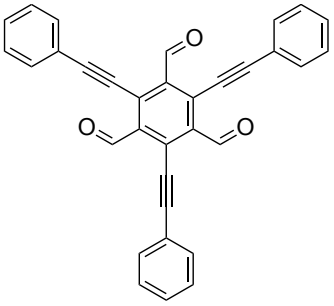
101



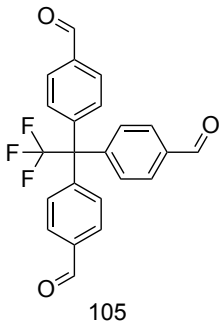
102



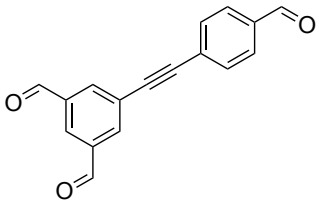
103



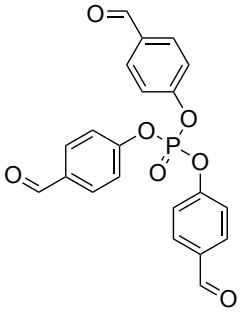
104



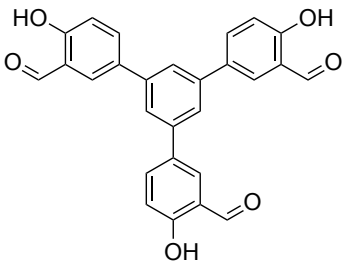
105



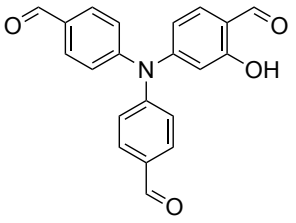
106



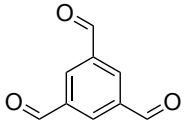
107



108

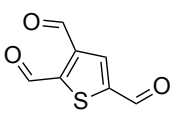


109

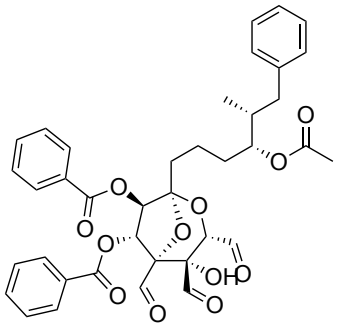


110

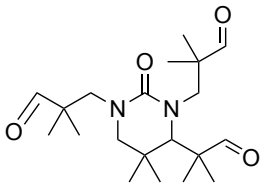
Highest MPScore Aldehydes



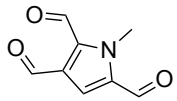
111



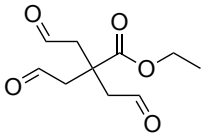
112



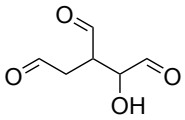
113



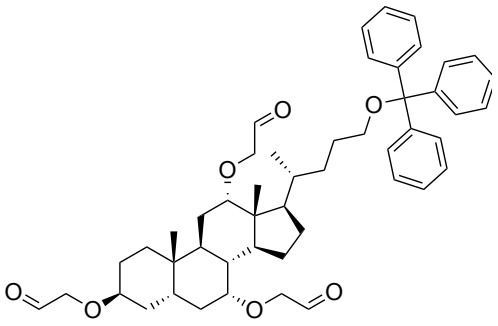
114



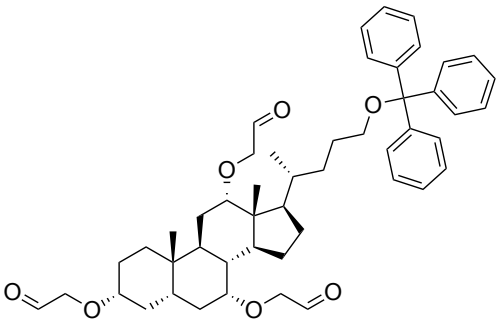
115



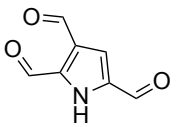
116



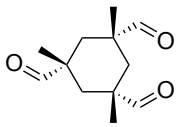
117



118



119



120

Table A.1: Synthetic difficulty scores of molecules provided to R.L.G for the blind classification task, calculated using the SAScore, SCScore and MPScore. The chemist answer column corresponds to whether R.L.G answered 'yes' or 'no' to the question: "Can you make 1 g of this compound in under 5 steps?" Molecules are categorised according to their functional group and whether they were scored the lowest or highest by their respective synthetic difficulty scoring method. The number column refers to the molecules in Figure A.1.

Number	Chemist Answer	SAScore	SCScore	MPScore
Lowest SCScore Amines				
1	Yes	0.53	0.064	0.10
2	Yes	0.75	0.08	0.05
3	No	0.74	0.08	0.23
4	Yes	0.35	0.08	0.35
5	Yes	0.49	0.09	0.13
6	Yes	0.57	0.09	0.11
7	Yes	0.78	0.09	0.72
8	Yes	0.57	0.09	0.42
9	Yes	0.72	0.09	0.60
10	Yes	0.68	0.09	0.44
Highest SCScore Amines				
11	No	0.97	1.0	0.94
12	No	0.97	1.0	0.92
13	No	0.97	1.0	0.93
14	No	0.97	1.0	0.90
15	No	0.97	1.0	0.88
16	No	0.97	1.0	0.92
17	No	0.97	1.0	0.91
18	No	0.97	1.0	0.91
19	No	0.97	1.0	0.88
20	No	0.96	1.0	0.90
Lowest SCScore Aldehydes				
21	No	0.56	0.09	0.89

Number	Chemist Answer	SAScore	SCScore	MPScore
22	No	0.51	0.11	0.82
23	Yes	0.40	0.11	0.53
24	Yes	0.56	0.11	0.16
25	Yes	0.54	0.11	0.30
26	Yes	0.38	0.11	0.80
27	Yes	0.62	0.12	0.27
28	Yes	0.68	0.13	0.18
29	No	0.62	0.14	0.81
30	No	0.57	0.14	0.38
		Highest SCScore Aldehydes		
31	No	0.67	0.57	0.41
32	No	0.83	0.57	0.49
33	No	0.80	0.57	0.92
34	No	0.94	0.58	0.94
35	No	0.91	0.59	0.91
36	No	0.94	0.60	0.94
37	No	0.88	0.60	0.63
38	Yes	0.83	0.66	0.20
39	Yes	0.93	0.70	0.43
40	No	0.76	0.73	0.90
		Lowest SAScore Amines		
41	Yes	0.08	0.40	0.52
42	Yes	0.08	0.24	0.79
43	No	0.08	0.53	0.73
44	No	0.18	0.48	0.95
45	No	0.20	0.58	0.94
46	No	0.20	0.37	0.57
47	No	0.20	0.58	0.62
48	Yes	0.22	0.27	0.04
49	No	0.24	0.45	0.49
50	No	0.24	0.83	0.93

Number	Chemist Answer	SAScore	SCScore	MPScore
Highest SAScore Amines				
51	No	1.0	0.93	0.93
52	No	1.0	0.76	0.94
53	No	1.0	0.68	0.94
54	No	1.0	0.92	0.91
55	No	1.0	1.00	0.90
56	No	1.0	0.82	0.94
57	No	1.0	0.80	0.94
58	No	1.0	0.97	0.94
59	No	1.0	1.00	0.93
60	No	1.0	0.75	0.94
Lowest SAScore Aldehydes				
61	Yes	0.20	0.33	0.54
62	Yes	0.21	0.28	0.57
63	Yes	0.38	0.12	0.80
64	Yes	0.39	0.31	0.57
65	Yes	0.40	0.11	0.53
66	No	0.40	0.43	0.46
67	Yes	0.41	0.15	0.50
68	No	0.42	0.41	0.58
69	No	0.46	0.23	0.80
70	Yes	0.47	0.23	0.55
Highest SAScore Aldehydes				
71	No	0.89	0.40	0.60
72	No	0.89	0.24	0.85
73	No	0.91	0.59	0.91
74	No	0.92	0.45	0.78
75	Yes	0.93	0.70	0.43
76	No	0.94	0.58	0.94

Number	Chemist Answer	SAScore	SCScore	MPScore
77	No	0.94	0.60	0.94
78	No	0.95	0.38	0.77
79	No	0.97	0.42	0.68
80	No	0.97	0.34	0.90
		Lowest MPScore Amines		
81	Yes	0.67	0.15	0.03
82	Yes	0.67	0.13	0.03
83	Yes	0.67	0.21	0.03
84	Yes	0.67	0.17	0.03
85	Yes	0.75	0.33	0.04
86	Yes	0.75	0.29	0.04
87	Yes	0.75	0.30	0.04
88	Yes	0.22	0.27	0.04
89	Yes	0.74	0.34	0.05
90	Yes	0.74	0.24	0.05
		Highest MPScore Amines		
91	No	0.86	0.50	0.97
92	No	0.86	0.62	0.97
93	No	0.88	0.57	0.97
94	No	0.85	0.86	0.97
95	No	0.83	0.56	0.97
96	No	0.88	0.73	0.97
97	No	0.78	0.57	0.97
98	No	0.86	0.45	0.97
99	No	0.86	0.38	0.97
100	No	0.85	0.66	0.97
		Lowest MPScore Aldehydes		
101	Yes	0.66	0.25	0.03
102	Yes	0.66	0.40	0.04

Number	Chemist Answer	SAScore	SCScore	MPScore
103	Yes	0.61	0.44	0.08
104	Yes	0.65	0.21	0.09
105	Yes	0.65	0.23	0.10
106	Yes	0.60	0.35	0.14
107	Yes	0.65	0.22	0.14
108	Yes	0.68	0.38	0.15
109	No	0.68	0.43	0.15
110	Yes	0.56	0.11	0.16
		Highest MPScore Aldehydes		
111	No	0.50	0.36	0.91
112	No	0.91	0.59	0.91
113	No	0.80	0.57	0.92
114	No	0.60	0.42	0.92
115	No	0.69	0.18	0.92
116	No	0.74	0.24	0.94
117	No	0.94	0.60	0.94
118	No	0.94	0.58	0.94
119	No	0.56	0.36	0.95
120	No	0.74	0.14	0.95

A.2 Chapter 5 Supplementary Data

A.2.1 Stacked NMR Spectra

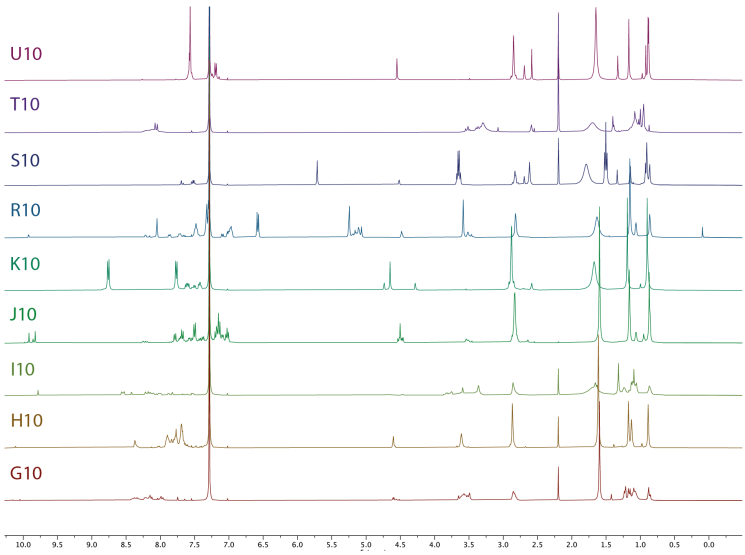


Figure A.2: Stacked ^{1}H NMR spectra ($CDCl_3$) of the crude reactions of diamine **10** with trialdehdyes **G-U**, taken directly from the high-throughput screen.

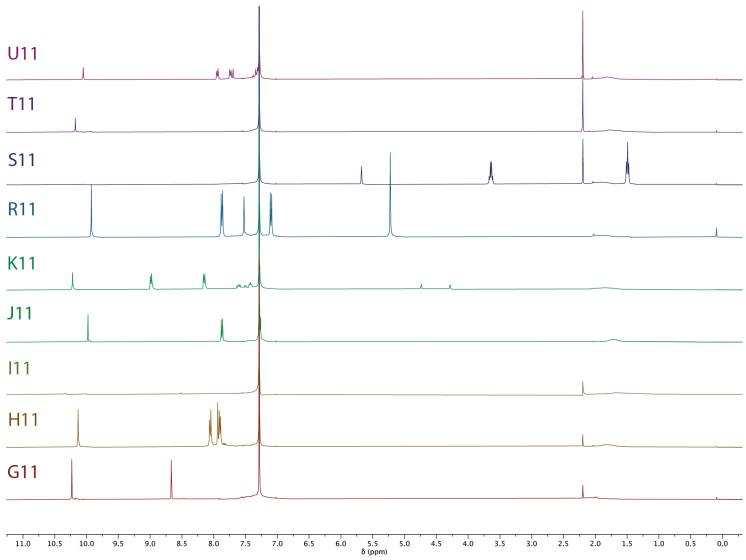


Figure A.3: Stacked ^{1}H NMR spectra ($CDCl_3$) of the crude reactions of diamine **11** with trialdehdyes **G-U**, taken directly from the high-throughput screen.

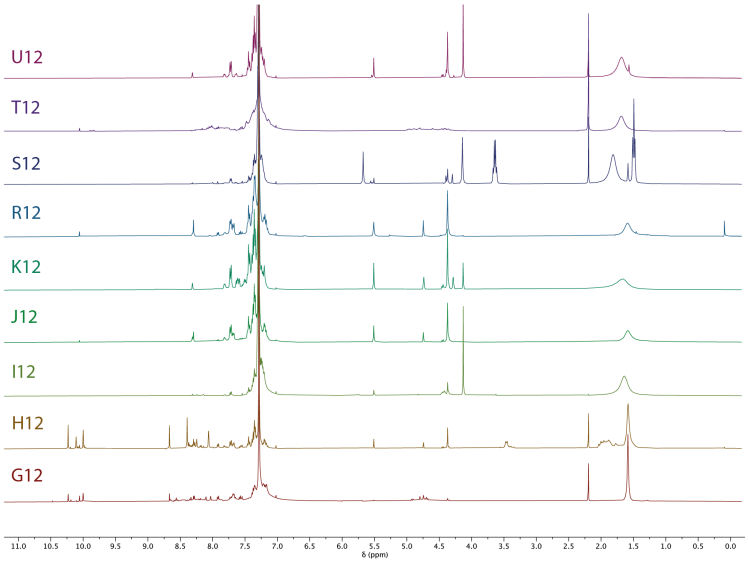


Figure A.4: Stacked ^{1}H NMR spectra ($CDCl_3$) of the crude reactions of diamine **12** with trialdehdyes **G**-**U**, taken directly from the high-throughput screen.

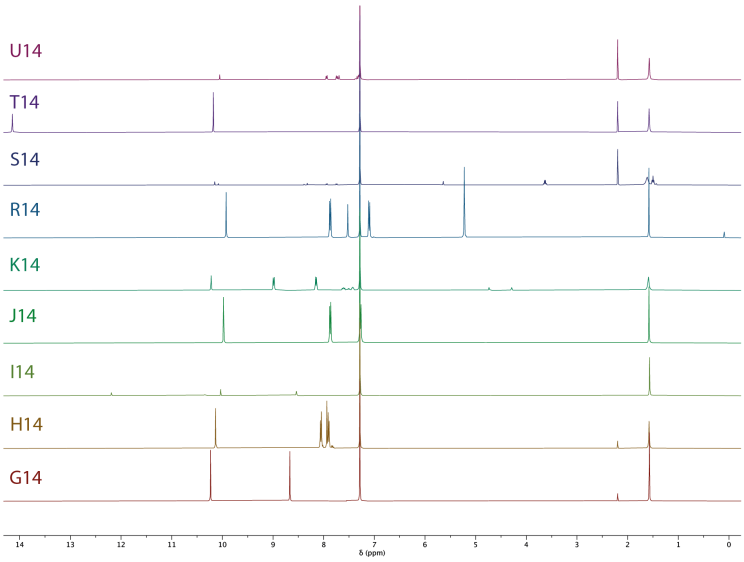


Figure A.5: Stacked ^{1}H NMR spectra ($CDCl_3$) of the crude reactions of diamine **14** with trialdehdyes **G**-**U**, taken directly from the high-throughput screen.

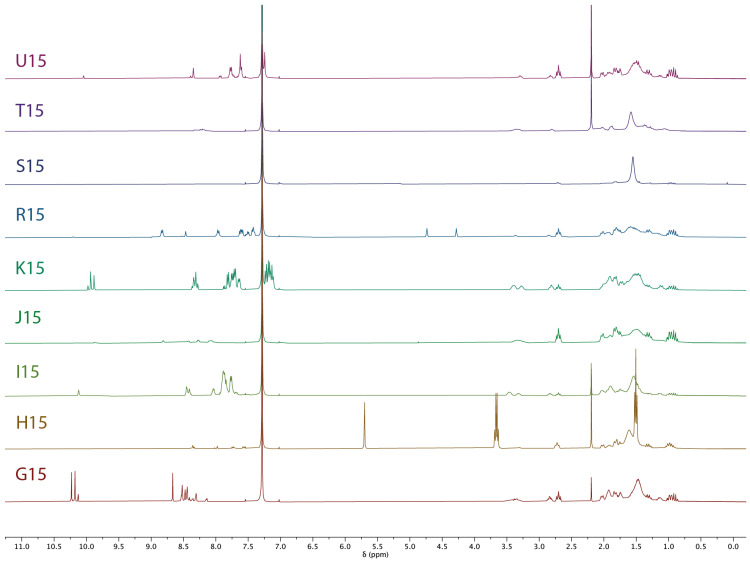


Figure A.6: Stacked ^{1}H NMR spectra ($CDCl_3$) of the crude reactions of diamine **15** with trialdehdyes **G-U**, taken directly from the high-throughput screen.

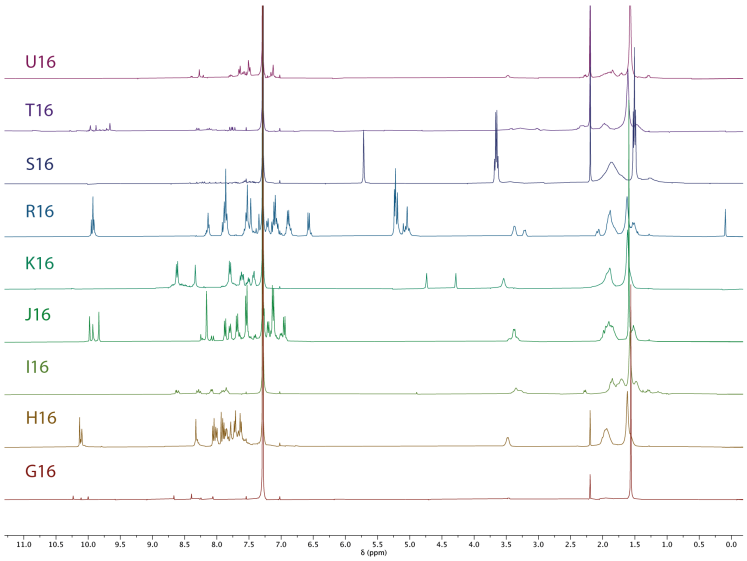


Figure A.7: Stacked ^{1}H NMR spectra ($CDCl_3$) of the crude reactions of diamine **16** with trialdehdyes **G-U**, taken directly from the high-throughput screen.

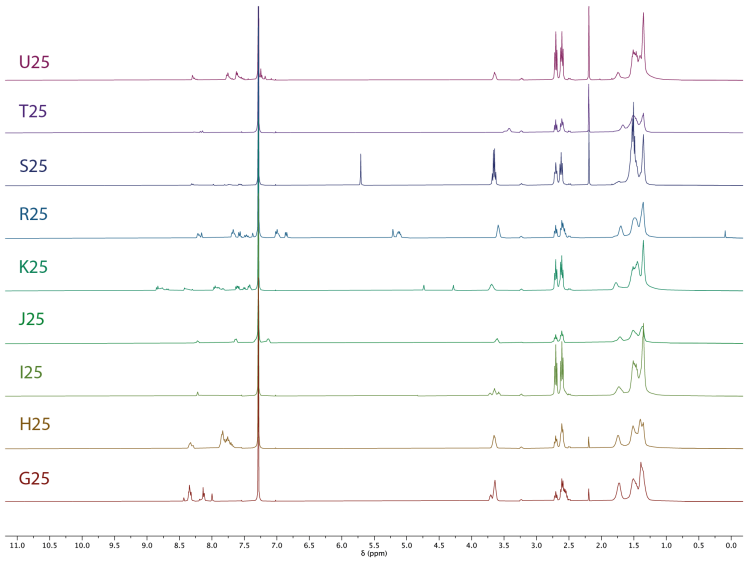


Figure A.8: Stacked ^{1}H NMR spectra ($CDCl_3$) of the crude reactions of diamine **25** with trialdehdyes **G**-**U**, taken directly from the high-throughput screen.

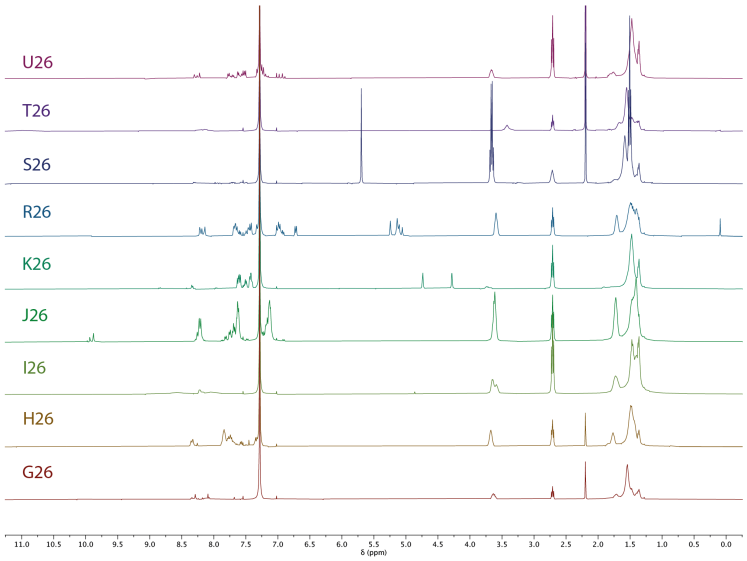


Figure A.9: Stacked ^{1}H NMR spectra ($CDCl_3$) of the crude reactions of diamine **26** with trialdehdyes **G**-**U**, taken directly from the high-throughput screen.

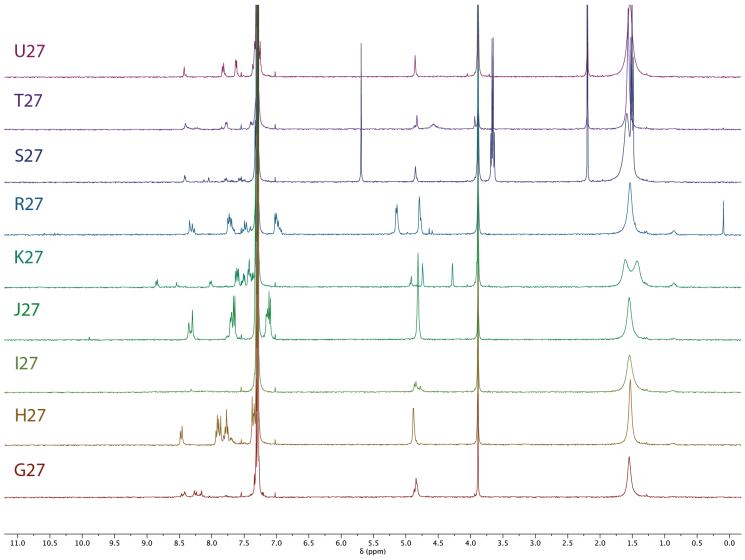


Figure A.10: Stacked ^{1}H NMR spectra ($CDCl_3$) of the crude reactions of diamine **27** with trialdehdyes **G**-**U**, taken directly from the high-throughput screen.

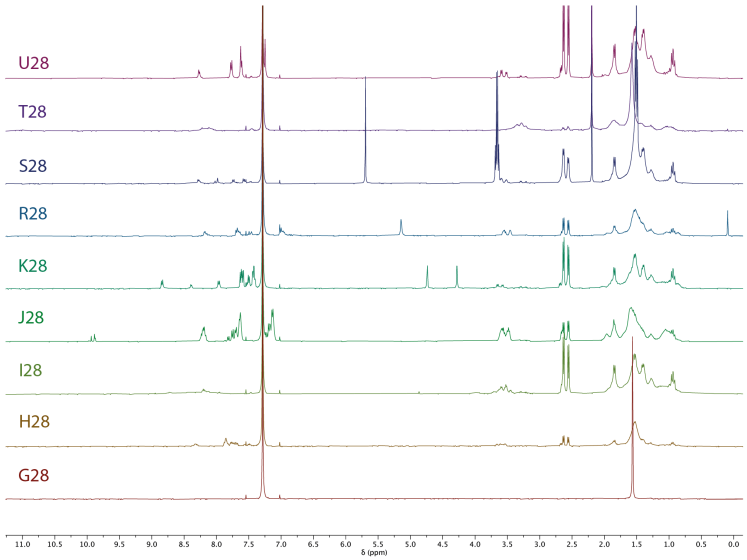


Figure A.11: Stacked ^{1}H NMR spectra ($CDCl_3$) of the crude reactions of diamine **28** with trialdehdyes **G**-**U**, taken directly from the high-throughput screen.

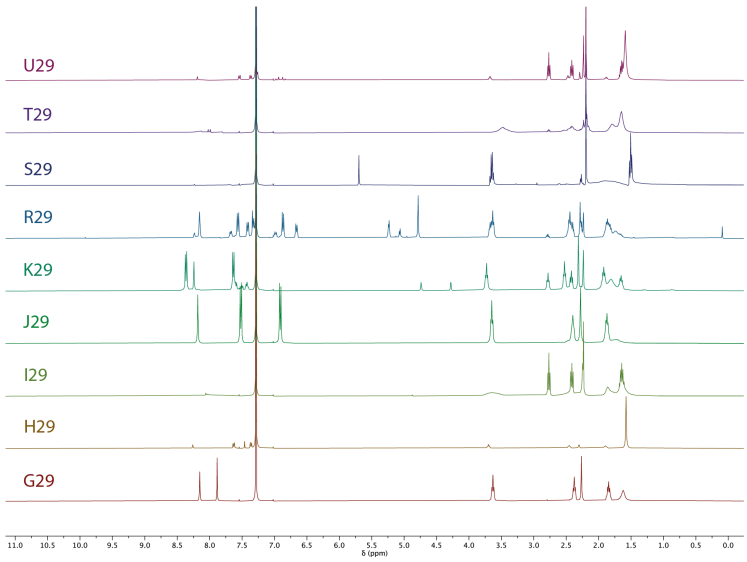


Figure A.12: Stacked ^{1}H NMR spectra ($CDCl_3$) of the crude reactions of diamine **29** with trialdehdyes **G-U**, taken directly from the high-throughput screen.

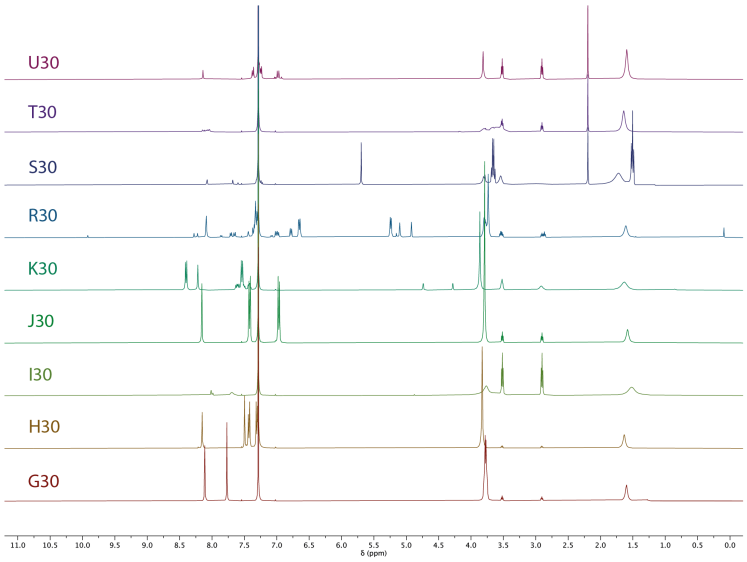


Figure A.13: Stacked ^{1}H NMR spectra ($CDCl_3$) of the crude reactions of diamine **30** with trialdehdyes **G-U**, taken directly from the high-throughput screen.

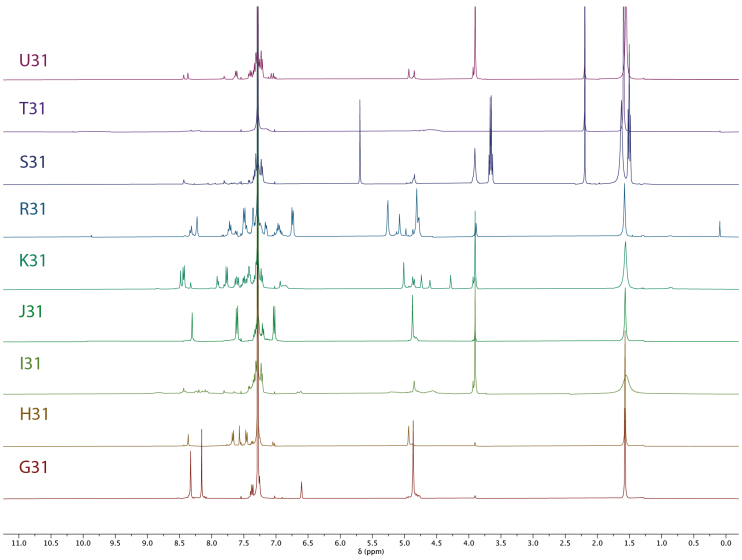


Figure A.14: Stacked ^{1}H NMR spectra ($CDCl_3$) of the crude reactions of diamine **31** with trialdehdyes **G-U**, taken directly from the high-throughput screen.

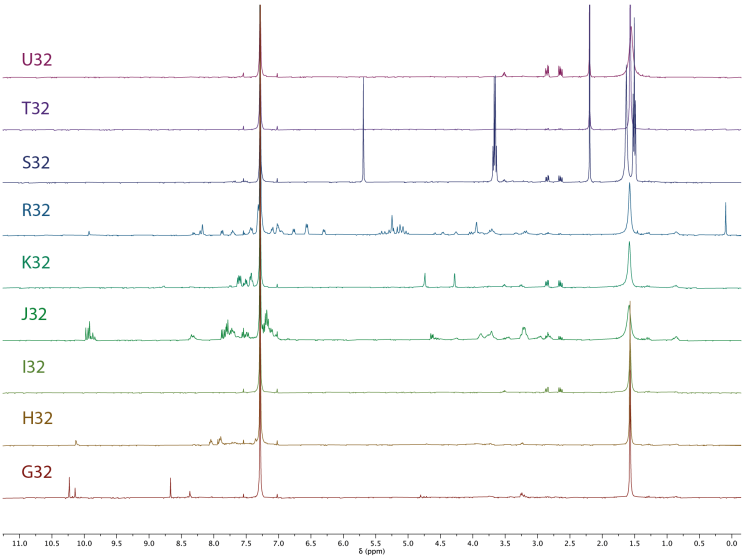


Figure A.15: Stacked ^{1}H NMR spectra ($CDCl_3$) of the crude reactions of diamine **32** with trialdehdyes **G-U**, taken directly from the high-throughput screen.

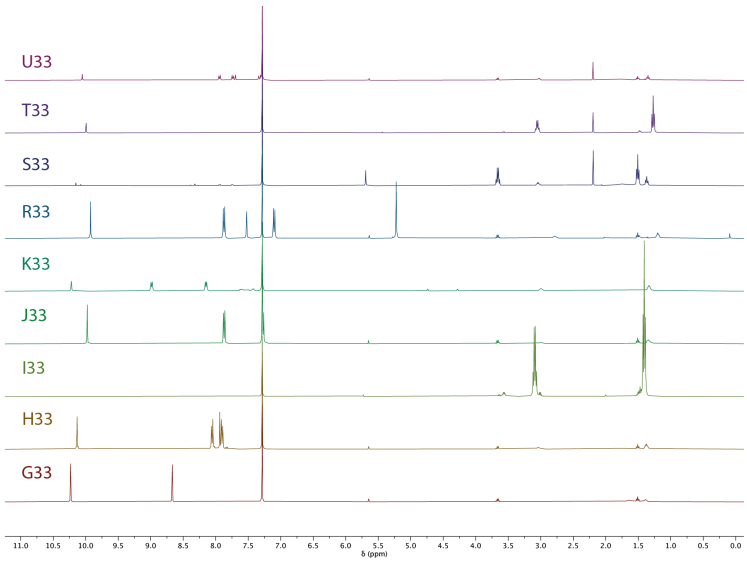


Figure A.16: Stacked ^{1}H NMR spectra ($CDCl_3$) of the crude reactions of diamine **33** with trialdehdyes **G-U**, taken directly from the high-throughput screen.

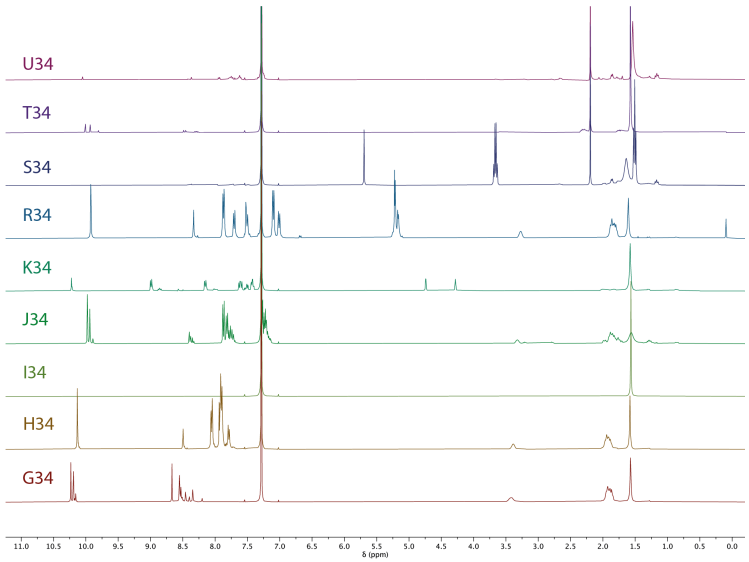


Figure A.17: Stacked ^{1}H NMR spectra ($CDCl_3$) of the crude reactions of diamine **34** with trialdehdyes **G-U**, taken directly from the high-throughput screen.

A.2.2 High-Throughput HRMS Data

Category 2

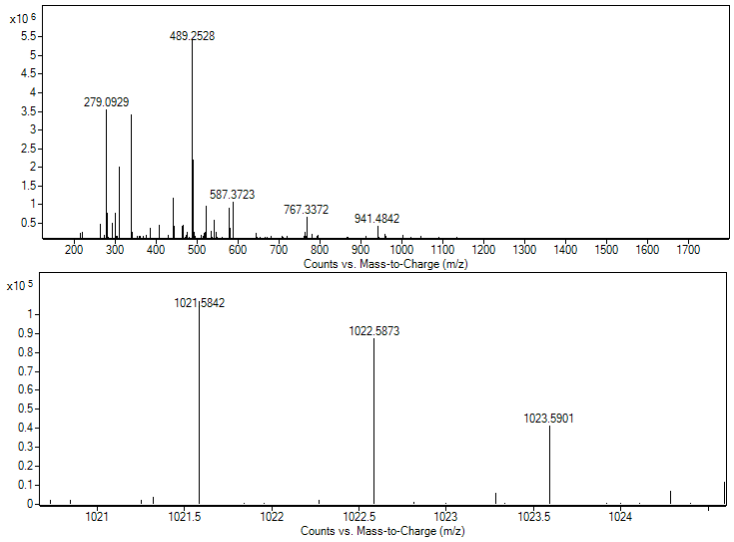


Figure A.18: HRMS spectra for crude reactions taken from the high-throughput screen for cage **H26**$_{[2+3]}$, indicating the formation of a [2 + 3] cage — calc. for $C_{72}H_{72}N_6$ 1020.5818, found $[M+H]^+$ 1021.584 and $[M+2H]^{2+}$ 511.2431.

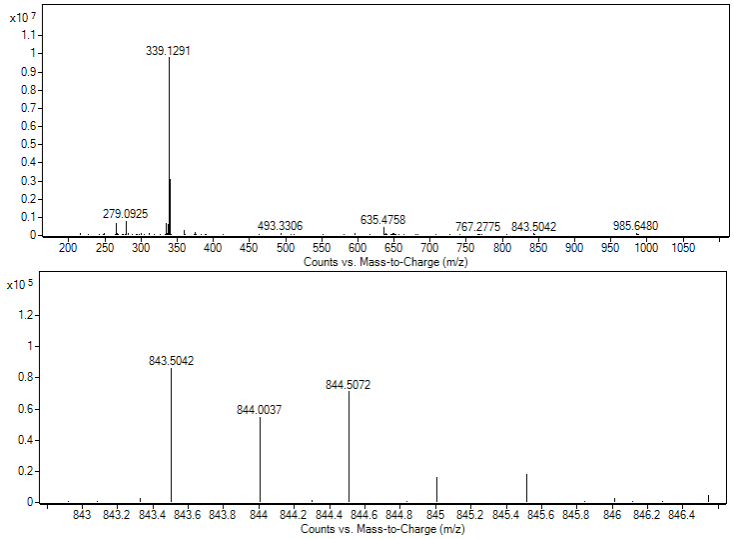


Figure A.19: HRMS spectra for crude reactions taken from the high-throughput screen for cage **S28**$_{[4+6]}$, indicating the formation of a [4 + 6] cage — calc. for $C_{116}H_{124}N_{12}$ 1685.0072, found $[M+2H]^{2+}$ 843.5042.

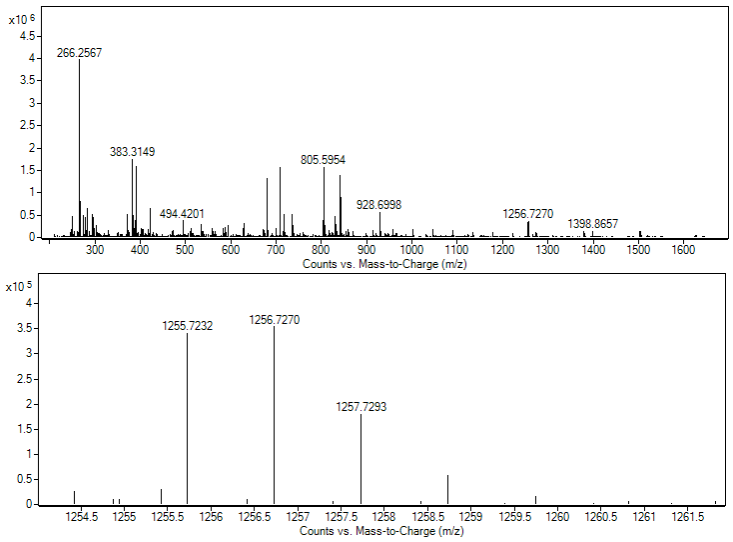


Figure A.20: HRMS spectra for crude reactions taken from the high-throughput screen for cage **U28**$_{[2+3]}$, indicating the formation of a [2 + 3] cage — calc. for $C_{90}H_{90}N_6$ 1254.7227, found $[M+H]^+$ 1255.7232 and $[M+2H]^{2+}$ 628.3694.

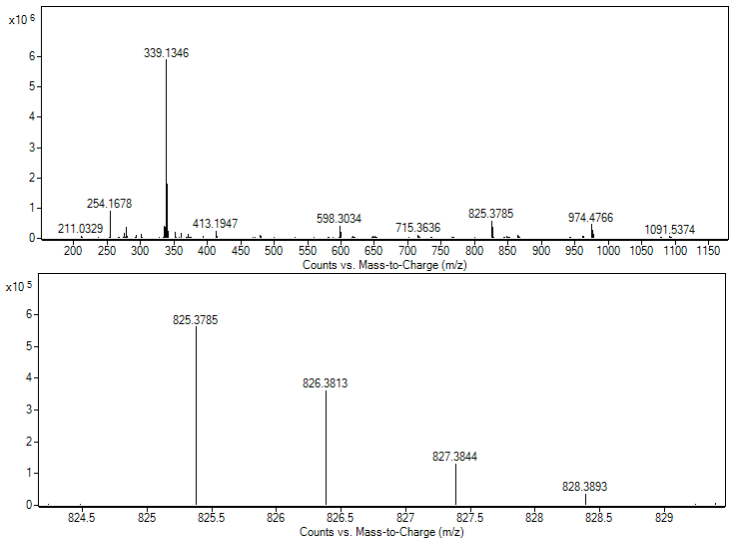


Figure A.21: HRMS spectra for crude reactions taken from the high-throughput screen for cage **S31**$_{[2+3]}$, indicating the formation of a [2 + 3] cage — calc. for $C_{58}H_{44}N_6$ 824.3627, found $[M+H]^+$ 825.3785 and $[M+2H]^{2+}$ 413.1920.

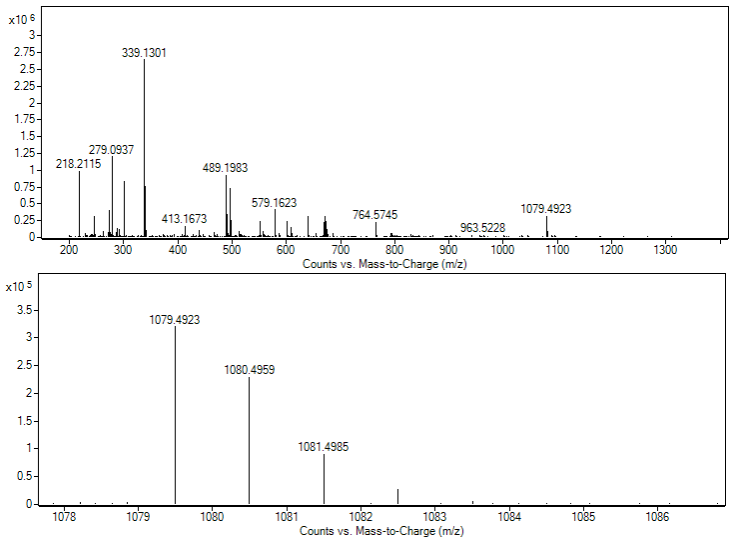


Figure A.22: HRMS spectra for crude reactions taken from the high-throughput screen for cage **H31**$_{[2+3]}$, indicating the formation of a [2 + 3] cage — calc. for $C_{78}H_{60}N_6$ 1080.4879, found $[M+H]^+$ 1081.4985.

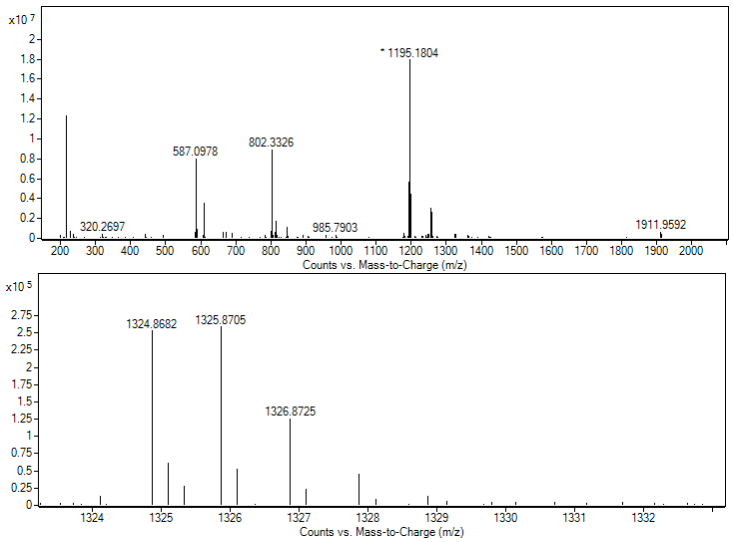


Figure A.23: HRMS spectra for crude reactions taken from the high-throughput screen for cage **K25**$_{[2+3]}$, indicating the formation of a [2 + 3] cage — calc. for $C_{84}H_{105}N_{15}$ 1323.8677, found $[M+H]^+$ 1324.8682 and $[M+2H]^{2+}$ 662.9391.

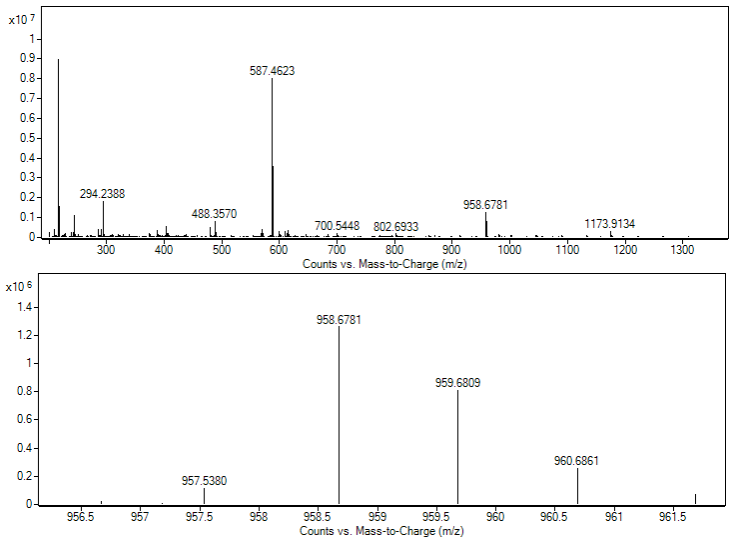


Figure A.24: HRMS spectra for crude reactions taken from the high-throughput screen for cage **T25**$_{[2+3]}$, indicating the formation of a [2 + 3] cage — calc. for $C_{54}H_{87}N_9O_6$ 957.6779, found $[M+H]^+$ 958.6781 and $[M+2H]^{2+}$ 479.8451.

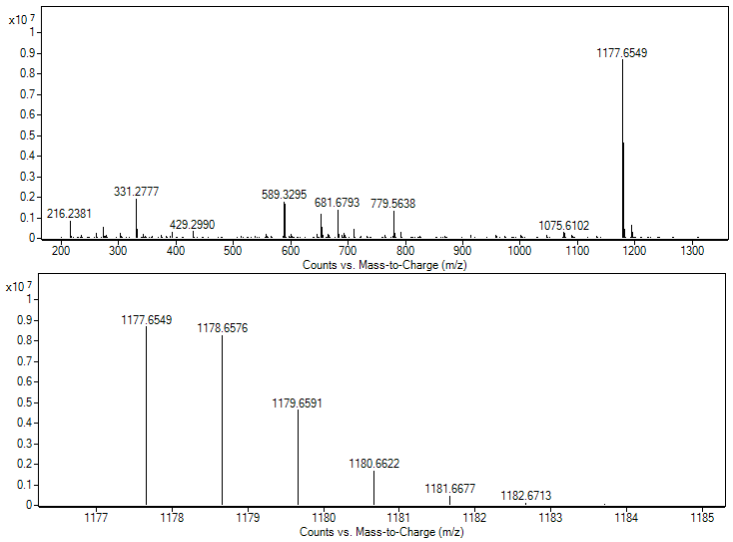


Figure A.25: HRMS spectra for crude reactions taken from the high-throughput screen for cage **U26**$_{[2+3]}$, indicating the formation of a [2 + 3] cage — calc. for $C_{84}H_{84}N_6$ 1176.6757, found $[M+H]^+$ 1177.6549 and $[M+2H]^{2+}$ 597.3598.

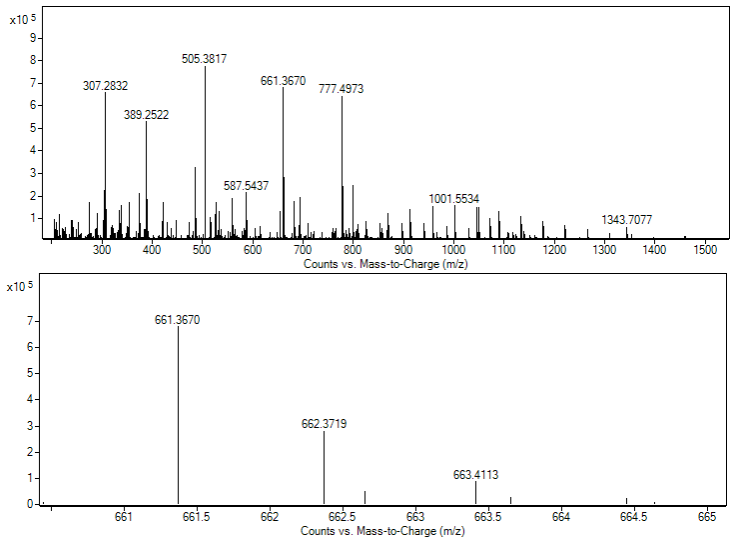


Figure A.26: HRMS spectra for crude reactions taken from the high-throughput screen for cage **T26**$_{[2+3]}$, indicating the formation of a [2 + 3] cage — calc. for $C_{36}H_{48}N_6O_6$ 660.3635, found $[M+H]^+$ 661.3672.

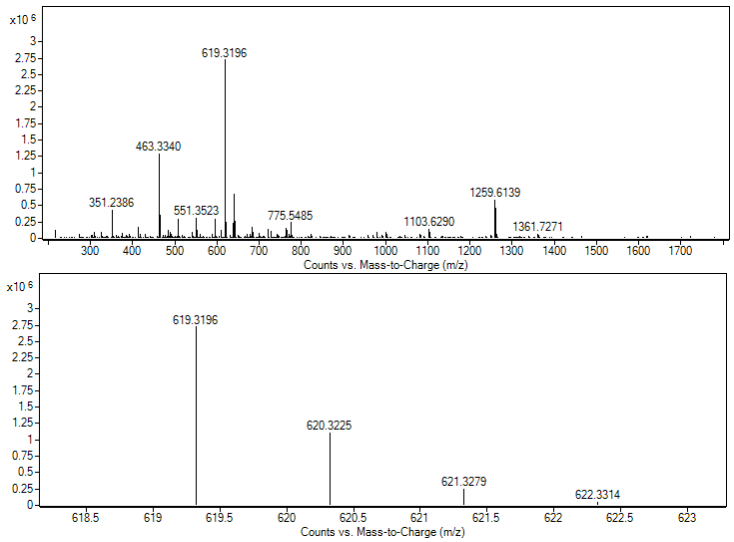


Figure A.27: HRMS spectra for crude reactions taken from the high-throughput screen for cage **T10**$_{[2+3]}$, indicating the formation of a [2 + 3] cage — calc. for $C_{33}H_{42}N_6O_6$ 618.31659, found $[M+H]^+$ 619.3196.

Category 3

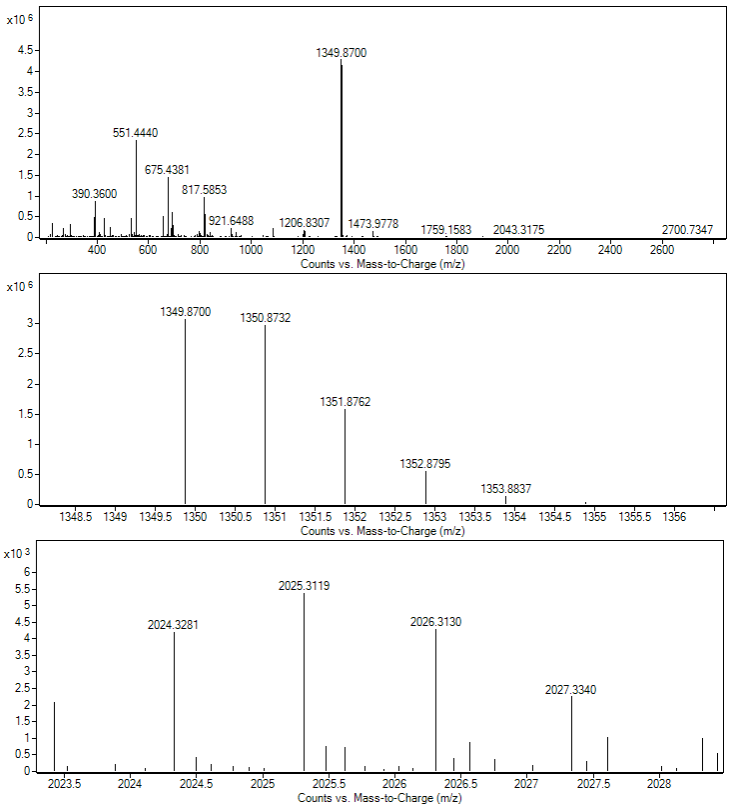


Figure A.28: HRMS spectra for crude reactions taken from the high-throughput screen for cage **I28**$_{[4+6],[6+9]}$, indicating the formation of [4 + 6] and [6 + 9] cages — calc. for $C_{84}H_{108}N_{12}O_4$ 1348.8617 and $C_{126}H_{162}N_{18}O_6$ 2023.2925, found $[M+H]_{[4+6]}^{+}$ 1349.8700, $[M+2H]_{[4+6]}^{2+}$ 675.4381, $[M+H]_{[6+9]}^{+}$ 2024.3281 and $[M+2H]_{[6+9]}^{2+}$ 1012.6731.

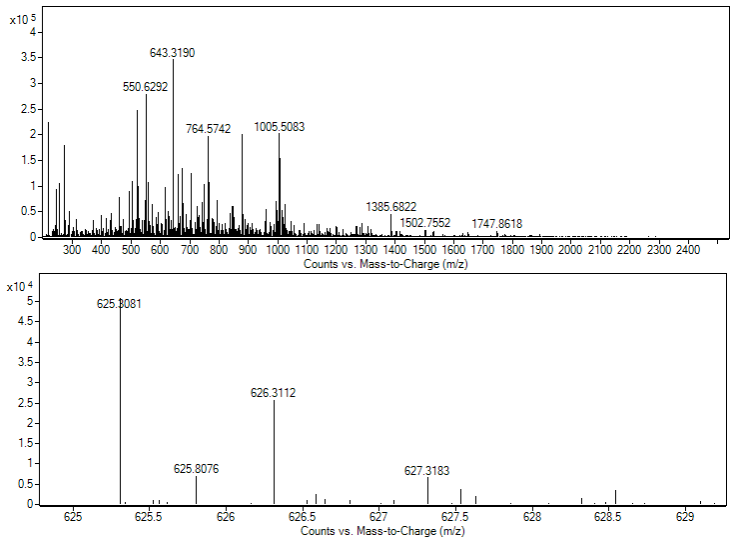


Figure A.29: HRMS spectra for crude reactions taken from the high-throughput screen for cage **G27**$_{[2+3],\,[4+6]}$, indicating the formation of [2 + 3] and [4 + 6] cages — calc. for $C_{42}H_{36}N_6$ 624.3001 and $C_{84}H_{72}N_{12}$ 1248.6003, found $[M+H]_{[2+3]}^{+}$ 625.3081, $[M+2H]_{[2+3]}^{2+}$ 313.1550 and $[M+2H]_{[4+6]}^{2+}$ 625.3081.

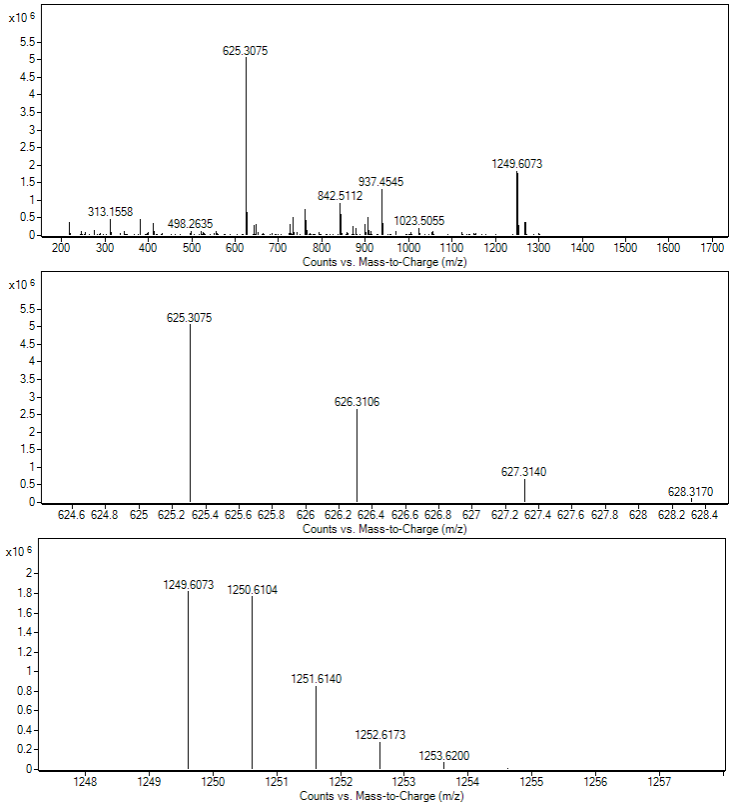


Figure A.30: HRMS spectra for crude reactions taken from the high-throughput screen for cage **G31**$_{[2+3],\,[4+6]}$, indicating the formation of [2 + 3] and [4 + 6] cages — calc. for $C_{42}H_{36}N_6$ 624.3001 and $C_{84}H_{72}N_{12}$ 1248.6003, found $[M+H]_{[2+3]}^{+}$ 625.3075, $[M+2H]_{[2+3]}^{2+}$ 313.1558 and $[M+H]_{[4+6]}^{+}$ 1249.6073.

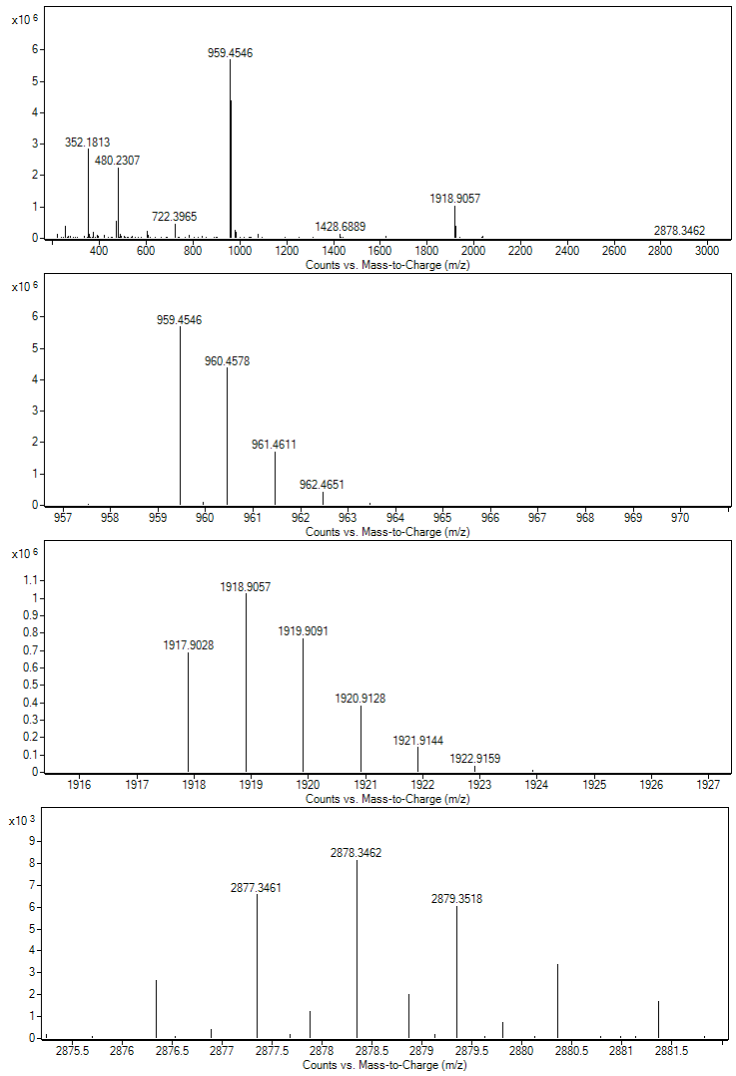


Figure A.31: HRMS spectra for crude reactions taken from the high-throughput screen for cage **J31**$_{[2+3],[4+6],[6+9]}$, indicating the formation of [2 + 3], [4 + 6], and [6 + 9] cages — calc. for $C_{66}H_{54}N_8$ 958.4471, $C_{132}H_{108}N_{16}$ 1916.8943 and $C_{198}H_{162}N_{24}$ 2875.3414, found $[M+H]_{[2+3]}^{+}$ 959.4546, $[M+2H]_{[2+3]}^{2+}$ 480.3207, $[M+H]_{[4+6]}^{+}$ 1917.9028 and $[M+H]_{[6+9]}^{+}$ 2877.3461.

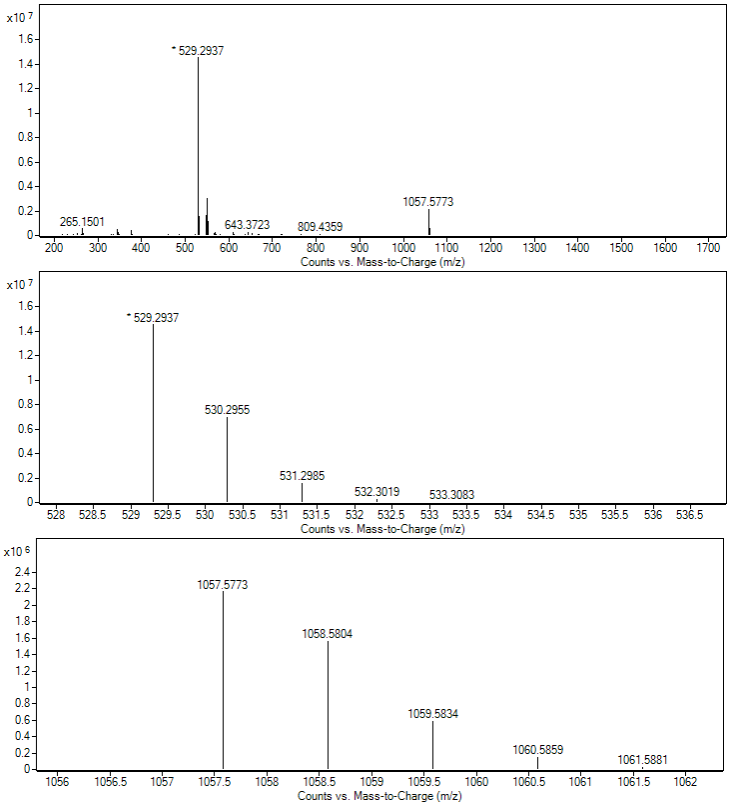


Figure A.32: HRMS spectra for crude reactions taken from the high-throughput screen for cage **G30**$_{[2+3],[4+6]}$, indicating the formation of [2 + 3] and [4 + 6] cages — calc. for $C_{30}H_{36}N_6O_3$ 528.2849 and $C_{132}H_{108}N_{16}$ 1056.5698, found $[M+H]_{[2+3]}^{+}$ 529.2937, $[M+2H]_{[2+3]}^{2+}$ 265.1501 and $[M+H]_{[4+6]}^{+}$ 1057.5773.

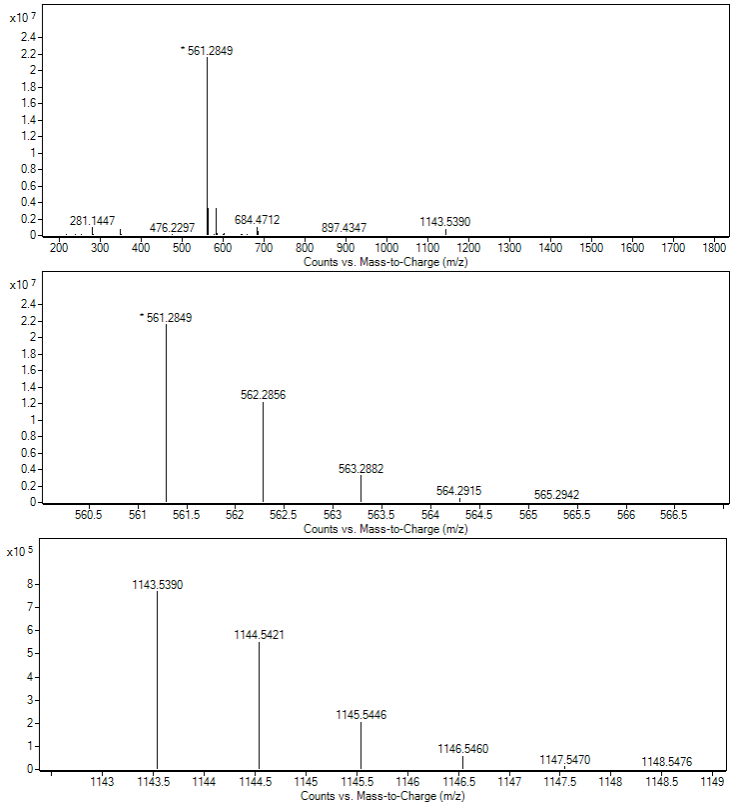


Figure A.33: HRMS spectra for crude reactions taken from the high-throughput screen for cage **I30**$_{[2+3],[4+6]}$, indicating the formation of [2 + 3] and [4 + 6] cages — calc. for $C_{30}H_{36}N_6O_5$ 560.2747 and $C_{60}H_{72}N_{12}O_{10}$ 1120.5494, found $[M+H]_{[2+3]}^+$ 561.2849 and $[M+Na]_{[4+6]}^+$ 1143.5390.

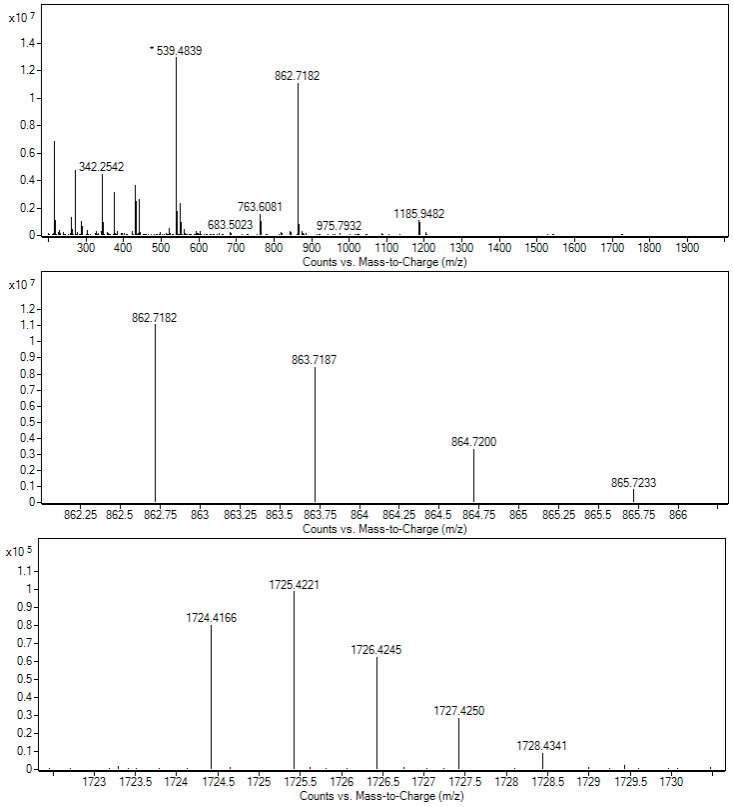


Figure A.34: HRMS spectra for crude reactions taken from the high-throughput screen for cage **G25**$_{[2+3],[4+6]}$, indicating the formation of [2 + 3] and [4 + 6] cages — calc. for $C_{54}H_{87}N_9$ 861.7084 and $C_{108}H_{174}N_{18}$ 1723.4169, found $[M+H]_{[2+3]}^{+}$ 862.7182, $[M+2H]_{[2+3]}^{2+}$ 431.8610 and $[M+H]_{[4+6]}^{+}$ 1724.4166.

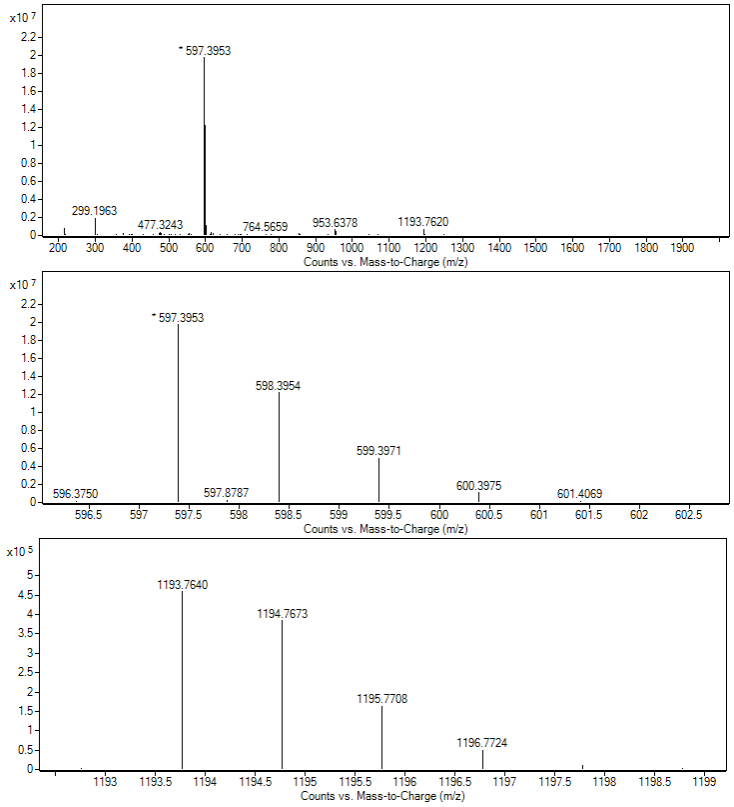


Figure A.35: HRMS spectra for crude reactions taken from the high-throughput screen for cage **I26**$_{[2+3],[4+6]}$, indicating the formation of [2 + 3] and [4 + 6] cages — calc. for $C_{36}H_{48}N_6O_2$ 596.3839 and $C_{72}H_{96}N_{12}O_4$ 1192.7678, found $[M+H]_{[2+3]}^{+}$ 597.3953, $[M+2H]_{[2+3]}^{2+}$ 299.1964 and $[M+H]_{[4+6]}^{+}$ 1193.7640.

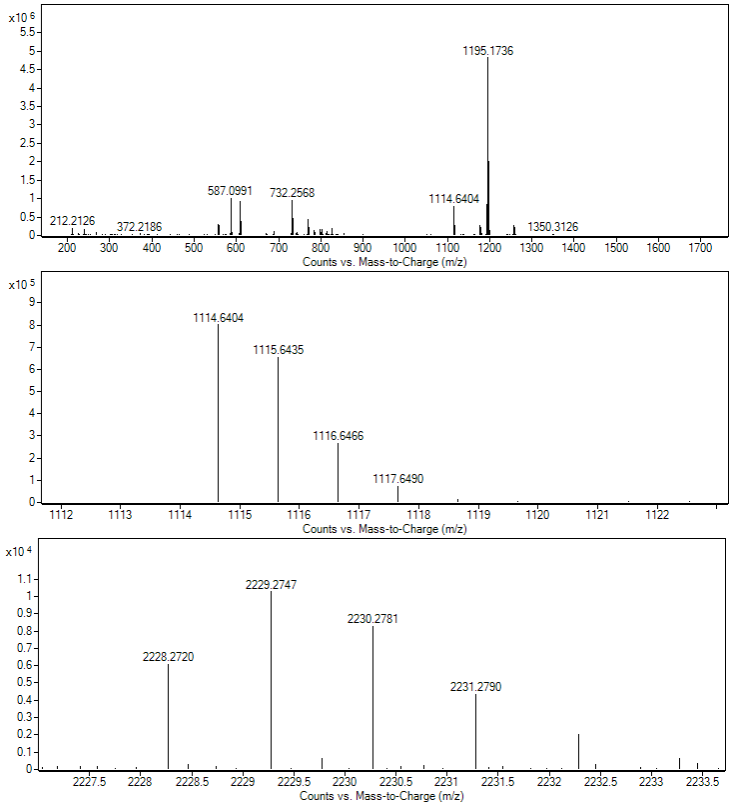


Figure A.36: HRMS spectra for crude reactions taken from the high-throughput screen for cage **K29**$_{[2+3],\,[4+6]}$, indicating the formation of [2 + 3] and [4 + 6] cages — calc. for $C_{69}H_{75}N_{15}$ 1113.6330 and $C_{138}H_{150}N_{30}$ 2227.2660, found $[M+H]_{[2+3]}^{+}$ 1114.6404, $[M+2H]_{[2+3]}^{2+}$ 557.8236 and $[M+H]_{[4+6]}^{+}$ 2228.2720.

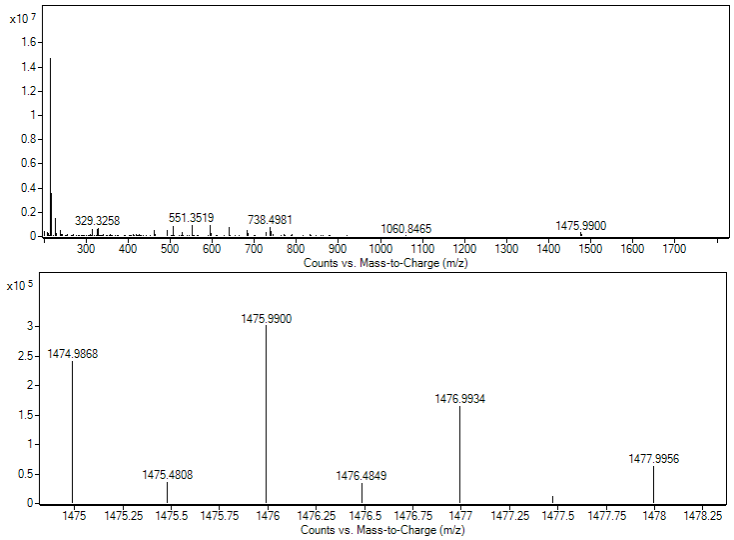


Figure A.37: HRMS spectra for crude reactions taken from the high-throughput screen for cage **U25**$_{[2+3],[4+6]}$, indicating the formation of [2 + 3] and [4 + 6] cages — calc. for $C_{102}H_{123}N_9$ 1473.9901 and $C_{204}H_{246}N_{18}$ 2947.9803, found $[M+2H]_{[2+3]}^{2+}$ 737.9961 and $[M+2H]_{[4+6]}^{2+}$ 1474.9868.

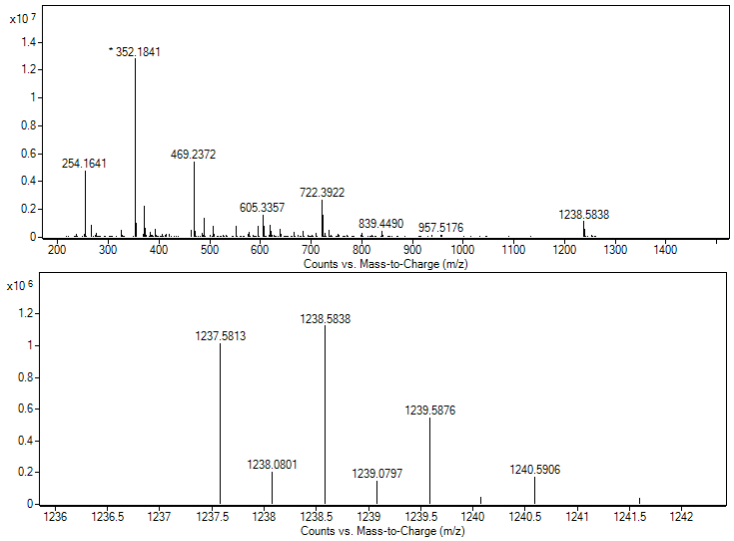


Figure A.38: HRMS spectra for crude reactions taken from the high-throughput screen for cage **U31**$_{[2+3],[4+6]}$, indicating the formation of [2 + 3] and [4 + 6] cages — calc. for $C_{90}H_{72}N_6$ 1237.58967 and $C_{180}H_{144}N_{12}$ 2474.1715, found $[M+2H]_{[2+3]}^{2+}$ 619.2955 and $[M+2H]_{[4+6]}^{2+}$ 1237.5813.

Category 4

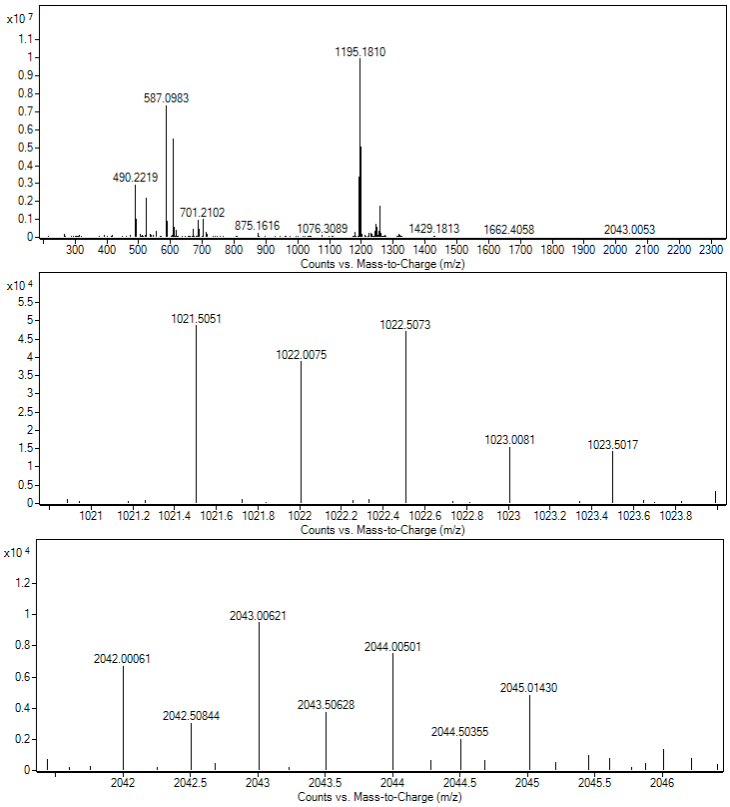


Figure A.39: HRMS spectra for crude reactions taken from the high-throughput screen for cage **K16$_{[4+6],[8+12]}$**, indicating the formation of [4 + 6] and [8 + 12] cages — calc. for $C_{132}H_{120}N_{24}$ 2041.0128 and $C_{264}H_{240}N_{48}$ 4082.0256, found $[M+H]_{[4+6]}^{+}$ 2042.0061, $[M+2H]_{[4+6]}^{2+}$ 1021.5051 and $[M+H]_{[8+12]}^{2+}$ 2042.0061.

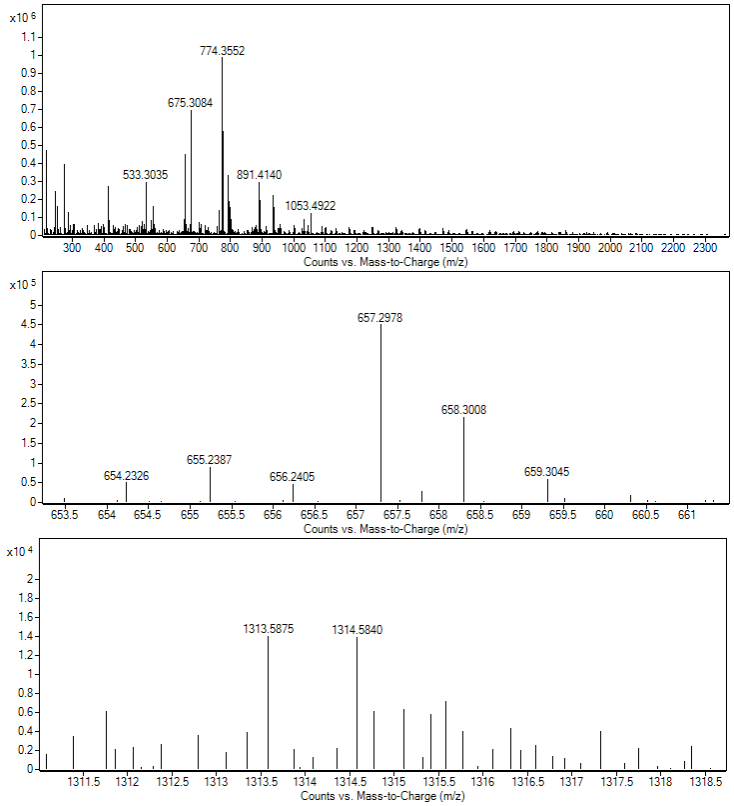


Figure A.40: HRMS spectra for crude reactions taken from the high-throughput screen for cage **I27**$_{[2+3],\,[4+6]}$, indicating the formation of [2 + 3] and [4 + 6] cages — calc. for $C_{42}H_{36}N_6O_2$ 656.2900 and $C_{84}H_{72}N_{12}O_4$ 1312.5800, found $[M+H]_{[2+3]}^{+}$ 657.2978 and $[M+H]_{[4+6]}^{+}$ 1313.5875.

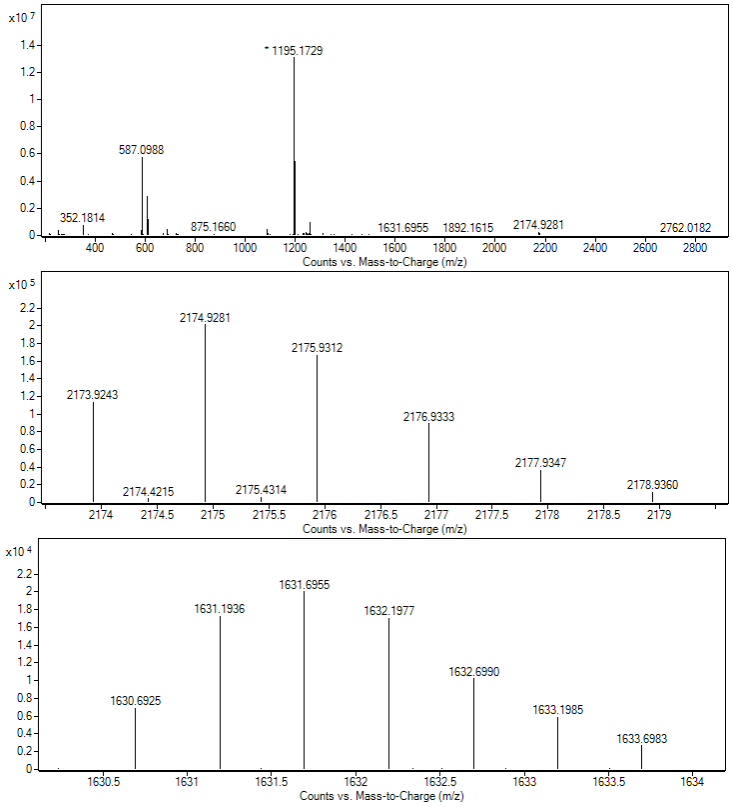


Figure A.41: HRMS spectra for crude reactions taken from the high-throughput screen for cage **K31**$_{[4+6],\,[6+9],\,[8+12]}$, indicating the formation of [4 + 6], [6 + 9], and [8 + 12] cages — calc. for $C_{144}H_{108}N_{24}$ 2172.9189, $C_{216}H_{162}N_{36}$ 3259.3783, and $C_{288}H_{216}N_{48}$ 4345.8378, found $[M+H]_{[4+6]}^{+}$ 2173.9243, $[M+2H]_{[6+9]}^{2+}$ 1630.6925 and $[M+2H]_{[8+12]}^{2+}$ 2173.9243.

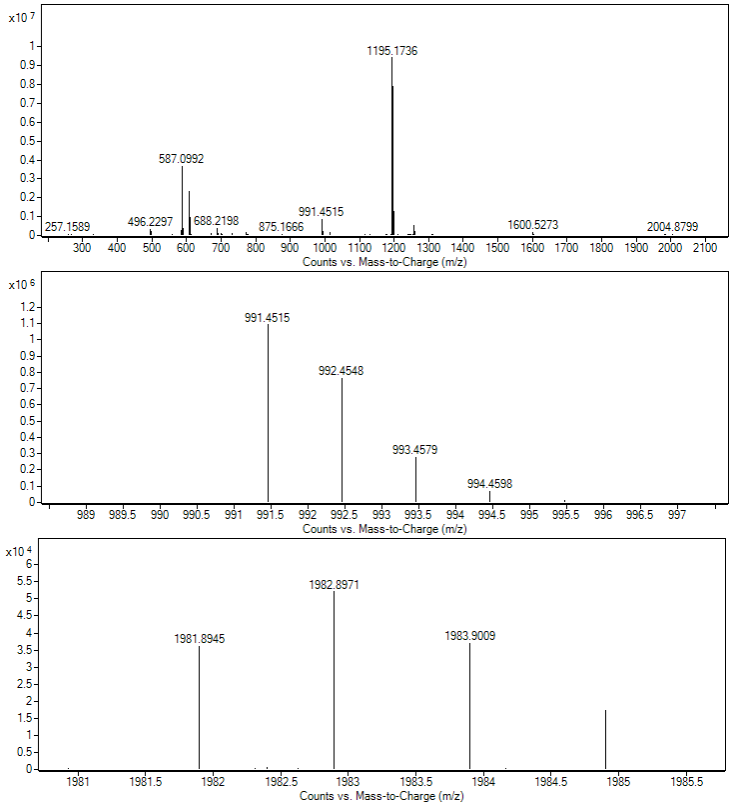


Figure A.42: HRMS spectra for crude reactions taken from the high-throughput screen for cage **K30**$_{[2+3],[4+6]}$, indicating the formation of [2 + 3] and [4 + 6] cages — calc. for $C_{60}H_{54}N_{12}O_3$ 990.4442 and $C_{120}H_{108}N_{24}O_6$ 1980.8884, found $[M+H]_{[2+3]}^{+}$ 991.4515, $[M+2H]_{[2+3]}^{2+}$ 496.2297 and $[M+H]_{[4+6]}^{+}$ 1981.8945.

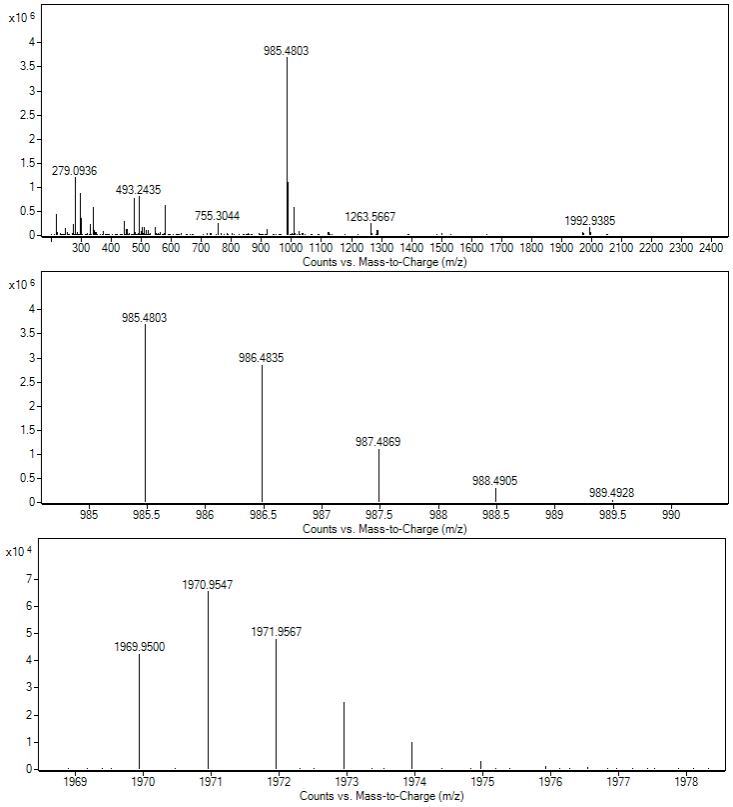


Figure A.43: HRMS spectra for crude reactions taken from the high-throughput screen for cage **H30**$_{[2+3],[4+6]}$, indicating the formation of [2 + 3] and [4 + 6] cages — calc. for $C_{66}H_{60}N_6O_3$ 984.4727 and $C_{132}H_{120}N_{12}O_6$ 1968.9454, found $[M+H]_{[2+3]}^{+}$ 985.4803, $[M+2H]_{[2+3]}^{2+}$ 493.2435 and $[M+H]_{[4+6]}^{+}$ 1969.9500.

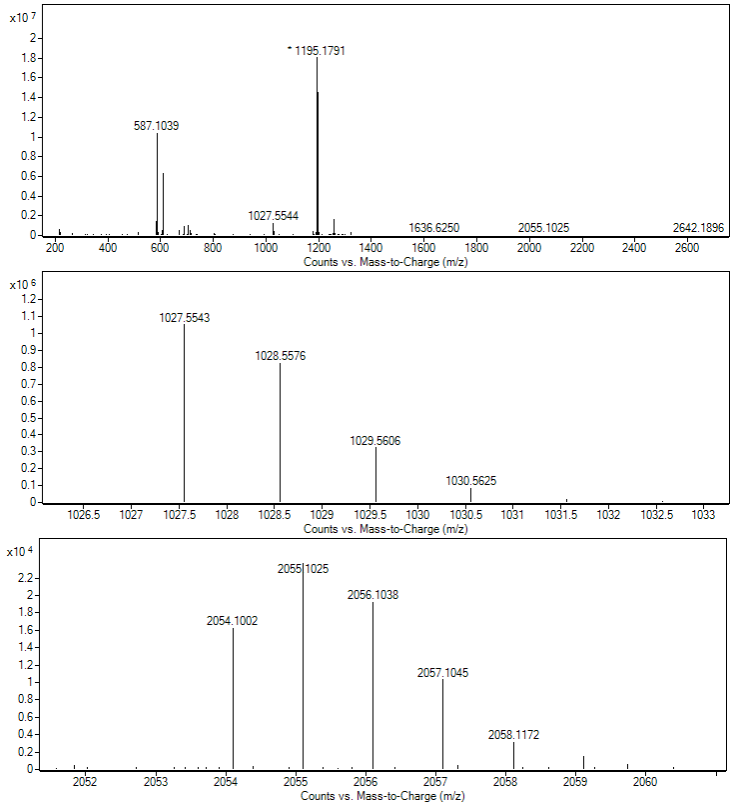


Figure A.44: HRMS spectra for crude reactions taken from the high-throughput screen for cage **K26**$_{[2+3],[4+6]}$, indicating the formation of [2 + 3] and [4 + 6] cages — calc. for $C_{66}H_{66}N_{12}$ 1026.5533 and $C_{132}H_{132}N_{24}$ 2053.1067, found $[M+H]_{[2+3]}^{+}$ 1027.5543, $[M+2H]_{[2+3]}^{2+}$ 514.2828 and $[M+H]_{[4+6]}^{+}$ 2054.1002.

Category 5

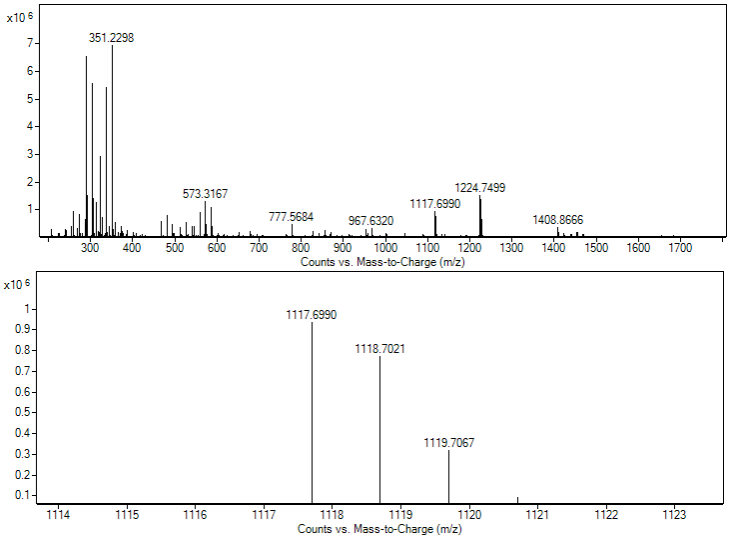


Figure A.45: HRMS spectra for crude reactions taken from the high-throughput screen for cage **G16**$_{[4+6]}$, indicating the formation of a [4 + 6] cage — calc. for $C_{72}H_{84}N_{12}$ 1116.6942, found $[M+H]_{[4+6]}^{+}$ 1117.6990 and $[M+2H]_{[4+6]}^{2+}$ 559.3094.

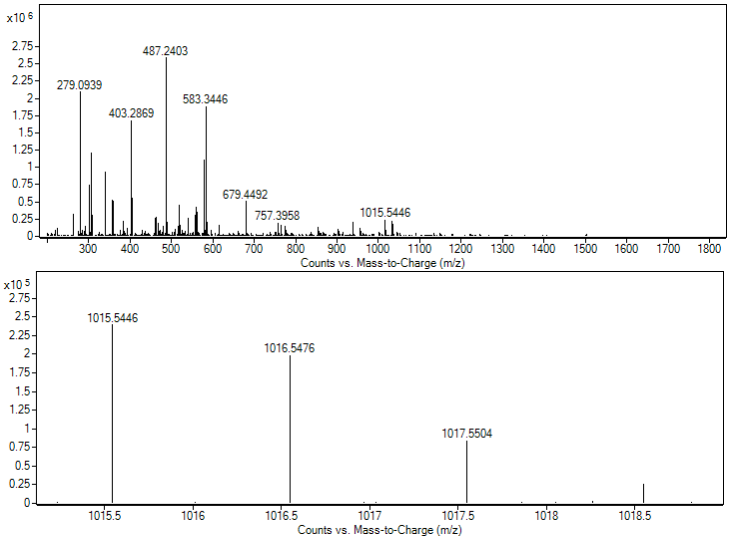


Figure A.46: HRMS spectra for crude reactions taken from the high-throughput screen for cage **H15**$_{[2+3]}$, indicating the formation of a [2 + 3] cage — calc. for $C_{72}H_{66}N_6$ 1014.5349, found $[M+H]_{[2+3]}^{+}$ 1015.5446 and $[M+2H]_{[2+3]}^{2+}$ 508.2866.

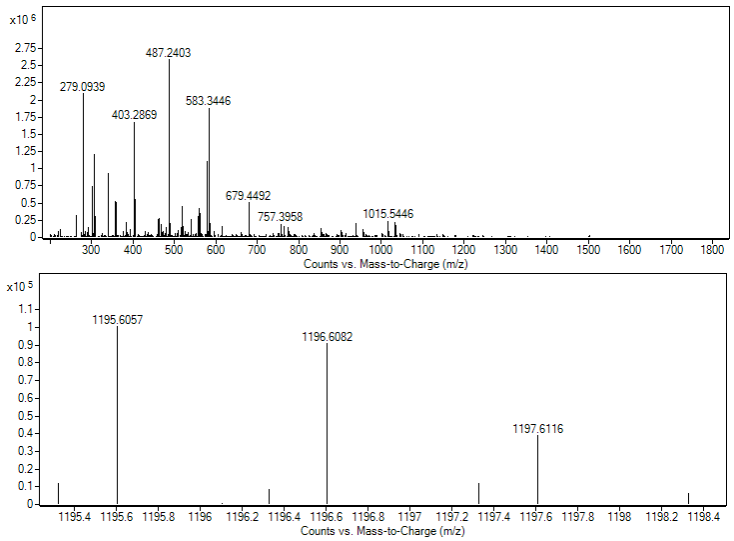


Figure A.47: HRMS spectra for crude reactions taken from the high-throughput screen for cage **R34**$_{[2+3]}$, indicating the formation of a [2 + 3] cage — calc. for $C_{78}H_{78}N_6O_6$ 1194.5983, found $[M+H]_{[2+3]}^{+}$ 1195.6055.

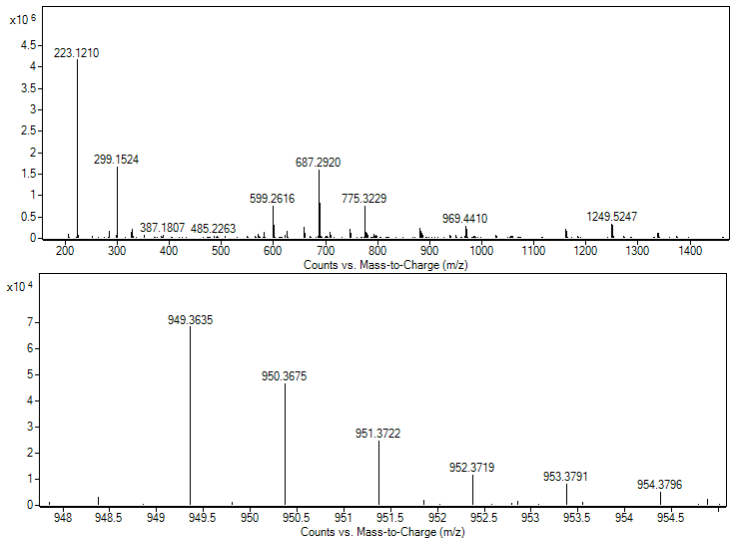


Figure A.48: HRMS spectra for crude reactions taken from the high-throughput screen for cage **T12**$_{[2+3]}$, indicating the formation of a [2 + 3] cage — calc. for $C_{60}H_{48}N_6O_6$ 948.3635, found $[M+H]_{[2+3]}^{+}$ 949.3635.

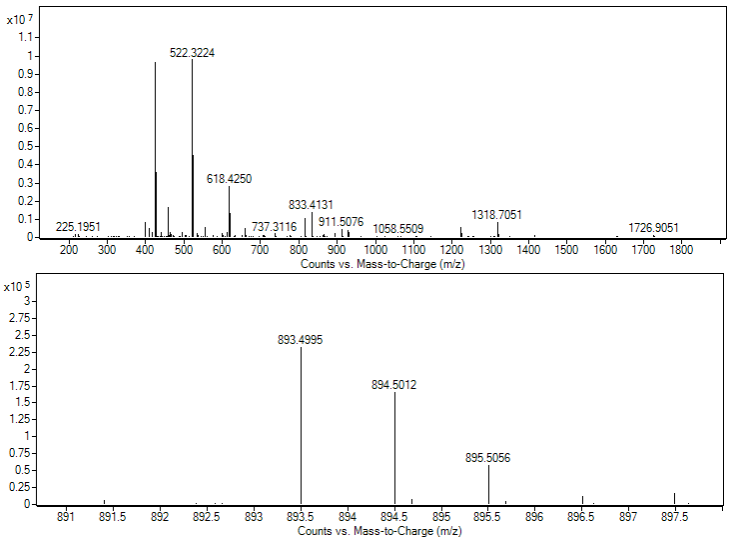


Figure A.49: HRMS spectra for crude reactions taken from the high-throughput screen for cage **J15**$_{[2+3]}$, indicating the formation of [2 + 3] cages — calc. for $C_{60}H_{60}N_8$ 892.4941, found $[M+H]_{[2+3]}^{+}$ 893.4995.

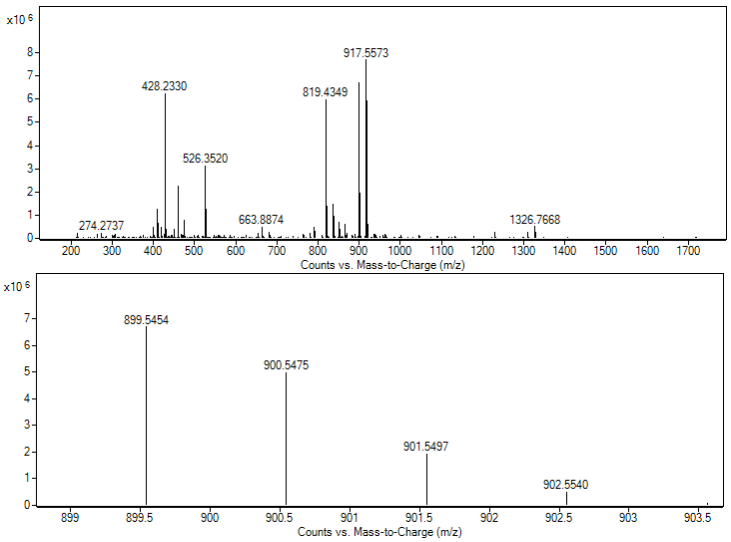


Figure A.50: HRMS spectra for crude reactions taken from the high-throughput screen for cage **J26**$_{[2+3]}$, indicating the formation of [2 + 3] cages — calc. for $C_{60}H_{66}N_8$ 898.5410, found $[M+H]_{[2+3]}^{+}$ 899.5454.

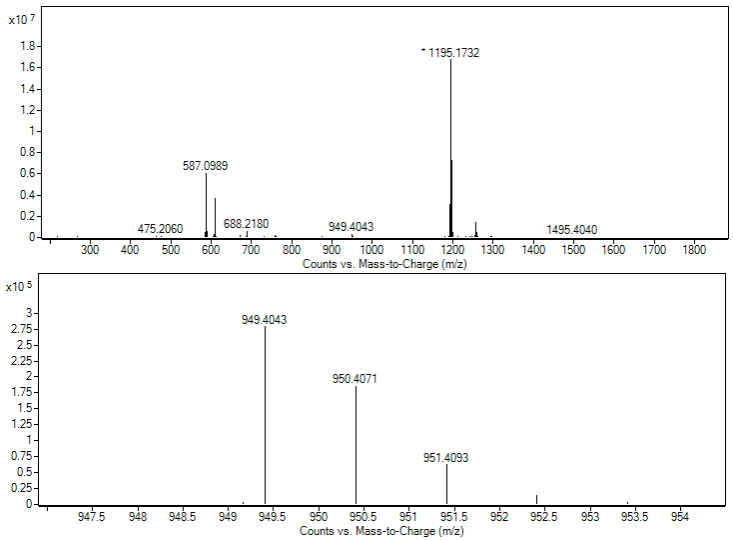


Figure A.51: HRMS spectra for crude reactions taken from the high-throughput screen for cage **K32**$_{[2+3]}$, indicating the formation of [2 + 3] cages — calc. for $C_{57}H_{48}N_{12}O_3$ 948.3972, found $[M+H]_{[2+3]}^{+}$ 949.4043.

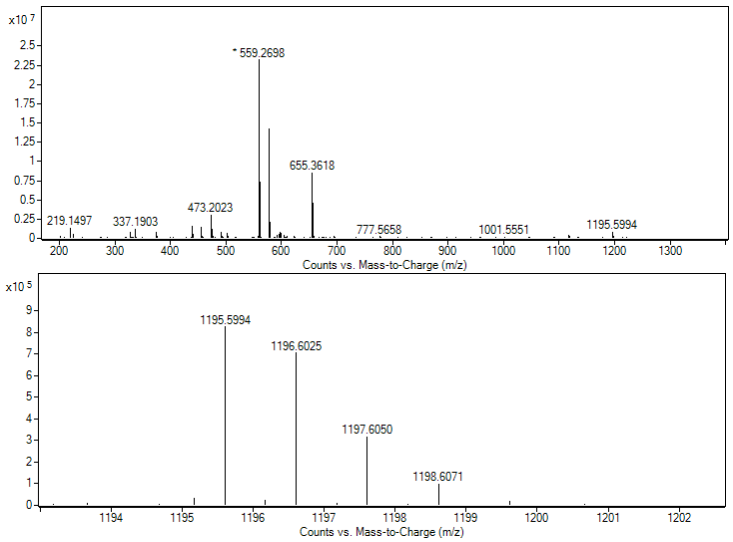


Figure A.52: HRMS spectra for crude reactions taken from the high-throughput screen for cage **R15**$_{[2+3]}$, indicating the formation of a [2 + 3] cage — calc. for $C_{78}H_{78}N_6O_6$ 1194.5983, found $[M+H]^{+}$ 1195.5994 and $[M+2H]^{2+}$ 598.2731.

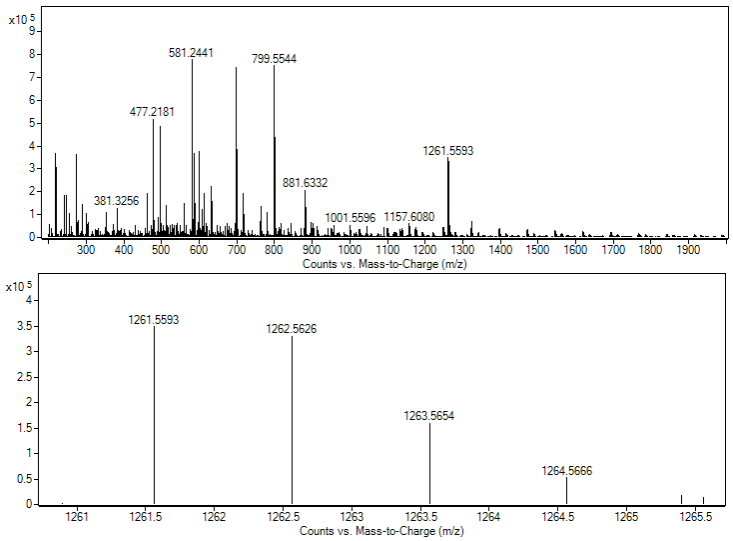


Figure A.53: HRMS spectra for crude reactions taken from the high-throughput screen for cage **R27**$_{[2+3]}$, indicating the formation of [2 + 3] cages — calc. for $C_{84}H_{72}N_6O_6$ 1260.5513, found $[M+H]_{[2+3]}^{+}$ 1261.5593.

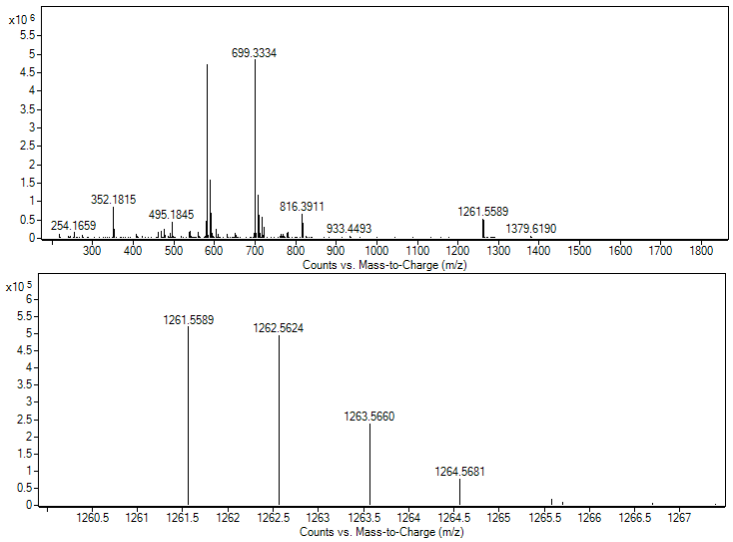


Figure A.54: HRMS spectra for crude reactions taken from the high-throughput screen for cage **R31**$_{[2+3]}$, indicating the formation of a [2 + 3] cage — calc. for $C_{84}H_{72}N_6O_6$ 1260.5513, found $[M+H]_{[2+3]}^{+}$ 1261.5593 and $[M+2H]_{[2+3]}^{2+}$ 631.2844.

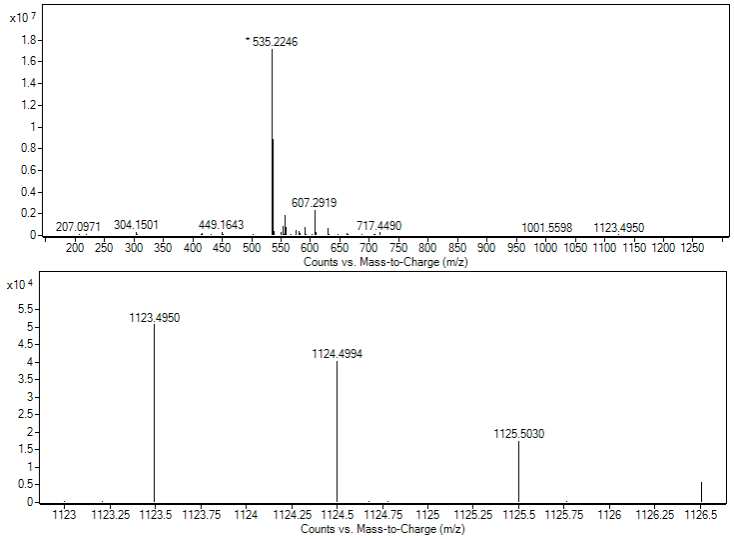


Figure A.55: HRMS spectra for crude reactions taken from the high-throughput screen for cage **R32**$_{[2+3]}$, indicating the formation of a [2 + 3] cage — calc. for $C_{60}H_{66}N_8$ 1122.4891, found $[M+H]_{[2+3]}^{+}$ 1123.4950.

Category 6

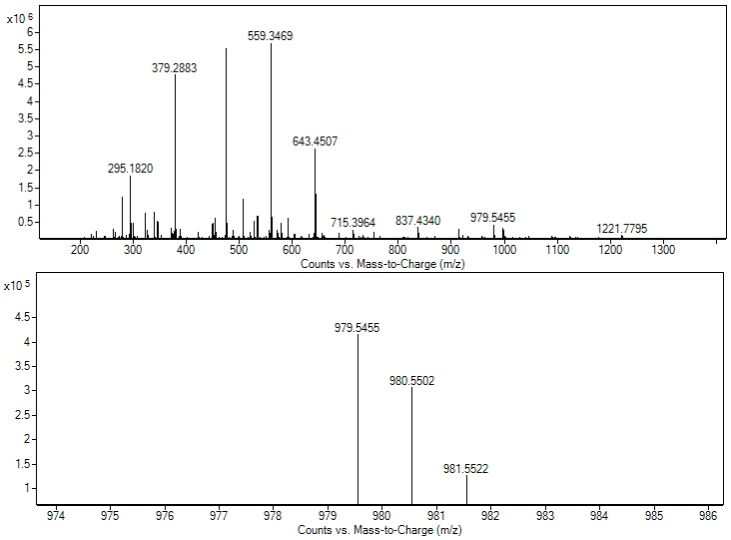


Figure A.56: HRMS spectra for crude reactions taken from the high-throughput screen for cage **H10**$_{[2+3]}$, indicating the formation of [2 + 3] cages — calc. for $C_{69}H_{66}N_6$ 978.5349, found $[M+H]_{[2+3]}^{+}$ 979.5455.

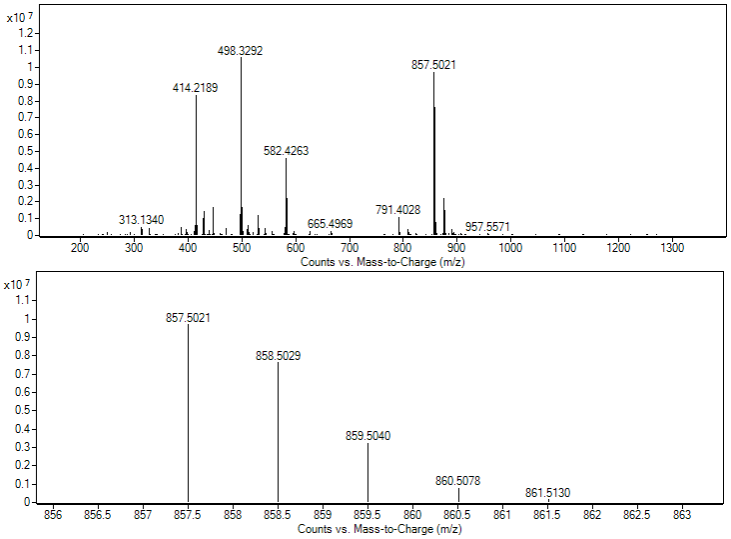


Figure A.57: HRMS spectra for crude reactions taken from the high-throughput screen for cage **J10**$_{[2+3]}$, indicating the formation of a [2 + 3] cage — calc. for $C_{57}H_{60}N_8$ 856.49409, found $[M+H]_{[2+3]}^{+}$ 857.5021.

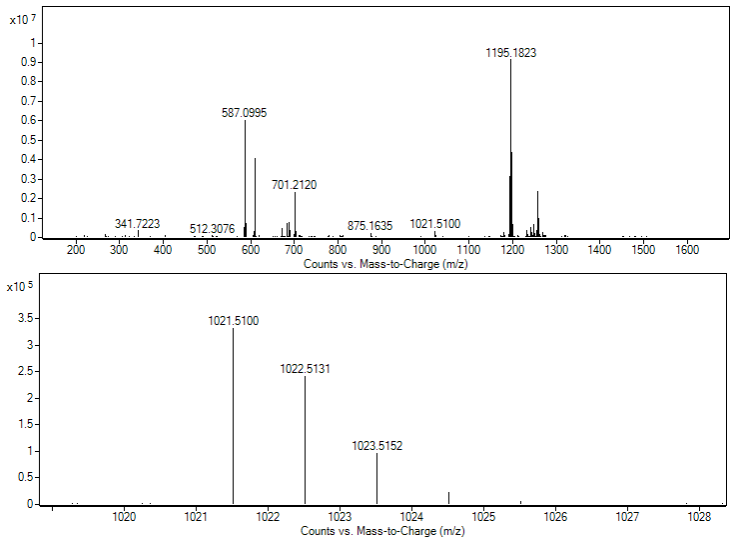


Figure A.58: HRMS spectra for crude reactions taken from the high-throughput screen for cage **K15**$_{[2+3]}$, indicating the formation of a [2 + 3] cage — calc. for $C_{66}H_{60}N_{12}$ 1020.5064, found $[M+H]_{[2+3]}^{+}$ 1021.5100 and $[M+2H]_{[2+3]}^{2+}$ 511.2623.

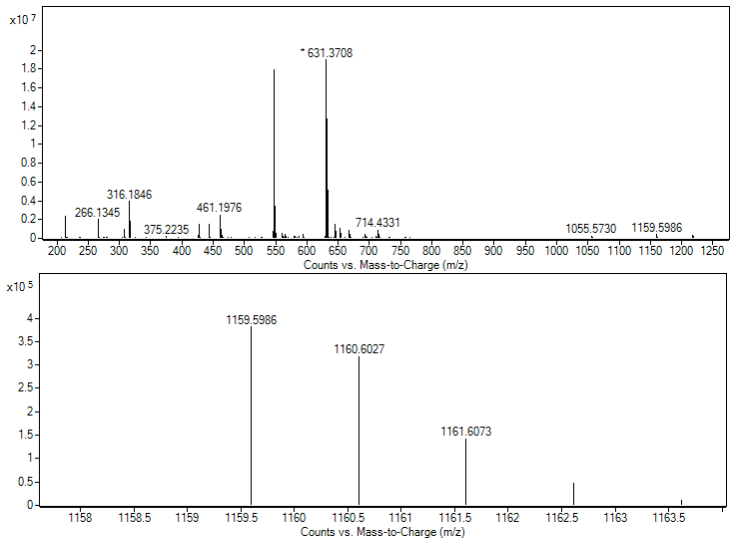


Figure A.59: HRMS spectra for crude reactions taken from the high-throughput screen for cage **R10**$_{[2+3]}$, indicating the formation of a [2 + 3] cage — calc. for $C_{150}H_{156}N_{12}O_{12}$ 1158.5983, found $[M+H]_{[2+3]}^{+}$ 1159.5986.

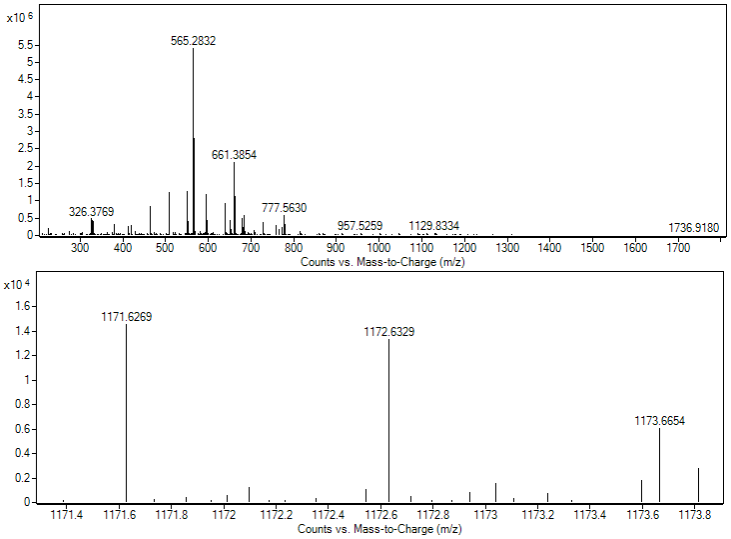


Figure A.60: HRMS spectra for crude reactions taken from the high-throughput screen for cage **U15**$_{[2+3]}$, indicating the formation of a [2 + 3] cage — calc. for $C_{84}H_{78}N_6$ 1170.6288, found $[M+H]_{[2+3]}^{+}$ 1171.6269.

Category 7

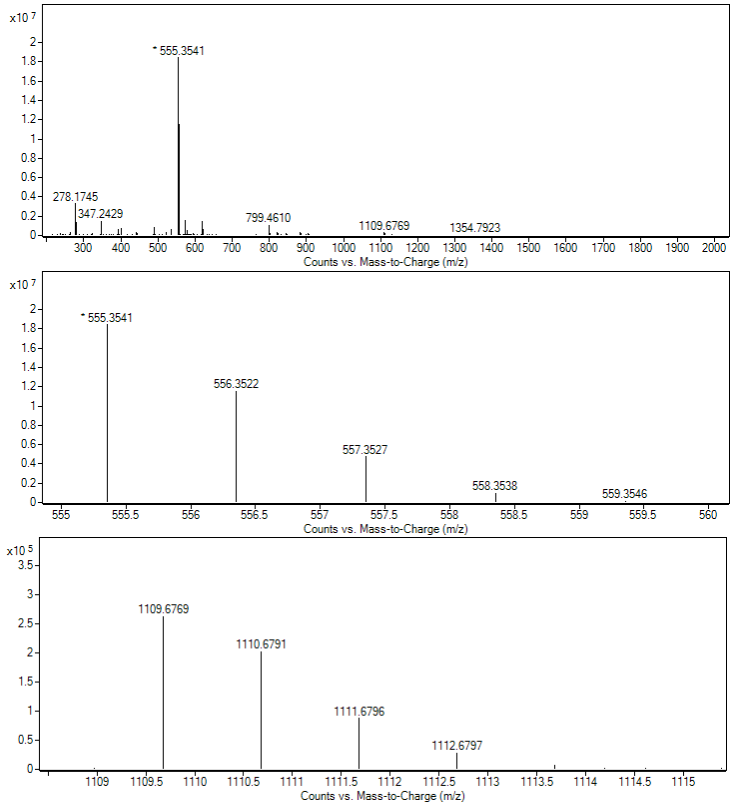


Figure A.61: HRMS spectra for crude reactions taken from the high-throughput screen for cage **I10**$_{[2+3],\,[4+6]}$, indicating the formation of [2 + 3] and [4 + 6] cages — calc. for $C_{33}H_{42}N_6O_2$ 554.3369 and $C_{66}H_{84}N_{12}O_4$ 1108.6739, found $[M+H]_{[2+3]}^{+}$ 555.3541 and $[M+H]_{[4+6]}^{+}$ 1109.6769.

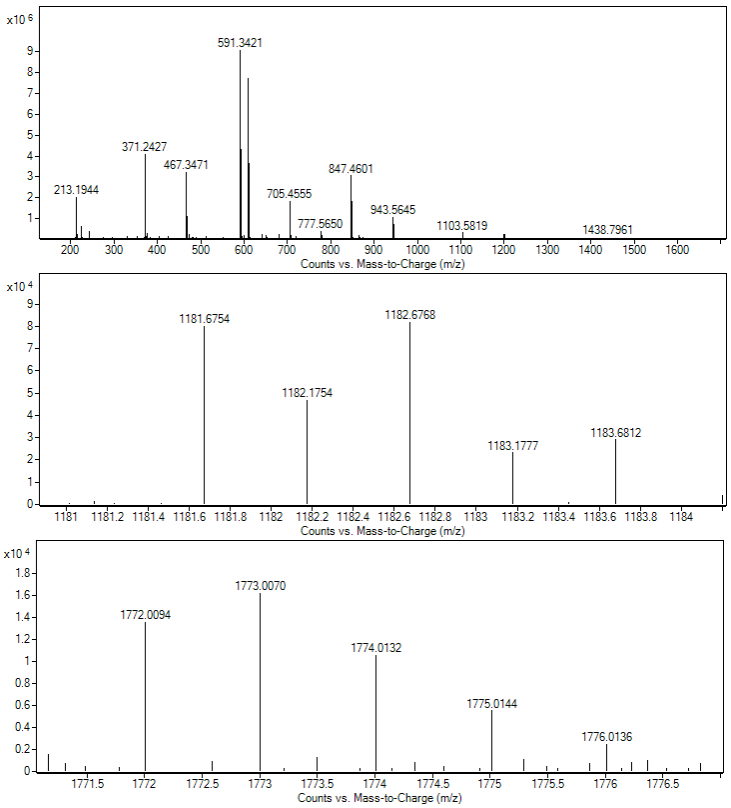


Figure A.62: HRMS spectra for crude reactions taken from the high-throughput screen for cage **I15**$_{\mathbf{[4+6],\,[6+9],\,[8+12]}}$, indicating the formation of [4 + 6], [6 + 9], and [8 + 12] cages — calc. for $C_{72}H_{84}N_{12}O_4$ 1180.6739, $C_{108}H_{126}N_{18}O_6$ 1771.0108, and $C_{144}H_{168}N_{24}O_8$ 2361.3477, found $[M+H]_{[4+6]}^{+}$ 1181.6754, $[M+2H]_{[4+6]}^{2+}$ 591.3421, $[M+H]_{[6+9]}^{+}$ 1172.0094 and $[M+2H]_{[8+12]}^{2+}$ 1181.6754.

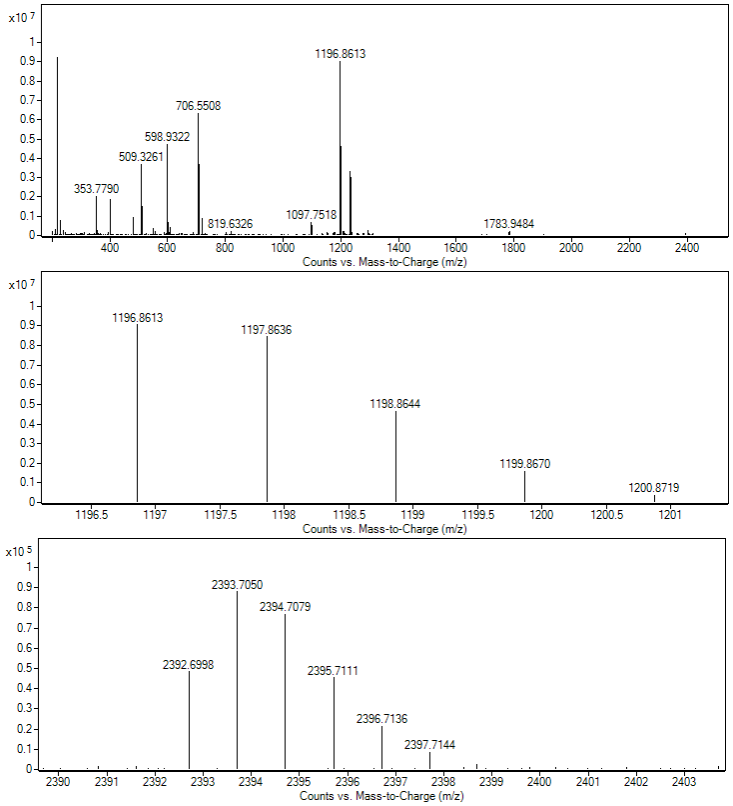


Figure A.63: HRMS spectra for crude reactions taken from the high-throughput screen for cage **J25**$_{[2+3],\,[4+6]}$, indicating the formation of [2 + 3] and [4 + 6] cages - calc. for $C_{78}H_{105}N_{11}$ 1195.8554 and $C_{156}H_{210}N_{22}$ 2391.7109, found $[M+H]_{[2+3]}^{+}$ 1196.8613 and $[M+H]_{[4+6]}^{+}$ 2392.6998.

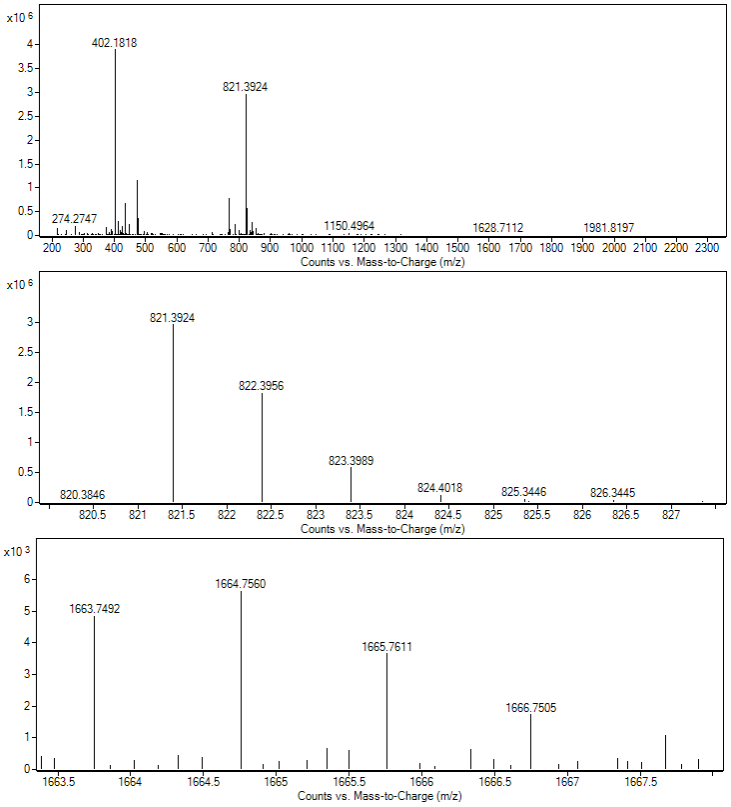


Figure A.64: HRMS spectra for crude reactions taken from the high-throughput screen for cage **J32**$_{[2+3],\,[4+6]}$, indicating the formation of [2 + 3] and [4 + 6] cages — calc. for $C_{51}H_{48}N_8O_3$ 820.3849 and $C_{102}H_{96}N_{16}O_6$ 1640.7699, found $[M+H]_{[2+3]}^{+}$ 821.3924 and $[M+Na]_{[4+6]}^{+}$ 1663.7492.

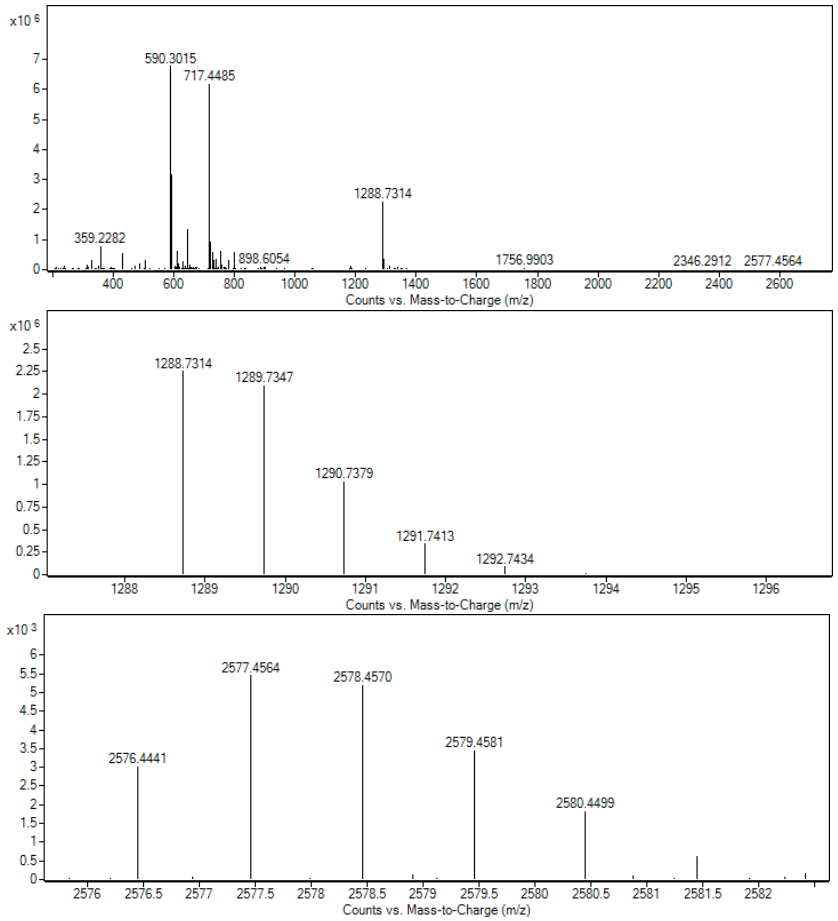


Figure A.65: HRMS spectra for crude reactions taken from the high-throughput screen for cage **R29**$_{[2+3],\,[4+6]}$, indicating the formation of [2 + 3] and [4 + 6] cages — calc. for $C_{81}H_{93}N_9O_6$ 1287.7249 and $C_{162}H_{186}N_{32}O_{12}$ 2575.4498, found $[M+H]_{[2+3]}^{+}$ 1288.7314, $[M+H]_{[4+6]}^{2+}$ 2576.4441.

Category 8

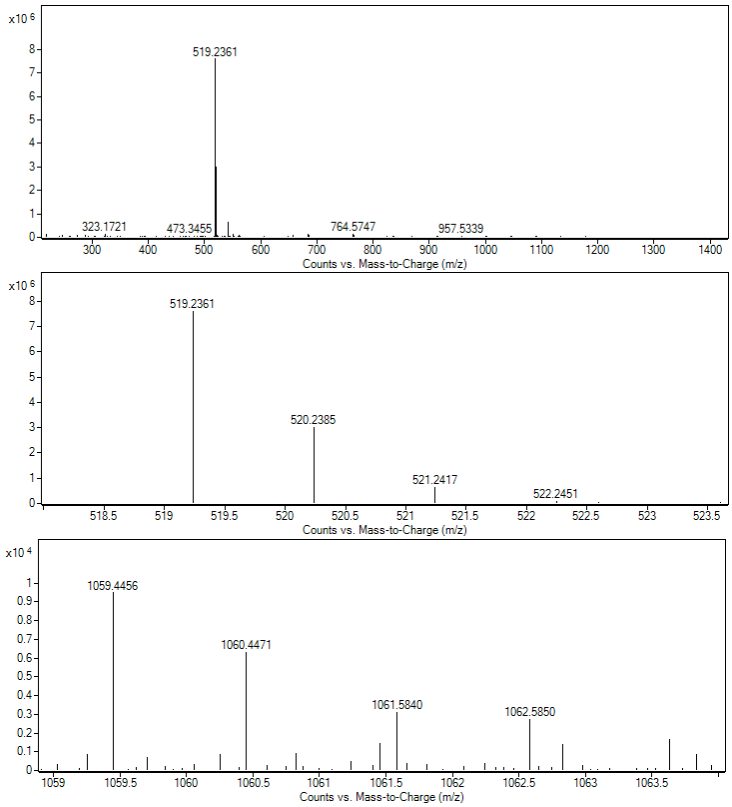


Figure A.66: HRMS spectra for crude reactions taken from the high-throughput screen for cage **I32**$_{[2+3],\,[4+6]}$, indicating the formation of [2 + 3] and [4 + 6] cages — calc. for $C_{27}H_{30}N_6O_5$ 518.2278 and $C_{54}H_{60}N_{12}O_{10}$ 1036.4555, found $[M+H]_{[2+3]}^{+}$ 519.2361 and $[M+Na]_{[4+6]}^{+}$ 1059.4456.

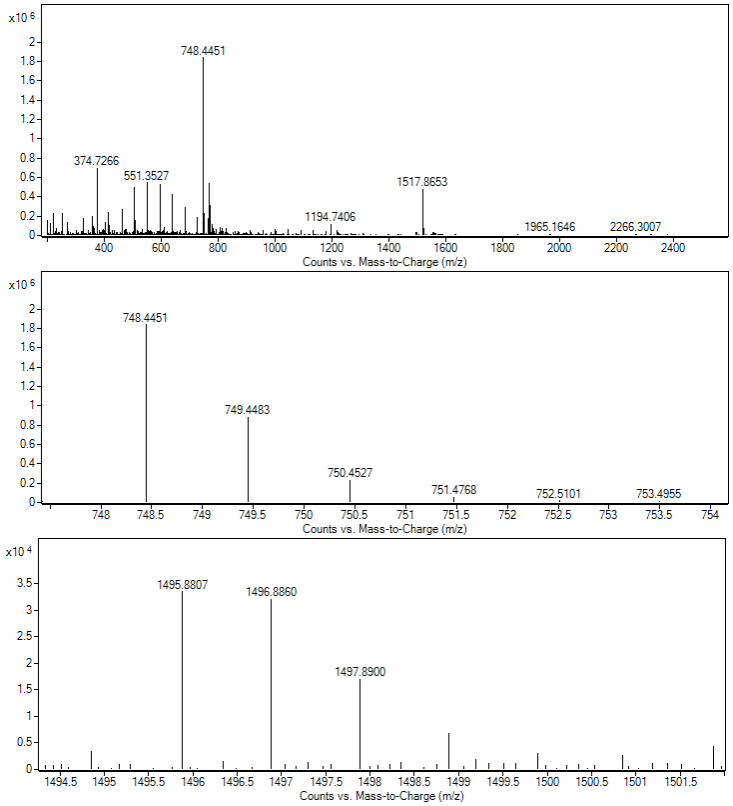


Figure A.67: HRMS spectra for crude reactions taken from the high-throughput screen for cage **T29**$_{[2+3],\,[4+6]}$, indicating the formation of [2 + 3] and [4 + 6] cages — calc. for $C_{39}H_{57}N_9O_6$ 747.4432 and $C_{78}H_{114}N_{18}O_{12}$ 1494.8864, found $[M+H]_{[2+3]}^{+}$ 748.4451 and $[M+H]_{[4+6]}^{+}$ 1495.8807.

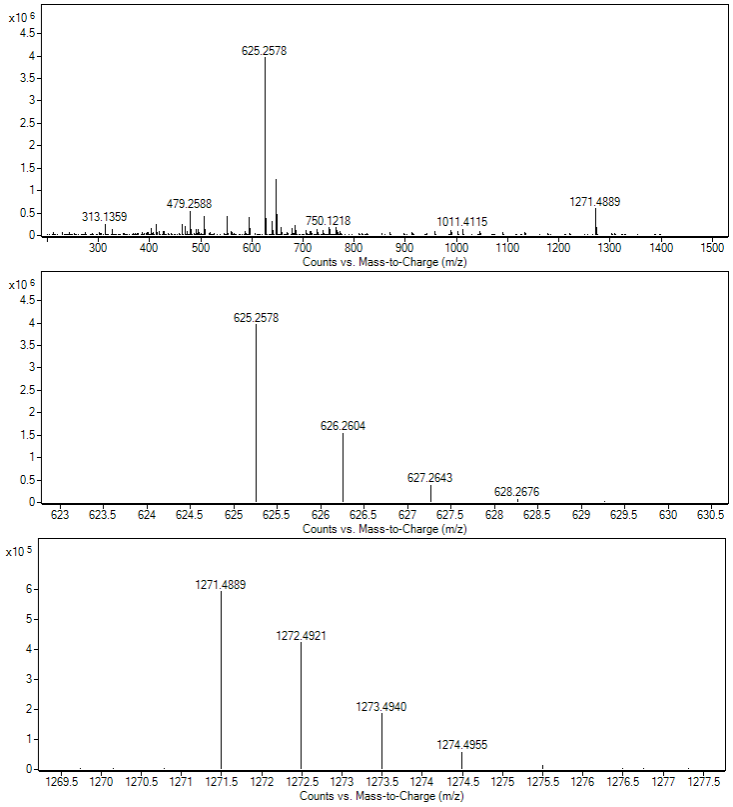


Figure A.68: HRMS spectra for crude reactions taken from the high-throughput screen for cage **T30**$_{[2+3],[4+6]}$, indicating the formation of [2 + 3] and [4 + 6] cages — calc. for $C_{30}H_{36}N_6O_9$ 624.2544 and $C_{60}H_{72}N_{12}O_{18}$ 1248.5088, found $[M+H]_{[2+3]}^+$ 625.2578 and $[M+Na]_{[4+6]}^+$ 1271.4889.

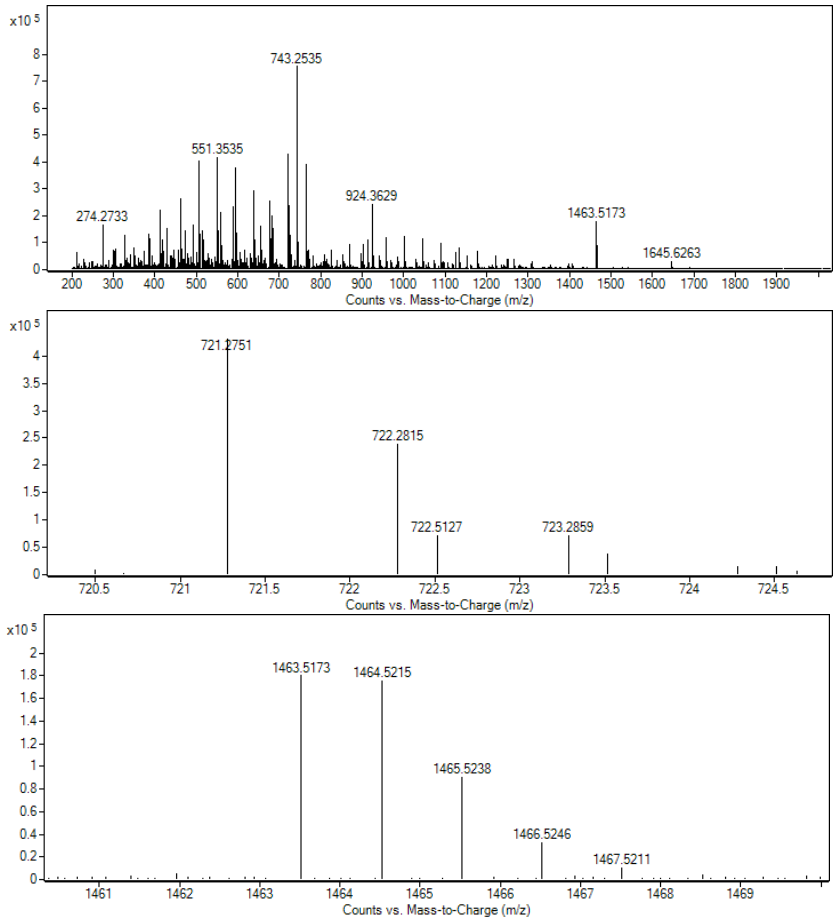


Figure A.69: HRMS spectra for crude reactions taken from the high-throughput screen for cage **T31**$_{[2+3],[4+6]}$, indicating the formation of [2 + 3] and [4 + 6] cages — calc. for $C_{42}H_{36}N_6O_6$ 720.2696 and $C_{84}H_{72}N_{12}O_{12}$ 1440.5393, found $[M+H]_{[2+3]}^+$ 721.2751 and $[M+Na]_{[4+6]}^+$ 1463.5173.

A.3 Copyright Permission Documents

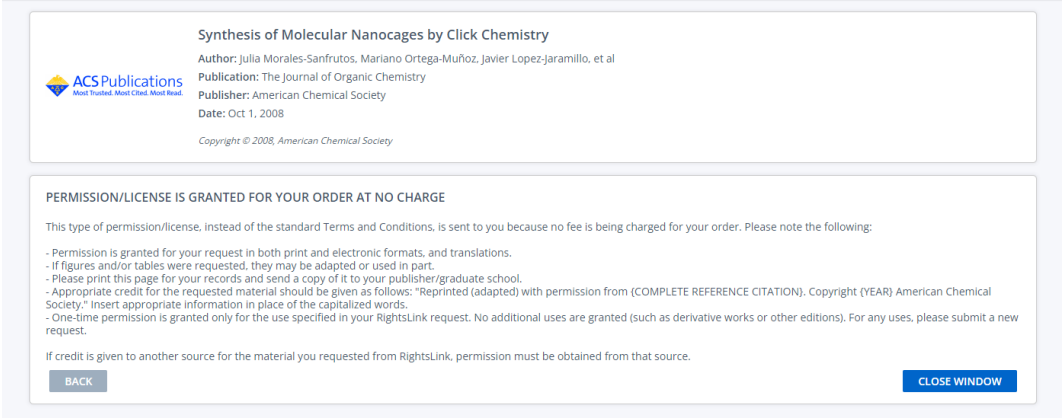


Figure A.70: The copyright permission document permitting re-use of content in reference (39).

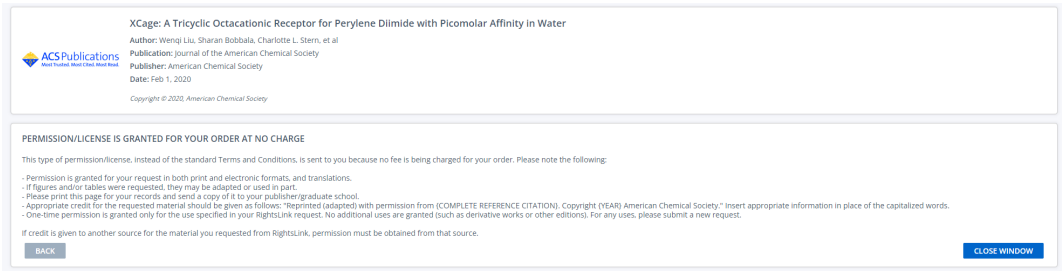


Figure A.71: The copyright permission document permitting re-use of content in reference (40).