Methods Towards Computational Catalyst Design

Tom Young

Balliol College

University of Oxford



A thesis submitted for the degree of

Doctor of Philosophy

at the University of Oxford

Trinity 2021

Methods Towards Computational Catalyst Design

Tom A. Young

Balliol College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

at the University of Oxford, 13th July 2021

Abstract

Identifying active, selective and stable catalysts computationally remains an unsolved problem with numerous benefits in chemistry and the wider world. Determining reaction free energy barriers enables such a classification and relies on predictions from calculations or chemical intuition. By understanding how a biomimetic metallocage achieves catalytic proficiency towards a [4+2] cycloaddition more active catalysts are proposed to require enhanced flexibility and electron deficiency (Chapter 3). In combination with automated approaches to construct host-guest geometries and generating transition states (Chapters 3 & 5) the efficient methodology enables screening new metallocages for a reaction to become routine. A method to improve the accuracy of calculated reaction free energy differences is proposed that uses a more realistic expression for the translational entropy contribution for solution-phase solutes (Chapter 4). Towards calculating accurate free energy differences by dynamics sampling, an efficient method for training machine learned potentials is proposed en-route to fast yet accurate molecular force fields (Chapter 6).

Statement of Originality

I, Tom Young, hereby confirm that this thesis and the work to which it refers are the results of my own efforts. Any ideas, data, images, or text resulting from the work of others (whether published or unpublished) are fully identified as such within the work and attributed to their originator in the text, bibliography, or in footnotes. This thesis has not been submitted in whole or in part for any other academic degree or professional qualification.

Date: 13th July 2021

Signature:

Acknowledgments

Seemingly no sooner than I started, here I am writing the acknowledgements section of my thesis – time really does fly when you're having fun. First and foremost I'd like to thank my supervisor, Prof. Fernanda Duarte for many interesting discussions and continued guidance. I'd also like to thank all the members of the Duarte group both past and present, without which the experience wouldn't have been the same: the OGs (Alistair and Matina/$\mu\alpha\lambda\acute{\alpha}\kappa\alpha$); the Part IIs, particularly Joe for all his work on *autodE* and his continued support of the project; Tristan for listening and questioning all my terrible ideas, and making them better; Tomasz for many engaging discussions; and the rotation students, including Kate, Callum and Matthew for making the office an enjoyable place to be. Although not a member, I'd also like to thank Harry as an honorary member of the group, whose ideas and discussions on unanswerable scientific questions were fun and enlightening (*what is a bond*?).

Thanks also to everyone in TMCS and the theoretical chemistry department: Izzy (or some spelling thereof); Hannah; Lachlan; Tom; Alex, for his trivial discussions; Mike; Ava; Tim, for his supervision during my TMCS rotation; Hugh for super helpful theory chats; and Max for his baked goods, and all the help teaching maths.

Many thanks to Dr. Rob Penfold and Prof. William Barford, and Balliol college generally, for giving me the opportunity to teach during my time here. Thanks to all the supervisors I've had along the way: Dr. Natalie Fey, who introduced me to computation; Profs. Paul Pringle and John Bower for fun collaborative projects and Prof. Sir David Clary. And thanks to many others for their guidance and suggestions: Dr. Kjell Jorner; Prof. Volker Deringer, Prof. Gábor Csányi, Dr. Paul Lusby and Prof. Tim Donohoe.

Finally, thanks to my family whose support made this possible.

Acknowledgments: Funding

I'd like to thank all the funding sources that have allowed me to undertake a DPhil. In particular, the TMCS CDT programme, AWE for the studentship, the Department of Chemistry in Oxford and Balliol College.

Data Availability

The computer code developed and itemised below is freely available on *GitHub*. Raw data is presented in the online supporting information of published manuscripts or in appendices where appropriate.

i. *cgbind*: `https://github.com/duartegroup/cgbind`

ii. *autodE*: `https://github.com/duartegroup/autodE`

iii. *molfunc*: `https://github.com/duartegroup/molfunc`

iv. *othem*: `https://github.com/duartegroup/otherm`

v. *gap-train*: `https://github.com/t-young31/gap-train`

vi. *soap-rust*: `https://github.com/t-young31/soap-rust`

Publications

This thesis is based in part on the following publications:

I. Young, T. A.; Martí-Centelles V.; Wang J.; Lusby P. J.; Duarte F. *J. Am. Chem. Soc.*, **2020**, *142*, 1300–1310.

Contribution: Methodology development, data generation and analysis, drafting.
Reprinted (adapted) with permission from the Journal of the American Chemical Society. Copyright 2020. American Chemical Society.

II. Young, T. A; Gheorghe R.; Duarte F. *J. Chem. Inf. Model.*, **2020**, *60*, 7, 3546–3557.

Contribution: Conceptualisation, methodology development, data generation and analysis, drafting.
Reprinted (adapted) with permission from the Journal of Chemical Information and Modeling. Copyright 2020. American Chemical Society.

III. Young, T. A.; Silcock J. J.; Sterling A. J.; Duarte F. *Angew. Chem .Int. Ed.*, **2021**, *60*, 4266–4274.

Contribution: Conceptualisation, methodology development, data generation and analysis, drafting.
Reprinted (adapted) with permission from Angewandte Chemie International Edition. Copyright 2021. John Wiley and Sons.

IV. Young, T. A.; Johnston-Wood T.; Deringer V.; Duarte F. *Chemical Science*, **2021**. DOI: 10.1039/d1sc01825f.

Contribution: Conceptualisation, methodology development, data generation and analysis, drafting.

The following contributing-author publications do not form part of this thesis.

V. Armstrong R. J.; Akhtar W. M.; Young, T. A., Duarte F.; Donohoe T. J. *Angew. Chem. Int. Ed.*, **2019**, *58*,12558–12562.

VI. Spicer R. L.; Stergiou A. D., Young, T. A.; Duarte F.; Symes M. D.; Lusby P. J. *J. Am. Chem. Soc.*, **2020**, *142*, 2134–2139.

VII. Wang G.-W.; Boyd O.; Young, T. A.; Bertrand S. M.; Bower J. F., *J. Am. Chem. Soc.*, **2020**, *142*, 1740–1745.

VIII. Wang J.; Young, T. A.; Duarte F.; Lusby P. J. *J. Am. Chem. Soc.*, **2020**, *142*, 17743–17750.

IX. Wang G.-W.; Sokolova O. O.; Young, T. A.; Christodoulou E. M. S.; Butts C. P.; Bower J. F., *J. Am. Chem. Soc.*, **2020**, *142*, 19006–19011.

X. Branfoot C.; Young, T. A.; Wass D. F.; Pringle P. G. *Dalton Trans.*, **2021**, *50*, 7094–7104.

XI. Aoki Y. *et. al*, Young, T. A.; *Faraday Discuss.*, **2019**, *220*, 282–316.

XII. Aoki Y. *et. al*, Young, T. A.; *Faraday Discuss.*, **2019**, *220*, 144–178.

XIII. Bauer M. *et. al*, Young, T. A.; *Faraday Discuss.*, **2019**, *220*, 464–488.

Contents

Abstract i

Statement of Originality ii

Acknowledgments iii

Data Availability v

Publications vi

List of Abbreviations xv

1 Introduction 1

1.1 Discovery 2

1.2 Direct-design and Optimisation 4

1.2.1 Ab initio Methods 4

1.2.2 Data-driven Methods 5

1.3 Inverse Design 5

1.4 Understanding 6

1.5 Summary and Outlook 8

2 Theoretical Background **9**

2.1 Electronic Structure . . . 9

2.1.1 Preamble . . . 9

2.1.2 Hartree-Fock . . . 10

2.1.3 Density Functional Theory . . . 12

2.1.4 Post Hartree-Fock Methods . . . 15

2.1.5 Molecular Implementation . . . 16

2.1.6 Periodic Electronic Structure . . . 19

2.1.7 Semi-empirical Methods . . . 20

2.1.8 Implicit Solvent Models . . . 22

2.1.9 Analysis Methods . . . 22

2.1.9.1 Partial Atomic Charges . . . 23

2.1.9.2 Frontier Molecular Orbital Theory . . . 24

2.1.9.3 Non-covalent Interactions . . . 25

2.1.9.4 Electrostatic Potentials . . . 25

2.2 Molecular Dynamics . . . 26

2.2.1 Preamble . . . 26

2.2.2 Integration Algorithms . . . 27

2.2.3 Thermostats . . . 28

2.2.4 Empirical Force Fields . . . 28

2.3 Machine Learned Potentials . . . 30

2.3.1 Preamble . . . 30

2.3.2 Representations . . . 31

2.3.3 Regression Algorithms . . . 33

3 Metallocage Catalysis **36**

3.1 Overview . . . 36

3.2 Manuscript I . . . 39

3.2.1 Abstract . . . 39

3.2.2 Introduction . . . 39

3.2.3 Results and Discussion . . . 43

3.2.3.1 Dynamic Properties . . . 43

3.2.3.2 Efficient Protocol for Binding Affinity Calculations . . . 45

3.2.3.3 Rationalizing Differences in Binding Affinities . . . 49

3.2.3.4 Rationalizing the Differences in Catalytic Activity . . . 51

3.2.3.5 Efficient Protocol for Calculating Catalytic Activity . . . 55

3.2.4 Conclusion . . . 57

3.2.5 Computational Methods . . . 58

3.3 Selected Supporting Information I . . . 61

3.3.1 Square-planar Metal Model . . . 61

3.3.2 Metallocage MD Modelling . . . 61

3.3.3 Thermodynamic Contributions . . . 72

3.4 Manuscript II . . . 94

3.4.1 Abstract . . . 94

3.4.2 Introduction . . . 94

3.4.3 Results and Discussion . . . 97

3.4.3.1 Software Outline . . . 97

3.4.3.2 Metallocage Construction . . . 99

3.4.3.3 Adding New Template Architectures . . . 102

3.4.3.4 Analysis Tools . . . 104

3.4.3.5 Substrate Encapsulation and Binding Affinity Prediction . . 107
3.4.3.6 Web app GUI 110
3.4.3.7 Examples 111
3.4.4 Conclusion 119
3.4.5 Computational Methods 119
3.5 Selected Supporting Information II 120

4 Solution Phase Entropy **128**

4.1 Abstract 128
4.2 Introduction 128
4.3 Theory 130
4.3.1 Classical Statistical Mechanics 131
4.3.2 Quantum Statistical Mechanics 132
4.3.3 Quantum: Gas Phase 134
4.3.4 Ideal Gas Method 134
4.3.4.1 Translation: Particle in a Box 134
4.3.4.2 Rotation: Rigid Rotor 135
4.3.4.3 Vibrational: Harmonic Oscillator 136
4.3.5 Alternative Methods 137
4.3.6 Concentration Corrections 140
4.4 Limitations of the IGM and Harmonic Approximation 141
4.4.1 Model Systems 141
4.4.2 Water-in-Water 143
4.5 Translational Surface 145
4.5.1 Methodology Optimisation 147
4.5.2 2D and 3D Surfaces 149

4.5.3 Solute and Solvent Dependence 149
4.6 Translational Entropy from Spherical Wells 152
4.6.1 Infinite Spherical Well 152
4.6.2 Exponential Well 154
4.6.3 Test Cases 163
4.7 Conclusion 168
4.8 Methods 169

5 autodE **170**

5.1 Overview 170
5.2 Manuscript III 172
5.2.1 Abstract 172
5.2.2 Introduction 172
5.2.3 Methodology 175
5.2.4 Results and Discussion 185
5.2.5 Conclusion 193
5.3 Selected Supporting Information III 195
5.3.1 Computational Methods 195
5.3.2 Compatibility of autodE with Different Electronic Structure Theory Codes 195
5.3.3 Metal-complex Conformers 199
5.3.4 Constructing Molecular Graphs 207
5.3.5 Non-covalent and Reactive Reactant Complexes 209
5.3.6 Algorithm Details 213
5.3.7 Reaction Classes 224

5.4 Further Developments . . . 230
5.4.1 Transition State Guesses . . . 230
5.4.2 Molecule Building . . . 237
5.4.3 Future Development . . . 238

6 Machine Learned Potentials 241

6.1 Overview . . . 241
6.2 Manuscript IV . . . 243
6.2.1 Abstract . . . 243
6.2.2 Introduction . . . 243
6.2.3 Results and Discussion . . . 246
6.2.3.1 A Prospective Error Metric . . . 246
6.2.3.2 Water Models . . . 248
6.2.3.3 Other Solvent Systems . . . 255
6.2.3.4 Aqueous Zn(II) . . . 256
6.2.3.5 Metallocage Dynamics . . . 257
6.2.3.6 Reaction Dynamics in Gas and Solvent Phase . . . 258
6.2.4 Conclusion . . . 266
6.2.5 Methods . . . 267
6.3 Selected Supporting Information IV . . . 270
6.3.1 Other Solvents . . . 284
6.3.2 Reactions . . . 288

7 Conclusions 295

7.1 Modelling Catalytic Processes . . . 295
7.2 Future Directions . . . 297

8 Appendices **298**

8.1 ML Representations 298

8.1.1 SOAP 298

8.2 Solution Phase Entropy 301

8.2.1 Partition Functions 301

8.2.2 Classical Truncated Harmonic Oscillator 302

8.2.3 Classical Morse Oscillator 303

8.2.4 Exponential Well 304

8.2.5 Test Sets 306

List of Abbreviations

Abbreviation	Meaning
ACSF	Atomic-centred symmetry functions
AE	Absolute error
AFIR	Artificial force induced reaction
AIMD	Ab initio molecular dynamics
AINR	Ab initio nanoreactor
AL	Active learning
ANN	Artificial neural network
AO	Atomic orbtial
API	Application programming interface
BFGS	Broyden-Fletcher-Goldfarb-Shanno optimisation
BSIE	Basis set incompleteness error
BSSE	Basis set superposition error
CC	Coupled cluster
CCDC	Cambridge Crystallographic Data Centre
CCSD	Coupled cluster singles and doubles
CCSD(T)	CCSD with perturbative triples
CI	Configuration interaction
COM	Centre of mass
DA	Diels-Alder

Abbreviation	Meaning
DCM	Dichloromethane
DFT	Density Functional Theory
DLPNO	Domain-based local pair natural orbital
ECP	Effective core potential
EDDM	Electron density difference map
ESP	Electrostatic potential
ETDG	Experimental-Torsion Distance Geometry
ETKDG	Experimental-Torsion Distance Geometry with Knowlege
EW	Exponential well
FEP	Free energy perturbation
FF	Force field
FSM	Freezing string method
GAP	Gaussian approximation potential
GDML	Gradient-domain machine learning
GGA	Generalized gradient approximation
GPR	Gaussian process regression
GSM	Growing string method
GTO	Gaussian-type orbtial
GUI	Graphical user interface

Abbreviation	Meaning
HF	Hartree-Fock
HO	Harmonic oscillator
HOMO	Highest occupied molecular orbtial
HVTS	High-throughput virtual screening
IGM	Ideal gas method
KS	Kohn-Sham
LCC	Local coupled cluster
LEPS	London-Eyring-Polanyi-Sato
LFM	Low frequency mode
LUMO	Lowest unoccupied molecular orbtial
MAD	Mean absolute deviation
MC	(Metropolis) Monte Carlo
MD	Molecular dynamics
MEP	Minimum energy pathway
ML	Machine learning
MLP	Machine learned potential
MO	Molecular orbtial
MOF	Metal organic framework
MP2	Møller-Plesset perturbation theory to 2nd order

Abbreviation	Meaning
MSD	Mean signed deviation
MSE	Mean squared error
NBO	Natural bond orbtial
NCI	Non-covalent interaction
NEB	Nudged elastic band
NN	[Artificial] neural network
PES	Potential energy surface
PF	Partition function
PIB	Particle in a box
PP	Pseudopotential
PT	Perturbation theory
QM	Quantum mechanics
QM/MM	Quantum mechanics in a molecular mechanics environment
QSAR	Quantitative structure activity relationships
RDF	Radial distribution function
RESP	Restrained electrostatic potential
RI	Resolution of the identity
RMSD	Root mean squared deviation
RMSE	Root mean squared error

Abbreviation	Meaning
RR	Ridgid rotor
SMILES	Simplified molecular-input line-entry system
SSD	Sum squared distance
STO	Slater-type orbtial
TS	Transition state
TSA	Transition state analogue
VRI	Valley-ridge inflection point
WF	Wave function
ZPE	Zero point energy

1 Introduction

Advances in the chemical sciences are essential if the world is to meet the United Nations Sustainable Development Goals by 2030[1] and prevent potentially catastrophic and irreversible climate change.[2] With most chemicals encountering a catalyst en-route to their final form,[3] enhancing catalytic activity and selectivity lowers the energy demand and associated waste products. Furthermore, 'circular' synthesis – reusing final chemical compounds by catalytic decomposition – will be required to reduce the ever-depleting fossil fuel feedstock.[4, 5] Optimising existing catalysts and developing new ones for as-yet inconceivable transformations is therefore of paramount importance.

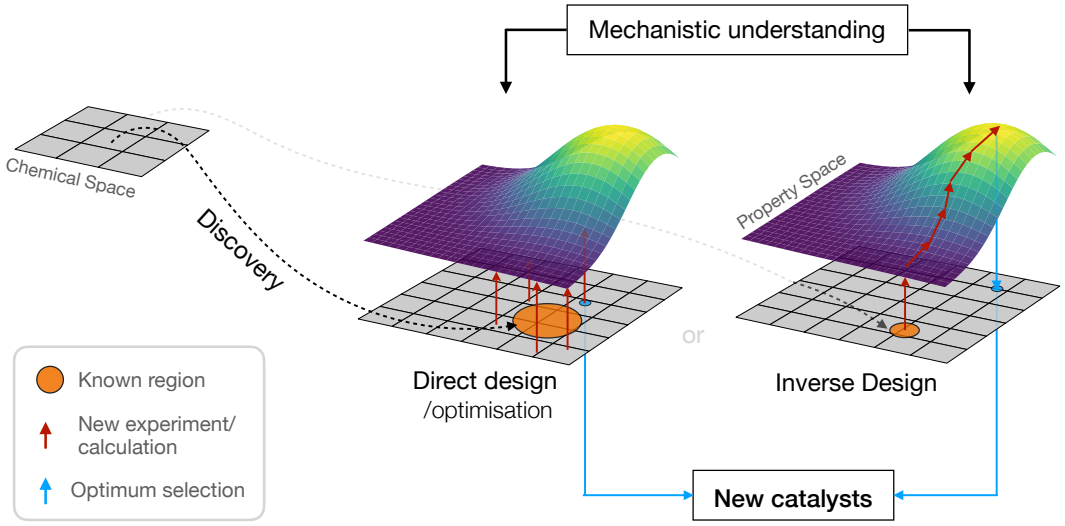


Figure 1.0.1. Schematic representation of the approaches to developing new catalysts.

Discovering and optimising catalysts *in-silico* is seen as somewhat of a *holy-grail* in computational chemistry.[6] This is because computation avoids costly experimental synthesis and testing. Strategies towards this goal will be the subject of this thesis following an overview of approaches often employed, while highlighting current limitations. Figure 1.0.1 outlines different approaches to catalyst design, where discovery of a hit leads from an unexplored chemical space to a known region. From there, either a direct or inverse approach is generally required to optimise the hit compound to an optimal point in property space, which is then selected as a potential new catalyst. The dimensionality and complexity of the property space means a targeted approach is required, which may be aided by mechanistic understanding. These stages are expanded upon in the following sections.

1.1 Discovery

Serendipity often plays a role in discovering catalysts and their pre-requisite reactions in the experimental laboratory.[7, 8] While high-throughput approaches are starting to be used,[8–10] a more targeted and less wasteful approach is highly desirable.

Computational methods that discover reactions and catalysts can be broadly divided into two categories: graph-based and simulation-based.[11, 12] In the former, a graph representation is defined using atoms as nodes and 'bonds' as edges, and new structures found from a reactant state by applying edge edits and searching for stable products (Figure 1.1.1).[13–18] Despite being an efficient way to explore new chemical space, they are limited to discovering reactions rather than catalysts due to an exponential scaling with system size.

Simulation-based approaches employ enhanced[a] reactive dynamics and are also currently lim-

[a]Standard equilibrium dynamics only sample close to the minima in currently accessible timescales (pico- to microseconds) and not over significant energy barriers ie. the rare event problem.[19]

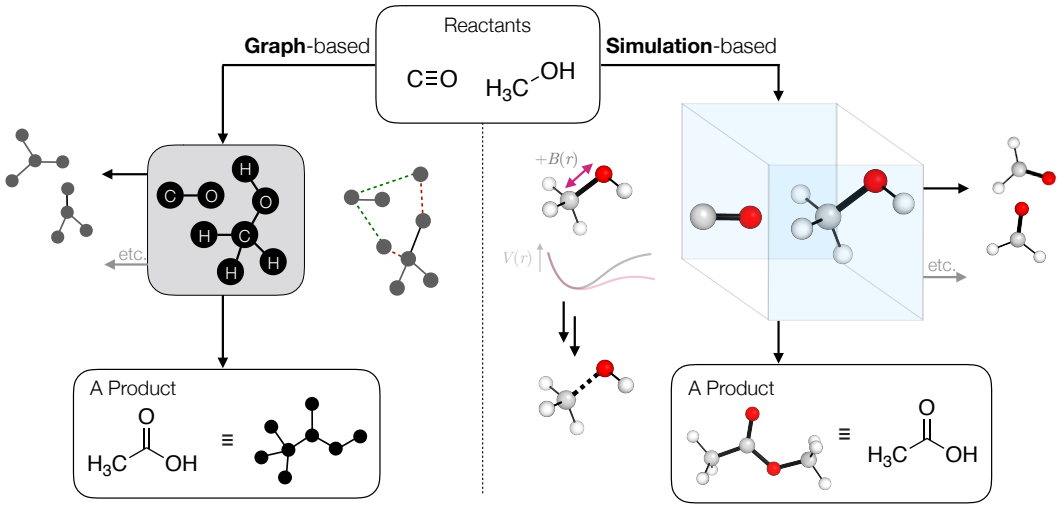


Figure 1.1.1. Representation of graph- and simulation-based approaches to discover new reactions between carbon monoxide and methanol. $B(r)$ is a biasing potential.

ited to reaction rather than catalyst discovery. Non-equilibrium molecular dynamics[20] such as meta-dynamics developed by Parrinello,[21] can traverse reaction barriers on the picosecond timescale (Figure 1.1.1b). These include the global reaction route mapping method,[22] which has been used to discover reaction intermediates[23, 24] reactions by biasing the total energy with harmonic-like terms to find new minima. Imposed activation and conformer searching has also been used to discover reactions.[25] These approaches apply some bias in the sampling to target 'active' atoms within bonds that break or form, thus narrowing the search space.

Methods that do not require defining 'active' atoms include the "*ab initio* nanoreactor"[26], which uses high temperature dynamics under an oscillating biasing potential to favour bond formation. Adding a biasing potential to already visited regions in the configuration space has also been used to discover plausible transformations.[27]

A comparison of these approaches is illustrated in Figure 1.1.1; from a carbon monoxide and methanol mixture both methods can generate an acetic acid product but neither can suggest

a catalyst. Including, for example, a transition metal and a section of ligands with a view to discovering a catalyst leads to an intractably complex problem.

Indeed, despite the efficiency of current computational approaches, to the best of my knowledge, a catalyst has yet to be discovered computationally without prior knowledge. Enhancing the speed and accuracy of computational methods will prove an important step towards tackle catalyst discovery, which will be the focus of this thesis.

1.2 Direct-design and Optimisation

Once a catalyst has been discovered, modification to enhance selectivity/activity is generally driven by experimental observation. Optimisation involves the reaction conditions (temperature/solvent etc.) and the catalyst identity, and while linked, the two are often considered separately. Selection of reaction conditions beyond the standard trial-and-error approach using previous experimental data is a promising approach,[28, 29] but is still limited by experimental throughput to at most hundreds of reactions.[30, 31] *In silico* optimisation of reaction conditions without prior knowledge is an extremely challenging problem, whereas catalyst optimisation by substructure modification is currently achievable. Methods towards this are outlined in the following sections.

1.2.1 Ab initio Methods

Perhaps the simplest approach to catalyst optimisation is to modify a canonical catalyst and enumerate all possibilities within a library, calculating a relevant observable or proxy, and selecting the optimum. Although more common in drug discovery,[32] endeavours along the same lines are being developed. For example, the *AARON* (An Automated Reaction

Optimizer for New Catalysts) tool has been used to exhaustively sample modified catalyst structures in organometallic complexes.[33] In a similar vein, the *CatVS* tool has been developed to predict enantioselectivities using enumeration over a library of ligands and ligand conformers.[34] While direct enumeration through a library is guaranteed to find the global minimum, exponential scaling of the library prevents a high diversity. However, with mechanistic understanding of the catalytic cycle targeted enumeration is possible. Towards using this approach in the context of supramolecular catalysis, a method for automated *in silico* construction of metallocages and their host-guest complexes is introduced in Chapter 3.

1.2.2 Data-driven Methods

Complementary to direct calculation *ab initio* are data-driven approaches, which incorporate experimental data to build regression models. These quantitative structure activity relationships (QSAR)[35] and machine learning[36–38] approaches have been employed widely to both understand and in some cases suggest catalyst modifications using generative models.[39] These complement classic genetic algorithms for exploring or traversing[40] a local space.[41] Along with suggesting perturbations to a catalyst, machine learning (ML) models also provide an avenue to select which experiments, thus observations to collect.[42, 43]

1.3 Inverse Design

Suggesting an optimal catalyst for a given transformation i.e. 'inverse design' is perhaps the ultimate goal of computational catalyst design. In contrast to optimisation, the end-to-end prediction is unobtainable using enumeration in a chemical space of $> 10^{60}$ just small organic molecules.[44] Indeed, inverse optimisation has only been used in small property spaces.[45,

46] Generally, this proceeds via optimisation in a continuous representation and conversion into a suggested structure. Notable examples include: application to nitrogen-fixating catalysts by optimising a 'jacket potential' of point charges surrounding a core,[47] and a linear combination of atomic potential approach, for optimising $CO \rightarrow CO_2$ catalysts.[48] In the latter case a calculated activation barrier was used as the objective function and the site occupancies for different groups on a chemical core the variable parameters.

The lack of examples using inverse design illustrates the difficulty of the problem; it thus remains as an ultimate goal with methods being developed towards its achievement.

1.4 Understanding

One of the most prevalent uses of computation in organo(metallic) chemistry is to understand a particular (catalytic) mechanism, with the hope that such knowledge can then be used to optimise and design new catalysts rationally.[49, 50] Given a set of experimental observations, a catalytic cycle may be calculated and an effective energy barrier quoted along with, perhaps, a rate limiting step.[51] From this knowledge, a human-guided approach to catalyst design can be employed using classical physical organic methods, e.g. considering inductive/resonance/steric/solvation effects to modulate kinetics, which in turn modify selectivity/activity/stability.[52]

A notable example of how understanding can lead to better catalysts is from the groups of Houk and List, who experimentally identified a highly selective catalyst for an asymmetric Mannich reaction,[53] based on computational modelling (see ref. [54] for further examples).[55] Understanding reaction mechanisms via computation is also crucial in organic chemistry, which was pioneered by Houk in the 1980s,[56] and more recently applied

in organometallic[57] and biocatalytic[58] settings.

In this context, a number of collaborative projects with experimental groups were undertaken, which sought to understand reactivity and catalysis. Broadly these projects are not included in this thesis (Manuscripts V–X) but have provided the motivation for methods developed here.

In particular, Manuscript I, which forms part of Chapter 3, centres on understanding the mechanism of (in)action of two supramolecular cage catalysts. This work attempts to address the lack of enzyme-mimicking (biomimetic) supramolecular catalysts, which is an emerging field following the early experimental work of Breslow,[59] Sanders[60] and Rebek[61] among others. Only a few years ago computational work lead by Harvey[62, 63], Himo[64] and Head-Gordon[65] has started to tackle supramolecular catalysts.

Gaining mechanistic understanding through computation is, however, often a slow process. A balance is required between speed/simplicity and accuracy, which is a focus of Manuscript I and II (Chapter 3). Methods to improve the accuracy of mechanistic predictions are presented in Chapter 4, wherein a novel method for calculating the entropic contribution to free energies is proposed. Automation also can lead to significant speed-ups, which is the subject of Manuscript III and is presented as Chapter 5. For modestly sized molecular systems of 50 atoms or fewer the whole process of finding minima and transition states from a 2D chemical representation accurately is automated. Finally, towards efficiently modelling dynamics and reactivity, Manuscript IV is presented as Chapter 6 and outlines a method to generate machine learned potentials efficiently for a variety of systems. These dynamical effects can be important in reactions not well described by transition state theory e.g. bifurcations.[66]

1.5 Summary and Outlook

Computational approaches have been, and continue to be, important tools in discovering, understanding and optimising catalysts. Their predictive ability is, however, far from optimal. Specifically, (1) commonly employed manual computation limits throughput; (2) observable properties are challenging to calculate, particularly solution-phase equilibria and rate constants, which rely on accurate solvent and thermal contributions to the energy; (3) gas-phase energetics have a speed accuracy trade-off that often preclude high-throughput enumeration.

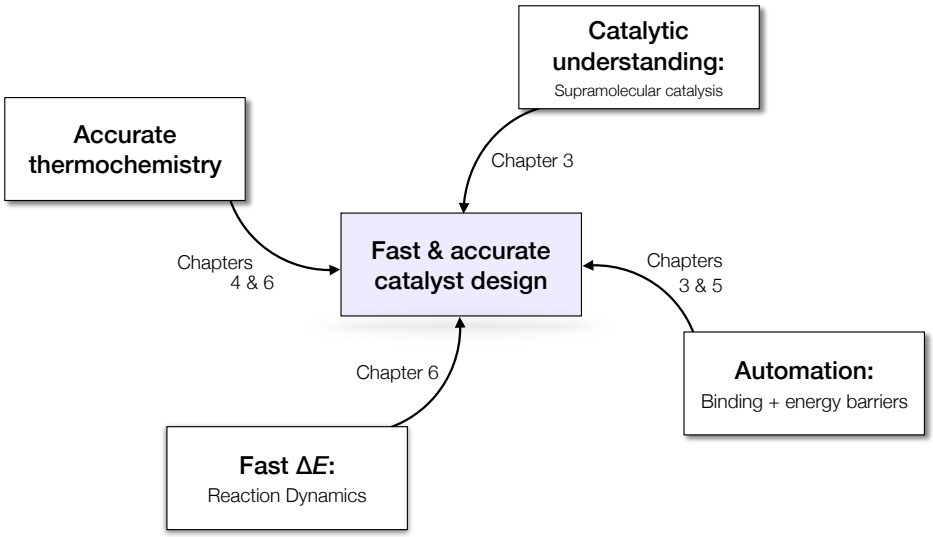


Figure 1.5.1. Schematic outline of this thesis.

Avenues in biomimetic catalysis, automation and obtaining more accurate energies from computation are the subject of this thesis and are discussed in the following chapters (Figure 1.5.1). From a brief theoretical background of the methods used in this thesis (Chapter 2) the first result chapter focuses on understanding a catalytic system and developing a method to enumeratively optimise it. The subsequent chapters are motivated by this work and focus largely on reactions as prerequisites to their catalytic counterparts for simplicity and speed. These methods are, however, applicable to simple catalytic processes and provide a route to developing complex catalysts computationally.

2 Theoretical Background

2.1 Electronic Structure

2.1.1 Preamble

Electronic structure theory plays a pivotal role in much of computational chemistry, providing energies and derivatives of the energy that can be used to predict and rationalise experimental observations. In particular, the solution of the time-independent Schröedinger equation,

$$\hat{H}\Psi = E\Psi \tag{2.1}$$

is sought; where $\hat{H}$ is the system's Hamiltonian describing the kinetic and potential energies of all particles described by the wave function (Ψ) with energy (E). Unfortunately, Eqn. (2.1) is insoluble for all but the most simple Hamiltonians; thus a series of approximations are therefore applied to reach a tractable problem for most chemical systems.

For polyatomic systems it is usually necessary to make the Born-Oppenheimer approximation, such that there is no coupling between electronic and nuclear wave functions, i.e. $\Psi = \psi(\boldsymbol{r};\boldsymbol{R})\phi(\boldsymbol{R})$, for electronic coordinates $\boldsymbol{r}$ and nuclear coordinates $\boldsymbol{R}$. The non-

relativistic Hamiltonian is then,

$$\hat{H}_{\text{elec}} = \hat{T}_{\text{elec}} + \hat{V}_{\text{elec-nuc}} + \hat{V}_{\text{elec-elec}} + \hat{V}_{\text{nuc-nuc}} \tag{2.2}$$

which, in first quantisation is,

$$\hat{H}_{\text{elec}} = -\frac{1}{2}\sum_{i}^{\text{elec}} \nabla_i^2 - \sum_{i}^{\text{elec}}\sum_{A}^{\text{nuc}} \frac{Z_A}{|\boldsymbol{r}_i - \boldsymbol{R}_A|} + \sum_{i}^{\text{elec}}\sum_{j<i}^{\text{elec}} \frac{1}{|\boldsymbol{r}_i - \boldsymbol{r}_j|} + \sum_{A}^{\text{nuc}}\sum_{B<A}^{\text{nuc}} \frac{Z_A Z_B}{|\boldsymbol{R}_A - \boldsymbol{R}_B|} \tag{2.3}$$

in atomic units, with nuclear charges Z_A, ∇ is the Laplacian operator and the final term is just a constant with respect to the electronic wave function ($\hat{V}_{\text{nuc-nuc}} = V_{\text{nn}}$). With an electronic Hamiltonian the ground sate energy may be obtained using the variational principle,

$$E_{\text{est.}} = \min_{\{\alpha\}} \left(\frac{\langle \psi(\alpha) | \hat{H}_{\text{elec}} | \psi(\alpha) \rangle}{\langle \psi(\alpha) | \psi(\alpha) \rangle} \right) \geq E_0 \tag{2.4}$$

for the best estimate of the energy given a wavefunction in parameters α, which is guaranteed to be higher or equal to the true ground state energy (E_0).

2.1.2 Hartree-Fock

Given Eqn. (2.1) with the electronic Hamiltonian [Eqn. (2.3)] is only analytically soluble for one-electron systems, e.g. the hydrogen atom, an approximate solution is required for poly-electron systems. Hartree-Fock (HF) theory is one such approximation and makes use of a wave function ansatz built from hydrogen-like electron eigenfunctions (orbitals), which

to enforce antisymmetry, may be written as a Slater determinant,

$$\psi_{\mathrm{HF}}(\boldsymbol{r}) = \frac{1}{n!}\begin{vmatrix} \varphi_1(\boldsymbol{r}_1) & \varphi_2(\boldsymbol{r}_1) & \varphi_3(\boldsymbol{r}_1) & \cdots \\ \varphi_1(\boldsymbol{r}_2) & \varphi_2(\boldsymbol{r}_2) & \varphi_3(\boldsymbol{r}_2) & \cdots \\ \varphi_1(\boldsymbol{r}_3) & \varphi_2(\boldsymbol{r}_3) & \varphi_3(\boldsymbol{r}_3) & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{vmatrix} \tag{2.5}$$

for n electrons in n spin orbitals. This representation is equivalent to an idealised system, where each electron interacts with an average of all the others, rather than explicitly, such that opposite spin electrons are 'uncorrelated'.[a]

From ansatz (2.5) the optimal HF wave function is obtained using the variational principle,[67] under the constraint of orbital orthonormality by minimising the Lagrangian function,

$$\mathcal{L}_{\mathrm{HF}}[\{\varphi\}] = \langle\psi_{\mathrm{HF}}|\hat{H}_{\mathrm{elec}}|\psi_{\mathrm{HF}}\rangle - \sum_{ij}\lambda_{ij}(\delta_{ij} - \langle\varphi_i|\varphi_j\rangle) \tag{2.6}$$

with respect to the orbitals, where $\langle\mu|\hat{O}|\nu\rangle = \int_{\mathrm{all}}\mu(\tau)^*\hat{O}\nu(\tau)\mathrm{d}\tau$ is an expectation value in Dirac bra-ket notation, $\delta_{mn} = 1$ if $n = m$ and 0 otherwise, is a Kronecker delta. The orthonormality constraint makes the expectation value the HF energy as a functional of the orbitals,

$$E_{\mathrm{HF}}[\{\varphi\}] = \sum_i \underbrace{\langle\varphi_i|\hat{h}_i|\varphi_i\rangle}_{h_i} + \frac{1}{2}\sum_i\sum_j \underbrace{\langle\varphi_i\varphi_i|r_{12}^{-1}|\varphi_j\varphi_j\rangle}_{J_{ij}} - \underbrace{\langle\varphi_i\varphi_j|r_{12}^{-1}|\varphi_j\varphi_i\rangle}_{K_{ij}} + V_{\mathrm{nn}} \tag{2.7}$$

where the first term includes the kinetic energy and electron-nuclear attraction of one electron and the second the electron-electron repulsion and exchange energy. The latter terms are

[a] Given a Slater product for simplicity $\psi = \varphi_1(\boldsymbol{r}_1)\varphi_2(\boldsymbol{r}_2)$ and so $P(\boldsymbol{r}_1 \in V \,|\, \psi) = \int_{\mathrm{all}}\int_V |\psi|^2\mathrm{d}\boldsymbol{r}_1\mathrm{d}\boldsymbol{r}_2 = \int_{\mathrm{all}}|\varphi_2(\boldsymbol{r}_2)|^2\mathrm{d}\boldsymbol{r}_2\int_V|\varphi_1(\boldsymbol{r}_1)|^2\mathrm{d}\boldsymbol{r}_1 = \int_V|\varphi_1(\boldsymbol{r}_1)|^2\mathrm{d}\boldsymbol{r}_1 = P_1$, which is not a function of $\boldsymbol{r}_2$.

written in 'physicists' notation.[b]

While HF theory provides a reasonable electronic density, indeed it can be used as an improvement over density functional theory (DFT) densities,[68] electron correlation is crucial in strongly correlated systems and important to achieve accuracy approaching 1 kcal mol^{-1} on energy differences.[69] The chemical systems in the following chapters have no low-lying excited states thus little 'strong'/static correlation, and a single electronic configuration is a qualitatively accurate description. Therefore, only methods to include dynamic[c] electron correlation will be introduced in the subsequent sections.

2.1.3 Density Functional Theory

Since its inception in the late 1920s by Thomas[70] and Fermi[71] DFT has become the one of the most widely used approaches to approximate the solution of Eqn. (2.1).[72] Indeed, 12 of the top 100 cited papers relate to DFT as of 2014.[73] This uptake has been a result of its improved accuracy over HF theory by inclusion of some dynamic electron correlation at a similar computational cost.

Although DFT is formally an exact theory for the ground state energy, with Hohenberg and Kohn deriving a one-to-one mapping from the density to the energy,[74] the exact functional has not, yet, been found.[75] Therefore, the development of approximate density functionals has been significant, making a veritable 'zoo' of functionals.[76, 77]

The general form of a functional is,

$$E[\rho(\boldsymbol{r})] = T_{\text{elec}}[\rho(\boldsymbol{r})] + V_J[\rho(\boldsymbol{r})] + V_{\text{ext}}[\rho(\boldsymbol{r})] + V_{\text{xc}}[\rho(\boldsymbol{r})] \tag{2.8}$$

[b] $\langle pq|r_{12}^{-1}|rs\rangle := \int \varphi_p(\boldsymbol{r}_1)^* \varphi_q(\boldsymbol{r}_2)^* |\boldsymbol{r}_1 - \boldsymbol{r}_2|^{-1} \varphi_r(\boldsymbol{r}_1) \varphi_s(\boldsymbol{r}_2) \mathrm{d}\boldsymbol{r}_1 \mathrm{d}\boldsymbol{r}_2$

[c] Conceptually, dynamic correlation is the correlation between opposite spin electrons within a single electronic configuration, although there is no clear distinction between static and dynamic correlation.

where T_{elec} is the total kinetic energy of the electrons, V_J the Coulomb repulsion between electrons, V_{ext} the 'external' electron-nuclear interaction and V_{xc} the exchange and correlation functional. Unfortunately, no good approximations to T_{elec} exist,[78] motivating the use of Kohn-Sham (KS) DFT, where ρ is expressed in an orbital basis.[79] This allows the exact expression for the kinetic energy for a fictitious non-interacting system with the same density as the full system to be used at the expense of increasing the dimensionality from 3 to $3N$ and the practical scaling from $O(1)$ to formally $O(N^4)$[d] in the system size N.[80]

With KS-DFT being the only practical type, the choice of exchange-correlation functional constitutes the DFT zoo's diversity. These are usually arranged according to a Jacobs Ladder hierarchy of increasing complexity, which may not necessarily correlate with accuracy.[81] The first rung up from the 'Hartree world' is the local density approximation, where V_{xc} is a function of the local density at a point; the second rung also depends on the gradient of the density, and includes generalised gradient approximations (GGAs, e.g. PBE,[82] BLYP[83, 84]); the third rung also the second derivative of the density (metaGGA, e.g. TPSS[85]); while the fourth and most commonly utilised rung uses exact HF exchange (hybrid, e.g. PBE0,[86] B3LYP,[87, 88] M06-2X[89]) to correct some of the self-interaction error encountered in DFT. Beyond the fourth rung post-HF correlation is introduced usually at a significant computational cost.

Despite their continued evolution, recent developments have focused on achieving accurate energetics on benchmark sets, rather than reducing density errors.[76] This reduces the generalisability of these functionals and makes 'old' functionals e.g. PBE0, or B3LYP still a good choice without a benchmarking study.[90, 91]

Another limitation of DFT is the description of dispersion, the inclusion of which is essential

[d] V_J is a 4-index quantity in atomic orbitals.

in weakly-bound systems as, for example, at PBE and B3LYP the benzene dimer is purely repulsive.[92] Although in principle the exchange correlation functional should account for dispersion, it is a non-local property and thus expensive to include into the functional.[93] Instead, semi-empirical corrections (e.g. D3[94]) are generally used to account for the absent dispersion contribution. These have the general form,

$$E_{\text{disp}} = -\sum_{A} \sum_{B>A} \sum_{n\in\{6,8..\}} s_n \frac{C_n^{AB}}{R_{AB}^n} f_{\text{damp}}(R_{AB}) \tag{2.9}$$

where the sum is over all unique atom pairs (A, B) with distances R_{AB} raised to even powers above 6 (empirically derived) and s_n is a scaling factor. To prevent the singularity as $R \to 0$ a damping function f_{damp} may be introduced, which has been found to increase accuracy on benchmark sets.[76] The latest generally implemented dispersion method is the damped D3 (D3BJ, $f_{\text{damp}}(x) = x^n/(x^n + c)$ for a constant c),[95, 96] with the most recent D4 method not yet wildly available.

With such a wide variety of functionals available a pragmatic approach is needed. Accepting an implicit error on reaction barriers and energies of ~ 5 kcal mol^{-1}[97] and ~ 1 kcal mol^{-1} on double differences for similar systems,[98] the transferability of the functional is more important than the absolute accuracy. For that reason, the non-empirical PBE functional and its hybrid variant PBE0 in combination with D3BJ semi-empirical dispersion will be used throughout this thesis unless greater accuracy is needed and benchmarks to post-HF methods have been performed. This method has been shown to be both transferable, accurate and is reasonably efficient.[76]

2.1.4 Post Hartree-Fock Methods

In contrast to DFT, *ab initio* approaches to include electron correlation provide a systematic way of improving the HF energy. The most simple of which uses perturbation theory (PT) in the electron-electron interaction. Møller-Plesset PT theory is the most common variant in which the first order term vanishes and has a second-order contribution,[99]

$$E_{\mathrm{MP2}} = -\frac{1}{4}\sum_{ij}^{\mathrm{occ.}}\sum_{ab}^{\mathrm{virt.}} \frac{|\,\langle\varphi_i\varphi_j|r_{12}^{-1}|\varphi_a\varphi_b\rangle - \langle\varphi_i\varphi_j|r_{12}^{-1}|\varphi_b\varphi_a\rangle\,|^2}{\epsilon_a + \epsilon_b - \epsilon_i - \epsilon_j} \tag{2.10}$$

where ϵ_k is a occupied or virtual orbital energy. The summation over two particle-hole excitations generally affords accuracy on-par with high-performing density functionals,[100, 101] but at a higher cost. Furthermore, the denominator in Eqn. (2.10) leads to instabilities (possible singularities) in systems with low-lying excited states and in orbital-optimised MP2.[102]

Simply increasing the expansion order of the perturbation does not generally result in increased accuracy, and can even diverge.[103] Therefore, alternative methods are required. The most conceptually simple being configuration interaction (CI) where the WF is built from a linear combination of Slater determinants with optimised coefficients. CI is, however, not size consistent and less applicable to calculating energy differences for association processes.[104] Coupled cluster (CC) theory on the other hand is size consistent and, given certain conditions, is referred to as the computational 'gold standard'.[105]

Coupled cluster theory makes use of a WF ansatz built from the HF wave function as,

$$|\psi_{\mathrm{CC}}\rangle = e^{\hat{T}}\,|\psi_{\mathrm{HF}}\rangle \tag{2.11}$$

where $\hat{T}$ is the cluster operator (rather than a kinetic energy),

$$\hat{T} = \underbrace{\sum_{i}^{\text{occ}} \sum_{a}^{\text{virt}} t_i^a \hat{a}_a^\dagger \hat{a}_i}_{\hat{T}_1} + \underbrace{\frac{1}{4} \sum_{ij}^{\text{occ.}} \sum_{ab}^{\text{virt.}} t_{i,j}^{a,b} \hat{a}_a^\dagger \hat{a}_b^\dagger \hat{a}_j \hat{a}_i}_{\hat{T}_2} + \cdots \tag{2.12}$$

t is an amplitude, $\hat{a}_p^\dagger$ and $\hat{a}_p$ are creation operator and annihilation operators, respectively, for an electron in orbital p. These operators generate excitations 'out of' the HF determinant by populating virtual orbitals. Without truncation and with infinitely many virtual orbitals the CC wave function is exact. To obtain a computationally tractable problem, the expansion is truncated at, for example, second order and is denoted coupled cluster singles and doubles (CCSD). The exponentiation means excitations far beyond singles and doubles are included and also makes it size consistent.

Inclusion of the triples cluster operator ($\hat{T}_3$) is usually required to obtain energy differences within chemical accuracy (1 kcal mol^{-1}),[106] but comes at a significant cost. For conventional CCSDT the cost scales at $O(N^8)$ in the number of basis functions and even for the perturbatively interoperated triples [CCSD(T)] scales as $O(N^7)$. Methods used to reduce this scaling and allow CC to be used for modestly sized molecular systems will be outlined in the following section. While many other post-HF methods exist,[107] CC remains the most applicable approach for ground state systems that are well described by a single electronic configuration and will, therefore, be the 'high-level' method of choice in this thesis.

2.1.5 Molecular Implementation

Employing numerical approaches to approximate E using WF or density ansatzes as (combinations of-) Slater determinants requires the orbitals to be expanded in a basis. Molecular

orbitals (MOs) are generally expressed as a linear combination of atomic orbitals (AOs) as,

$$\varphi(\boldsymbol{r}_1) = \phi(\boldsymbol{x}_1)\sigma(\omega_1) \qquad ; \qquad \phi(\boldsymbol{x}_1) = \sum_{k}^{\text{AOs}} c_k \phi_k^{\text{AO}}(\boldsymbol{x}_1) \tag{2.13}$$

where the spin orbital φ is a product of spatial (ϕ) and spin (σ) components. The AOs are hydrogen-like and, in molecular systems, usually expressed using exact spherical harmonic angular parts and a radial part. The latter is often composed of as a sum of Gaussian primitives to form Gaussian type orbitals, (GTOs), which approximate the correct exponential decay in electron-nucleus distance (Figure 2.1.1). GTOs are preferred over Slater type functions (STOs, $\exp(-r)$) as the resulting integrals (overlap, kinetic, 2-electron) are analytic and therefore much faster to evaluate.

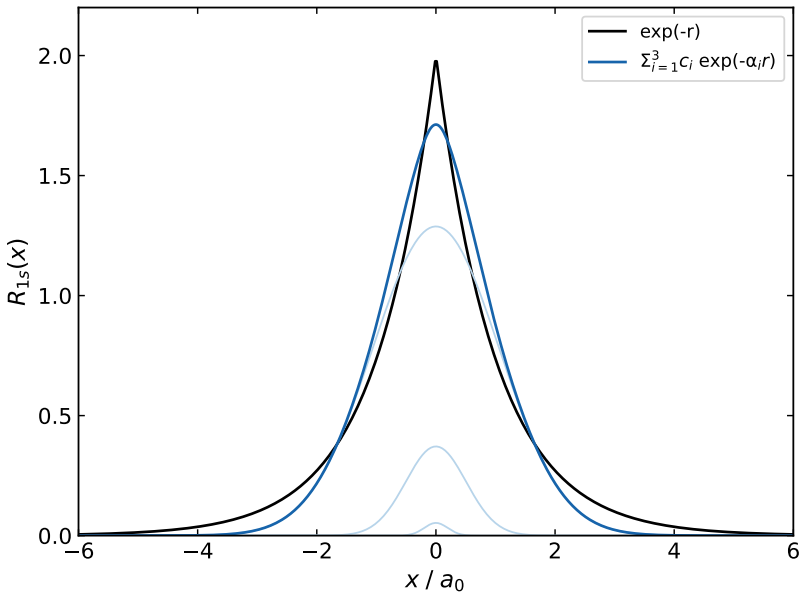


Figure 2.1.1. Exact (black line) and approximate (blue line) radial part of the hydrogen 1s orbital. Exponents and coefficients of primitive GTOs obtained from the def2-SVP basis set.[108, 109]

For multielectron species a better approximation to the true AOs can be obtained using a

combination of several STOs, themselves comprised of Gaussian primitives, as

$$\phi_l^{\mathrm{AO}}(\boldsymbol{x}_1) = Y(\theta,\phi)[k_1 r^l \exp(-\zeta_1 r) + k_2 r^l \exp(-\zeta_2 r) + \cdots + k_n r^l \exp(-\zeta_n r)] \tag{2.14}$$

for a n-ζ basis set with angular quantum number l. High kinetic energy close to nuclei means the curvature is important to the energy so several primitives are often combined for a single-ζ, and fewer with multiple$-\zeta$ for valance electrons to form a 'split-valance' basis set.

With enough GTOs the AOs can be well approximated, but at a higher computational cost. Therefore, there is a balance between acceptable accuracy and cost in each calculation. DFT has the advantage of faster basis set convergence,[110] with acceptable forces using a double-ζ quality basis and energies at triple-ζ.[111] WF methods are slower to converge with at least a triple-ζ basis set required for accurate energetics, with a CCSD(T) approximation energy differences are often within 'chemical accuracy'.[112] Therefore, the split valance def2-SVP basis set will be used for geometry optimisations and the def2-TZVP (and variants thereof) for single point energy evaluations using both DFT and CC methods. These choices have been validated previously.[113]

Two types of error emerge using finite basis sets: basis set incompleteness error (BSIE) and basis set superposition error (BSSE). The former is intrinsic for any calculation but particularly prominent when, for example, calculating anion energies without slowly decaying radial components (diffuse basis functions). The latter appears only when energy differences for association/dissociation processes are calculated, where one molecule can 'steal' basis functions from another (often) lowering the energy more than would be expected at the basis set limit, thus artificially stabilising the complex.[e][114]

[e]Note that the distinction is somewhat arbitrary but BSIE is often taken to be the residual error to the complete basis set value once the BSSE has been corrected with a scheme. Both errors tend to zero at an infinitely large basis set.

The basis set size required for calculations of synthetically interesting molecules (10s-100s atoms) makes the formal scaling of DFT $O(N^3$–$N^4)$ and *ab–initio* $O(N^{>4})$ too severe. Asymptotically, the scaling is reduced to $O(N^2)$ as the overlap between significantly separated AOs becomes negligible.[115] A resolution of the identity (RI) approximation builds on this to reduce the scaling to $O(N^2)$ for all system sizes by approximating two-electron integrals in an auxiliary basis.[116] However, as most methods rely on a matrix diagonalisation, the scaling is only reduced to $O(N^3)$ if conventional (non-sparse) diagonalisation routines are employed. Multiple algorithms for implementing RI are available, with the chain-of-spheres (COS) for exact exchange terms being particularly efficient for large molecules.[407]

CCSD(T) calculations on large systems requires accelerations beyond RI in the Hartree-Fock step to combat the prohibitive $O(N^7)$ scaling. Most of these rely on the locality of electron correlation (i.e. local coupled cluster methods, LCC) to truncate the number of amplitudes included in the CC WF.[117] Of particular note is the domain-based local pair natural orbital DLPNO method, which uses this truncation to approach a linear $[O(N^1)]$ scaling method for CCSD,[118] and perturbative triples.[119]

2.1.6 Periodic Electronic Structure

Molecular finite size calculations allow gas phase or implicit-environment energies to be calculated. However, to include the environment explicitly without unphysical boundary effects would make the system size unmanageable. Periodic electronic structure calculations allow for a quasi-infinite system with repeated periodic images using DFT most commonly.[120] Periodic DFT usually makes use of a 'plane wave' basis set,

$$\phi_{\boldsymbol{k}}^{\mathrm{PW}}(\boldsymbol{r}_1) = \sum_{i,j} c_{i,j} \exp(\mathrm{i}(\boldsymbol{k}_i + \boldsymbol{G}_j) \cdot \boldsymbol{r}_1) \tag{2.15}$$

as a complex Fourier series, which satisfies translational symmetry, and is orthogonal. Here, i is the imaginary unit, $\boldsymbol{k}$ a point in reciprocal space and $\boldsymbol{G}_i$ a quantised reciprocal lattice vector with $|\boldsymbol{k}_i + \boldsymbol{G}_i|^2/2 < E_{\text{cut}}$ for a cut-off PW energy of E_{cut}.[121] For sufficiently large super cells sampling just at the origin in k-space (the Γ-point) is sufficient.[122] To overcome the intrinsically shallow curvature of plane waves a large E_{cut} is required along with treating core electrons, with high kinetic energy, using pseudopotentials (PP). These PPs seek to replace the nuclear charge and thus diverging Coulomb attraction and the tightly-bound electrons with an effective ionic potential and are essential in periodic calculations. A similar strategy is employed in molecular systems for heavy elements (greater than period 4), where relativistic effects need to be included and are done so implicitly in effective core potentials (ECPs).

2.1.7 Semi-empirical Methods

For systems of more than a few hundred atoms the computational demands of DFT preclude its routine use. Semi-empirical methods on the other hand, sacrifice accuracy for speed and the ability to easily perform calculations with hundreds or thousands of atoms.[123] HF-derived methods including the PMx (parametrisation method number x) family rely on several approximations:

1. Only valance orbitals are included in the calculation (i.e. ECPs are employed for period 2 elements and above).

2. Core-core repulsion is parametrised rather than the true Columb interaction between the nuclei.

3. Two index electron repulsion integrals are parametrised.

4. Atomic orbitals on different atoms are assumed to have no overlap.

The most recent PM7 method makes use of the modified neglect of diatomic overlap (MNDO) method in which (1) two-electron integrals are approximated by a classical expansion of multipole moments (2) a dispersion correction is included and (3) adds a Gaussian repulsion between cores.

Tight binding (TB) DFT, in contrast, makes use of a Taylor expansion in the density and a minimal basis set to arrive at a general expression,

$$E_{\text{TB-DFT}} = \sum_a f_a \sum_{i,j} c_i^a c_j^a H_{ij} + \sum_{A>B} \gamma_{AB}(R_{AB}) \Delta q_A \Delta q_B + \sum_{A>B} V_{\text{rep}}(R_{AB}) \tag{2.16}$$

where f_a is the occupation number of MO a, c_i is an atomic orbital coefficient, H is the parametrised Hamiltonian matrix, q are partial charges and V_{rep} is a repulsive function depending on internuclear distances. This simple functional form for the energy is fast to evaluate but is highly parametrised, thus can be non-transferable.[124]

Recently, Grimme and co-workers have developed a variant of DFTB denoted GFNx-XTB,[125] parametrised on elements up to radon, which has been successful in a variety of contexts including molecular geometries and non-covalent interactions. The GFN2-xTB method uses a minimal basis with polarisation without pair-specific parameters, making it more applicable than standard DFTB.

These methods are, however, not without their limitations. For example, employing a minimal basis set generally leads to exaggerated anion instability and limited polarisation.[126] As such, quantitative predictions from derived energy differences are generally not possible.

2.1.8 Implicit Solvent Models

Without modification molecular electronic structure calculations are performed in an effective vacuum. Chemical reactions, however, are generally conducted in the condensed phase, e.g. solution, melt or on surface.[127] For non-polar reactions in apolar solution the solvent effect may be small,[128] but for highly charged species the stabilisation of charge separation by the solvent is critical to obtaining even modest accuracy (error < 10 kcal mol^{-1}).[129]

The most efficient method of including solvent effects is to model them implicitly, wherein the electronic Hamiltonian is modified to include effective point charges describing the solvent plus other contributions. Their aim is to calculate the solvation energy by splitting the free energy into electrostatic, cavitation, dispersion and repulsion terms as,[130]

$$\Delta G_{\text{solv}}(\text{solute}) = \Delta G_{\text{ele}} + \Delta G_{\text{cav}} + \Delta G_{\text{disp}} + \Delta G_{\text{rep}}. \tag{2.17}$$

In PCM and derivatives thereof (including CPCM) the electrostatic contribution involves interactions of point charges on the tessellated solute surface with a continuum model of the solvent (depending on its dielectric). The cavitation term depends on the spherical size of the solute and the number density of the solvent, the repulsion just on the surface, while the dispersion uses an effective solute density matrix – as in the SMD solvent model[131] – or set of point charges.[132]

2.1.9 Analysis Methods

Once a calculation has been performed and an energy (difference) obtained, analysis methods can be used to try rationalise a experimental or computational energy difference. These are based on the density or molecular orbitals of the system with those pertinent to this thesis

summarised in Figure 2.1.2 for a prototypical S_N2 reaction.

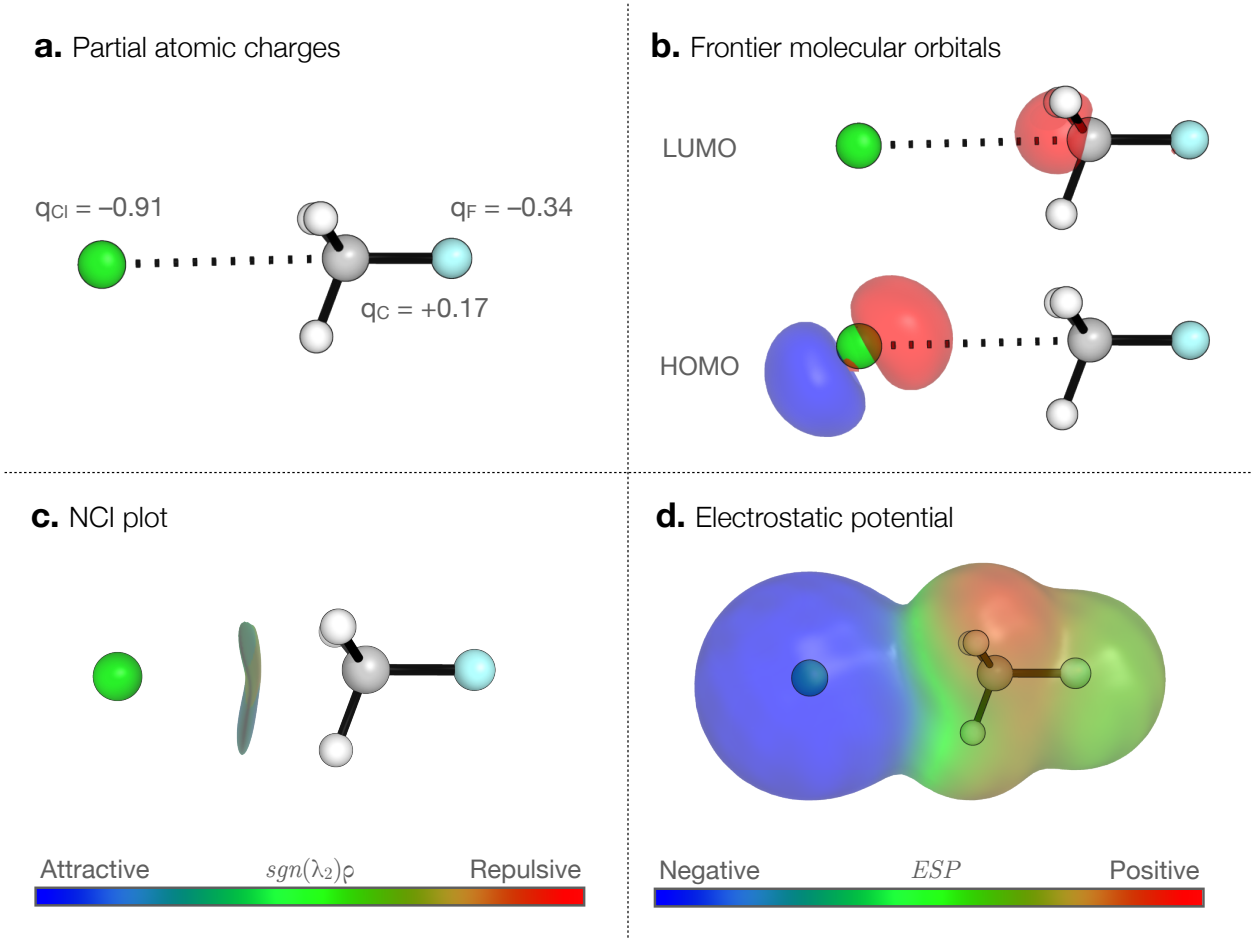


Figure 2.1.2. Electron density and orbital analysis methods applied to an S_N2 association complex $Cl^-.CH_3F$. DFT (SMD(water)-PBE0-D3BJ/def2-SVP) orbitals and corresponding density used in each case.

2.1.9.1 Partial Atomic Charges

Partial atomic charges may be calculated from a set of orbitals or a density and can provide useful chemical insights. For example, the atomic charges within the $Cl^-.CH_3F$ association complex (Figure 2.1.2a) suggest the flow of electron density from Cl^- to C in the reaction.

As they are not observable quantities many schemes exist; the most simple being the Mulliken

scheme, where an atomic charge on an atom A is defined as,[133]

$$q_A = Z_A - \sum_{\mu \in A} (\boldsymbol{PS})_{\mu\mu} \tag{2.18}$$

where $\boldsymbol{P}$ and $\boldsymbol{S}$ are density and overlap matrices respectively (in the AO basis) and the sum is over orbitals centred on atom A. However, the basis set dependence of Mulliken partial charges is significant,[134] and so have been superseded by other schemes. Hirshfeld charges are derived from the density and converge as the basis set increased in particular,[135]

$$q_A = Z_a - \int w_A(\boldsymbol{x}) \rho^{\text{mol}}(\boldsymbol{x}) - \rho_A^{\text{atom}}(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \tag{2.19}$$

where w_A is a weighting function defined by $w_A = \rho_A^{\text{atom}} / \sum_B \rho_B^{\text{atom}}$ and ρ_A^{atom} is the isolated atomic density of atom A. Although sensitive to the choice of reference atomic density, the trends between systems are generally considered to be robust compared to basis-set dependent schemes.[136]

2.1.9.2 Frontier Molecular Orbital Theory

While not physical observables, canonical frontier orbitals – those close in energy to the first unoccupied – can be predict where a reaction will occur.[137–139] Interaction between the highest occupied MO (HOMO) and lowest unoccupied MO (LUMO) can lead to a stabilising interaction,[140] and used to rationalise attractive interactions. For example, types of halogen bonding (B→X–E, where B is a base, X a halogen and E another element) may be attributed to a $n \rightarrow \sigma^*$ interaction, where high-lying filled MO has lone pair (n)-like character and a low-lying unoccupied orbital is a σ^* orbital.[141] Pictorially, Figure 2.1.2b shows the HOMO of $Cl^-.CH_3F$ as a chloride lone pair that can 'donate into' the LUMO as

a σ^* of the C–F bond, thus suggesting the classic backside attack in the S_N2 reaction.

2.1.9.3 Non-covalent Interactions

The identification of non-covalent interactions (NCIs) may be located quantitatively using the reduced density gradient (RDG),[142]

$$\mathrm{RDG}(\boldsymbol{x}) \propto \frac{|\nabla\rho(\boldsymbol{x})|}{\rho(\boldsymbol{x})^{4/3}} \tag{2.20}$$

which is the normalised magnitude of the density gradient. This definition affords moderate $\mathrm{RDG}(\boldsymbol{x})$ with low densities but smaller values at higher density, such as within a chemical bond. Attractive interactions (e.g. H-bonds) have $\mathrm{sgn}(\lambda_2)\rho(\boldsymbol{x}) < 0$, while repulsive (steric clash) have $\mathrm{sgn}(\lambda_2)\rho(\boldsymbol{x}) > 0$, where λ_2 is the second largest eigenvalue of the Hessian matrix of the density and measures the charge accumulation in the perpendicular plane of the interaction.[143] In the $Cl^-.CH_3F$ association complex there is a moderate attractive interaction between the chloride and the slightly positively charged carbon (Figure 2.1.2c), as would be expected.

2.1.9.4 Electrostatic Potentials

An electrostatic potential (ESP, $\Phi(\boldsymbol{x})$) is defined as the work required to add a unit positive charge at a point $\boldsymbol{x}$. In a molecule,

$$\Phi(\boldsymbol{x}) = \sum_A \frac{Z_A}{|\boldsymbol{R}_A - \boldsymbol{x}|} - \int \frac{\rho(\boldsymbol{x'})}{|\boldsymbol{x'} - \boldsymbol{x}|}\mathrm{d}\boldsymbol{x}' \approx \sum_A \frac{q_A}{|\boldsymbol{R}_A - \boldsymbol{x}|} \tag{2.21}$$

where the sum runs over nuclei A. Plotting the ESP over a molecular surface (e.g. Van der

Walls, using some radii) allows for the rapid identification of areas of charge accumulation and where, for example, hydrogen bonds are likely to form.[144] For example, the ESP of $Cl^{-}.CH_3F$ over a density isosurface is shown in Figure 2.1.2d, clearly showing the location of the negative charge 'located' on Cl^{-}.

2.2 Molecular Dynamics

2.2.1 Preamble

So far the methods outlined above have focused on calculating energies and analysing compounds at zero temperature at a single geometry. However, sampling of a system's potential energy surface (PES, $E(\boldsymbol{R})$) is essential to obtain observable properties. In the most simple case, the sampling corresponds to a geometry optimisation; where from a single point, optimisation affords a stationary point on the hypersurface at which energies may be calculated.

Sampling more than just a set of points en-route to a stationary point is sometimes essential. Most molecules can exist in multiple conformations, disregarding some of which may lead to incorrect reaction energetics.[145] Moreover, considering only stationary points can be a severe approximation, as chemistry is conduced at non-zero temperatures where nuclei can move. To calculate free energies and sample from the correct ensembles requires dynamics. For example, in free energy perturbation (FEP) theory the difference is given by the Zwanzig equation,

$$\Delta F_{\mathrm{X}\to\mathrm{Y}} = -RT \left\langle \exp\left(-\frac{E_{\mathrm{Y}} - E_{\mathrm{X}}}{RT}\right)\right\rangle_{\mathrm{X}} \tag{2.22}$$

where R is the gas constant, T the temperature, F a free energy (Helmholtz or Gibbs) and the angle brackets indicate an ensemble average over state X. A Helmholtz free energy (A)

is defined within an NVT ensemble where the number of particles (N), volume (V) and temperature is constant, while the Gibbs free energy replaces constant volume with pressure NPT ensemble.

The average in (2.22) may be computed using molecular dynamics (MD), or Monte Carlo (MC, e.g. Metropolis Monte Carlo) both of which *can* sample from the correct ensembles.

2.2.2 Integration Algorithms

MD simulations require propagating equations of motion forwards using the energy and curvature of the PES. This may be using quantum mechanics, but is prohibitively expensive for large molecules ($\gtrsim$10 atoms).[146] Classical MD is, therefore, the choice method for sampling, where Newton's equations of motions ($\boldsymbol{F} = \boldsymbol{ma}$) are integrated with, for example, a velocity Verlet algorithm for updated positions and velocities,

$$\begin{aligned} \boldsymbol{R}(t + \Delta t) &= \boldsymbol{R}(t) + \boldsymbol{v}(t)\Delta t + \frac{\boldsymbol{F}(t)}{2\boldsymbol{m}}(\Delta t)^2 \\ \boldsymbol{v}(t + \Delta t) &= \boldsymbol{v}(t) + \frac{\boldsymbol{F}(t) + \boldsymbol{F}(t + \Delta t)}{2\boldsymbol{m}}\Delta t \end{aligned} \tag{2.23}$$

for a collection of nuclear coordinates $\boldsymbol{R}$ with velocities $\boldsymbol{v}$, masses $\boldsymbol{m}$ and forces $\boldsymbol{F}$. This just corresponds to a Taylor expansion in the position and velocities, and the velocity update uses the average of the force between the current step and the next.

Iterating these equations requires a time-step (Δt) small enough for the expansion to be accurate but large enough to be computationally efficient. In practice, this is chosen to be $\sim$ 1 fs and corresponds to the fastest motion (bond stretching), while fixing E-H bonds allows for a larger 2 fs time-step.[147]

In a closed NVE ensemble the energy is conserved, thus the temperature $T \sim 2T_{\mathrm{KE}}/3Nk_{\mathrm{B}}$

(for N particles) fluctuates as the system accesses different regions of the PES. To simulate at constant temperature requires thermostats.

2.2.3 Thermostats

Constant simulation temperature can be achieved using stochastic; strong or weak coupling methods; additional degrees of freedom to the system can also be added. Langevin dynamics makes use of a velocity damping along with stochastic Gaussian random noise with an equation of motion,

$$\ddot{\boldsymbol{R}}(t) = \frac{\boldsymbol{F}}{\boldsymbol{m}} - \gamma \boldsymbol{v}(t) + \sqrt{\frac{2\gamma k_{\mathrm{B}} T}{\boldsymbol{m}}}\vartheta(t) \tag{2.24}$$

where γ is a damping parameter and $\vartheta(t)$ a stationary distribution with zero mean and zero covariance with any other time. From the strong coupling thermostats the velocity rescale thermostat in its most simple form multiplies all the velocities by a factor $\alpha = \sqrt{\bar{T}_{\mathrm{KE}}/T_{\mathrm{KE}}}$, where $\bar{T}_{\mathrm{KE}}$ is the desired kinetic energy given a temperature.[148] As most condensed phase molecular systems undergo little volume change through reactions etc. constant pressure simulations (NPT) will not be discussed; likewise for unconstrained particle number simulations.

2.2.4 Empirical Force Fields

With equations of motion to propagate for sampling within an ensemble depending on forces, a method of calculating them efficiently is required. While QM methods (DFT, MP2 etc.) can be used to calculate forces, often many thousands of gradient evaluations are required and can be too computationally demanding. Collectively known as *ab initio* molecular dynamics (AIMD) is generally only suitable for sampling on picosecond timescales.[149] To calculate

forces efficiently and access timescales of nano or microseconds requires a parametrisation of the PES. Empirical force-fields for molecular systems rely on a bonded + non-bonded decomposition of the energy,

$$E_{\text{FF}} = E_{\text{bonded}} + E_{\text{non-bonded}} \tag{2.25}$$

where the respective terms usually have the form,

$$E_{\text{bonded}} = \underbrace{\sum_{i,j}^{\text{bonds}} k_{ij}(r_{ij} - r_{ij}^0)^2}_{\text{bond stretching}} + \underbrace{\sum_{i,j,k}^{\text{angles}} k_{ijk}^{\theta}(\theta_{ijk} - \theta_{ijk}^0)^2}_{\text{angle bending}} + \underbrace{\sum_{m}^{\text{dihedrals}} k_m(1 + \cos(n^m \chi_m - \mu^m))}_{\text{dihedral rotation}} \tag{2.26}$$

for distances (r), angles (θ) and dihedrals (χ)[f] where superscript 0 denotes an equilibrium value while i, j, k enumerate over atoms m over dihedrals, n is an integer while μ is a shift factor. The non-bonded component often takes the form of a Lennard-Jones term for dispersion and repulsion, and an electrostatic monopole interaction,

$$E_{\text{non-bonded}} = \sum_{i,j}^{\text{pairs}} \underbrace{\epsilon_{ij}\left(\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma'_{ij}}{r_{ij}}\right)^{6}\right)}_{\text{LJ 12-6 potential}} + \underbrace{\frac{q_i q_j}{r_{ij}}}_{\text{electrostatic}} \tag{2.27}$$

where ϵ is proportional to the well depth and q a partial atomic charge.[150] In combination with neighbour lists the complexity scales as roughly $O(N)$ for large systems and, because only simple mathematical operations are required, the execution is rapid. While minimal, this type of description is employed in almost all classical biomolecular MD simulations.[151]

This description of the PES is, however, not without its limitations, particularly important are: (1) bonds are harmonic and cannot break, making reactive events impossible to

[f]Improper dihedrals between four non-contiguous atoms employ a harmonic term but are omitted for clarity.

study; (2) static charges are used with no polarisation/charge transfer possible; (3) dispersion is not pairwise additive; (4) closed shell repulsion is exponential in distance not r^{-12}; (5) angle potentials are fully harmonic and cannot have two minima (e.g. trigonal bi-pyramidal and square planar metal complexes); (6) fixed atom types are used to define ϵ, σ for a pair of atoms using combination rules, which are an approximation to the optimum values.[152] Therefore, employing these empirical potentials requires consideration of the resulting (somewhat unknowable) errors, potential lack of transferability and may not facilitate in quantitative comparisons.[153, 154]

2.3 Machine Learned Potentials

2.3.1 Preamble

Machine learning concerns the development and study of computer algorithms that improve with data.[155] Categorised into supervised and unsupervised flavours depending on whether the target values are known or not, ML has been widely adopted in the sciences since its inception in the 1960s.[156] Within the chemical domain applications include advanced quantitative structure activity relationships (QSAR) and interatomic potentials (ML force-fields), amongst many others.[157] An ML or 'data-driven' approach to a potential may be written as,

$$E(\boldsymbol{R}) = f(\boldsymbol{R}; \boldsymbol{\alpha}) \tag{2.28}$$

where the potential energy E is given by a function (the ML potential) of the systems' nuclear coordinates and a set of parameters ($\boldsymbol{\alpha}$). The parameters are obtained using a set of prior observations $\{\boldsymbol{R}_{\text{train}}\}$ and output labels $\{E_{\text{train}}\}$ with, perhaps, forces ($\{-\partial E_{\text{train}}/\partial R_i\}$). Exhaustively generating points in the space is, however, impossible for more than a few

atoms as $M^{3N_{\text{atoms}}}$ points are needed for a grid over the whole space with M points in each dimension. To somewhat aliviate the exponential scaling, generally the total energy is generally written as a sum of atomic contributions as,

$$E(\boldsymbol{R}) = \sum_{i}^{\text{atoms}} f_i(\boldsymbol{R}_i; \boldsymbol{R}_{\neq i}, \tilde{\boldsymbol{\alpha}}) \tag{2.29}$$

where now the atomic energy given by f_i depends on its position ($\boldsymbol{R}_i$), and of those atoms around it (within some cut-off) along with a set of parameters. With this atomic expansion of the energy, which is assumed to be possible, the space to fit is much smaller if the cut-off does not include the whole system and symmetries in the system can be exploited.[g]

2.3.2 Representations

To minimise the quantity of data required to obtain a ML potential at some accuracy requires encoding all the relevant symmetries into a 'representation' i.e. $\boldsymbol{R} \mapsto \boldsymbol{R}_F$ for a feature vector $\boldsymbol{R}_F$. The vector $\boldsymbol{R}_F$ should be invariant to translation, rotation and permutation of like nuclei, as the Hamiltonian is invariant to all of those operations.

The initial work of Behler and Parrinello showed that PES fitting using ML approaches was possible using artificial neural networks (ANNs, or NNs) and employed atomic-centred symmetry functions (ACSF).[158, 159] While many possible radial and angular functions were proposed, the most frequently employed is a set of shifted Gaussian functions as,

$$G_i^{R_s,\eta}(\boldsymbol{R}) = \sum_{j}^{\text{neighbours}} \exp\left[-\eta(R_{ij} - R_s)^2\right] f_c(R_{ij}) \tag{2.30}$$

[g]For example, two well separated water molecules,are essentially non-interacting thus the energy is a simple sum of the two components.

where $R_{ij} = |\boldsymbol{R}_i - \boldsymbol{R}_j|$, η, R_s are parameters and $f_c(R_{ij}) = 0.5[\cos(\pi R_{ij}/R_c)+1]$ for $R_{ij} < R_c$ and zero otherwise, given a cut-off distance R_c. The cut-off function (f_c) ensures a zero output outside of the cut-off region so that the derivatives are continuous. A similar function to Eqn. (2.30) is defined for angular components using a cosine function damped with Gaussians and cut-off functions. Concatenating elements of radial and angular functions with varying parameters affords $\boldsymbol{R}_F$. These functions and their variants have been used extensively in bespoke and general (e.g. ANI[160]) machine learnt potentials.

Alternatively, an atomic environment may be expressed using an effective atomic density and $\boldsymbol{R}_F$ constructed from the set of expansion coefficients of that density, as in the smooth overlap of atomic positions (SOAP) descriptor.[161] Specifically, for an atomic neighbour density,

$$\rho_i(\boldsymbol{x}) = \sum_{j}^{\text{neighbours}} w_i \delta(\boldsymbol{x} - \boldsymbol{R}_j) \tag{2.31}$$

with a set of weights w_i, $\boldsymbol{x}$ is a position in space centred at $\boldsymbol{R}_i$, and δ a Dirac delta centred on each neighbouring atom. To ensure the descriptor is smooth under coordinate perturbations Gaussians are used in place of the delta functions and the density expanded in a radial (g) and spherical harmonic (Y) basis as,

$$\rho_i(\boldsymbol{x}) = \sum_{j}^{\text{neighbours}} \exp(-\sigma_{\text{atom}}|\boldsymbol{x} - \boldsymbol{R}_j|^2) = \sum_{nlm} c_{nml} g_n(r) Y_{lm}(\theta, \phi) \tag{2.32}$$

where σ_{atom} controls the atomic density extent. See Appendix 8.1.1 for a derivation of the coefficients. The SOAP vector is then defined from the power spectrum elements as

$$\boldsymbol{p}_i = \{p_{nn'l}\} \qquad ; \qquad p_{nn'l} = \frac{1}{\sqrt{2l+1}} \sum_{m=-l}^{l} c^*_{nlm} c_{n'lm} \tag{2.33}$$

to afford a rotationally-invariant description of the neighbour density of atom i, schematically shown in Figure 2.3.1.

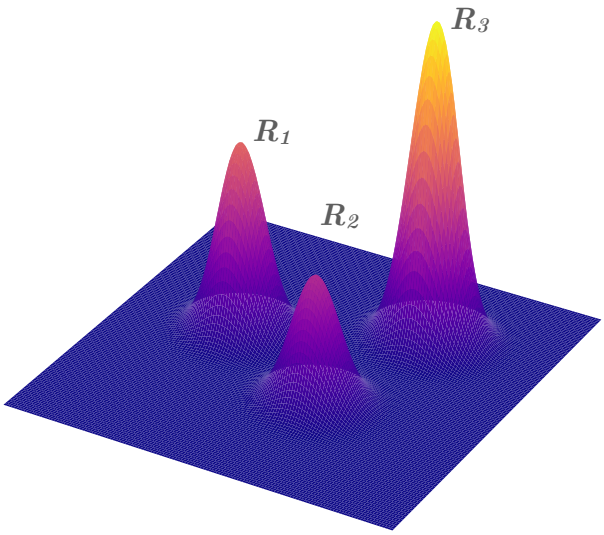


Figure 2.3.1. Smooth overlap of atomic positions density of three neighbours with differing weights.

Using the default parametrisations both ACSFs and the SOAP power spectrum seem to perform similarly at representing atomic environments.[162, 163] The choice of atomic description amongst the wide variety (also including Chebyshev polynomials[164], many-body tensor representation[165], permutationally invariant polynomials[166, 167], amongst others) is then just dependent on the implementation and computational efficiency.

2.3.3 Regression Algorithms

From a compact and representative $\boldsymbol{R}_F$ an atomic energy prediction is made using a regression algorithm. Of the wide variety (linear, polynomial, support vector machine (SVM), decision trees etc.[168]) ANNs and Gaussian process regression (GPR) are the most widely employed,[169] and so are briefly outlined below.

A simple 'feed-forward' neural network consists of a set of layers, each containing nodes which transform an input x as,

$$y = \sigma(wx + b) \tag{2.34}$$

where w is a weight and b a bias, and σ an activation function e.g. $\sigma = \tanh$ is generally introduced to apply a non-linear transform enabling a ANN to approximate any function.[170] Introducing n layers each with multiple nodes the output is,

$$y = \sigma_n \circ \cdots \circ \sigma_1(\boldsymbol{w}_1 \boldsymbol{X} + \boldsymbol{b}_1) \tag{2.35}$$

where at each layer the input $\boldsymbol{X} \in \mathbb{R}^{N_{\text{in}}}$, $\boldsymbol{b} \in \mathbb{R}^{N_{\text{out}}}$ and $\boldsymbol{w} \in \mathbb{R}^{N_{\text{out}} \times N_{\text{in}}}$ is a weight matrix with $\circ$ denoting function composition. The weights and biases are then optimised by minimising a smooth loss function $[L(y, y_{\text{target}})]$ with respect to the parameters over many target values within a 'batch'.

Gaussian process regression starts by defining the output in a linear way,

$$y = \sum_i w_i \phi_i(\boldsymbol{X}) \tag{2.36}$$

where again w are weights but now ϕ defines a basis function. By using weights initialised from Gaussians with zero mean an variance, a new value can be predicted at its mean,[171, 172]

$$y^* = \boldsymbol{k}^T(\boldsymbol{X}^*) K^{-1} \boldsymbol{y} \tag{2.37}$$

where $\boldsymbol{k}^T(\boldsymbol{X}^*) = (k(\boldsymbol{X}_1, \boldsymbol{X}^*), ..)$ and K is the kernel matrix with elements $K_{ij} = k(\boldsymbol{X}_i, \boldsymbol{X}_j)$, and $\boldsymbol{y}$ a column vector of target values for each $\boldsymbol{X}$. In the context of a machine learnt potential i.e. a Gaussian approximation potential (GAP) the observations are total energies[h]

[h] Atomic forces can also be used as additional labels and, almost always, improve the fit.

and $\boldsymbol{X} = \boldsymbol{R}_F$ with the canonical SOAP kernel having elements

$$k(\boldsymbol{X}_i, \boldsymbol{X}_j) = \sigma_w^2 \int_{\text{space}} \rho_i(\boldsymbol{x})\rho_j(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} = \sum_{n,n',l} p^{(i)}_{nn'l} p^{(j)}_{nn'l} \tag{2.38}$$

which just measures the similarity between two atomic neighbour densities, such that similar environments have kernel values close to one and, thus, likely similar energies. GAP implementations may also include two and three-body components such that the total energy is then a sum of terms, often employing a radial basis function kernel,

$$k_{\text{RBF}}(\boldsymbol{X}_i, \boldsymbol{X}_j) = \exp\left(-\frac{|\boldsymbol{X}_i - \boldsymbol{X}_j|}{2\theta^2}\right) \tag{2.39}$$

where θ controls the interpolation range and $\boldsymbol{X}_i = d_i$ is the pairwise distance in a two-body descriptor.

With generally fewer parameters to define a GAP than an ANN they are the choice method for sparse datasets, but the evaluation scales (without other approximations) as $O(N^3)$ for N training points due to the matrix inversion.

3 Metallocage Catalysis

3.1 Overview

Arguably, biological catalysts – referred to as enzymes – are the best catalysts known. They operate in safe and abundant solvent (water), are highly active at room temperature or close to it and are highly selective. While *de-novo* enzyme design is a promising and active area of research, laborious directed-evolution – random mutation of amino acids and selection – is generally required.[173] Metallocages (Figure 3.1.1) offer a promising alternative, with defined cavities, broadly predictable self-assembly and, therefore, easily tunable micro-environments, they provide a simpler approach to achieve enzyme-like catalytic properties.[174]

Understanding how these metallocages can act as catalysts is thus an important step to rationally design catalytic metallocages with improved activity and selectivity. Manuscript I focuses on understanding how an M_2L_4 construct catalyses a broad class of Diels-Alder reactions as shown experimentally by Lusby and co-workers in 2018.[175] The cages' flexibility, their cationic charge positioning, and ability to accommodate a guest are found to afford catalytic activity. En-route to these findings, the absolute and relative binding affinities and the catalytic ranking for a set of cages and substrates were all accurately reproduced by the

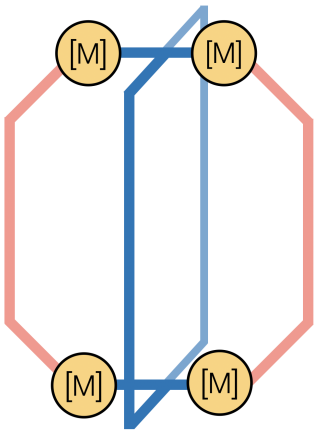


Figure 3.1.1. Example metallocage structure with an $M_4L_1L'_1$ topology, where [M] is a metal fragment e.g. Au(I) or PdX_2.

computational approaches. These results suggest that more active catalysts could be developed by (1) increasing the effective charge of the metal sites; (2) reducing the steric clash between the substrate and cage upon translation to the exterior of the cavity and perhaps, (3) developing a cage large enough to accommodate both substrates (while still maintaining low product affinity, thus inhibition).

With an understanding of the catalytic mechanism for the metallocage a strategy to propose new catalysts is required. One such approach is to use chemical intuition and invest resources in setting up and analysing the calculations to minimise both the transition state barrier from the bound substrate and the binding energy of the product. An alternative approach would be to employ high-throughput virtual screening (HVTS). Towards this goal, Manuscript II proposes a method and presents command-line (*cgbind* Python module) and graphical (`http://cgbind.chem.ox.ac.uk/` web app) implementations.

The graphical web-app was developed as a simple user-facing front-end to the Python module which currently provides only a subset of features but, nevertheless, is used for rapid

visualisation of potential new metallocages for five architectures.[a] Behind the scenes, *cgbind* attempts to construct a metallocage by efficiently searching the space of donor atoms and conformations for possible metallocages. Distinct from other approaches[176] no hard-coded templates are defined but are rather automatically extracted from crystal structures. The Python module has been used within the Duarte group for a year or so for constructing, analysing and has started to be used to perform cage screening for catalytic proficiency.

[a]Private correspondence with the Lusby (University of Edinburgh), Craig (University of Strathclyde) and Zysman-Colman (University of St. Andrews) groups.

3.2 Manuscript I

Rationalizing the Activity of an "Artificial Diels-Alderase": Establishing Efficient and Accurate Protocols for Calculating Supramolecular Catalysis

3.2.1 Abstract

Self-assembled cages have emerged as novel platforms to explore bio-inspired catalysis. While many different size and shape supramolecular structures are now readily accessible, only a few are known to accelerate chemical reactions under sub-stoichiometric conditions. These limited examples point to a poor understanding of cage catalysis in general, limiting the ability to design new systems. Here we show that a simple and efficient density functional theory-based methodology, informed by explicitly solvated molecular dynamics and coupled cluster calculations is sufficient to accurately reproduce experimental guest binding affinities (MAD = 1.9 kcal mol^{-1}) and identify the catalytic Diels-Alder proficiencies (>80 % accuracy) of two homologous Pd_2L_4 metallocages with a variety of substrates. This analysis reveals how subtle structural differences in the cage framework affect binding and catalysis. These effects manifest in a smaller distortion and more favourable interaction energy for the catalytic cage compared to the inactive structure. This study gives a detailed insight that would otherwise be difficult to obtain from experiments, providing new opportunities in the design of catalytically active supramolecular cages.

3.2.2 Introduction

Self-assembly of molecular building blocks is ubiquitous in nature. Examples of this appear in protein cages,[177, 178] enzyme complexes,[179] and bacterial nanocompartments.[180] The

remarkable properties of these structures have inspired the design of biomimetic systems that exhibit similar properties at a minimalistic scale. In this context, a large variety of self-assembled cages constructed from simple building blocks via metal–ligand interactions,[181, 182] hydrogen bonds,[183–185] and other non-covalent interactions[186] have been reported in recent decades.

While many cages have been used successfully for recognition and sensing, far fewer have mimicked the catalytic proficiency and selectivity observed in enzymes. More than twenty years have passed since Sanders raised the question regarding the scarcity of effective supramolecular catalysts.[60] Still, his words remain a timely reminder of both the challenges and opportunities within this field. To date, only a handful of self-assembled catalytic metallocages exist,[187] most notably the $[Ga_4L_6]^{12-}$ tetrahedron originally developed by Raymond.[188] Through a fruitful collaboration with Bergman and recently Toste, this water-soluble anionic assembly has been shown to catalyse a number of transformations, including aza-Cope[189, 190] and Prins rearrangements,[191, 192], the Nazarov cyclisation,[193] and hydrolysis of acid-labile compounds under basic conditions,[194] among others.[187] The prototypical self-assembled metallocages developed in Makoto Fujita's group have also been shown to catalyse Diels-Alder and Knoevenagel reactions,[195, 196] while other water soluble systems can promote different hydrolysis reactions.[197–199] Several trends have emerged from these investigations. First, substrate encapsulation occurs due to the hydrophobic effect. Second, many of the reactions produce water soluble products to avoid inhibition. Thirdly, acceleration is commonly a consequence of high effective concentration of substrates or functional groups, either through co-encapsulation, ion-pairing or constrictive binding, and/or coulombic interactions between the charged host and electrostatically matched intermediates.[200, 201] These approaches while effective in providing acceleration, in many cases suffer from product inhibition (turnover) or are specific to water-soluble cages. Despite notable examples of

cage-catalysis, there has been a distinct lack of complementary computational investigation, probably due to their large size and the presence of multiple metal centres. Indeed, only recently, the catalytic activity of the $[Ga_4L_6]^{12-}$ cage has been explored computationally by Head-Gordon and Ujaque, respectively.[202, 203] Other catalytic activity studies have almost exclusively focused on organic systems.[62, 63, 183, 200, 201] In this work, we focus on the $[Pd_2L_4]^{4+}$ architecture, which occupies a prominent place in supramolecular chemistry.[204–208] Originally reported by Steel and co-workers,[209] the $[Pd_2L_4]^{4+}$ topology is one of the simplest and most versatile architectures.[205, 207] To date, several homo- and heteroleptic variants (consisting of one or different types of ligand, respectively) have been reported by Hooley,[210] Shionoya,[211] Clever,[212–215] Crowley,[216, 217] and others,[218–220] with applications in molecular recognition,[220] drug delivery,[204] stabilization of reactive species,[221] and recently catalysis.[175] Despite their promising applications and scope for redesign, no computational investigation has been carried out to explore their catalytic power. The few prior computational studies on this system have only focused on structural[206] and spectroscopic analysis.[222]

In 2018, Lusby and co-workers evaluated the ability of simple $[Pd_2L_4]^{4+}$ capsules C-1 ($[Pd_2L_4^{CH}]^{4+}$) and C-2 $[Pd_2L_4^{N}]^{4+}$) to catalyse Diels-Alder (DA) reactions, using quinone substrates as dienophiles in dichloromethane (DCM) solvent at room temperature (Figure 3.2.1).[175] Unlike most other capsule-catalysed reaction, this method principally exploits enthalpic activation, where C–H hydrogen bond interactions were postulated to activate the dienophile. It was observed that while C-2 was catalytic (rate accelerations (k_{cat}/k_{uncat}) of up to 10^3), the homologous C-1 cage was inactive, despite C-1 binding quinones in the same mode. The contrasting catalytic ability was postulated to arise as a result of two factors: weakened substrate binding in C-2 due to a repulsive interaction in the ground-state between the nitrogen lone pair (n) and the π-bond of the guest, and stabilization of the TS

through the formation of favourable N···HC hydrogen bonds. The latter was inferred from the stronger binding of a TS mimic (the DA adduct of benzoquinone and cyclopentadiene) in C-2 compared to C-1.[175] However, the difference in catalysis has been purely rationalized on the basis of thermodynamic parameters, such that the precise origin of acceleration, and by extension lack of activity for C-1, are still poorly understood. This substantially limits our ability to design supramolecular catalysts for new chemical reactions.

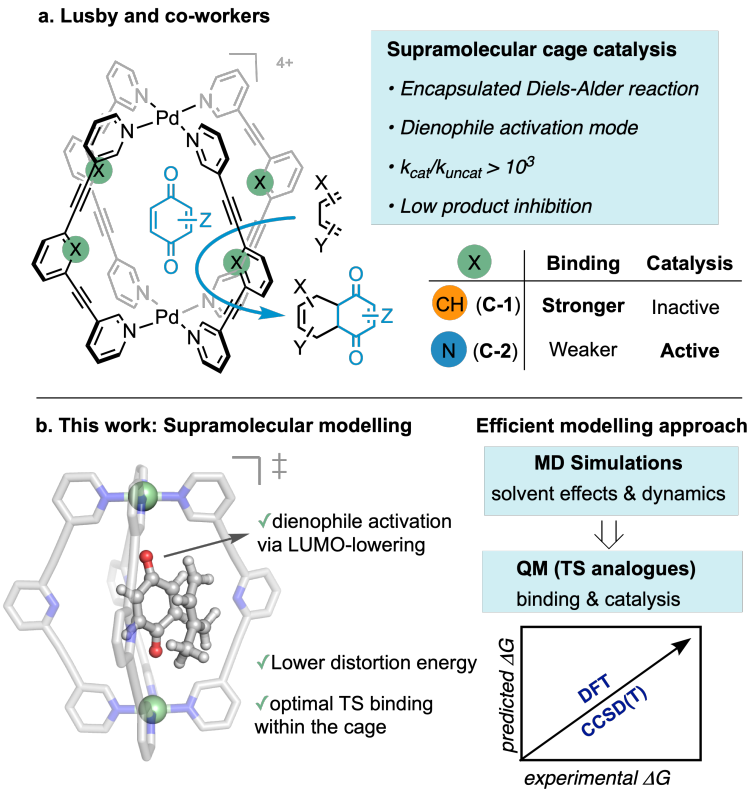


Figure 3.2.1. (a) Diels-Alder reaction studied by Lusby and co-workers employing the [Pd2L4]2+ capsules, C-1 and C-2. (b) Schematic representation of our computational approach.

Computational studies that could quantitatively rationalize such observations are highly desirable to optimize and design of novel catalytic cages. However, the ability to routinely study and predict binding and catalysis in solution remains an open challenge. This is particularly true for the latter, where both thermodynamic and kinetic aspects must be considered. In this context, this work aims to rationalize the observed differences in binding and catalysis

between metallocages C-1 and C-2, paving the way to an efficient computational protocol to understand related systems. To achieve these goals, we employ explicitly solvated molecular dynamics (MD) and density functional theory (DFT) methods, which are validated against experiments and coupled cluster [CCSD(T)] calculations. Using this protocol, we elucidate the effect of intermolecular interactions, solvent and structural flexibility on the guest binding, and provide molecular-level insight into the catalytic properties of these systems. Our approach provides an affordable route to explore novel metallocage designs as non-covalent catalysts for applications in synthetic methodology.

3.2.3 Results and Discussion

3.2.3.1 Dynamic Properties

The self-assembled structures of C-1 and C-2 have been assumed to be highly rigid, even though they contain several potential rotatable σ-bonds (sixteen C-C bonds at either side of the acetylene group and eight Pd-pyridine coordination interactions).[a] To examine the flexibility of these systems, we performed explicitly solvated (DCM) molecular dynamics (MD) simulations in the presence and absence of a quinone guest. A modified version of the Pd^{2+} dummy-model – where the square planar geometry is obtained by adding four dummy atoms each with +0.5 e charge around the Pd centre – was used (Table 3.3.1).[223] This model provided Pd–N distances in the cages C-1 (avg. = 1.95 Å) and C-2 (avg. = 1.95 Å) similar to those found for Pd–pyridyl containing cages in the Cambridge Structural Database (2.02±0.01 Å)[204, 224] and also those obtained from DFT calculations (2.02 Å, Table 3.3.2). However, this model was found to substantially underestimate the Pd–O

[a]We have found that the barrier to rotation around the acetylene bond is negligible (0.4 and 1.1 kcal mol^{-1} for L^{CH} and L^{N} respectively, cf. ethane $\Delta E^{\ddagger}$ = 2.7 kcal mol^{-1} at the M06-2X/def2-TZVP level of theory, Figure 3.3.7).

distances in aqua-complexes (avg. 1.85 Å cf. expt.[225] 2.01 Å), a measurement that is usually used to parametrize metal complexes and for which extensive amount of experimental data is available (Figure 3.3.1).[226] This demonstrates the challenges of directly transferring these parameters between different solvent environments. It is also important to note that while standard soft-sphere models (with no dummy atoms) have been used to model aqua-complexes, it was found that they failed to maintain stable metallocage assemblies in DCM beyond the picosecond time-scale.

With suitable force field parameters to describe the metal and organic building blocks, we then assessed the flexibility of the two cages by monitoring several geometric parameters (Figures 3.2.2 and 3.3.2-3.3.6). They include the Pd–Pd and Pd–N(pyr) distance, twist angle (θ) and a 'squareness' (Δl) estimate (see SI §3.3.2). Despite the subtle difference between ligands L^{CH} and L^{N}, slightly different Pd–Pd distances are observed, which vary between 11.3–12.9 Å for C-1 and 10.9 –12.5 Å for C-2 (avg. 12.11 and 11.74 Å respectively). The slightly shorter Pd–Pd distance of C-2 has also been observed in several crystal structures. We hypothesize this difference arises from a subtle variation in ligand geometries, with L^{N} having a more concave angle than L^{CH} (Figure 3.3.7). This also manifests in contrasting helical flexibility, as defined by the twist angle, wherein $\theta \sim 0°$ corresponds to the highest symmetry conformation and non-zero values are twisted structures. MD simulations show that this twisting can be as high $\pm 60°$ for both cages. However, the distribution of twisted states is different. For the vacant cages, C-1 is biased towards the symmetric structure ($\theta \sim 0°$, Figure 3.2.2a) while C-2 shows a spread of states ($\pm 30°$, Figure 3.2.2b) with similar energy. DFT geometry optimization of representative conformations for C-1 found during MD simulation, confirm several local minima that could be thermally accessible at room temperature. They include structures with almost perfect D_{4h} symmetry (E_{rel}(i) = 0.0 kcal mol^{-1}), a slight distortion (E_{rel}(ii) = 0.7 kcal mol^{-1}) and a pronounced helical twist (E_{rel}(iii)

= 2.6 kcal mol^{-1}, Figure 3.2.2). Overall, this analysis shows that correlated motion of several partially rotatable bonds gives rise to an overall macromolecular flexibility. The accessibility of helical geometries also provides potential new opportunities in enantioselective catalysis, which have recently started to be explored in alternate metallocage (M_4L_6) assemblies.[227]

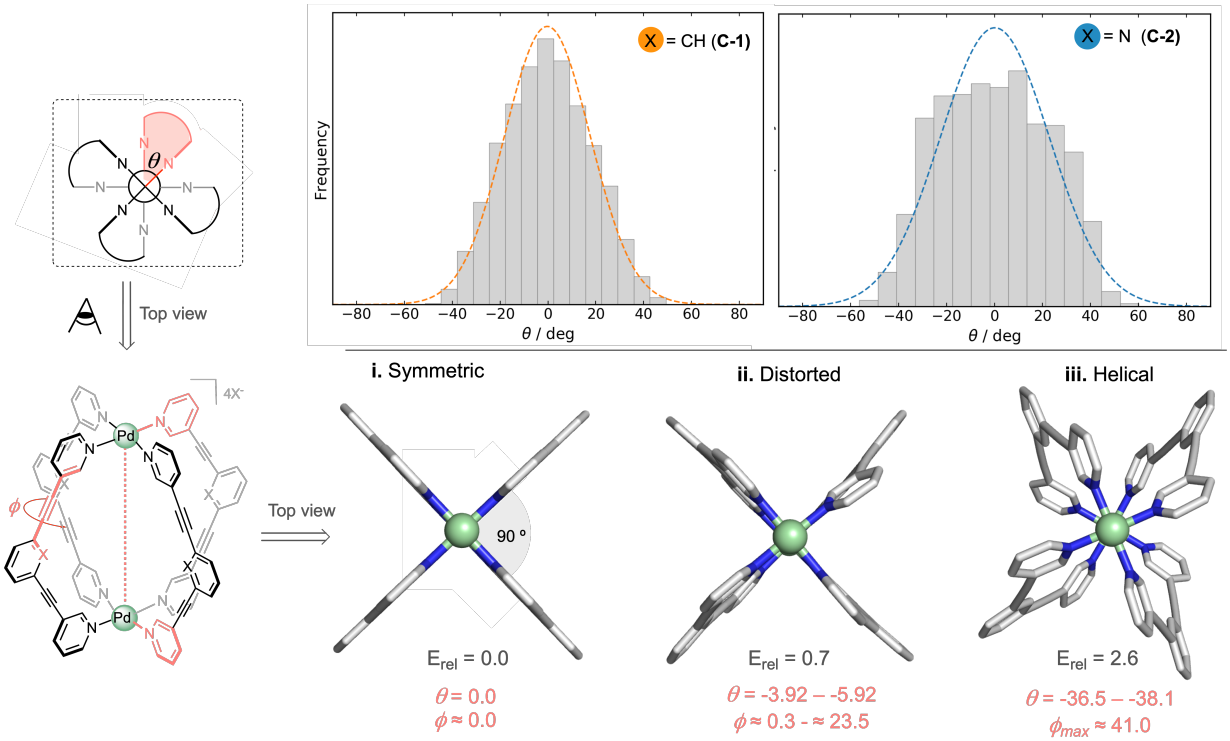


Figure 3.2.2. (a) Twist angle (θ) frequency for C-1 and C-2 calculated in explicit DCM solvent, over 30 ns of cumulative MD simulations and (b) local minima for C-1 (i-iii) calculated at the PBE0-D3BJ/def2-SVP level of theory. Relative energies (E_{rel}) in kcal mol^{-1}.

3.2.3.2 Efficient Protocol for Binding Affinity Calculations

To quantitatively rationalize catalysis and binding for C-1 and C-2, each with ∼150-200 atoms, a computationally efficient protocol is required. Here, we targeted a total time less than the corresponding experiments would take (∼1 day). We selected the M06-2X/def2-TZVP (M2) level of theory, which has been shown to provide accurate association energies

for large supramolecular systems.[228] Furthermore, to compare the binding affinity across a range of quinone/Pd_2L_4 cages, and considering the challenges associated with entropy calculations,[229–231] we analysed relative potential energy differences ($\Delta\Delta E_{\text{bind}}$) rather than Gibbs free energy differences ($\Delta\Delta G_{\text{bind}}$), i.e without considering entropic or zero-point-energy corrections, which cancel when comparing both cages (see detailed discussion in SI).

The binding energy of a general quinone (qn) is defined as ΔE_{bind}(qn⊂C-X) = E(qn⊂C-X) – E(qn) – E(C-X), with negative values suggesting favourable binding; the relative binding affinity between the two cages is then defined as $\Delta\Delta E_{\text{bind}}$(qn) = ΔE_{bind}(qn⊂C-2) – ΔE_{bind}(qn⊂C-1), where positive values indicate a preference of the guest to bind within C-1 over C-2. From Figure 3.2.3a and Table 3.3.3, it can be observed that for benzoquinone (bq), in the gas phase, binding is preferred in C-2 over C-1 ($\Delta\Delta E_{\text{bind}} = -4.5$ kcal mol^{-1}), in contrast to the solution-phase experimental observation. The gas-phase C-2 preference arises from the H-bond interactions between the nitrogen lone pairs (n) and the anti-bonding σ^*(C–H) orbitals in bq, which are stronger than the interaction between the C-H group of the meta-substituted benzene and the π bond in bq (see below). However, when implicit DCM solvent is introduced, these interactions are almost entirely masked, leading to a preference for C-1 over C-2, in agreement to experimental results (calcd. $\Delta\Delta E_{\text{bind}} = +1.9$ kcal mol^{-1} cf. expt.[175] $\Delta\Delta G_{\text{bind}} = +1.2$ kcal mol^{-1}). For the larger guest anthraquinone (aq), for which the $n \rightarrow \sigma^*$(C–H) interactions are not possible, binding in C-1 is favoured in both gas and solvent phase, in good agreement with experiment (calcd. $\Delta\Delta E_{\text{bind}}= +3.7$ kcal mol^{-1} cf. expt. $\Delta\Delta G_{\text{bind}}= +5.1$ kcal mol^{-1}). Pleasingly, the absolute binding affinity of bq was also predicted to within chemical accuracy (Figure 3.2.3), which suggest that entropic corrections to the free energy of binding are negligible for this system i.e. calcd. $\Delta E_{\text{bind}} \approx$

expt. ΔG_{bind}. (see detailed discussion in SI §3.3.3).[b]

The absolute binding affinities of aq are overestimated ($\Delta E_{\text{bind}} = -16.8(-13.1)$ vs $\Delta G_{\text{bind}} = -10.5(-5.4)$ kcal mol^{-1} for C-1(C-2), Table 3.3.3), potentially arising from the significant loss in rotational entropy of aq compared to bq upon binding, in which case ΔS is no longer negligible. Nevertheless, the relative binding energies for both cages compare well to experiment. These results demonstrate that the predicted relative binding affinities for bq and aq between C-1 and C-2 are accurate to within $\sim$2 kcal mol^{-1}. To probe whether this result was general for this class of system, we considered 14 experimental binding affinities (guest molecules shown in Figure 3.3.15) for C-1 and C-2, including a new data point corresponding to pentacenedione (q4) bound to C-2. Bearing in mind the computational expense of performing geometry optimizations for this set at the SMD(DCM)-M2 level of theory (five days on a CPU with 8 cores), calculations were performed at the more efficient M1 level, followed by single point energy evaluations at SMD(DCM)-M2. We refer to this approach as SMD(DCM)-M2//M1 (Table 3.3.7).

In these systems, both relative ($\Delta\Delta E_{\text{bind}}$) and absolute binding energies (ΔE_{bind}) were calculated (Figure 3.2.4). Once again, not only are the relative binding affinities obtained to within a reasonable mean absolute deviation (MAD = 2.2 kcal mol^{-1}), the absolute binding affinities are close to chemical accuracy (MAD = 1.7 kcal mol^{-1}, Figure 3.2.4, Table 3.3.7), with calculations taking less than a day on a CPU with 8 cores. Compared to predictions in organic host-guest complexes, e.g. via the Statistical Assessment of the Modelling of Proteins and Ligands (SAMPL) challenge, where errors of at least 1–4 kcal mol^{-1} and correlations (R^2) below 0.5 have been reported, our correlation to experimental binding affinities ($R^2 = 0.759$) can be considered very good.[228, 232, 233]

[b] $G_{\text{bind}} = E_{\text{bind}} + \text{ZPE} + E_{\text{therm}} + P\Delta V – T\Delta S$. The last three terms are obtained from frequency calculations assuming an ideal gas model (IGM)

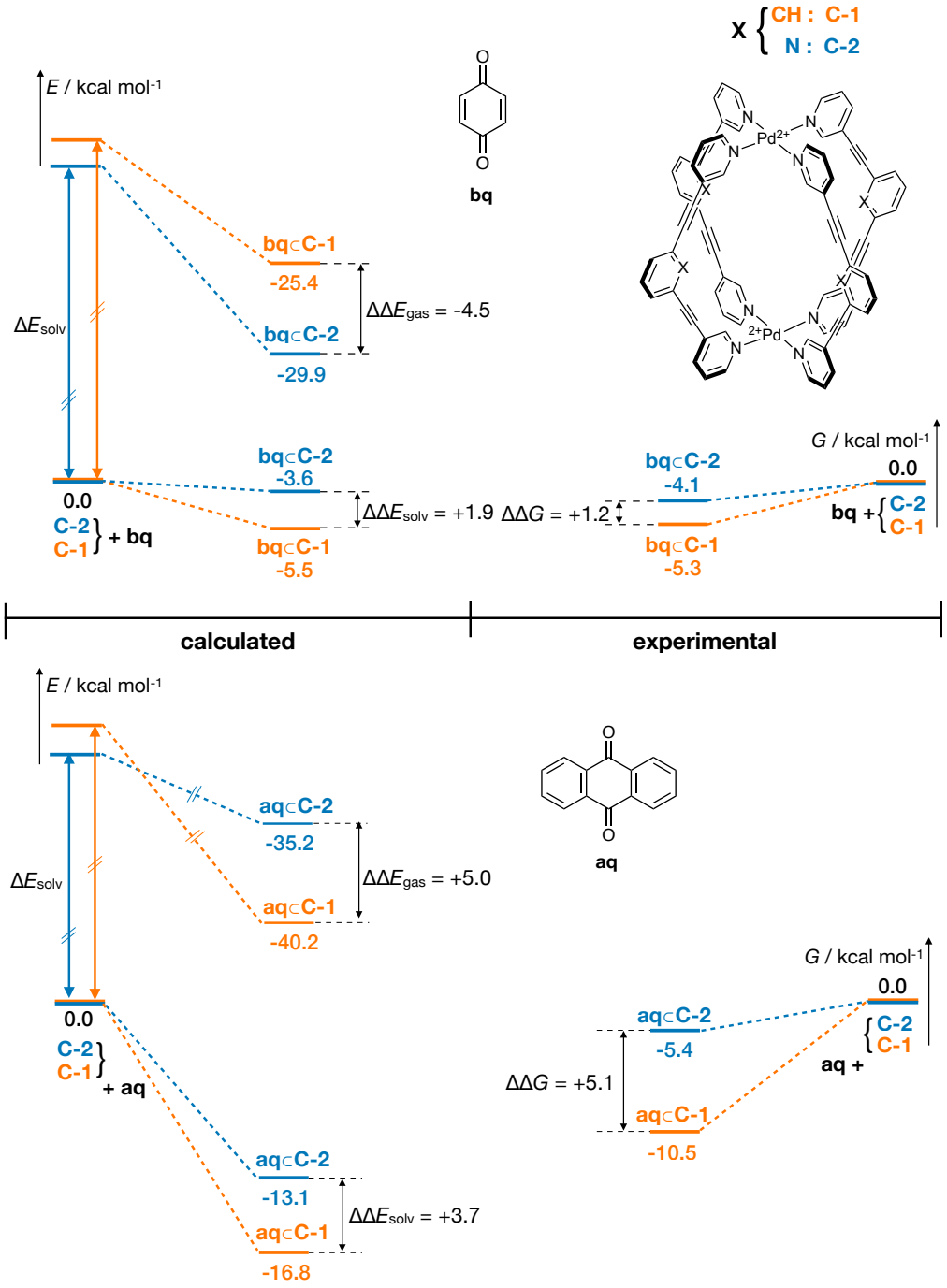


Figure 3.2.3. Calculated and experimental absolute and relative binding affinities for (a) bq⊂C-1 and bq⊂C-2. Experimental values were obtained in dichloromethane (DCM) solvent at room temperature. Gas phase and solution phase calculations performed at the M06-2X/def2-TZVP and SMD(DCM)-M06-2X/def2-TZVP level of theory, respectively. Values reported in Table 3.3.3.

The binding affinity outlier is tert-butyl benzoquinone (top right in Figure 3.2.4a), for which we find binding within the cavity to be unfavourable. Experimentally, this quinone has been hypothesized to bind to the outer 'pocket' of the cage, i.e. expt. ΔG_{bind} of encapsulation may, in fact, be positive. We also hoped to further accelerate the methodology by obtaining structures at a semi-empirical (PM7[234]) or tight-binding DFT (GFN-xTB[125]) level of theory. Unfortunately, these methods lead to relatively poor correlation (r^2(SMD(DCM)-M2//PM7) = 0.309, r^2(SMD(DCM)-M2//GFN-xTB) = 0.173, Figures 3.3.9–3.3.11). Note the latter is in spite of the relatively small root mean squared deviation (RMSD) to M1 geometries (0.3 ± 0.2 Å).

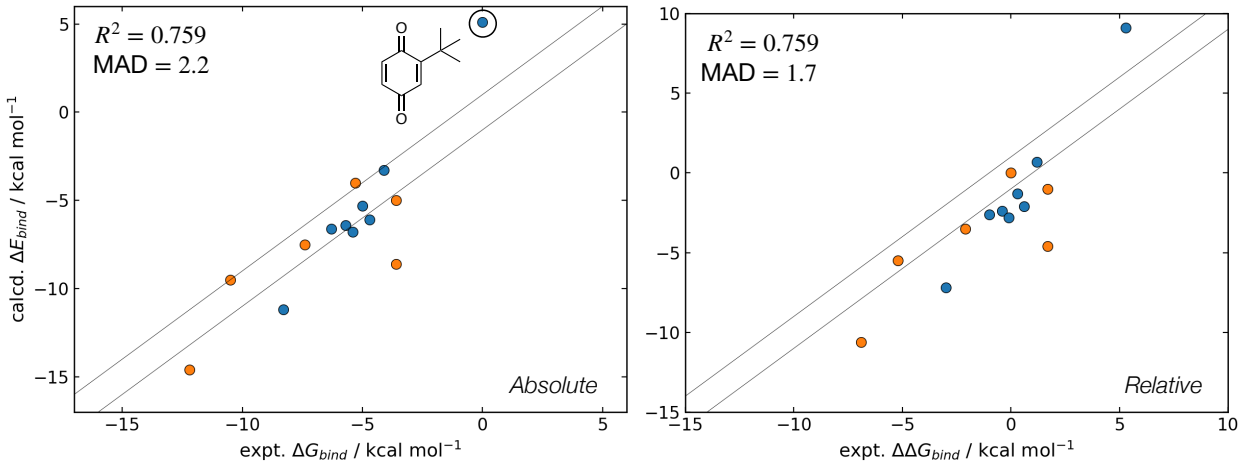


Figure 3.2.4. Correlation plots of (a) $\Delta\Delta E_{\text{bind}}$ = ΔE_{bind}(qn⊂C-X) – ΔE_{bind}(bq⊂C-1). and (b) ΔE_{bind} calculated at the SMD(DCM)-M06-2X/def2-TZVP//PBE0-D3BJ/def2-SVP level of theory to experimental free energies. Orange and blue markers corre-spond to C-1 and C-2 cages, respectively. The different diagonals bracket the ±1 kcal mol^{-1} area of accuracy.

3.2.3.3 Rationalizing Differences in Binding Affinities

To elucidate the nature of the non-covalent interactions driving quinone binding, and the differences between C-1 and C-2, we constructed two reduced models. The first uses $[\text{Pd(pyridine)}_4]^{2+}$

as a model of the "top" and "bottom" of the cage (model 1, Figure 3.2.5a), while the second describes the interaction between the quinone guest with the central meta-substituted benzene or 2,6-pyridyl groups in C-1 and C-2, respectively (model 2, Figure 3.2.5b).

Using model 1 we analyse the effect of the metal and the +2 charge it introduces Figure 3.2.5a, by comparing the canonical complex ($[Pd(pyridine)_4]^{2+}$, light blue) to a complex without metal and an overall +2 charge ($[(pyridine)_4]^{2+}$, cyan), and without metal and an overall zero charge ($[(pyridine)_4]$, green). This model demonstrates that the metal influences binding by mainly introducing a positive charge that polarizes the adjacent *o*-pyridine (C)H donor groups (Figure 3.2.5a and 3.3.16). The second model, model 2, describes the interaction between the quinone guest with the 'equatorial' meta-substituted phenyl or pyridyl moieties. In this system, rotation of bq around the z-axis (defined by the angle χ, Figure 3.2.5b-c) shows a minimum at $\chi \approx 0°$ for the bq⊂$[(pyridine)_4]$ system (blue line), while for the bq⊂$[(benzene)_4]$ system (orange line), minima are found at $\chi \pm 45°$. In the former, electron donation from the nitrogen lone pair (n) to the anti-bonding σ^*(C–H) orbital leads to the formation of weak hydrogen bonds at $\chi = 0°$ (Figure 3.3.17, 3.3.18).[216] In contrast, for the bq⊂$[(benzene)_4]$ system, repulsion between CH groups produces a maximum at the same position (Figure 3.2.5b). This explains why in C-1, bq sits out of the plane of opposing ligands in ($\chi \approx 45°$), while in C-2, bq lies in the plane ($\chi \approx 0°$, Figure 3.2.5c). The $n \rightarrow \sigma^*$(C–H) interactions are strong in the gas phase; however, due to their electrostatic nature, they are lost once implicit solvent is included. In contrast, while the $\pi \rightarrow \sigma^*$(C–H) interactions are relatively weak, they are less dependent on the solvent and contribute to the binding of the guest within C-1 (Figure 3.3.19). These results explain the observed preference for binding qn within C-1 over C-2, which originates from differences in solvation energy, with C-2 being the intrinsically stronger binder.

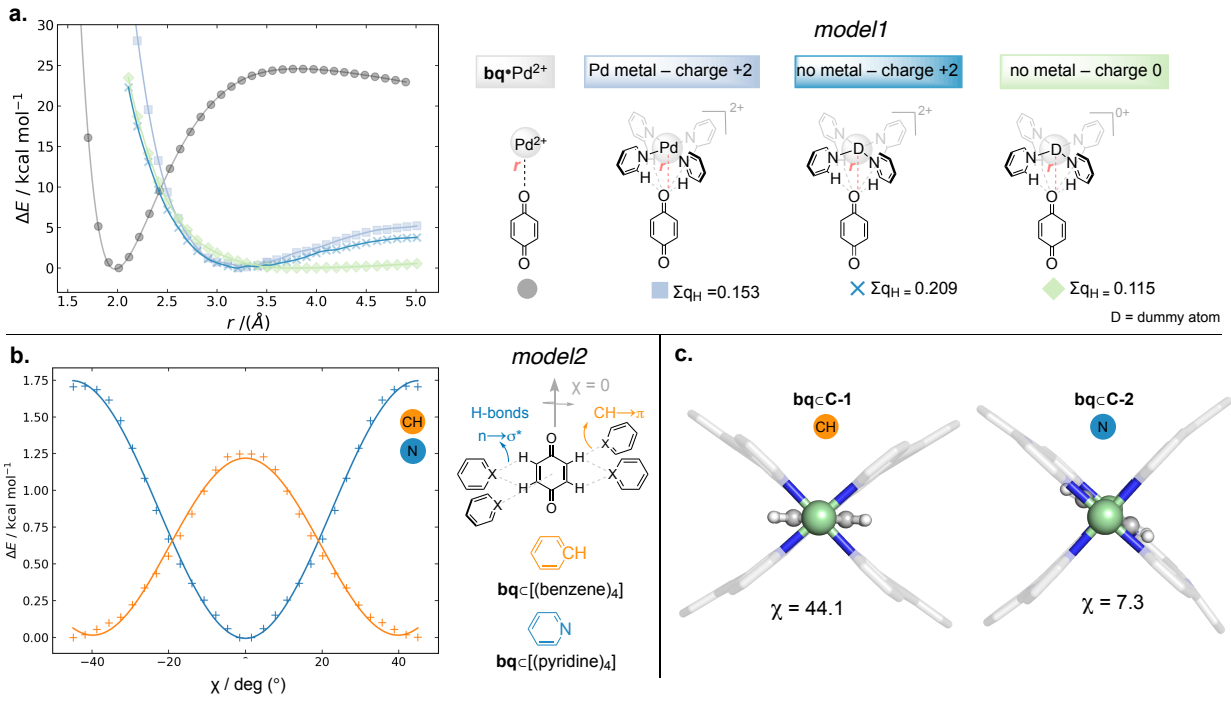


Figure 3.2.5. (a) Comparative PES for bq with Pd(II) (grey), the canonical complex ($[Pd(pyridine)_4]^{2+}$, light blue) complex without metal and an overall +2 charge ($[(pyridine_4]^{2+}$, cyan), and complex without metal and an overall zero charge ($[(pyridine)_4]^0$, green). The sum of atomic partial charges on the (C)H atoms, using the Hirshfeld scheme, is shown in each case (b) Non-covalent interaction in model 2, including benzenes/pyridines and bq. (c) Global minima of bq⊂C-1 and bq⊂C-2. All calculated at the M06-2X/def2-TZVP level of theory.

3.2.3.4 Rationalizing the Differences in Catalytic Activity

To assess the influence of the cages on DA reactivity, the activation energies for a set of 10 uncatalysed reactions (Figure 3.3.20) were calculated using a highly accurate quantum chemical method. Pleasingly, the error to experiments[175] was within ±1 kcal mol^{-1} (MAD = 0.9 kcal mol^{-1}). Among the different DFT functionals tested, the M2 method was found to have the lowest error relative to the *ab initio* results (MAD 1.2 kcal mol^{-1} for reaction barriers and 1.7 kcal mol^{-1} for reaction energies, Figures 3.3.21, 3.3.22). In all cases, a highly synchronous transition state (TS) was observed, with the endo cycloadduct being preferred (Table 3.3.8). These results are in agreement with previous computational studies

on the uncatalysed reaction of para-quinone imine derivatives using the M06-2X functional in vacuum.[235]

With a chemically accurate computational methodology in hand, the reaction barriers for the DA reaction between bq and isoprene (d1) in the presence of C-1 or C-2 were calculated. Formation of an association complex between d1 and bq⊂C-1/C-2 is unfavourable. For example, with C-2 it leads to a complex 8.8 kcal mol^{-1} higher in free energy than bq⊂C-2 and isoprene separately (Table 3.3.10). This result is in line with ^{1}H NMR data of the catalytic process, which shows that neither the cage nor the quinone signals shift upon addition of diene to the reaction,[175] and contrast with most supramolecular host-guest complexes exhibiting DA catalytic activity, where a so-called "ternary Michaelis complex" is observed.[62] As suggested experimentally, our calculations confirm that the complex bq⊂C-2 is first formed and then reacts intermolecularly with an incoming isoprene molecule.

Defining the catalytic activity as $\Delta\Delta E_{\mathrm{CA}} = \Delta E^{\ddagger}_{\mathrm{uncat}}$ - $\Delta E^{\ddagger}_{\mathrm{cat}}$, such that positive values correspond to effective catalysis we calculate $\Delta\Delta E_{\mathrm{CA}}$ (d1+bq⊂C-2) = +5.4 kcal mol^{-1} (expt. $\Delta\Delta G_{\mathrm{CA}}$ = +3.6 kcal mol^{-1}) while $\Delta\Delta E_{\mathrm{CA}}$ (d1+bq⊂C-1) = −0.3 kcal mol^{-1} (expt. catalytically inactive, Table 3.2.1). As shown in Figure 3.2.6a, while C-1 binds the substrate more strongly in the reactant state, C-2 provides a better stabilisation at the TS and slightly better stabilization of the product state (Figure 3.3.24). Analysis of the encapsulated TS geometries indicate that, for asymmetric dienes, the cage increases the asymmetric nature of the TSs compared to the uncatalysed analogues, making it less pericyclic and more stepwise (Figure 3.2.6b). The active C-2 cage affords a slightly more asynchronous reaction (Δd = 0.28/−0.17, 0.36/−0.23 Å for TS[d1+bq⊂C-1] and TS[d1+bq⊂C-2], respectively). To rationalize these differences, we explored the electronic and steric contributions to the reaction energy in each system. In both cases, and as previously seen for Lewis acid-catalysed

[4+2] cycloadditions,[236–238] the cages enhance the electrophilic character of the dienophile bound in the cavity. Polarization of the dienophile within the cage leads to a net charge of +0.25 and +0.16 in bq for C-1 and C-2, respectively,[239] leading to a lowering the LUMO energy by 1.5 and 1.6 eV in C-1 and C-2, respectively (see full SI). These results, demonstrate that both C-1 and C-2 can activate the dienophile by reducing its LUMO energy. However, as discussed below, other interactions in C-1 offset this activating effect. To analyse this contrasting behaviour between C-1 and C-2, we performed an energy decomposition analysis using the distortion/interaction framework.[240] Within this framework, the activation energy of a reaction ($\Delta E^{\ddagger}$) is partitioned into the energy required to distort the reactants from their ground-state to their transition state geometry ($\Delta E^{\ddagger}_{\text{dist}}$) and the energy of interaction between these fragments ($\Delta E^{\ddagger}_{\text{int}}$). For the reaction under study, the optimal interaction between the activated dienophile and the incoming diene takes place within one of the cage portals (Figure 3.2.6b); this requires the ligand to distort significantly, thus suggesting that distortion of the cage and the substrate is required to reach the TS in the cage-catalysed process.

The energetic cost of distorting each component of the reaction is shown in Table 1. Here, we separated each TS structure into three fragments ($\Delta E^{\ddagger}_{d}$[C-X], $\Delta E^{\ddagger}_{d}$[diene], and $\Delta E^{\ddagger}_{d}$[bq]) and also evaluated the bq⊂C-X complex as a single fragment ($\Delta E^{\ddagger}_{d}$[bq⊂C X]). In C-2 a smaller distortion compared to C-1 is obtained for both the diene (1.3 kcal mol^{-1}) and the cage-dienophile complex (4.5 kcal mol^{-1}). For C-1, the cost of aligning the dienophile bq with the cage portal is found to be twice the cost seen in C-2 (Figure 3.3.27). This can be associated to steric clashes arising between the central moiety of L^{CH} and the dienophile. This effect can be visually understood through the use of non-covalent interaction (NCI) plots, which in C-1 shows steric clashes between the dienophile and the cage, while in C-2 these regions are smaller and compensated by a favourable N$\cdots$HC hydrogen interaction

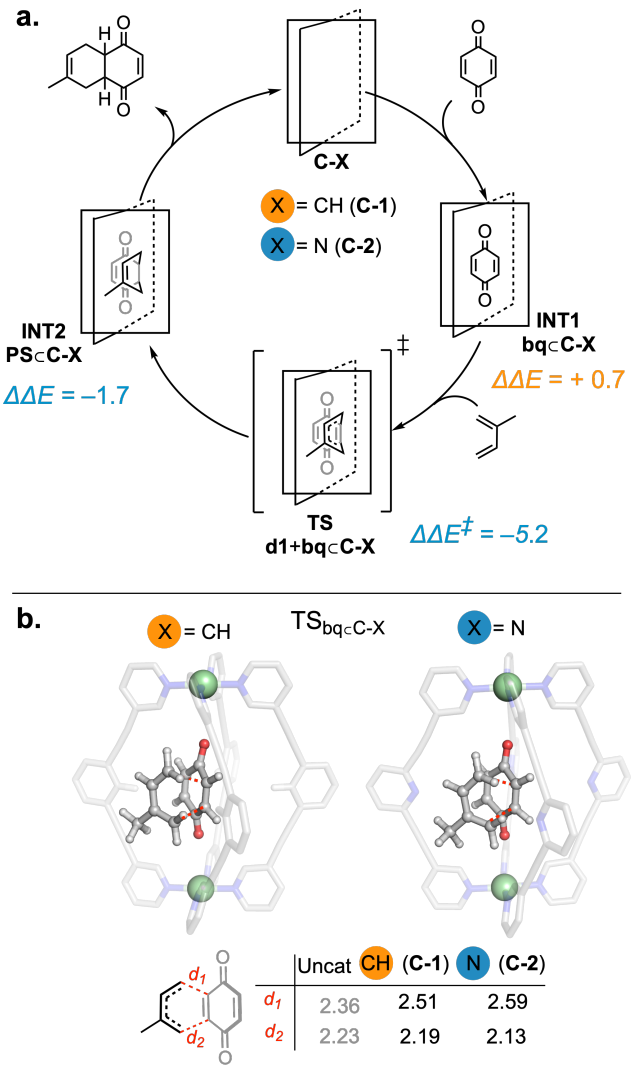


Figure 3.2.6. (a) Catalytic cycle for the reaction between benzoquinone and isoprene ($\Delta\Delta E$ = ΔE(C-2) – ΔE(C-1) in kcal mol^{-1} at SMD(DCM)-M06-2X/def2-TZVP//PBE0-D3BJ/def2-SVP, negative(positive) values in blue(orange) refer to favourable stabilization for C-2 (C-1)). (b) Optimized TS geometries; Distances (Å) of the forming bonds at the TS for the uncatalysed (grey) and encapsulated (black) reaction.

between the diene and the cage. The latter may also lead to the slightly more positive interaction term for this cage (Figure 3.3.27). For both cages, the interaction energy ($\Delta E^{\ddagger}_{\text{int}}$) is more favourable compared to the uncatalysed reaction, which is in line with lowering of dienophile LUMO energy. However, distortion effects render C-1 uncatalytic. These results demonstrate that small differences in cage sterics and flexibility play an important role in

	$\Delta E_d^{\ddagger}$ [C-X]	$\Delta E_d^{\ddagger}$ [diene]	$\Delta E_d^{\ddagger}$ [bq]	$\Delta E_d^{\ddagger}$ [bq⊂C-X]	$\Delta E_d^{\ddagger}$[bq⊂C-X +diene]	$\Delta E_{\text{int}}^{\ddagger}$	$\Delta E^{\ddagger}$
Uncat.	–	15.8	7.1	–	22.9[a]	-13.0	9.9
C-1	5.2	16.2	8.8	13.2	29.4	-19.2	10.2
C-2	5.1	14.9	9.3	10.0	24.9	-20.4	4.5
$\Delta\Delta E_{\text{C2-C1}}$	-0.1	-1.3	0.5	-3.2	-4.5	-1.2	-5.7

Table 3.2.1. Distortion/Interaction Analysis for the [4+2] cyclisation of isoprene with benzoquinone calculated at the SMD(DCM)-M2 level of theory. Shown here are the distortion energies of the cage ($\Delta E_d^{\ddagger}$[cage]), substrate ($\Delta E_d^{\ddagger}$[diene]), dienophile ($\Delta E_d^{\ddagger}$[bq]), and bq⊂C-X complex ($\Delta E_d^{\ddagger}$[bq⊂C-X]). The interaction energy is defined as $\Delta E_{\text{int}}^{\ddagger} = \Delta E^{\ddagger} - \Delta E_d^{\ddagger}$, where $\Delta E^{\ddagger}$ is the activation energies and total $\Delta E_d^{\ddagger} = \Delta E_d^{\ddagger}$[bq⊂C-X] + $\Delta E_d^{\ddagger}$[diene]. All energies are given in kcal mol^{-1}. *a*. Sum of diene and quinone distortion.

determining the potential catalytic activity of very similar metallocages towards a given reaction. Such energetic costs will vary depending on the activation mode required and the nature of the substrates involved but can be easily quantified. Finally, while our analysis in Table 3.2.1 included only potential energies, Gibbs free energies can be obtained by adding thermal and entropic contributions (Table 3.3.11). By doing so, we obtain $\Delta G_{\text{uncat}}^{\ddagger}$ = 23.5 kcal mol^{-1} and $\Delta G_{\text{cat}}^{\ddagger}$ (C-1/C-2) = 23.6/18.8 kcal mol^{-1}, which are in good agreement with experimental results and confirm that catalysis in C-2 is enthalpic in origin.

3.2.3.5 Efficient Protocol for Calculating Catalytic Activity

Having rationalized the catalytic properties of the two cages for a single cycloaddition reaction, we sought to validate our methods with a wider range of substrates. Although fully optimized TSs for the DA reactions outlined in Figure 3.2.7 can be located within these cages, the computational demand is significant, even using a low level of theory (> 3 days

on a CPU with 8 cores). With the aim to decrease this time, we proposed to evaluate catalytically promising cages considering the following approximations: (1) To rely on potential energy rather than free energies differences, and (2) to calculate the activation energy of 'TS analogues' rather than the true transition state. We use the term 'TS analogues' to indicate that the bonds being formed and broken are constrained to the distance values found in the uncatalysed TS.

The catalytic activity ($\Delta\Delta E'_{\mathrm{CA}}$) is then found by performing a constrained minimization and comparing to the uncatalysed variant (Figure 3.2.7). This computationally inexpensive approach can achieve 80% accuracy in predicting the catalytic proficiency towards the DA reactions tested (Figure 3.2.7), while providing a ten-fold reduction in computational time.

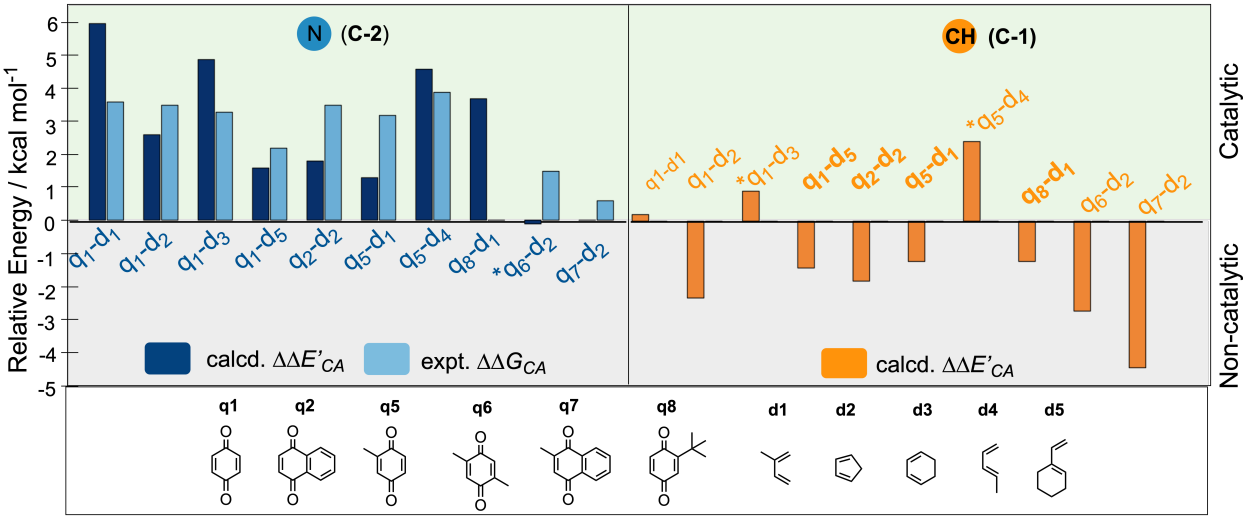


Figure 3.2.7. Calculated analogue activation energy ($\Delta\Delta E'_{\mathrm{CA}}$, dark color) and experimentally observed catalytic efficiency ($\Delta\Delta G_{\mathrm{CA}}$, light blue, only available for C-2). For C-1, the experimentally untested quinone-diene reactions are shown in bold. Two false positive for C-1 are indicated with an asterisk *. Values tabulated in Table 3.3.11.

Using this computationally inexpensive approach cage C-1 was found to be non-catalytic for seven out of ten reactions tested, while q1-d1 and q1-d3 combinations were within computational error. However, the substrate combination q5-d4 is clearly predicted to lead to

catalysis. As this cage-substrate combination was not reported in our original work,[175] we have now monitored this reaction and found no obvious acceleration compared to the cage-free control experiment. This suggests that there are other subtleties involved in promoting catalysis, which will need to be explored in further studies.

We have compared these results to those obtained using full optimization (Figure 3.3.28). Selecting the set of substrates with the largest discrepancy to experiments (q1+d1 q5+d4) and q7+d2 within C-1 and C-2, we observed that the full transition states provided almost identical results to those obtained using TS analogues ($R^2 = 0.94$, MAD = 0.6 kcal mol^{-1} , Figure 3.3.28). Moreover, the number of CPU-hours (CPUh) required to calculate a single $\Delta\Delta E$ value using TS analogues (∼100 CPUh, up to 12 hours on 8 CPUs) is at least an order of magnitude lower than when using fully optimised TS (∼1000 CPUh, 5 days on 8 CPUs using the TS analogues as starting point), thus demonstrating the substantial computational cost saved by our approach without compromising accuracy.

3.2.4 Conclusion

Supramolecular metallocages have emerged as promising biomimetic catalysts. However, a theoretical understanding of the structural and electronic factor that determine efficient binding and catalysis has been lacking. Here, we have rationalized the binding and significantly different Diels-Alder catalytic activity for two highly-homologous metallocages (C-1 and C-2) through complementary classical simulations and quantum calculations. This is the first computational study that simultaneously explores binding and catalysis with multiple cage and substrate combinations. For the DA reactions studied here, we find that electronic activation of dienophile is observed for both cages. However, in the case of inactive C-1, this is offset by significant distortion energy which inhibits the interaction with the incoming

diene. In contrast, catalytically active C-2 is much more able to accommodate distortion approaching the transition state. This suggests that, in addition to the catalytic machinery required to activate the substrates undergoing chemical reactions, novel metallocages designs should consider the plasticity of the building blocks and the cage as a whole. While these features are now understood in enzyme catalysis, they have been much less studied in biomimetic catalyst design.

In addition to rationalising the catalytic activity, this work introduces an efficient computational protocol to model metallosupramolecular catalysts in solution. Further work is underway to automate this process within an open-source Python module (cgbind, `https://github.com/duartegroup/cgbind`). We expect this to motivate a closer interaction between experimentalists and computational chemists in the quest towards the discovery of novel catalysts.

3.2.5 Computational Methods

MD Simulations. Configurational sampling of the cage structures in dichloromethane (DCM) was performed using classical MD simulations. For ligands and substrates, OPLS-AA-compatible force-field parameters were generated using the Macromodel ffld14 version (Schrödinger LCC).[241] Restrained electrostatic potential (RESP) charges were derived by fitting partial charges to HF/6-31G(d) electrostatic potentials calculated using Gaussian 09 (version D.01).[242] Geometries for these calculations were obtained from optimisations at the TPSS-D3BJ/def2TZVP/fit level of theory. The topology for DCM was taken from the GROMACS Molecule & Liquid Database.[243] All simulations were performed using the GROMACS package (version 2019.1)[244] using three-dimensional periodic boundary conditions. Long-range electrostatic interactions were treated using the particle mesh Ewald

(PME) approach[245] with a cut-off length of 1.0 nm. A dispersion correction was applied to energy and pressure terms to account for truncation of van der Waals terms. The systems were immersed in a box of solvent with a distance from the border of at least 15 Å. Following steepest descent minimization, the systems were equilibrated in two steps; the first phase involved simulating for 100 ps under a constant volume (NVT) ensemble with position restraints applied to heavy atoms. The temperature was maintained at 298 K using the velocity-rescale method. This was followed by 100 ps of constant-pressure (NPT) equilibration using the Parrinello–Rahman pressure coupling algorithm with the compressibility set to 4.5×10^{-5} bar^{-1} and the time constant set to 5 ps. A 1 fs time step was used during these stages to allow potential inhomogeneities to self-adjust. Each system was then equilibrated initially for 10 ns with a 1 fs time step at constant pressure. Three different runs, differing in the initial random seed were run per system. The first 1 ns was removed from the analysis.

Electronic Structure Calculations. All density functional theory (DFT) calculations were performed in the ORCA v. 4.1 software package.[246] Initial geometry optimisations for cages (C-1, C-2) and cage–substrate complexes (bq⊂C-1, bq⊂C-2) were performed with the M06-2X functional,[89] the Ahlrichs def2-TZVP[247, 248] basis set which includes the ECP28MWB81 on Pd. Solvent effects were accounted for with the SMD[131] solvent model with parameters appropriate for DCM. Subsequent optimisations were performed with the PBE0[86] functional with the Becke-Johnson damped D3 dispersion correction (D3BJ)[249] in combination with a def2-SVP basis set. Where stated vibrational frequencies were computed at the optimization level of theory to confirm whether the structures correspond to minima or a transition states and to evaluate the zero-point vibrational energy (ZPE) and thermal corrections at 298 K. Gibbs free energies in which a change of molecularity occurs include a 1 atm to 1 M standard state correction. Single-point energies were obtained at the M06-2X/def2-TZVP level of theory in DCM solvent using the SMD solvent model. Single

point energies for non-catalysed DA reactions have also been carried out using PBE0,[86] TPSS,[85] PW6B95,[250] ωB97X-D3[251] and M06-2X[89] functionals to analyse the influence of the functional and dispersion upon optimized structures and stabilities when compared to the DLPNO-CCSD(T)[252] linear-scaling coupled cluster implementation, which was used in combination with the minimally augmented doubly polarized ma-def2-TZVPP basis set. The chain of spheres resolution of identity (RIJCOSX in ORCA) approximation was used for the Hartree-Fock exchange component of hybrid functionals, with the default fitting bases. NBO charge and second order perturbation analyses were carried out in NBO[253] v. 6.0. Multiwfn v. 3.6[254] has been used for electron density analysis. Unless specified otherwise, all energy differences are calculated at SMD(DCM)-M06-2X/def2-TZVP//PBE0-D3BJ/def2-SVP level of DFT theory.

3.3 Selected Supporting Information I

Full Supporting Information including raw data can be found at:

`https://pubs.acs.org/doi/abs/10.1021/jacs.9b10302`

3.3.1 Square-planar Metal Model

The square planar palladium dummy atom model originally presented by Fujita et al. was used for all MD simulations.[255] This model is based on previous work on transition metal parameterization by Aqvist and Warshel.[256] The model consists of four particles, each of them possessing a charge of +0.5, while the central particle possesses zero charge. The geometry of the dummy complex itself is kept rigid by the imposition of large force constants on the metal–dummy bonds (Table 3.3.1). However, as there are no bonds between the dummy complex and the surrounding ligands, overall rotation is allowed, and no internal forces are associated with such rotation. This means that the system is free to exchange ligands on the relevant time scale.

The modified Fujita parameters a Pd(II) were initially tested in water, employing the three-point transferable intermolecular potential (TIP3P).[257] The system was simulated for 5 ns and the Pd–O distance obtained from analysis of the radial distribution function (RDF, Figure 3.3.1). The modal distance is somewhat shorter than the experimental value (1.85 Å cf. expt.[225] 2.01 Å).

3.3.2 Metallocage MD Modelling

van der Waals				
name	Mass / amu	Charge / e	σ / nm	ϵ / kJ mol^{-1}
M	98.336	0.0	0.1718	2.89951
$D_{i/j}$	2.016	0.500	0.0	0.0
	Bond			
	r_0 / nm	K_b / kJ mol^{-1}		
$D_{i/j}$–M	0.09	669440		
	Angle			
	θ / degrees	K_θ / kJ mol^{-1} rad^{-2}		
D_i-M-D_j(trans)	180	2092		
D_i-M-D_j(cis)	90	2092		

Table 3.3.1. Force Field Parameters for the Pd dummy model used in this work. Definitions appropriate for the GROMACS simulation package.

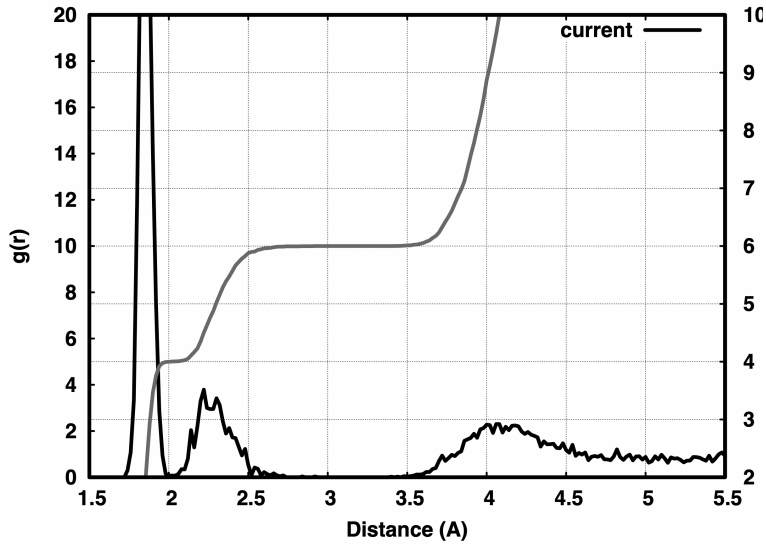


Figure 3.3.1. Pd–O radial distribution function [$g(r)$] for a Pd(II) ion in a cubic box balanced with two Cl^- ions at 298.15 K from a 5 ns MD trajectory. Simulation set up by Prof. Fernanda Duarte.

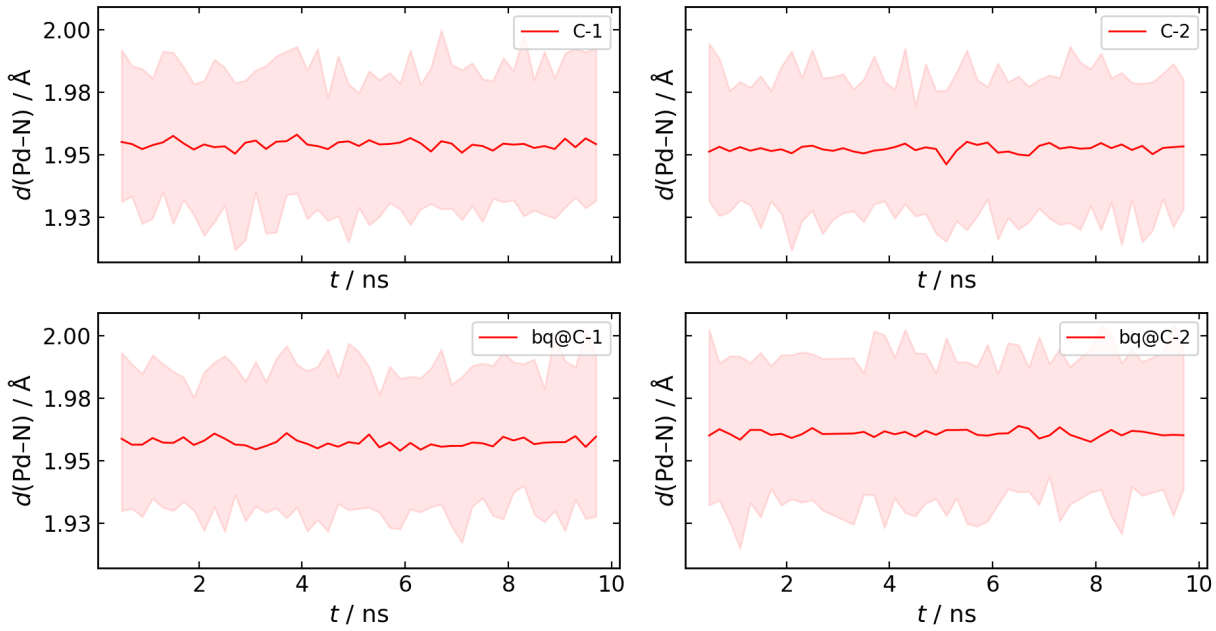


Figure 3.3.2. Temporal Pd–N(donor) distance variation. The solid line corresponds to a 20-step block average over 3 replicas and the shaded area the range of the block.

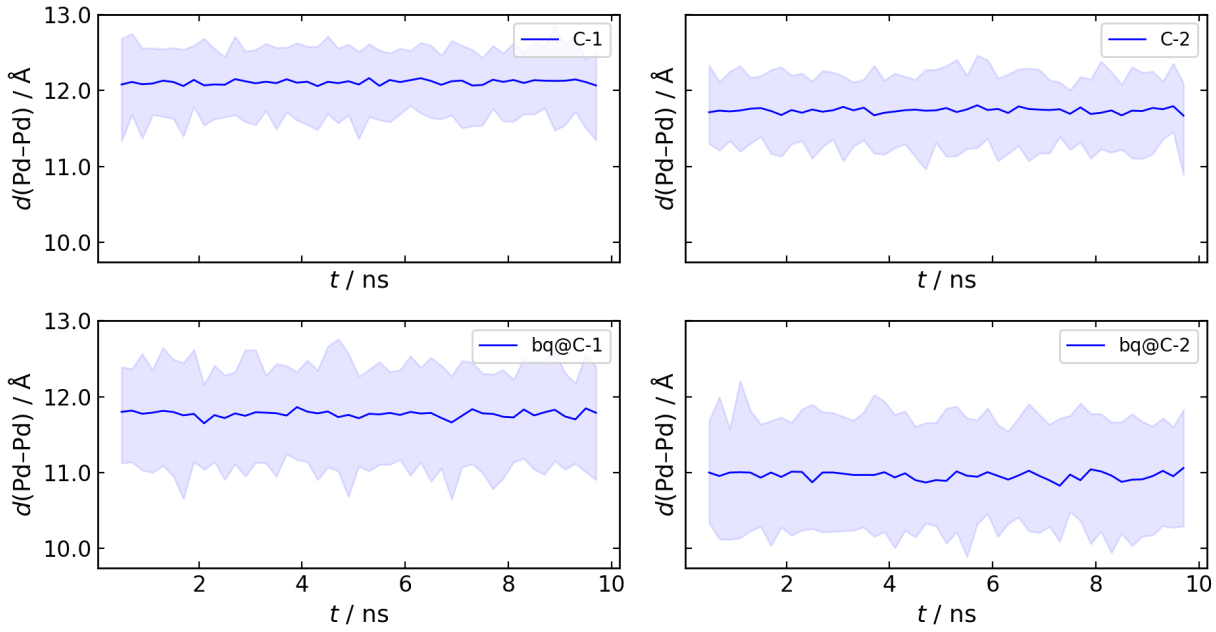


Figure 3.3.3. Temporal Pd–Pd distance variation. The solid line corresponds to a 20-step block average over 3 replicas and the shaded area the range of the block.

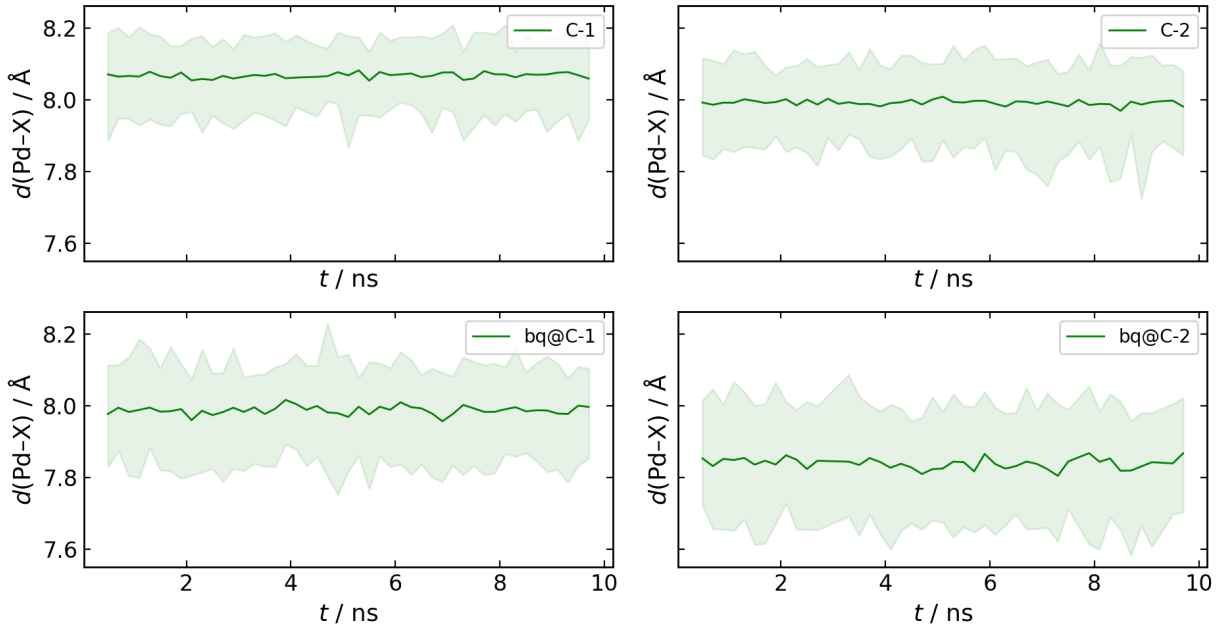


Figure 3.3.4. Temporal Pd–X distance variation, where X=CH, N for C-1 and C-2 respectively. The solid line corresponds to a 20-step block average over 3 replicas and the shaded area the range of the block.

Species	Description	Distance / Å	Ref. distance / Å
	Pd–N	1.954±0.002	2.02±0.01[a]
C-1	Pd–Pd	12.111±0.029	11.85[a]
	Pd–X	8.067±0.007	7.99±0.11[a]
	Pd–N	1.957±0.002	2.02±0.00[b]
bq⊂C-1	Pd–Pd	11.771±0.043	12.21[b]
	Pd–X	7.987±0.011	8.02±0.17[b]
	Pd–N	1.952±0.002	2.02±0.01[c]
C-2	Pd–Pd	11.738±0.032	11.50[c]
	Pd–X	7.991±0.007	7.96±0.17[c]
	Pd-N	1.961±0.001	
bq⊂C-2	Pd–Pd	10.960±0.052	NA
	Pd–X	7.83±0.015	

Table 3.3.2. Average key distances for C-1, C-2, bq⊂C-1 and bq⊂C-2 calculated from MD simulations. Standard deviations in the block averages through the trajectory are given. NA = not available. a) OTf⊂C-1[OTf]. CCDC: 768969. Ref [210]; b) [pentacenedione⊂C-1][OTf]4. CCDC: 1492902. Ref. [224]; c) [C-2][SbF_6^-]$_2$. CCDC: 853226. Ref. [204].

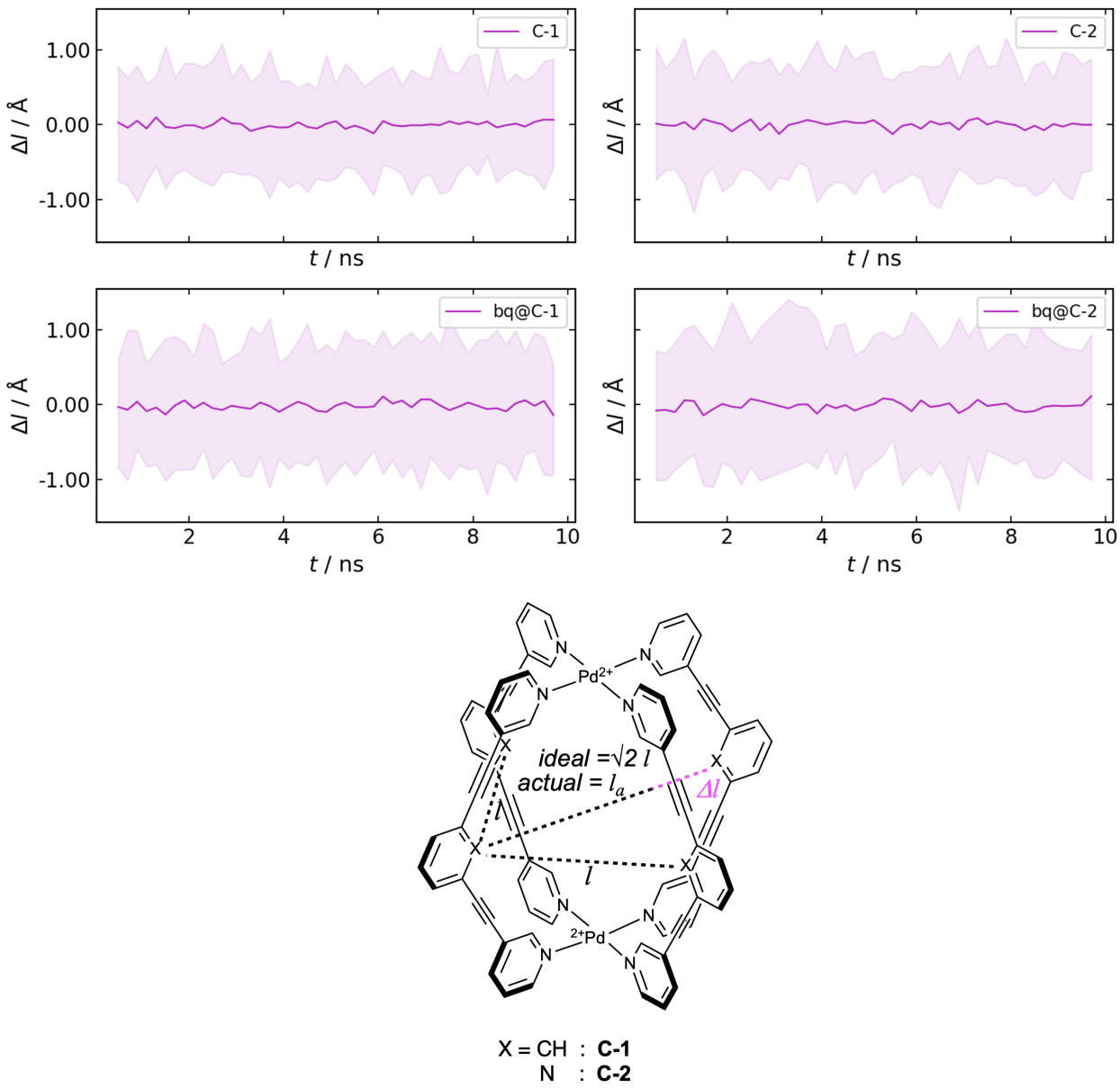


Figure 3.3.5. Temporal variation of the absolute value of Δl, a measure of squareness of the cavity.

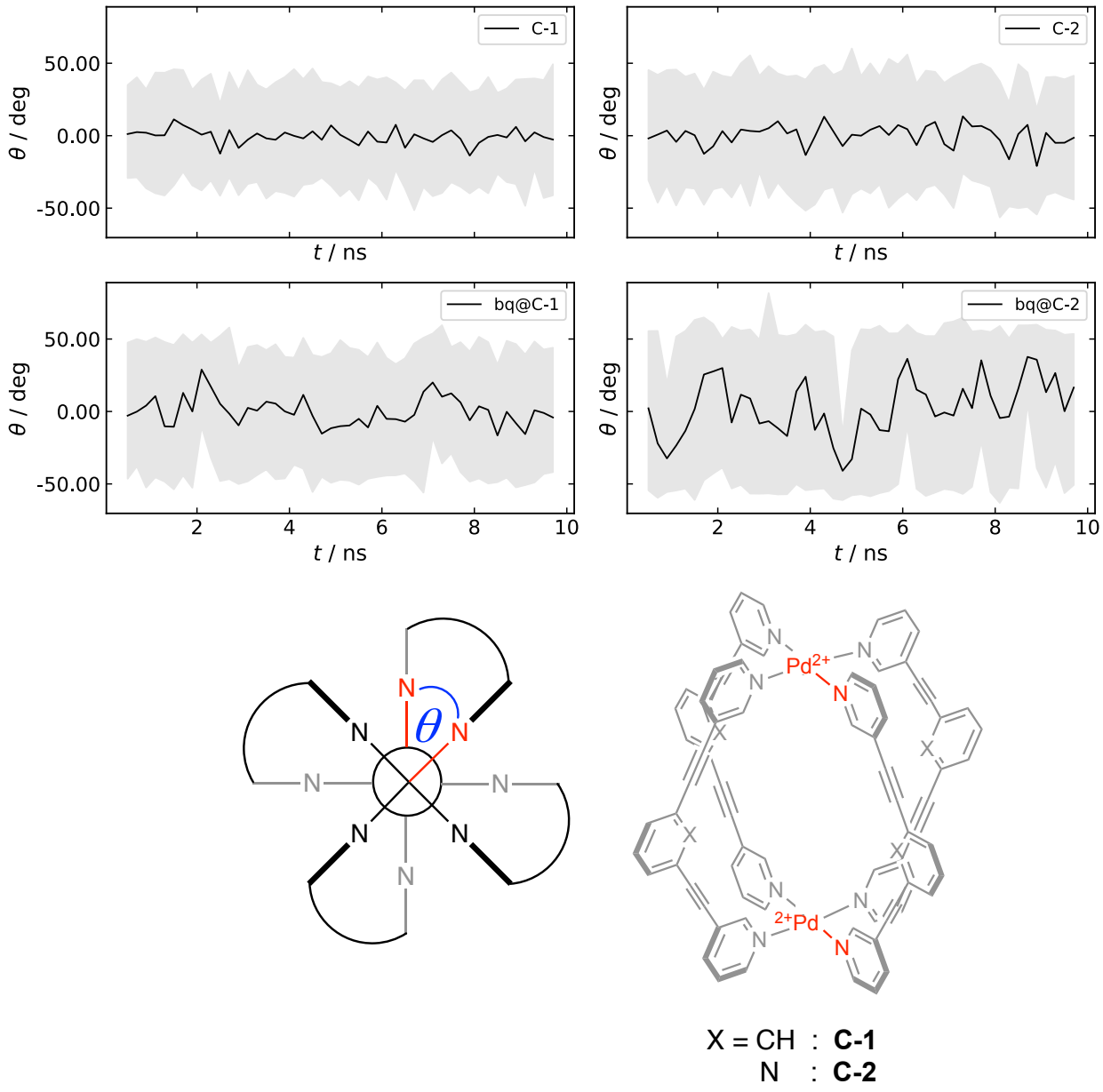


Figure 3.3.6. Temporal variation of the absolute deviation of θ.

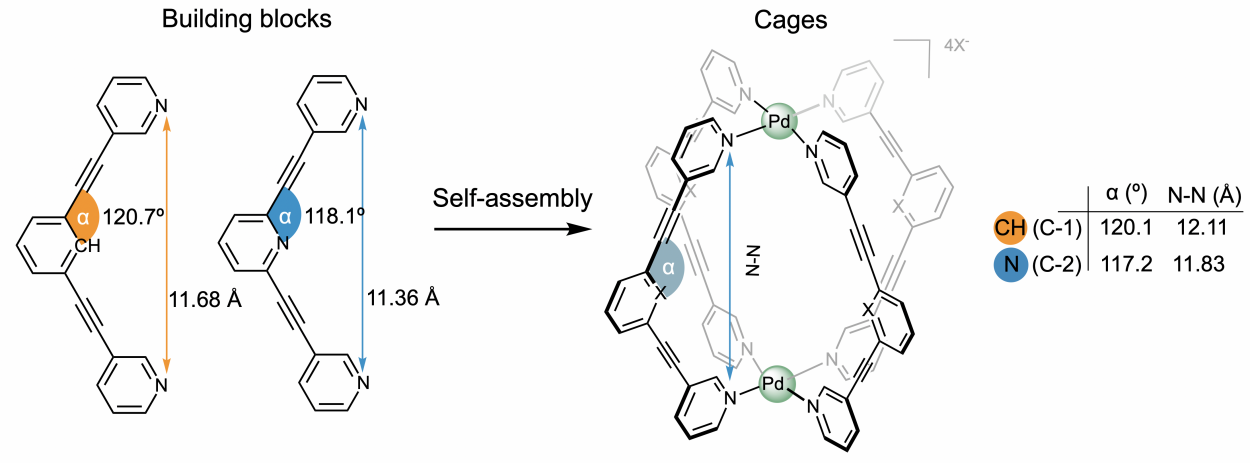


Figure 3.3.7. Illustration of the interior angles on the N-N distance in the D_{4h}-symmetric conformation.

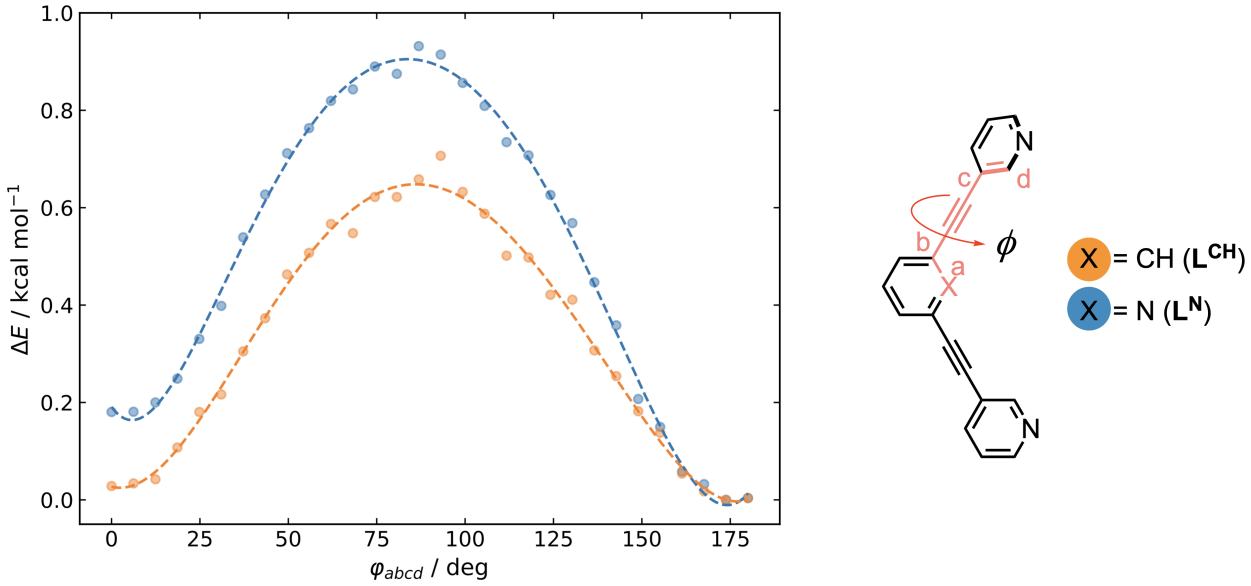


Figure 3.3.8. PES calculated at the M06-2X/def2-TZVP level of theory for rotation about the abcd dihedral for the linker in C-1 and C-2. This very small barrier is retained at a different level of DFT theory (PBE0-D3BJ/def2-TZVP) with $\Delta E^{\ddagger}$= 2.1, 1.3 kcal mol^{-1} for L^{CH} and L^{N} respectively. Note that these barriers are smaller than that for rotation of ethane ($\Delta E^{\ddagger}$ = 2.7 kcal mol^{-1} calculated at the same M2 level, which compares well to the expt. [258] value of 2.9 kcal mol^{-1}).

Opt. Method	B3LYP/DZ	PBE/DZ	PBE-D3BJ/DZ	PBE0/DZ	PBE0-D3BJ/DZ	PBE-D3BJ/TZ	PBE0-D3BJ/TZ	M06-2X/TZ	M06-2X/TZ/SMD(DCM)
ΔE_{bind}(bq⊂C-1)	-19.7	-23.0	-35.2	-27.1	-22.7	-35.5	-29.4	-25.4	-5.5
ΔE_{bind} (bq⊂C-2)	-28.1	-31.1	-43.6	-34.1	-30.9	-43.8	-35.9	-29.5	-3.6
$\Delta\Delta E_{\text{bind}}$(bq)	-8.4	-8.1	-8.3	-7.1	-8.2	-8.3	-6.6	-4.0	2.0
ΔE_{bind}(aq⊂C-1)	-20.4	-26.4	-52.4	-43.8	-26.7	-53.8	-47.1	-40.2	-16.8
ΔE_{bind}(aq⊂C-2)	-21.5	-27.3	-52.6	-43.5	-26.9	-53.4	-45.6	-35.2	-13.1
$\Delta\Delta E_{\text{bind}}$(aq)	-1.1	-0.9	-0.3	0.3	-0.2	0.4	1.5	5.1	3.7
			SMD(DCM)-M06-2X/def2-TZVP//Opt. Method						
ΔE_{bind}(bq⊂C-1)	-4.7	-4.1	-3.3	-4.1	-4.0	-3.7	-4.7	-8.1	-5.5a
ΔE_{bind}(bq⊂C-2)	-5.7	-4.9	-4.3	-5.3	-3.3	-4.4	-4.9	-3.8	-3.6
$\Delta\Delta E_{\text{bind}}$(bq)	-1.0	-0.8	-1.1	-1.2	0.6	-0.7	-0.2	4.2	1.9
ΔE_{bind}(aq⊂C-1)	-8.6	-8.7	-10.3	-10.1	-9.5	-9.3	-11.9	-14.2	-16.8[a]
ΔE_{bind}(aq⊂C-2)	-7.6	-6.8	-7.7	-8.2	-6.8	-7.7	-9.2	-8.3	-13.1
$\Delta\Delta E_{\text{bind}}$(aq)	1.0	1.9	2.6	1.9	2.8	1.6	2.7	5.9	3.7

Table 3.3.3. Effect of utilizing different optimization levels (MX) on ΔE_{bind} and $\Delta\Delta E_{\text{bind}}$ (kcal mol^{-1}). All optimizations apart from PBE0-D3BJ/DZ were initialized from the PBE0-D3BJ/DZ-optimized geometries and single point energies calculated at SMD(DCM)-M06-2X/def2-TZVP where indicated. DZ = def2-SVP, TZ = def2-TZVP. *a*. Optimization of C-1 required loose optimisation criteria and a quasi-Newton step method (rather than the default Rational function).

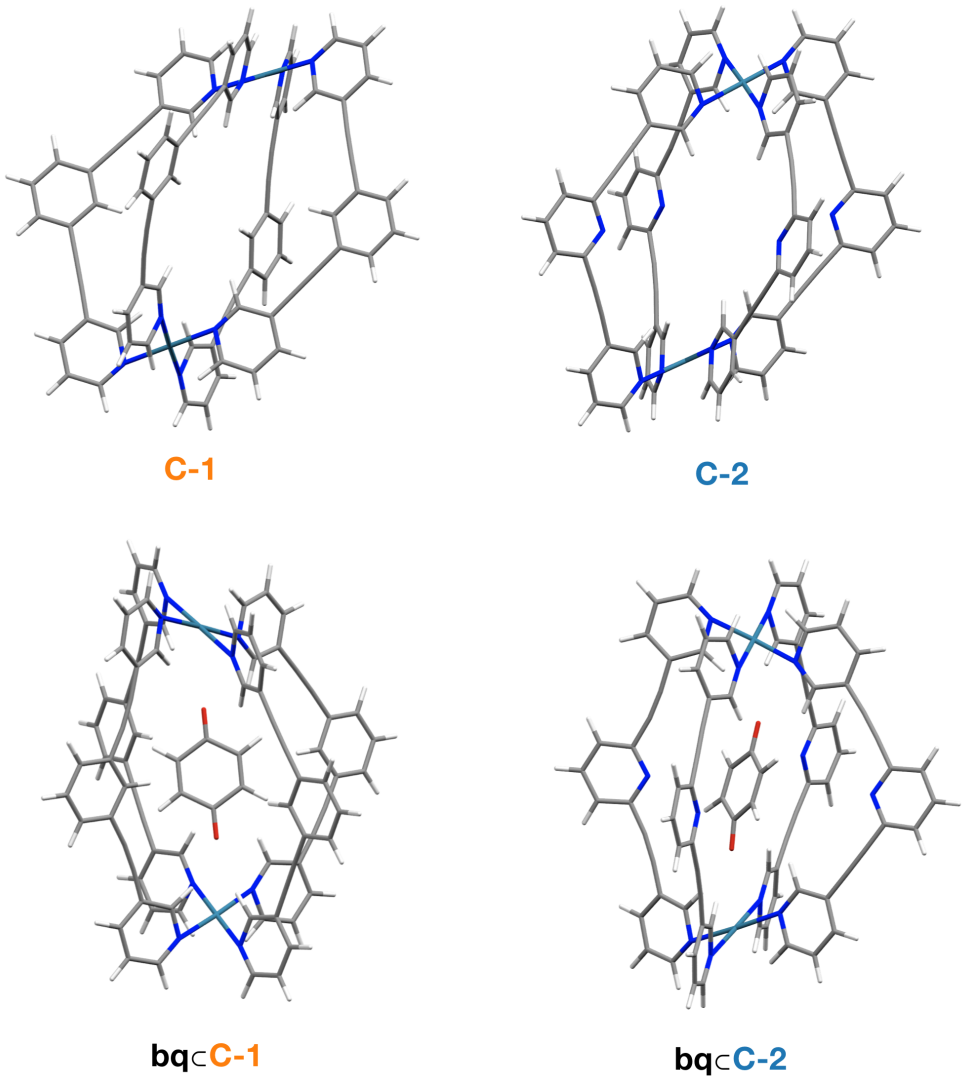


Figure 3.3.9. PM7 optimized geometries of cages and cage-substrate complexes.

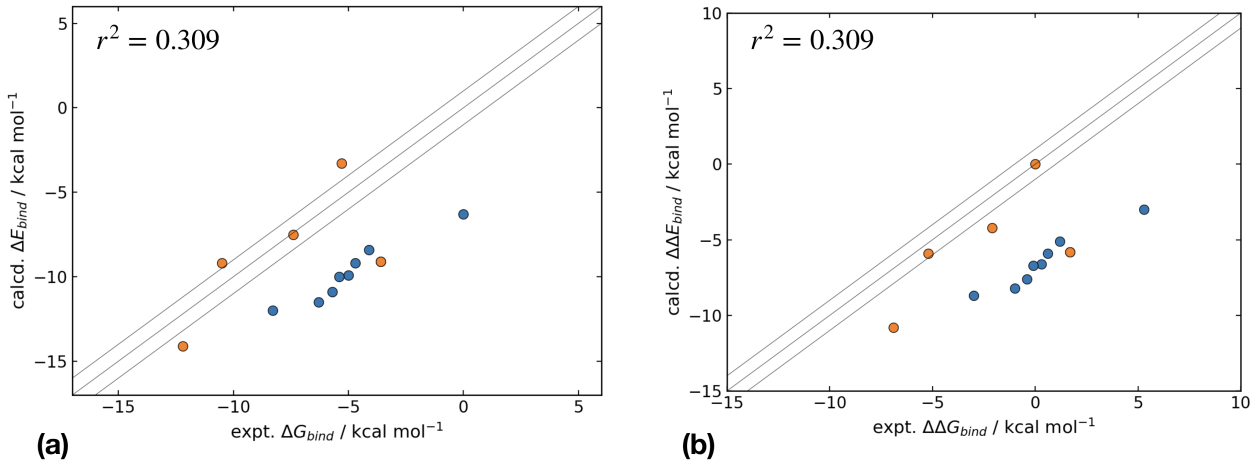


Figure 3.3.10. Correlation plots of (a) ΔE_{bind} and (b) $\Delta\Delta E_{\text{bind}}$ calculated at the PM7 level of theory. Orange and blue markers correspond to C-1 and C-2 cages, respectively. The different diagonals bracket the ± 1 kcal mol^{-1} area of accuracy.

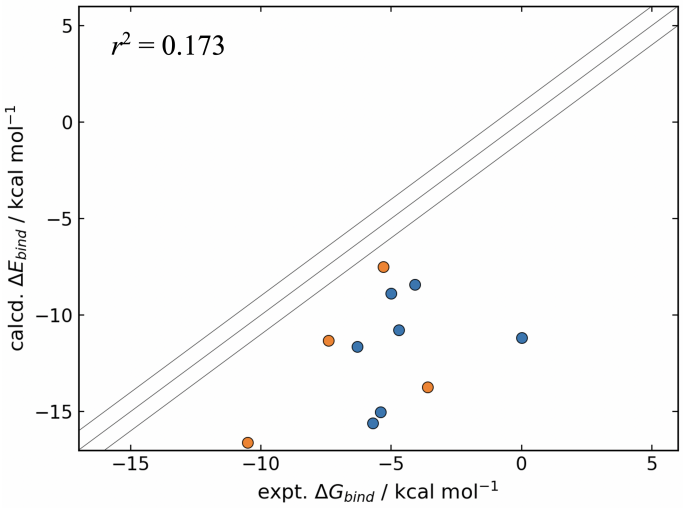


Figure 3.3.11. Correlation plots of ΔE_{bind} calculated at the SMD(DCM)-M06-2X/def2-TZVP//tight-binding DFT (GFN-xTB) level of theory. Orange and blue markers correspond to C-1 and C-2 cages, respectively. The different diagonals bracket the ± 1 kcal mol^{-1} area of accuracy.

3.3.3 Thermodynamic Contributions

While thermodynamic contributions are required to obtain binding free energies (ΔG_{bind}) and to compare to experiments directly, the accurate quantification remains challenging. Furthermore, the accuracy (in principle) gained by including these contributions comes with a much higher computational cost. For large systems, as the one considered here, the computational cost increase is from a few hours to a couple of days.[c]

The free energy of binding qn in C-X is given by,

$$\Delta G_{\text{bind}}(\text{qn} \subset \text{C-X}) = \Delta E_{\text{bind}}(\text{qn} \subset \text{C-X}) + \Delta G_{\text{cont}}(\text{qn} \subset \text{C-X}) \tag{3.1}$$

and the relative free energy of binding qn is defined in,

$$\begin{aligned}\Delta\Delta G_{\text{bind}} &= [\Delta E_{\text{bind}}(\text{qn} \subset \text{C-2}) + \Delta G_{\text{cont}}(\text{qn} \subset \text{C-2})] - [\Delta E_{\text{bind}}(\text{qn} \subset \text{C-1}) + \Delta G_{\text{cont}}(\text{qn} \subset \text{C-1})] \\ &= \Delta\Delta E_{\text{bind}} + \Delta\Delta G_{\text{cont}}\end{aligned} \tag{3.2}$$

where ΔG_{cont} are the thermodynamic corrections to the potential energy difference (ΔE_{bind}) given by the inverse of (3.1).

These corrections include an enthalpic contribution (H_{cont}) containing the zero-point energy (ZPE), thermal energy plus $k_B T$, and an entropic contribution (TS_{cont}), which contains contributions from translation, rotation, vibration and electronic degrees of freedom. Table 3.3.4 outlines these components retrieved directly from frequency calculations from the ORCA package for C-1/C-2, bq and bq⊂C-1/bq⊂C-2. Note that the electronic entropy is

[c]For example, the analytical hessian required to compute the vibrational frequencies for the zero-point energy and vibrational entropy components for bq⊂C-1 took 808 CPUh (M1) while the optimisation (M1) and single point (SMD(DCM)-M2) calculation took 36 CPUh.

zero.

As can be seen in Table 3.3.4, the thermodynamic contributions to the binding energy (ΔG_{cont}) of bq in C-1 and C-2 are large in magnitude but very similar (ΔG_{cont}(bq⊂C-1) = 14.9 kcal mol^{-1}, ΔG_{cont}(bq⊂C-2) = 16.0 kcal mol^{-1}) due to the almost identical enthalpic and entropic contributions. This explains why the calculated relative binding affinities ($\Delta\Delta E_{\text{bind}}$) closely match the experimental binding free energies (i.e. $\Delta\Delta E_{\text{bind}} \sim \Delta\Delta G_{\text{bind}}$). Secondly, the large positive free energy contribution (ΔG_{cont}) to the binding energy for both C-1 and C-2 would suggest a very unfavourable binding process (i.e. ΔG_{bind}(bq⊂C-1) and ΔG_{bind}(bq⊂C-2) > 0), which is contrary to experimental observations. These results give rise to two possibilities, either (1) ΔG_{cont} is incorrect (mostly due to entropic contributions) or (2) ΔE_{bind} is underestimated when calculated using the SMD(DCM)-M06-2X/def2-TZVP methodology. The following sections explain how the former, given the literature available, is the most plausible scenario.

To do so, we analysed the entropic components ($T\Delta S_{\text{total}}$), which is the dominant contrition to ΔG_{cont}). The partition function used to calculate the translational entropy derives from a particle in a box (PIB) treatment of the potential, within an ideal gas model of the system. In this scenario, at standard-state, the effective volume of a molecule is $V = RT/p$ ($p = 1$ atm), which is an overestimation of the 'real' volume available, and therefore an overestimation of the translational entropy upon binding. A more appropriate effective volume can be obtained by moving to a 1 M 'standard state',[259] which when calculated with $TS_{1atm\rightarrow 1M} = RT\ln(c^{\ominus}_{1\text{ atm}}/c^{\ominus}_{1\text{ M}})$ leads to a contribution of 1.9 kcal mol^{-1} at 298 K for each species considered.

This correction reduces the translational entropy term; for a two to one change in molecularity (e.g. A+B → C) this reduction is 1.9 kcal mol^{-1}. Using this correction, ΔG_{cont}(bq⊂C-1)

	H_{cont}	TS_{trans}	TS_{rot}	TS_{vib}	TS_{tot}	G_{cont}
C-1	712.0	14.1	12.8	79.6	106.6	605.5
bq⊂C-1	773.1	14.2	12.9	90.7	117.8	655.3
C-2	681.6	14.1	12.8	79.1	106.0	575.6
bq⊂C-2	742.9	14.2	12.8	89.4	116.5	626.4
bq	58.4	11.9	8.3	3.4	23.5	34.8
Δ(bq⊂C-1)	2.6	-11.8	-8.3	7.8	-12.3	14.9
Δ(bq⊂C-2)	2.8	-11.8	-8.3	7.0	-13.1	16.0
	$\Delta\Delta H_{\text{cont}}$	$T\Delta\Delta S_{\text{trans}}$	$T\Delta\Delta S_{rot}$	$T\Delta\Delta S_{\text{vib}}$	$T\Delta\Delta S_{\text{tot}}$	$\Delta\Delta G_{cont}$
	-0.2	0.0	0.0	0.8	0.8	-1.0

Table 3.3.4. Thermodynamic contributions (kcal mol^{-1}) to the potential energy calculated at the PBE0-D3BJ/def2-SVP level of theory obtained directly from ORCA. ΔX_{bind} = bq⊂C-X – (bq + C-X); X=H, S, G.

= 13.1 kcal mol^{-1} and ΔG_{cont}(bq⊂C-2) = 14.1 kcal mol^{-1} (Table 3.3.5).

	1 atm, qRRHO	1 M, qRRHO	1 M, HO	1 M, HO, real freqs
ΔG_{cont}(bq⊂C-1)	14.9	13.0	11.2	11.6
ΔG_{cont}(bq⊂C-2)	16.0	14.1	12.5	12.3

Table 3.3.5. Free energy contributions (kcal mol^{-1}) calculated at the PBE0-D3BJ/def2-SVP level of theory using different treatments of the translational entropy and low frequency modes.

However, despite this correction, there is still an inconsistency in the treatment of the translational entropy. In solution the translation of bq and C-1 are treated as a particle in a box, while within C-1, the translation of bq is treated by an interpolation of a harmonic oscillator (HO) and rigid rotor (RR), this quasi-RRHO approximation is the default in the ORCA package. While Grimme's quasi-RRHO approximation[230] is suitable for treating

low-frequency normal modes corresponding to rotations, obviously it is inappropriate for treating a translational mode such as those depicted in Figure 3.3.12, which are treated as >90% free rotors within the default weighting scheme used by Grimme and implemented in ORCA (Figure 3.3.13).

Treating all the modes harmonically partially removes the dampening of the low frequency modes contribution to the entropy and again reduces ΔG_{cont}. Using the harmonic approximation along with the 1 M standard stare correction within our otherm code (`https://github.com/duartegroup/otherm`) we obtain $\Delta G_{\text{cont}}(\text{bq}\subset\text{C-1}) = 11.2$ kcal mol^{-1}and $\Delta G_{\text{cont}}(\text{bq}\subset\text{C-1}) = 12.5$ kcal mol^{-1} (Table 3.3.5). Of course, this then removes the more correct treatment of modes which correspond to rotations.

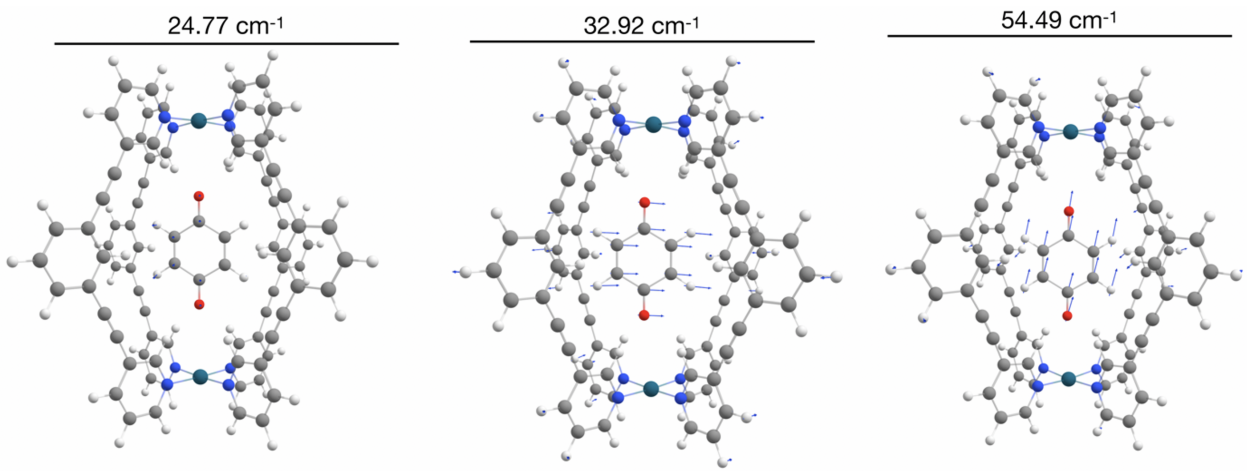


Figure 3.3.12. Normal modes of bq in C-1 dominated by x, y, z translation calculated at M1 annotated with scaled displacement vectors calculated in Chemcraft v. 1.8.

Furthermore, we note that with the exception of bq both cages and cage-substrate complexes contain small imaginary modes (all < 20i cm^{-1}) due to one or more of (1) premature geometry convergence, (2) numerical errors in the Hessian, (3) the resolution of identity (RI-JCOSX) approximation or (4) convergence to a saddle point. This observation is common in large supramolecular structures,[62] and poses another issue when calculating the vibra-

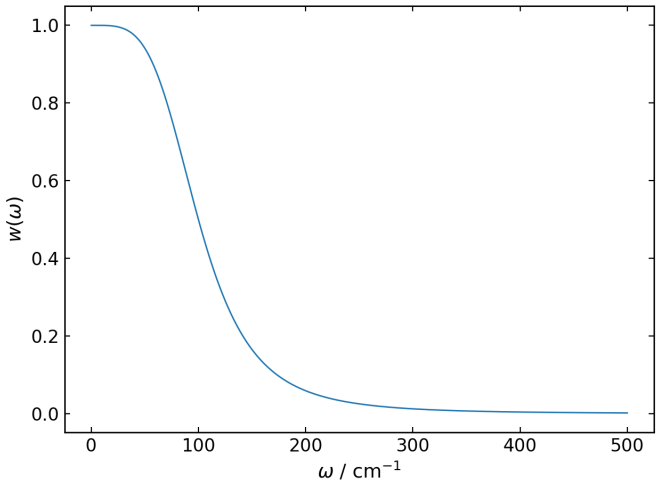


Figure 3.3.13. Weighting utilized within the q-RRHO approximation to interpolate between the free rotor and harmonic oscillator expressions of the entropy as a function of normal mode frequency (ω).

tional contribution to the free energy (both entropic and enthalpic). Harvey et al. proposed treating these modes as real rather than discarding them as both ORCA and Gaussian do by default. Doing so we have ΔG_{cont}(bq⊂C-1) = 11.6 kcal mol^{-1} and ΔG_{cont}(bq⊂C-1) = 12.3 kcal mol^{-1} (Table 3.3.5). These differences are still very similar to the treatment without these modes, due to their being a fortuitous cancellation between the reactant (C-1 + bq, C-2 + bq) and product (bq⊂C-1, bq⊂C-2) as both C-1 and bq⊂C-1 possessed two imaginary frequencies and C-2 and bq⊂C-2 both contained a single imaginary frequency.

These results show that, despite moving from an almost entirely RR approximation to the HO approximation for the low frequency modes and treating the imaginary modes as real, the difference in free energy contribution on binding is still large and positive ($\sim$ 10 kcal mol^{-1}). New methods have been recently developed to improve the estimation of the effective volume based on the HO approximation.[260, 261]

Finally, we note that from the molecular dynamics calculations the cavity is not empty prior to substrate binding, containing in average one solvent molecule in the cavity (Figure 3.3.14).

Himo and Harvey have previously shown that consideration of solvent occupying the internal volume of supramolecular systems is important to reach quantitative results.[62] In fact, considering the cage to be occupied by a single DCM solvent molecule, the binding process can be represented as: DCM⊂C-X + qn → qn⊂C-X + DCM In this case the translational entropy component of the free energy change quenches dramatically ($T\Delta S_{\text{trans}}$(bq⊂C-1)= −0.2 kcal mol^{-1}, Table 3.3.6). This in turn has the effect of reducing the free energy contribution difference by about 15 kcal mol^{-1} for both bq⊂C-1 and bq⊂C-2.

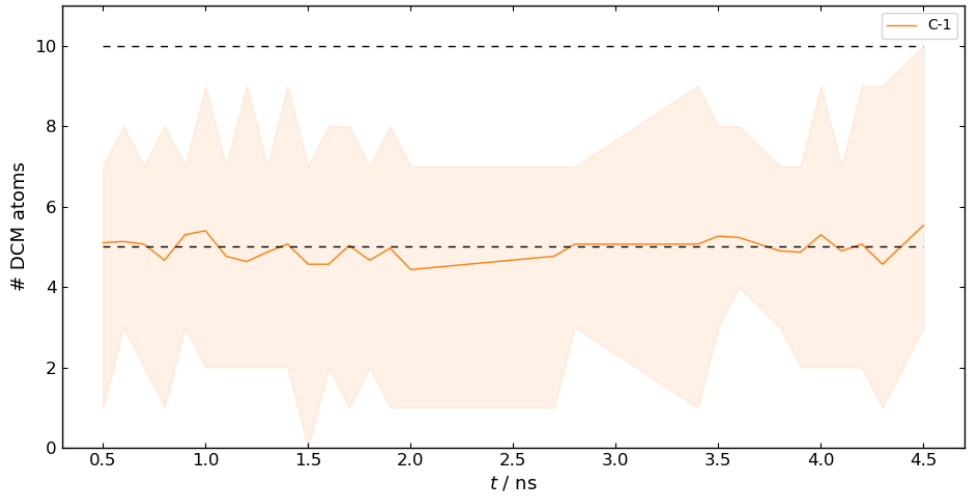


Figure 3.3.14. Temporal variation of the number of DCM solvent atoms ($n = 5$ for one molecule of DCM) within the cavity of C-1 defined by the distance from the Pd–Pd midpoint to the nearest H atom (minus the vdW radius of H). Dashed lines correspond to one and two solvent molecules respectively.

From the proceeding discussion we have shown there to be differences with the treatment of the translational entropy and treatment of the low frequency modes, which dominate the free energy contribution. However, if the cage is considered to be occupied by a single solvent molecule, as confirmed by performing classical MD simulations in explicit DCM solvent, then the problematic treatment of the translational entropy is subject to a cancellation of errors, with no change in molecularity in the reaction: bq + DCM⊂C-X → bq⊂C-X + DCM.

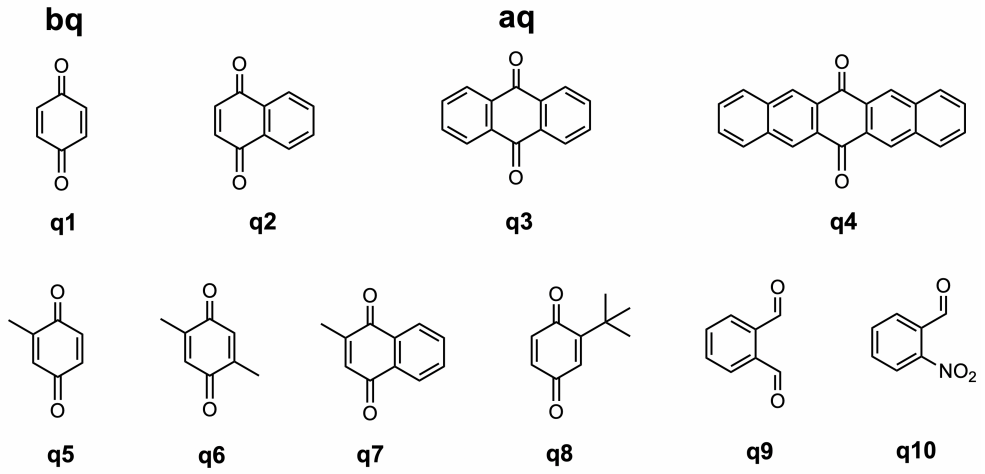


Figure 3.3.15. Quinones utilized in the binding affinity study.

	H_{cont}	TS_{trans}	TS_{rot}	TS_{vib}	TS_{tot}	G_{cont}	E_{bind}	G_{bind}
DCM⊂ C-1	734.9	14.2	12.9	87.4	114.4	620.5		
DCM⊂ C-2	703.8	14.2	12.8	85.2	112.2	591.6		
DCM	21.3	11.7	7.3	0.7	19.6	1.7		
Δ_{bind} (bq⊂C-1)	1.1	-0.2	-1.0	0.7	-0.5	1.6	-3.0	-1.4
$\Delta\Delta_{\text{bind}}$ (bq⊂C-2)	2.0	-0.2	-1.0	1.5	0.3	1.6	-4.8	-3.2
$\Delta\Delta_{\text{bind}}$(bq)	-0.9	0.0	0.0	-0.9	-0.9	0.0	-1.8	-1.8

Table 3.3.6. Thermodynamic contributions (kcal mol^{-1}) to the potential energy calculated at the PBE0-D3BJ/def2-SVP level of theory. Δ_{bind} = (bq⊂C-X + DCM) - (bq + DCM⊂C-X). Energies calculated at the SMD(DCM) M2/M1 level of theory. Note that the geometry optimized structures of DCM⊂C-1 and DCM⊂C-2 contained two and three imaginary modes respectively (all < 20i cm^{-1}).

Cage	Quinone	calcd. ΔE_{bind}	expt. ΔG_{bind}	AE
	q1	-4.0	-5.3	1.3
	q2	-7.5	-7.4	0.1
	q3	-9.5	-10.5	1.0
	q4	-14.6	-12.2	2.4
C-1	q5	-5.3	-	-
	q6	-6.8	-	-
	q7	-8.4	-	-
	q8	8.1	-	-
	q9	-8.6	-3.6	5.0
	q10	-5.0	-3.6	1.4
	q1	-3.3	-4.1	0.8
	q2	-6.1	-4.7	1.4
	q3	-6.8	-5.4	1.4
	q4	-11.2	-8.3	2.9
C-2	q5	-5.3	-5.0	0.3
	q6	-6.6	-6.3	0.3
	q7	-6.4	-5.7	0.7
	q8	5.1	∼0.0	5.1
	q9	-7.1	-	-
	q10	-3.4	-	-
MAE				1.7

Table 3.3.7. Calculated and experimental binding affinities for quinone substrates in cages C-1 and C-2 along with absolute errors (AE), and the mean absolute error. Experimental data from ref. [224] for C-1, and ref. [175] for C-2. this work. Absolute errors (AE) are calculated to experiment. All values in kcal mol^{-1}. $\Delta E_{\text{bind}} = E_{\text{M2/SMD(DCM)}}(\text{qn}\subset\text{C-X}) - [E_{\text{M2/SMD(DCM)}}(\text{C-X}) + E_{\text{M2/SMD(DCM)}}(\text{qn})]$. bq ≡ q1 and aq ≡ q3.

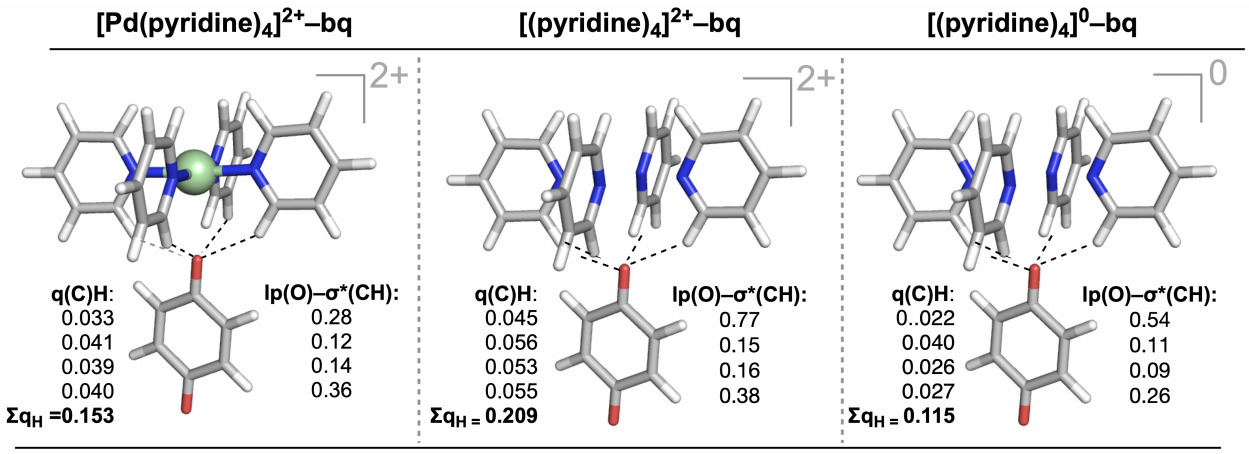


Figure 3.3.16. Comparison of atomic partial charges calculated using the Hirshfeld scheme and NBO second order perturbation decomposition calculated at the M06-2X/def2-TZVP level of theory. Charges are quoted in e and interaction energies in kcal mol^{-1}.

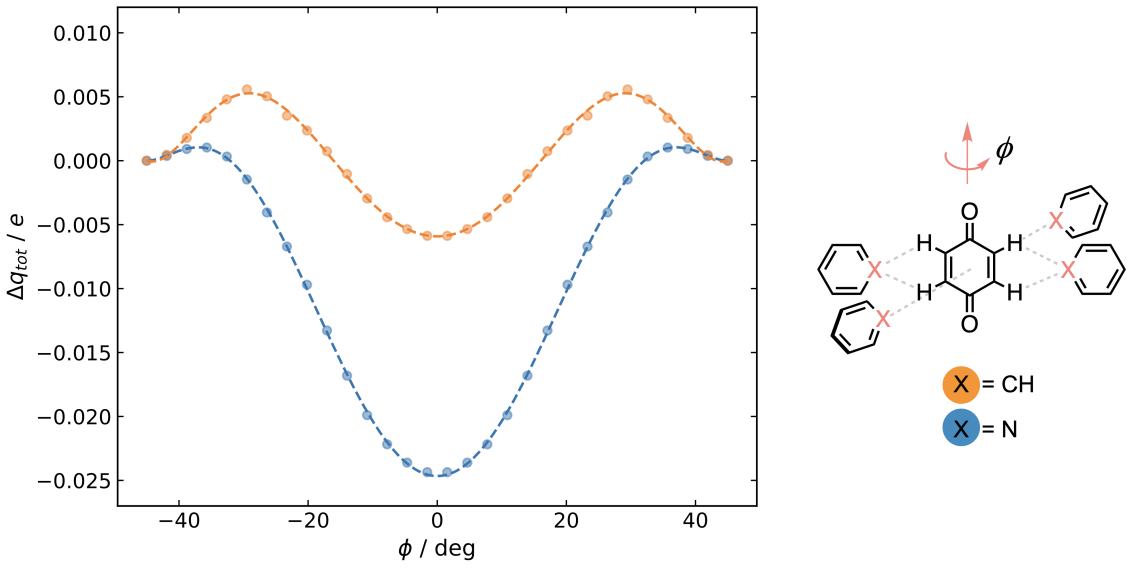


Figure 3.3.17. Total partial atomic charge on bq relative to $\phi = -45°$ (Δq_{tot}) calculated with the Hirschfeld scheme at the M06-2X/def2-TZVP level of theory as a function of rotation around the z axis (ϕ). $q_{\mathrm{tot}} = \sum_i q_i$, where i enumerates quinone atoms.

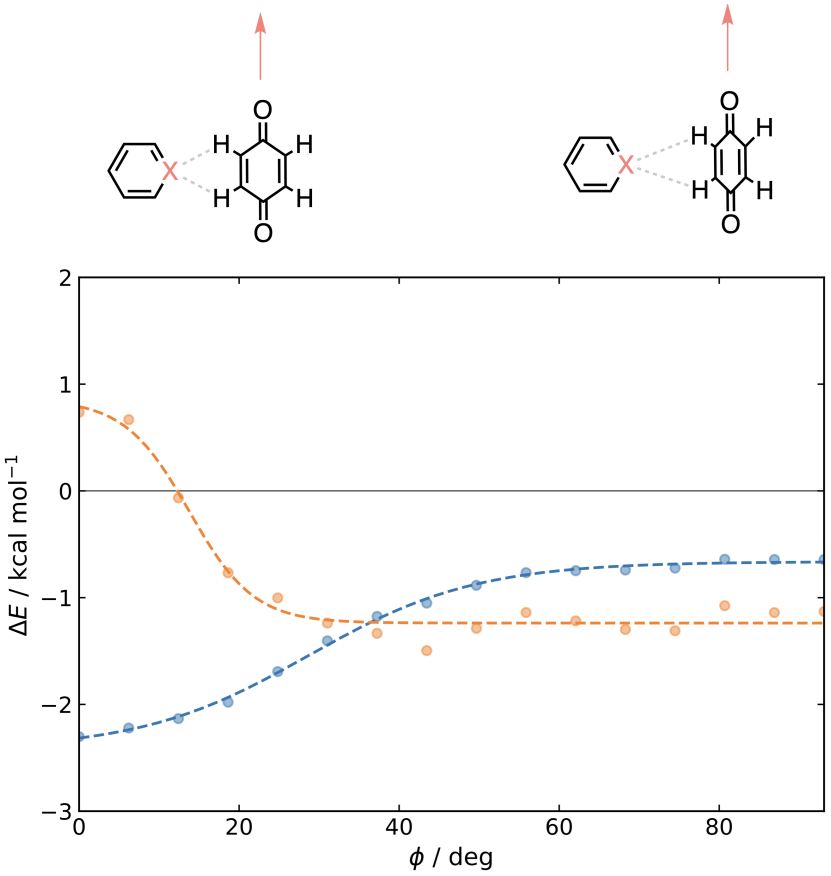


Figure 3.3.18. Relative energy (kcal mol^{-1}), calculated at the M06-2X/def2-TZVP level of theory, as a function of ϕ for a bq⊂benzene (orange) and bq⊂pyridine (blue) system.

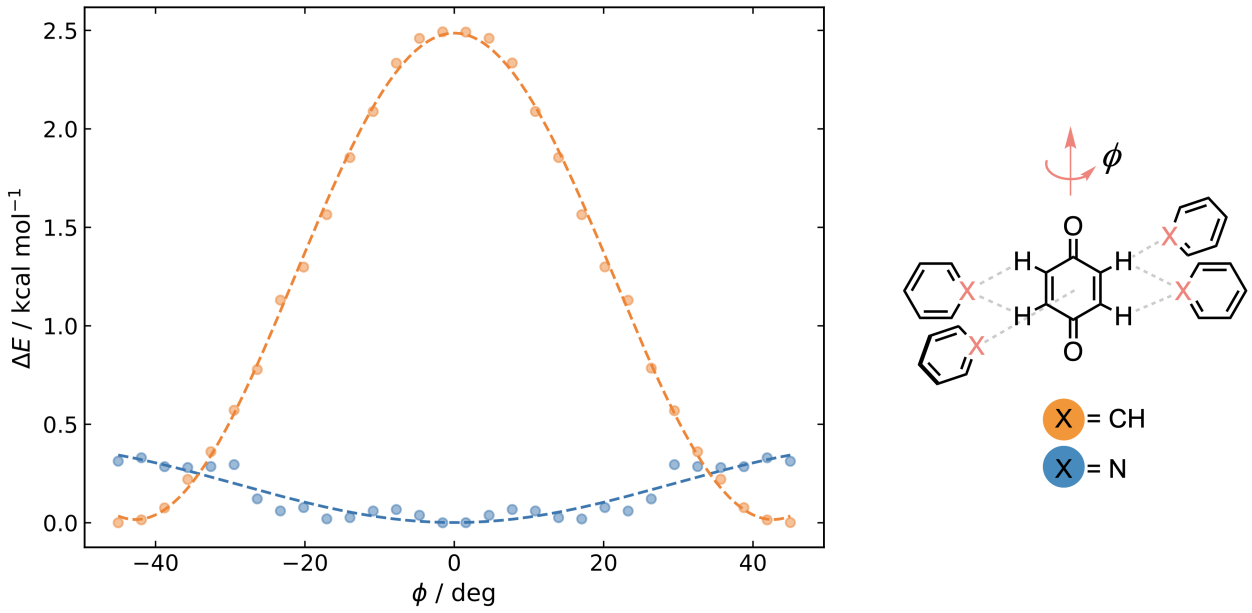


Figure 3.3.19. Relative energy (kcal mol^{-1}), calculated at the SMD(DCM)-M06-2X/def2-TZVP level of theory, as a function of ϕ for a bq⊂benzene (orange) and bq⊂pyridine (blue) system. Implicit solvent considerably reduces the pyridine–bq interaction while the Pauli repulsion is not modulated by the solvent.

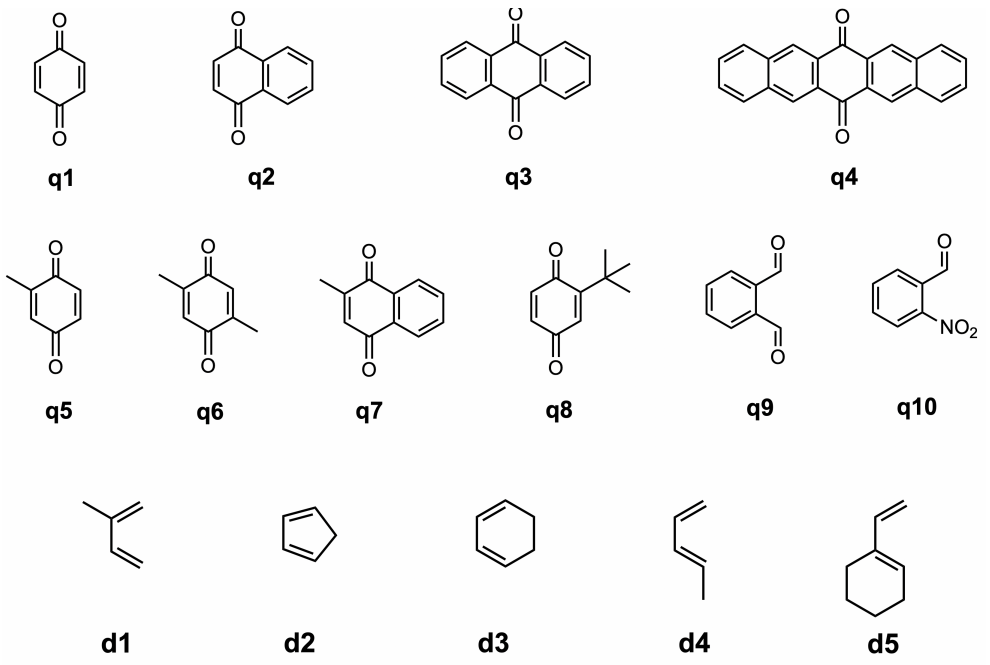


Figure 3.3.20. Quinones and dienes for a representative set of 10 uncatalysed Diels–Alder (DA) reactions reported in ref. [175].

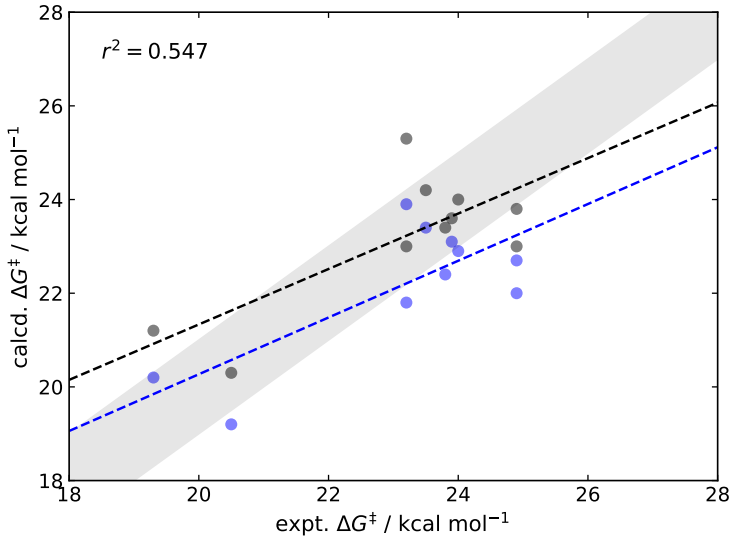


Figure 3.3.21. Correlation plot of the gas phase $\Delta G^{\ddagger}$ (black) and solvated $\Delta G^{\ddagger}_{\text{solv.}}$ (blue) activation energies for the uncatalysed endo-DA reactions (data reported in Table 3.3.8). Correlation coefficients are identical. Grey diagonal area indicates the ± 1 kcal mol^{-1} area of accuracy.

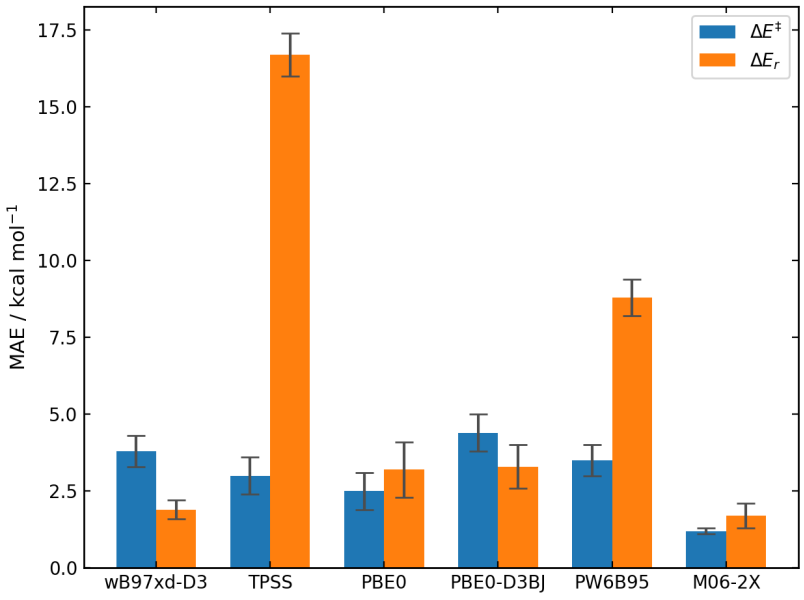


Figure 3.3.22. Mean absolute deviations to DLPNO-CCSD(T)/ma-def2-TZVPP potential energy differences for reaction barriers ($\Delta E^{\ddagger}$), and reaction energies (ΔE_r) for a set of 10 Diels Alder reactions outlined in Table 3.3.3. Error bars are std. error in the mean.

Diene	Dienophile	$\Delta E^{\ddagger}$	$\Delta H^{\ddagger}$	$\Delta G^{\ddagger}$	$\Delta G^{\ddagger}_{\text{solv}}$	$\Delta G^{\ddagger}_{\text{expt}}$
d1	q1	11.0	11.9	24.0	22.9	24.0
d1	q5	10.7	11.5	23.8	22.7	24.9
d1	q8	9.8	10.6	23.0	22.0	24.9
d2	q1	8.4	9.4	21.2	20.2	19.3
d2	q2	7.2	8.2	20.3	19.2	20.5
d2	q6	10.9	11.9	24.2	23.4	23.5
d2	q7	10.2	11.2	23.6	23.1	23.9
d3	q1	12.5	13.5	25.3	23.9	23.2
d4	q5	9.8	10.6	23.4	22.4	23.8
d5	q1	9.0	9.9	23.0	21.8	23.2

		ΔE_r	ΔH_r	ΔG_r	ΔG_{solv}
d1	q1	-45.0	-41.1	-28.1	-27.3
d1	q5	-45.1	-41.2	-28.1	-27.3
d1	q8	-45.1	-41.2	-27.9	-27.1
d2	q1	-27.5	-23.9	-11.2	-10.4
d2	q2	-27.2	-23.6	-10.7	-10.1
d2	q6	-25.1	-21.7	-8.4	-7.3
d2	q7	-24.4	-21.0	-7.7	-6.4
d3	q1	-34.4	-30.8	-17.8	-16.9
d4	q5	-41.2	-37.3	-23.7	-22.7
d5	q1	-41.0	-37.1	-23.4	-22.5

Table 3.3.8. Gas phase and solvated($\Delta G^{\ddagger}_{\text{solv}}$) activation energies (kcal mol^{-1}) for the uncatalysed endo-DA reactions calculated at the DLPNO-CCSD(T)/ma-def2-TZVPP//PBE0-D3BJ/def2-SVP level of theory with thermodynamic and solvation contributions calculated at PBE0 D3BJ/def2-SVP with the 1M standard state correction, and SMD(DCM)-PBE0 D3BJ/def2 SVP/ respectively.

Diene	Dienophile	$\Delta E^{\ddagger}$	$\Delta H^{\ddagger}$	$\Delta G^{\ddagger}$	$\Delta G^{\ddagger}_{\text{solv}}$
d1	q1	29.2	26.3	-25.2	-28.1
d1	q5	29.4	26.2	-25.1	-28.1
d1	q8	29.0	25.1	-25.1	-28.9
d2	q1	27.2	24.2	-7.0	-9.4
d2	q2	26.5	21.5	-6.2	-10.6
d2	q6	30.1	27.0	-4.3	-6.9
d2	q7	29.1	24.3	-3.4	-7.4
d3	q1	32.4	29.2	-13.8	-16.7
d4	q5	28.8	26.1	-23.1	-25.8
d5	q1	28.2	23.7	-23.9	-28.2

Table 3.3.9. Gas phase and solvated($\Delta G^{\ddagger}_{\text{solv}}$) activation energies (kcal mol^{-1}) for the uncatalysed exo-DA reactions calculated at the DLPNO-CCSD(T)/ma-def2-TZVPP//PBE0-D3BJ/def2-SVP level of theory with thermodynamic and solvation contributions calculated at PBE0 D3BJ/def2-SVP with the 1M standard state correction, and SMD(DCM)-PBE0 D3BJ/def2 SVP/ respectively.

First, the most stable species or set of species available to the C-2, bq, isoprene in DCM system is found, from which the TS likely occurs. We assume that a pre-equilibrium is established quickly, with all energetic barriers being smaller than the for the DA reaction between isoprene and bq. Due to the conformational flexibility of L^{CH}/L^{N}, other arrangements of the reactants inside the cage, including binding of both the diene and dienophile is also possible. However, they were found to have higher energies both at the complex and TS stage.

The increased electrophilicity of the quinone within the cage can be quantified by the total effective charge of bq within C-1/C-2 (relative to its neutral form in solution). Within C-1/C-2 bq has a slightly positive charge – indicating that the +4 cage removes some electron density from bq, thereby activating it as a dienophile. This effect is slightly more pronounced in C-1 across the different decomposition schemes.

Species	ΔE	ΔH	ΔG
bq⊂C-2	0.0	0.0	0.0
d1⊂C-2a	3.7	2.3	1.3
bq.d1⊂C-2	-6.4	-4.4	8.8

Table 3.3.10. Energies of possible complexes relative to bq⊂C-2.

To assess the effect the C-H$\cdots$O interactions play in enhancing the electrophilic character an electron density difference map (EDDM) is generated between the electron density of bq⊂C-2 and the electron density of bq (neutral) and C-2 (4+ charge) at the PBE0-D3BJ/def2-SVP level of theory. The resultant EDDM is shown in Figure 3.3.23 and shows how the C(2)–Hs present in the donor pyridines remove electron density away from the carbonyls.

For the full catalytic cycle presented in the manuscript the corresponding enthalpy and free energies are presented in Table 3.3.11 and the corresponding structures in Figure 3.3.25. Note that the free energies are much too positive, as the resting state is taken to be an empty

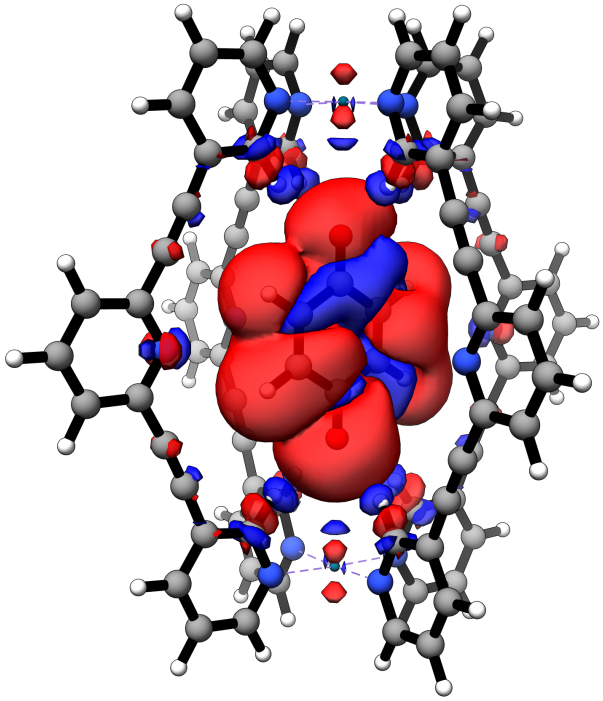


Figure 3.3.23. Electron density difference map for bq⊂C-2 showing regions depleted (red) and enhanced (blue) in electron density when the interaction between cage and substrate is activated. Density difference plotted at an isovalue of 0.001 au.

cage, rather than occupied by a solvent molecule, thus there is a change in molecularity and a dominant translational entropy component. Energy differences balanced against DCM⊂C-2 are also presented in Table 3.3.11.

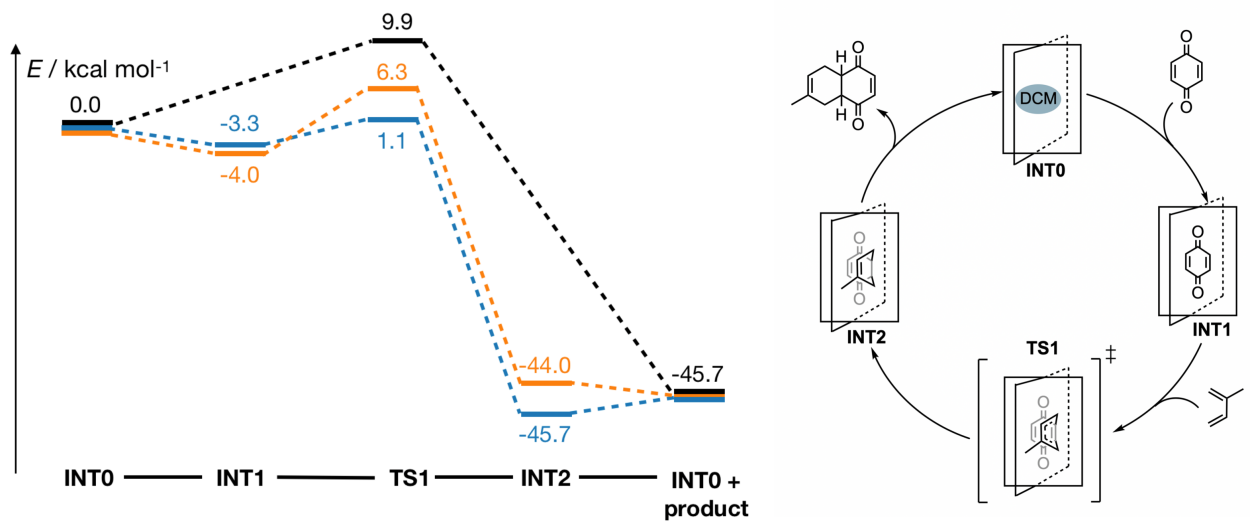


Figure 3.3.24. Reaction profile for the C-1/C-2 (orange/blue) encapsulated DA reaction between benzoquinone and isoprene.

		Relative to empty cage			Relative to DCM⊂C-X		
Cage	Species	ΔE	ΔH	ΔG	ΔE	ΔH	ΔG
	INT0	0.0	0.0	0.0	0.0	0.0	0.0
	INT1	-4.0	-1.9	9.8	-10.2	-9.7	-7.9
C-1	TS1	6.3	9.9	33.4	0.0	2.1	15.7
	INT2	-44.0	-37.9	-11.6	-50.2	-45.7	-29.3
	INT0 + product	-45.7	-41.7	-28.7	-45.7	-41.7	-28.7
C-2	INT0	0.0	0.0	0.0	0.0	0.0	0.0
	INT1	-3.3	-0.5	10.7	-7.3	-5.4	-5.7
	TS1	1.1	4.9	29.6	-2.9	0.0	13.1
	INT2	-48.2	-41.4	-16.6	-52.2	-46.2	-33.0
	INT0 + product	-45.7	-41.7	-28.7	-45.7	-41.7	-28.7
Cage	Species	ΔE	ΔH	ΔG	$\Delta H^{\ddagger}$	$\Delta H^{\ddagger}$	$\Delta G^{\ddagger}_{\text{expt}}$
uncat	TS1			9.9	10.7	23.5	24.0
C-1	TS1			10.2	11.8	23.6	>.0
C-2	TS1			4.4	5.4	18.8	20.4

Table 3.3.11. Energetics for the full cycle for the uncatalysed and catalysed endo-reaction between benzoquinone and isoprene in the presence of C-1 and C-2. All calculated at the SMD(DCM)-M06-2X/def2-TZVP//PBE0-D3BJ/def2-SVP level of theory. Free energies are in the 1 M standard state. Experimental data from in ref [175].

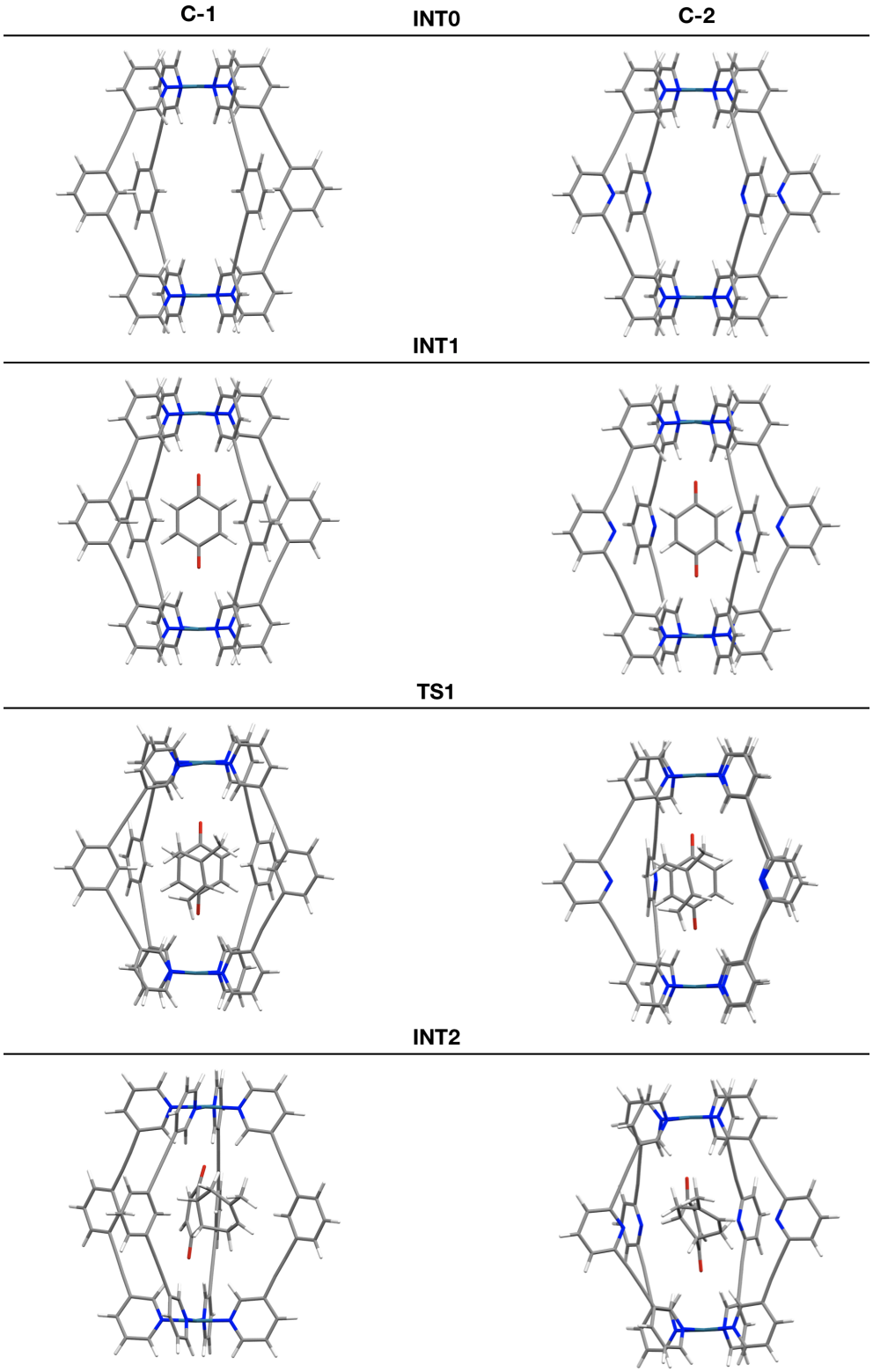


Figure 3.3.25. Optimized structures of the minima and TSs along the catalytic cycle at the at the PBE0-D3BJ/def2-SVP level of theory

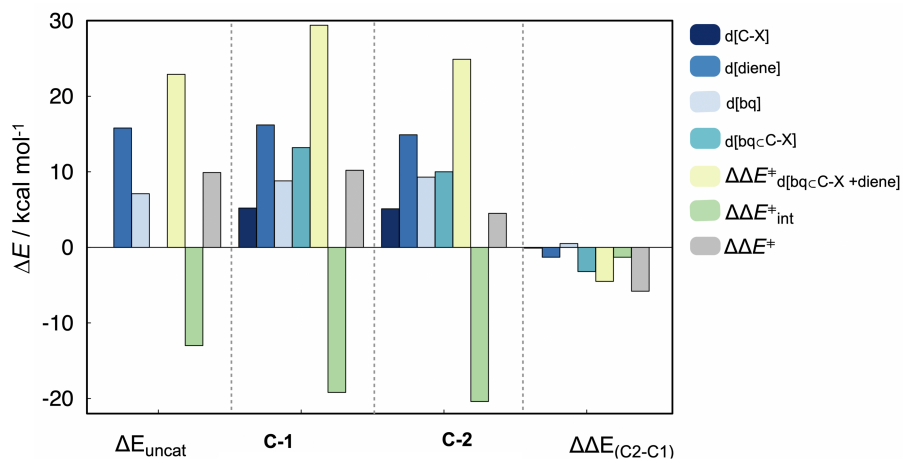


Figure 3.3.26. Distortion interaction analysis.

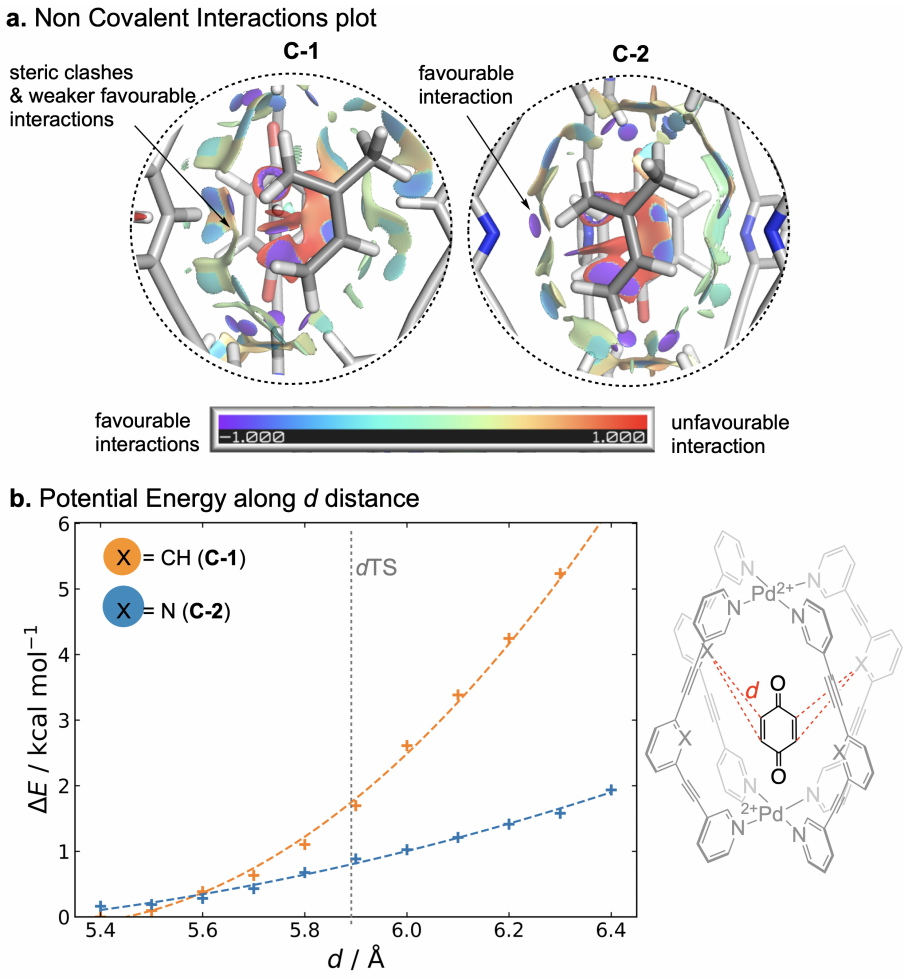


Figure 3.3.27. (a) Non-covalent interaction (NCI) isosurface (0.05 au) at the TS geometry for C-1 and C-2. An identifiable CH···N interaction (purple circles) between the CH group of the diene and the catalyst nitrogen is observed in C-2. NCI calculated at the PBE0-D3BJ/def2-SVP level of theory. (b) Potential energy as a function of distance from 2 rearmost X (C/N) atoms to C_αs on bq calculated at the PBE0-D3BJ/def2-SVP level of theory. Distance at the TS is indicated with a vertical grey line.

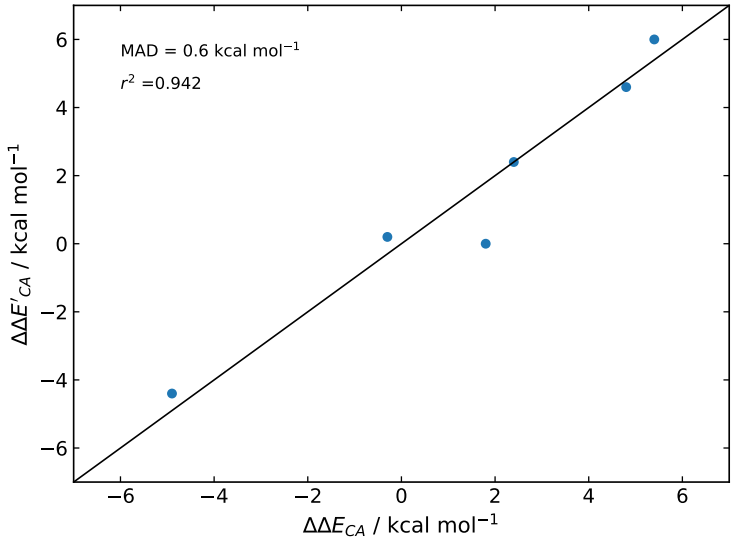


Figure 3.3.28. Calculated catalytic activity ($\Delta\Delta E_{\mathrm{CA}} = \Delta E^{\ddagger}_{\mathrm{uncat}} - E^{\ddagger}_{\mathrm{cat}}$) using fully characterised transition states ($\Delta\Delta E_{\mathrm{CA}}$) and transition state analogies ($\Delta\Delta E'_{\mathrm{CA}}$).

3.4 Manuscript II

cgbind: A Python Module and Web App for Automated Metallocage Construction and Host-Guest Characterization

3.4.1 Abstract

Metallocages offer a diverse and underexplored region of chemical space to search for novel catalysts and substrate hosts. However, the ability to tailor such structures towards applications in binding and catalysis is a challenging task. Here, we present an open-source computational toolkit, *cgbind* , that facilitates the construction, characterization and prediction of functional metallocages. It employs known structural scaffolds as starting points and computationally efficient approaches for property evaluation. We demonstrate the ability of *cgbind* to construct libraries of cages with varied topologies and linker functionalities, generate accurate geometries (RMSD <1.5 Å to crystal structures) and predict substrate binding with accuracy on-par with semi-empirical QM, all in seconds. The *cgbind* code presented here is freely available at `github.com/duartegroup/cgbind` and also via a web-based graphical user interface at `cgbind.chem.ox.ac.uk`. The protocol described here paves the way for high-throughput virtual screening of potential supramolecular structures, accelerating the search for new hosts and catalysts.

3.4.2 Introduction

Metal-mediated self-assembly has emerged as a promising approach to develop complex biomimetic systems from simple building blocks. The synthetic modularity and tunability, achievable by modifying the metal centre or organic linkers, has led to the discovery of

an immense diversity of supramolecular metallocages with applications in molecular separation[262, 263] medical imaging,[264, 265] drug-delivery[174, 204, 266–269] and catalysis.[187] Moreover, the dynamic nature of metal–ligand bonds, intermediate in strength between covalent bonds and weak non-covalent interactions (NCI), facilitates error correction and reversibility.[270, 271] The latter being more challenging to achieve with covalent organic cages.[272]

Despite these attractive attributes, the development of new metallocages and their use in real-life applications has been slow, and in many cases driven by trial-and-error rather than rational design. To date, most reported binding studies have been limited to a relatively small set of guests and host combinations. This is primarily due to current synthetic techniques being time-consuming and expensive; while also limited in environmental sustainability.[273] To fully exploit their potential, the development of efficient methods to identify structures with desirable binding or catalytic properties, without having to synthesize all possible structures, is necessary.

Computational screening offers a path to overcome these limitations and accelerate the discovery process.[6] While reaching chemical accuracy ($<$1 kcal mol^{-1} error), generally required for quantitative comparison to experiments remains a challenge,[274] current computational and cheminformatic approaches already allow a fast sift of the desired space. For example, molecular descriptors[275–277] and inverse design methods[45] can be used to select (with reasonable accuracy) promising regions of the chemical space for further computational and eventually experimental study.[46] Similarly, efficient electronic structure methods, in particular density functional theory (DFT), have enabled the calculation of structure and properties for hundreds of compounds in a single study.[33, 278] While these protocols are accessible for small organic molecules and periodic structures, extension of similar ap-

proaches in supramolecular assemblies has been much less explored.[279] We have recently shown that an efficient DFT-based approach can reach a ∼2 kcal mol^{-1} level of accuracy in quantifying binding affinities for M_2L_4 metallocage assemblies.[280] However, with a sufficiently fast computational approach, the required time to set up and analyse a new set of metallocages rapidly increases as the size of the building block library grows. Therefore, methods to automate these steps are becoming increasingly valuable.

In the field of metal-organic frameworks (MOFs), several computational screening tools have been developed in recent years.[281, 282] They include open-source tools to screen existing MOF libraries (Pymatgen[283] and Zeo++[284]), deconstruct known structures in their building blocks,[285] and evaluate their catalytic proficiency for specific reaction (PyMOFScreen[286]). In the space of supramolecular metallocages, however, no large already synthesized libraries exist; thus, tools that both generate and rank structures for desired properties are required. For porous organic cages, *in silico* toolkits have been developed for constructing and predicting their adsorption properties via a combination of building blocks and high-throughput screening (e.g. stk, pywindow, Eigencages).[176, 277, 287–290], Current toolkits to generate structures are, however, not directly transferable to metal-containing systems. Therefore, a different methodology is required to further enable high-throughput analysis of supramolecular metallocages. Here, we describe *cgbind*, an open-source Python API and its web-based graphical user interface (GUI). We first describe the methodology and its implementation and proceed to demonstrate the ability of *cgbind* to explore the space of metallocages by combinatorial cage generation from a building block library. Furthermore, we show how *cgbind* can be used to generate accurate geometries and obtain estimated binding affinities.

3.4.3 Results and Discussion

3.4.3.1 Software Outline

cgbind is written in object-oriented Python and employs scientific programming libraries, including Numpy, Scipy and NetworkX.[291, 292] The code is built around four key objects (`Linker`, `Cage`, `Substrate` and `CageSubstrateComplex`), the relation and inheritance between which, their attributes, and key methods are outlined in Figure 3.4.1.

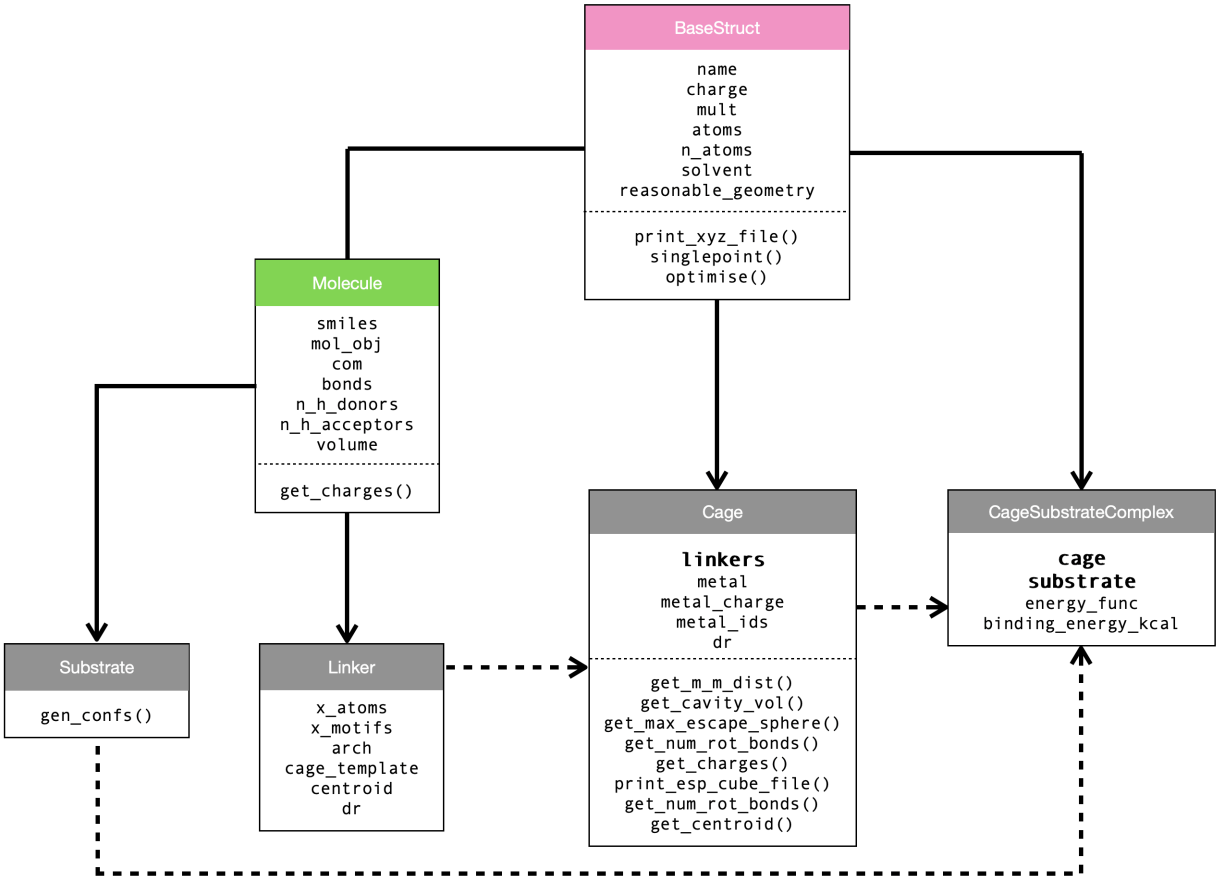


Figure 3.4.1. Class structure of *cgbind*. A solid line from A to B indicates that B inherits from A and a dashed line B is built from A.

A homoleptic `Cage` (M_xL_y with x metals and y linkers) is initialised from a single `Linker` object, while heteroleptic (e.g. $M_2L_2L'_2$) variants from a set of different Linkers. All `Molecule`

objects and their subclasses are designed to be initialized from SMILES strings,[293] with their corresponding 3D structures generated using conformer generation algorithms implemented in RDKit.[294, 295] While this affords rapid access to 3D structures, conformers are occasionally not adequately generated (see below). Molecules can also be initialized from a variety of molecular file formats (.xyz, .mol, .mol2) using the input_output module. Upon `Cage` construction a number of analysis tools are available as class methods. The analysis of cage-substrate complexes (`CageSubstrateComplex` objects) is also possible. They are built from a `Cage` and one `Substrate`, where a substrate may contain more than one molecule, thus multiple guests are supported. Furthermore, a `Substrate` can either be a minimum or a transition state (/analogue), providing a route to predicting catalytic activity. Binding affinities between a substrate (S) and cage (C) [$\Delta E_{\mathrm{BA}} = E_{\mathrm{C\text{-}S}} - (E_{\mathrm{C}} + E_{\mathrm{S}})$] are calculated in the first instance using a simple and general forcefields developed here (see below). More accurate estimates can be obtained from electronic structure codes, by invoking `singlepoint()` and `optimise()` functions, which deliver energies (`Molecule.energy`) and optimized structures (`Molecule.atoms`). These functions are currently handled through a `Calculation` object in our Python API *autodE*,[296] which facilitates the automated generation of reaction profiles. This enables *cgbind* to utilize electronic structure packages with currently implemented wrappers (inc. GFN2-XTB,[297] MOPAC,[298] ORCA[246] and NWChem[299]).

The structure of *cgbind* benefits from modularity, allowing analysis functions to be implemented by simply adding methods to the `Cage` or `CageSubstrateComplex` objects. The fragment approach to generating a `CageSubstrateComplex` allows for all the properties of the cage, and constituent linkers to be contained within a single object. Linker, Cage and Substrate objects may therefore be reloaded from a saved (e.g. Pickled) `CageSubstrateComplex` object, allowing for data mining within a saved library. Independent calculations are parallelized up to the user-defined number of cores set in the `Config` object. The software is

distributed under the MIT license and is provided with unit tests and comprehensive online documentation including examples.

3.4.3.2 Metallocage Construction

Broadly, metallocage structures are generated from a SMILES or 3D representation of the linkers, which are templated onto known or user defined metallocage templates. The methodology employed is not template specific, such that extension to architectures other than those provided with the code (M_2L_4, M_4L_6, M_6L_8, $M_{12}L_{24}$) is straightforward. Figure 3.4.2 provides an overview of the workflow behind *cgbind*, which can be broken down into six key steps:

I. Generate n linker (L) conformers from a SMILES string, a common way of representing molecules, which can be obtained from online databases (e.g. PubChem) or chemical software (e.g. ChemDraw™).

II. The donor atoms bonded to the metal atoms (X-atoms) are located in L.

III. Locate possible X-motifs, as all combinations of one or more X-atoms and their nearest neighbours in L.

IV. Discard any X-motifs with a different number of atoms to the template X-motifs.

V. Minimize the fitting cost with respect to linker and template parameters. The fitting cost is defined as the sum squared distance (SSD) between the X-motif atoms in the template and L.

VI. Fit linkers to the expanded/contracted template, controlled by a distance Δr, while keeping steric clashes below a threshold.

These steps are discussed in detail below for two architectures, M_2L_4 and M_4L_6, already

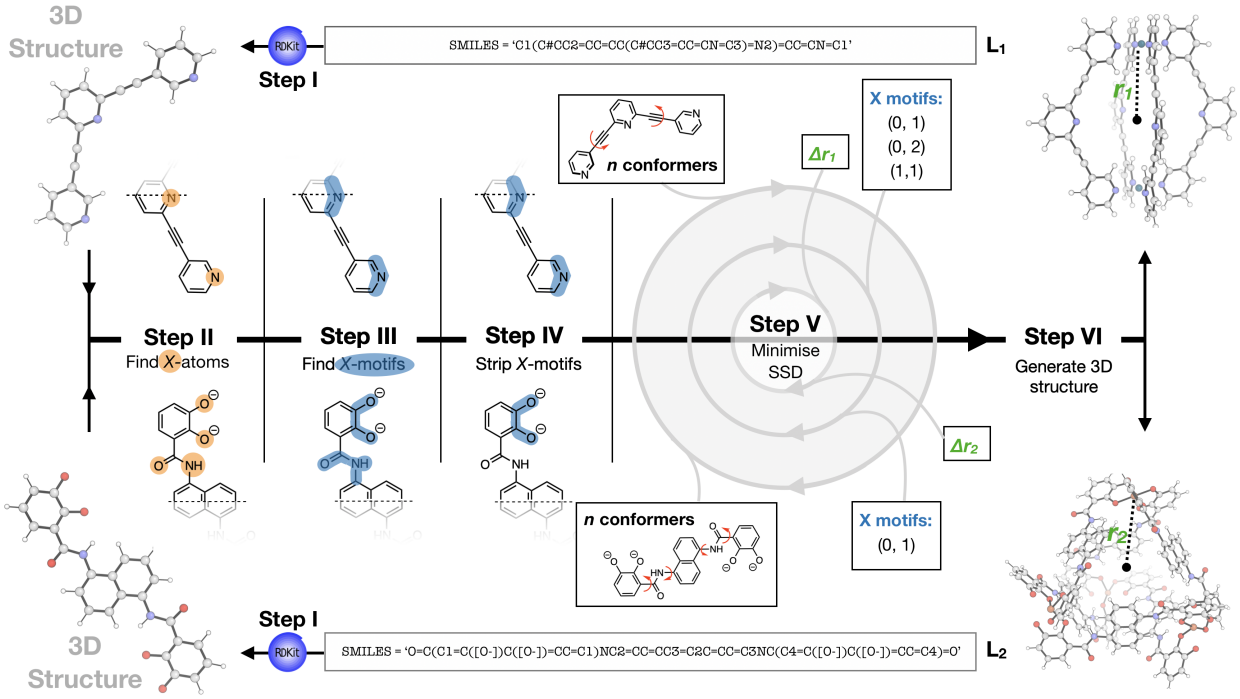


Figure 3.4.2. Methodology used to construct metallocages in *cgbind* . Centroid–metal distances are denoted as r and shifts Δr. X-motifs are indexed from zero to the total number in the linker.

saved in the template library. Note that, depending on the input provided by the user, not all steps must be executed. For example, if a `Linker` is initialized from a 3D geometry then the Step I can be skipped. Moreover, if the X-motifs are specified by the user then steps III and IV may be skipped, as well as minimization with respect to X-motif sets in Step V.

For a bis-pyridyl linker found experimentally to form a M_2L_4 assembly (L_1, top Figure 3.4.2),[175] the first step involves generation of 3D conformers using the Experimental-Torsion Distance Geometry with Knowledge (ETKDG) algorithm[295] implemented in RDKit[294] (Step I). For this linker, requesting 300 conformers and using a 0.3 Å root mean square displacement (RMSD) threshold to compare them, renders 50 conformers. This provided a good balance between speed and high probability of locating a suitable conformer. Subsequently, X-atoms are identified as the heteroatoms with at least one 'lone

pair' capable of metal donation (here the nitrogen atoms, Step II). Combining X-atoms and their nearest neighbours affords three X-motifs containing CNC atoms (Step III). In the M_2L_4 template, the X-motifs also contain three atoms (see §3.4.3.3 for how these motifs are found); therefore, all X-motifs are kept (Step IV). Minimization of the fitting cost is then performed exhaustively with respect to all conformers and X-motif sets, and minimization with respect to expansion and contraction of the template is achieved using the Broyden–Fletcher–Goldfarb–Shanno (BFGS) optimisation algorithm implemented in *SciPy*[292] (Step V). The template is expanded/contracted by shifting the X-motifs closest to metal i by the M_i–centroid vector a distance Δr. For L_1, there are three X-motifs (top: 0, middle: 1 and bottom: 2, with 0 and 2 being equivalent), while there are only two X-motifs in the template linker; therefore, the three X-motifs in L_1 can be fitted in three different ways, i.e. (0, 1), (0, 2) and (1, 2). X-motif atoms are fit to the template using an implementation of the Kabsch algorithm.[300] Step V correctly identifies the linker with the most coplanar pyridyls as the optimum conformer, the two terminal CNC atom sets (0, 2) as the optimum selection of X-motifs and a small shift required in the template ($\Delta r_1 = 0.0037$ Å). Upon sequentially fitting the optimum X-motifs in the linker to the adjusted template and adding two metal ions, a metallocage geometry is generated (Step VI).

In an M_4L_6 structure formed from identical catechol-derived linkers and Fe(II) (L_2, Figure 3.4.2),[188] steps I/II proceed as for L_1, while identification of X-motifs is slightly more complex. X-atoms are joined by their nearest neighbours to generate X-motifs for C-O-, C=O and CN(H)C and those separated by a one bond joined generating -OCCO- and O=CN(H)C motifs to give a total of 12 motifs in L_2 (Step III). The template contains X-motifs with four atoms, thus only the -OCCO- X-motifs remain upon discarding those with greater and fewer than four atoms (Step IV). Steps V and VI proceed as for L_1 but requires no enumeration over X-motif sets as both L_2 and the template contain two X-motifs. This strategy is amenable

to heteroleptic cage generation; different linkers are generated within the same architecture and fitted in the order defined by the template. The template shift distance Δr is taken as an average over all linkers, as to accommodate linkers with (slightly) differing lengths. For an experimentally accessible *cis*-$M_2L_2L_2$' metallocage[213] the process is shown in Figure 3.4.3. We note that the trans isomer may be generated by simply reordering the linkers in the `Cage` initialization (i.e. `linkers=[linker1, linker2, linker1, linker2]` in Figure 3.4.3).

For a linker with n conformers, j X-motifs in the template and i metallocage X-motifs, the minimisation is conducted in a $\sim n \times {}^{i}C_{j}$ dimensional space. However, despite this large space, generally it is the conformer generation in RDKit, rather than minimization which dominates the execution time. Construction of cages with non-planar and or conformationally flexible linkers is facilitated by checking on the ordering of X-motifs and linker conformer as to achieve a threshold in the linker–partial cage repulsion during the build algorithm.

3.4.3.3 Adding New Template Architectures

Extensibility to new architectures is an essential feature of the code. The inclusion of new architectures by the user is achieved as follows: A 3D structure file (.mol2) is used for initialization of a `Template`. This structure can be downloaded from the CSD, or provided by the user (from non-deposited crystal structures or e.g. Spartan[301] generated constructs).

I. Utilizing NetworkX a molecular graph of the system is generated.

II. A metallocage is identified as the molecular component with the largest number of metal atoms. All other molecules (ions, substrates) are stripped out; this step is required if the template is initialized from a crystal structure.

```python
from cgbind import Linker, Cage, xtb, Config

# Use 8 cores throughout the construction and optimisation
Config.n_cores = 8

# Initialise a cage from two linkers to generate a M2L2L'2 cage
linker1 = Linker(smiles='O=C1C2=C(C=CC(C#CC3=CC=C(C)N=C3)=C2)NC4=C1C=C(C#CC5=CN=C(C)C=C5)C=C4',
                 name='L_A', arch_name='m2l4')
linker2 = Linker(smiles='O=C1C2=C(C=CC(C#CC3=CC=CN=C3C)=C2)NC4=C1C=C(C#CC5=C(C)N=CC=C5)C=C4',
                 name='L_B', arch_name='m2l4')

# Construct a metallocage, optimise with XTB and print the geometry as a .xyz file
cage = Cage(linkers=[linker1, linker1, linker2, linker2], metal='Pd', metal_charge=2)
cage.print_xyz_file()
cage.optimise(method=xtb)
```

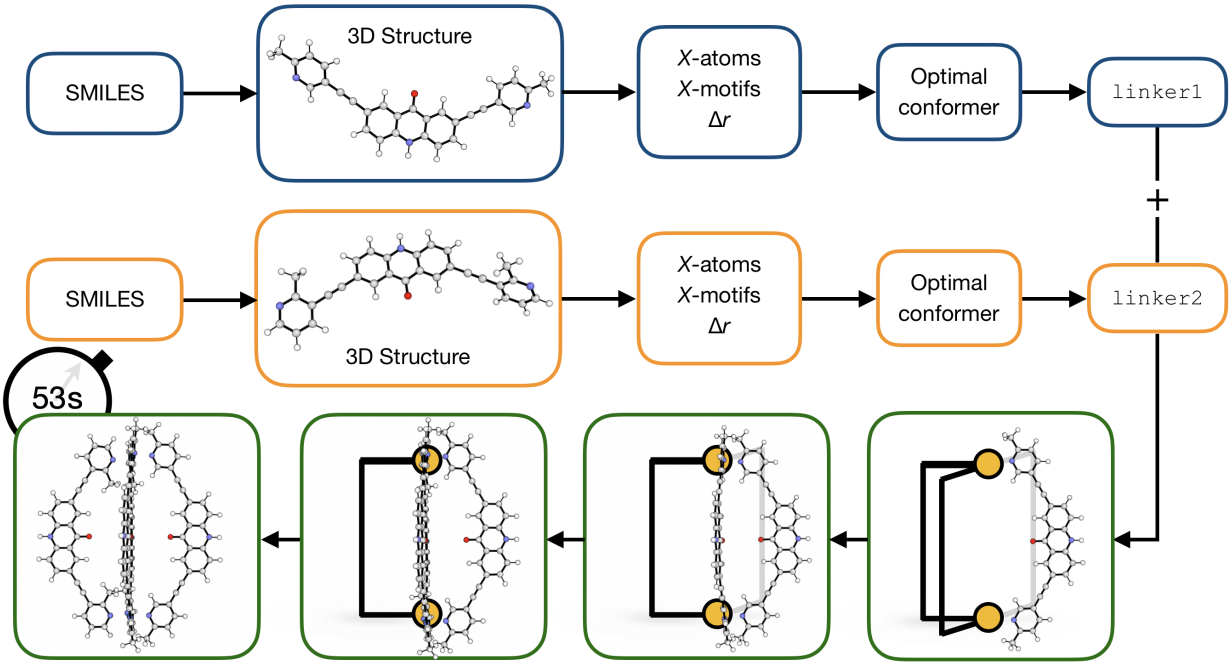


Figure 3.4.3. Formation of a heteroleptic cage within *cgbind* . Total execution time includes XTB optimization.

III. The donor (X-) atoms bonded to the metal atoms are located.

IV. Linkers are identified by deleting metal atoms and determining its connected components using NetworkX.

V. Within the linkers, and from the previously located X-atoms, X-motifs are located as the maximally connected set of X-atoms and their nearest neighbours.

VI. The shift vector for each X-motif is calculated to enable symmetric expansion of the template controlled by Δr as above.

Overall, this workflow facilitates the addition of novel architectures into the code with minimal human intervention (i.e. a single command) and without structures needed to be hardcoded.

3.4.3.4 Analysis Tools

Upon generation of a metallocage structure, *cgbind* provides several geometry-based parameters that can be used to analyse and select metallocages for further binding and/or catalysis studies (Figure 3.4.4a). These include: (1) Number of rotatable bonds, which may be used as a proxy for cage flexibility (by default, the M-L bonds are not included). (2) Total number of H-bond donors/acceptors, which can be used as a proxy for hydrophilicity. (3) Cage size metrics including (i) Average M–M distance. (ii) Maximum enclosed sphere: defined as the largest sphere centred on the cage centroid that does not contain any atomic (van der Waals) volume. The cage centroid is defined as the mean position of the metal atom centres. (iii) Maximum escape sphere: defined the largest sphere that may be removed from the cavity. This volume is determined by maximizing the sphere size at a distance r from the cage centroid and taking the minimum sphere size along the path to the outside of the cage.

The latter process can be performed with a basin-hopping algorithm from the *SciPy* library, which provides a more accurate estimate of the extrema but is computationally demanding.[302] If the cage is symmetric, initializing with a random vector generally affords the largest escape sphere. However, for asymmetric cages, the use of a basin-hopping algorithm is necessary. This methodology is similar to that employed in pywindow[290] and may be

used as a proxy for window size, and consequently, substrate diffusion rates into/from the interior of the cage.

Finally, (4) electrostatic potentials (ESP), which can be used to identify favourable guest-host complexes by maximising electrostatic interactions between them. By interfacing *cgbind* with the open-source tight binding DFT code XTB,[297] ESPs can be obtained in seconds in the standard Gaussian cube format and visualized in PyMOL or other graphical software (3.4.4b). This is achieved by performing a single point energy evaluation on the structure, retrieving the partial atomic charges (Mulliken population analysis) and constructing the ESP on a uniform grid as $\mathrm{ESP}(\boldsymbol{R}) = \sum_i q_i/|\boldsymbol{R} - \boldsymbol{R}_i|$ in atomic units, where i enumerates over atoms. Note that this approximation does not result in any loss in qualitative accuracy compared to the true integral over the electron density (Figure 3.5.1).

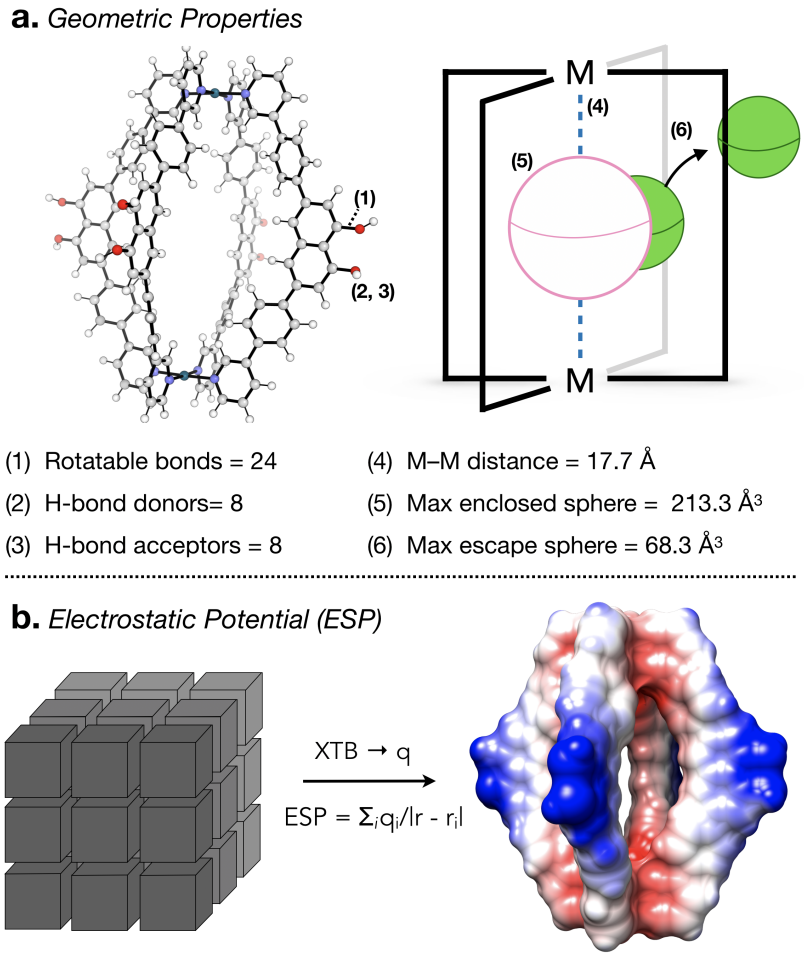


Figure 3.4.4. Calculated metallocage properties for an archetype M_2L_4 cage. (a) Geometrical properties: (1) rotatable bonds, (2) H-bond donors, (3) metal-metal distance, (4) maximum enclosed sphere and (5) maximum escape sphere. (b) Electrostatic potential (ESP) calculated from XTB partial atomic charges enumerated over a grid where blue/red corresponds to low/high values of the ESP.

3.4.3.5 Substrate Encapsulation and Binding Affinity Prediction

Once a metallocage is constructed, a substrate can be added to generate a cage-substrate complex. This is achieved by adding the substrate such that its centre of mass (COM) is located at the centroid of the cage in a similar way to ridged protein docking. The system is then energy minimized with respect to the substrate rotation and conformer, i.e. the global minimum on the rigid-body PES for each conformer is located (Figure 3.4.5). To obtain the global minimum a simple (random + BFGS) optimization algorithm is employed,[292] requiring $\sim n \times m \times o$ energy evaluations, where n is the number of conformers, m the number of initial rotations and o the number of BFGS optimization steps (~ 20). We have found that $m = 50$ affords the global minimum with probability ~ 0.8 for the systems tested (Figure 3.5.2). The number of conformers to screen is of course substrate dependent and is, as such, user specified. Considering the large number of energy evaluations that would be required to evaluate binding of each conformer, electronic structure theory is not viable. To speed up this process we instead resort to empirical forcefields (FF). While a number of general FFs are available, their use is labour-intensive, requiring the tabulation of parameters and/or the use of specific molecular simulation packages. Furthermore, inherent approximations in their development limit their accuracy. To circumvent and simplify this problem, and assuming that substrate-cage interactions are dominated by sterics, we employed the following empirical energy function,

$$E_{\text{int}}(r+k) = \sum_{i,j}^{\text{pairs}} c \exp(a - |\boldsymbol{R}_i - \boldsymbol{R}_j|/b) + k \times S_{\text{atoms}} \tag{3.3}$$

where the first half is a pairwise repulsive term, similar to the one found in a Buckingham potential, with $c = 1$ and with units of energy. Using noble gas dimers as as a model for

closed shell repulsion, we found that a and b are roughly linearly dependent on the sum of van der Waals radii (Figure 3.5.3). The second term is a constant attractive contribution (k) based on the number of substrate atoms (S_{atoms}). $k = 0.4$ kcal mol^{-1} atom^{-1} was derived as the optimum classifier for the set of binary binding data (see full SI). This approximation is considerable, as the attractive terms are dependent on the nature of non-covalent interactions (NCI) between the substrate and the cage (dispersion, H-bonding, etc.). Therefore, we only suggest using this FF to generate cage-substrate complexes where there the attractive term is expected to be small. We refer to this method as 'repulsive' ($r + k$) and the binding energies arising from it as $E_{\text{int}}(r + k)$ (`energy_method='repulsion'`).

We have also considered a more physical FF, which utilises repulsive and electrostatic terms. The latter is calculated using either Gasteiger atomic partial charges from RDKit, which do not account for polarization of the atomic centres due to the surrounding environment (fe, `energy_method='electrostatic_fast'`), or alternatively, charges derived from an XTB single point calculation which account for polarization effects
(e, `energy_method='electrostatic'`). These methods more accurately account for NCI between the substrate and the cage. For example, within $[Pd(II)_2L_4]^{4+}$ metallocages, both schemes deliver the expected binding modes, which involves hydrogen bond interactions between the quinone oxygen and the C_α-H groups at the top/bottom of the cage (Figure 3.4.5).

Quantifying the binding affinity of a given cage for a substrate is key in determining their applicability in real-world applications. In *cgbind* , once the cage-substrate complex has been generated, an approximate gas-phase interaction energy (E_{int}) based on a simple FF above is available immediately, as it is minimised in the construction. This gives a rough estimate of the binding affinity and the likelihood of a cage to bind/not-bind a given substrates. A more

```python
from cgbind import Cage, Linker, Substrate, CageSubstrateComplex, xtb

# Generate a cage and optimise with XTB
linker = Linker(name='L', smiles='C1(C#CC2=CC=CC(C#CC3=CC=CN=C3)=C2)=CC=CN=C1', arch_name='m2l4')
cage = Cage(linker, metal_charge=2, metal='Pd')
cage.optimise(method=xtb)

# Generate a substrate and optimise with XTB
substrate = Substrate(name='benzoquinone', smiles='O=C1C=CC(C=C1)=O')
substrate.optimise(method=xtb)

# Build the substrate@cage complex
cs_complex = CageSubstrateComplex(cage, substrate, energy_method='electrostatic_fast')
cs_complex.print_xyz_file()
```

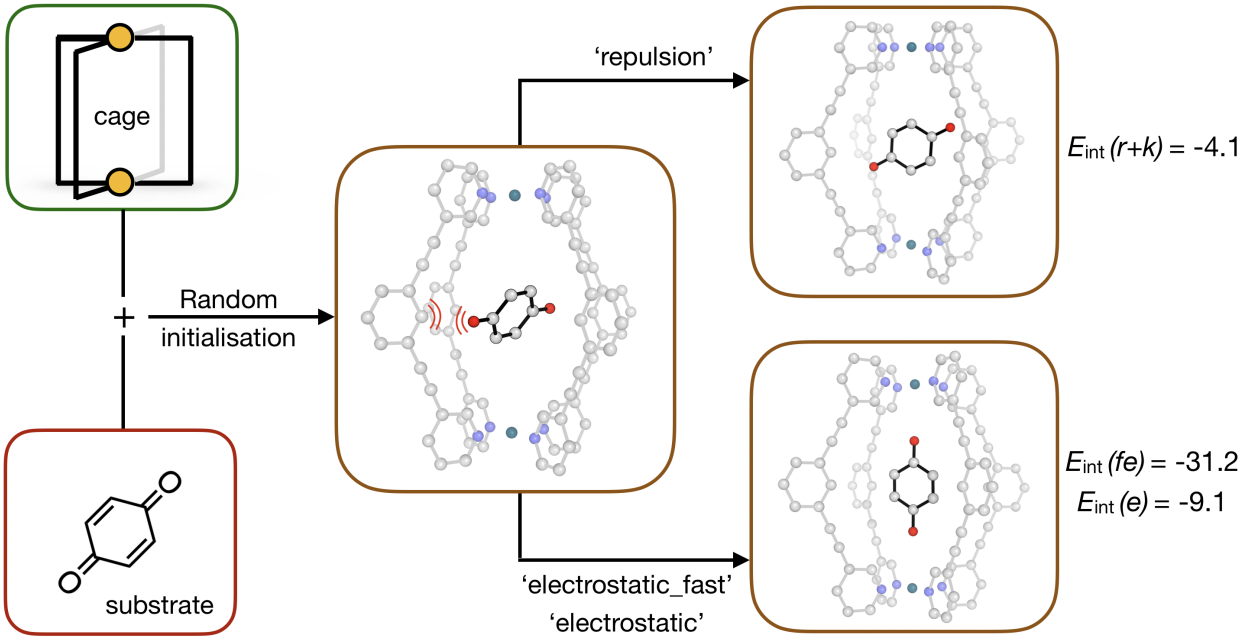


Figure 3.4.5. Metallocage-substrate complex construction and interaction energy calculation for benzoquinone inside a $[Pd(II)_2L_4]^{4+}$ cage using the simple repulsive (r+k) and electrostatic forcefields (fe and e). Interaction energy predictions are quoted in kcal mol^{-1} and do not include solvent, thermal or entropic contributions. Hydrogen atoms omitted for clarity.

accurate estimate can be obtained from electronic structure theory calculations, calculating either potential or Gibbs free energies in implicit solvent. The former we recently showed is sufficient to obtain quantitative agreement with experimental binding free energies for $[Pd_2L_4]^{4+}$ metallocages.[280] Key sources of error in the calculation of free energies include inadequate quantification of entropy, sampling and errors in the methodology used.[303] Sol-

vent effects can be accounted for implicitly using implicit solvent models implemented in electronic structure theory packages and are essential in modulating interaction energies, particularly in polar solvents. Moreover, *cgbind* provides a starting point for free energy calculations required to treat explicitly solvated cavities. The free energy for cavity desolvation and subsequent substrate binding can then be obtained using all-atom molecular dynamics (MD) methods or by adding explicit solvent molecules to the cavity and using the ideal gas model (available in most electronic structure codes).

3.4.3.6 Web app GUI

Alongside the Python interface we have also developed a web application as a GUI to the module. This approach allows a user to interact with key components of the code within a web browser, thereby reducing the activation barrier to its utilisation. Furthermore, by developing a web app the user does not need to download or install any packages/dependencies, which in many cases can be limited to specific platforms (Windows/Mac/Linux). The app is currently available at `cgbind.chem.ox.ac.uk` and provides the ability to: (a) generate and visualize a selection of metallocages, (b) perform analysis of metallocage size, (c) compute and visualize electrostatic potential maps, and (d) perform substrate binding calculations. SMILES strings of linkers can be pasted from chemical software packages, or generated in the *cgbind* GUI using the 2D drawing tool Kekule.js.[304] Interactive molecular visualization of the generated structures is enabled using the open-source 3Dmol.js package, allowing useful actions such as rotating and zooming.[305] The whole app is distributed in a microservice architecture [front-end (Flask), web-server (Gunicorn/nginx), message-broker (Redis), database (PostgreSQL) and task queue (Celery)] using Docker.

3.4.3.7 Examples

Geometric Accuracy. To explore the ability of *cgbind* to generate cages with varied topologies and linker functionalities we generated nine metallocages (a–i, Figure 3.4.6) and computed the root mean squared displacement (RMSD) of the heavy atoms to known crystal structures. With the exception of L_h (see below) all the metallocages were successfully generated from just the linker SMILES string all with RMSD < 1.5 Å. Geometry optimization of the structures at the tight-binding DFT level did, in some instances, improve RMSDs while for others (d, g, i) a larger deviation was obtained with the structure falling into a different local minimum. Similarly, optimization at the DFT level of theory (PBE-D3BJ/def2-SVP) provides no improvement in RMSD. This observation highlights the importance of accounting for the conformational flexibility of the linker and cage when constructing and analysing metallocages and their properties. We believe that this conformational variability provides an interesting avenue to explore substrate-dependent metallocage structure and function as well as a variety of allosteric effects.

To generate a cage, the algorithm requires a conformer with donor atoms in the correct orientation. While conformational generation algorithms available in RDKit (ETKDG, ETDG) are generally adequate, for L_h (Figure 3.4.6), we found that they do not afford a conformer with the torsion parameters required to generate a reasonable M_6L_8 metallocage, even with 10^4 conformers requested. Similarly, the open-source conformer generation implemented in OpenBabel (CONFAB),[306] which performs well at generating structures close to that found in the crystal structure, does not afford the correct conformer from the RDKit initialized 3D structure.[307] We note that the required conformer of L_h is 0.6 kcal mol^{-1} more stable than the first RDKit generated analogue (Figure 3.5.5) and therefore should be generated in the conformer ensemble. This highlights the limitation of current conformer-generation

algorithms. Therefore, Cage h was constructed from linker L_h which was built manually in the correct conformation and a Linker object initialized from an xyz file. Nevertheless, we consider our methodology to be robust in generating metallocages for varying architectures and multiple possible donor atoms. We also note that *cgbind* provides reasonable structures in seconds rather than in hours (XTB) or weeks (DFT) for the largest metallocages shown in Figure 3.4.6.

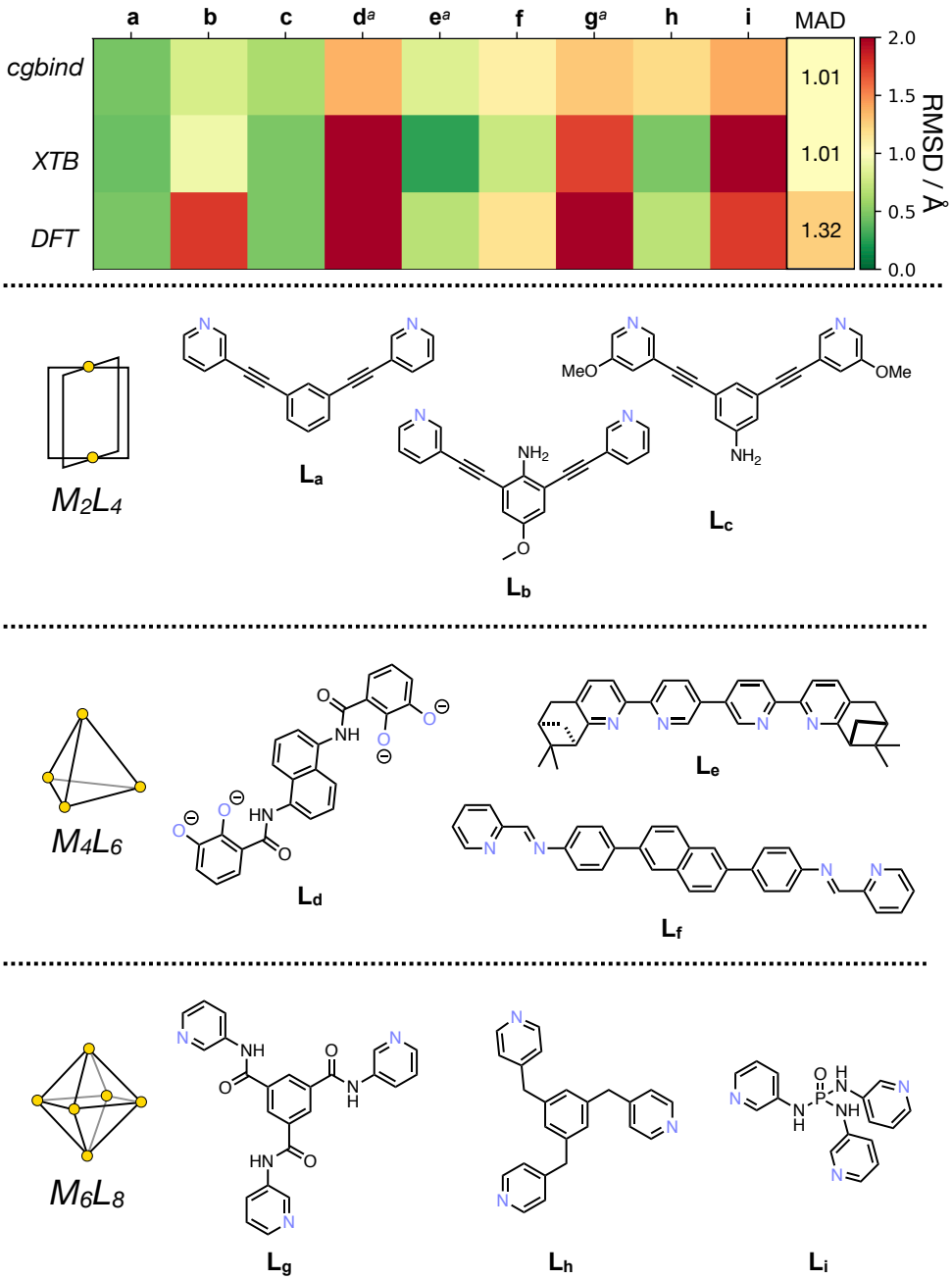


Figure 3.4.6. Root mean squared deviation (RMSD, Å) and their mean absolute deviations (MAD) of *cgbind*-generated metallocage structures relative to available crystal structures in the CSD as a function of increasing computational expense, only heavy atoms were considered for the RMSD calculation. Counterions/solvent molecules in the crystal structures were removed for the analysis. Data in Table 3.5.1. *a.* DFT geometries converged to a criteria only on forces (RMS(grad) < 5 mHa a_0^{-1}).

Binding Accuracy. For 102 host-guest complexes for which either quantitative or qualitative binding information exist in literature we explored the ability of *cgbind* to generate reasonable host-guest complexes and predict their binding ability. Those systems include M_2L_4 and M_4L_6 host architectures, and a range of polar and apolar substrates and solvents (dichloromethane, acetonitrile and water; see full SI). Specifically, the dataset includes: (1) 12 quinone-type guests and two different $Pd(II)_2L_4$ cages (24 host-guest complexes in total). Their host–guest properties in dichloromethane and acetonitrile have been studied by Lusby et al.[224] Depending on the nature of the quinone guest, binding constants of up to 10^8 M^{-1} have been observed for these systems. Many of them have also been found to be useful substrates in Diels-Alder reactions and cofactors in radical-cation cycloadditions.[175, 308] (2) Six different guests, including polycyclic aromatic hydrocarbons and steroids, with a series of nine aromatic-panelled $Fe(II)_4L_6$ tetrahedral cages in acetonitrile studied by Nitschke et al. (54 hots-guest complexes in total).[309] In this cage series, the size and arrangement of the aromatic panels were suggested to dictate guest binding propensity. (3) 24 neutral and charged guests bound to $Fe(II)_4L_6$ tetrahedral cages studied by Nitschke et al.,[310] of which 21 were experimentally found to bind with binding constants ranging from 3 up to 10^4 M^{-1}. The host-guest properties of these systems in aqueous solution were found to strongly dependent on the hydrophobicity of the guest.

Cage substrate complexes were initialized using a single conformer in most cases, with the exception of cholesterol where 50 conformers were considered. In all cases reasonable host-guest complexes were obtained. To qualitatively explore binding affinity, and as a first approximation, we consider the experimental data and our prediction as binary classifiers, i.e. either the guest is observed to bind or not within the cage. The binding affinity is considered to be accurate if $E_{\text{int}} < 0$ for substrates experimentally reported to bind. This approximation assumes that these experimental data can be directly compared and that only

host-binding interactions determine binding, omitting the effect of solubility of the substrates and building blocks and the effect of entropy upon binding.

Interestingly, the interaction energies calculated using our simple repulsive $(r + k)$ and both fast and normal electrostatic forcefields (fe and e, respectively) are comparable to those achieved at the tight binding DFT level of theory (XTB) with implicit solvent, and considerably better than single point calculations at the semi-empirical PM7-COSMO level of theory, which provides a statistically different prediction and is considerably worse than our forcefields and XTB (Figure 3.4.7). Optimizations at the PM7 level were not performed as we have previously found that this leads to unphysical metallocage geometries.[280] Overall our accuracy is modest, ∼60 %. This includes excellent (> 80 %) accuracy for M_2L_4 cages with quinone guests, while for M_4L_6 cages with aromatic guests the accuracy is limited to ∼40 % (Figure 3.5.8). We hypothesised that the low accuracy of our FF and single point energy evaluations in the M_4L_6 instance was due to the lack of cage relaxation not captured by our ridged body docking approach. To test this, XTB geometry optimisation were performed for the cage, substrate and cage-substrate complexes and the binding affinities evaluated again. In this case, all guests are predicted to bind, leaving the obtained accuracy static at ∼ 40% (Figure 3.5.8). For a representative example that does not bind experimentally we performed optimisations at the DFT (SMD(MeCN)-M06-2X/def2-TZVP//PBE-D3BJ/def2-SVP) level and found that the interaction energy was still negative ($E_{\text{int}} = -5$ kcal mol^{-1}) suggesting that XTB it is not the origin of the limited accuracy (Figure 3.5.9). Instead, thermal and entropic contributions, not accounted for here, are expected to shift the binding preference for this series. However, their consideration is known to be costly and subject to major errors. Moreover, the assumption that both host/guest are soluble and therefore equally available in solution should be considered. Nevertheless, our FFs provide accuracy on par with optimisation at the tight binding DFT level in a matter of seconds, with a confidence

estimate that suggests true binding substrates are unlikely to be missed (Figure 3.4.7b).

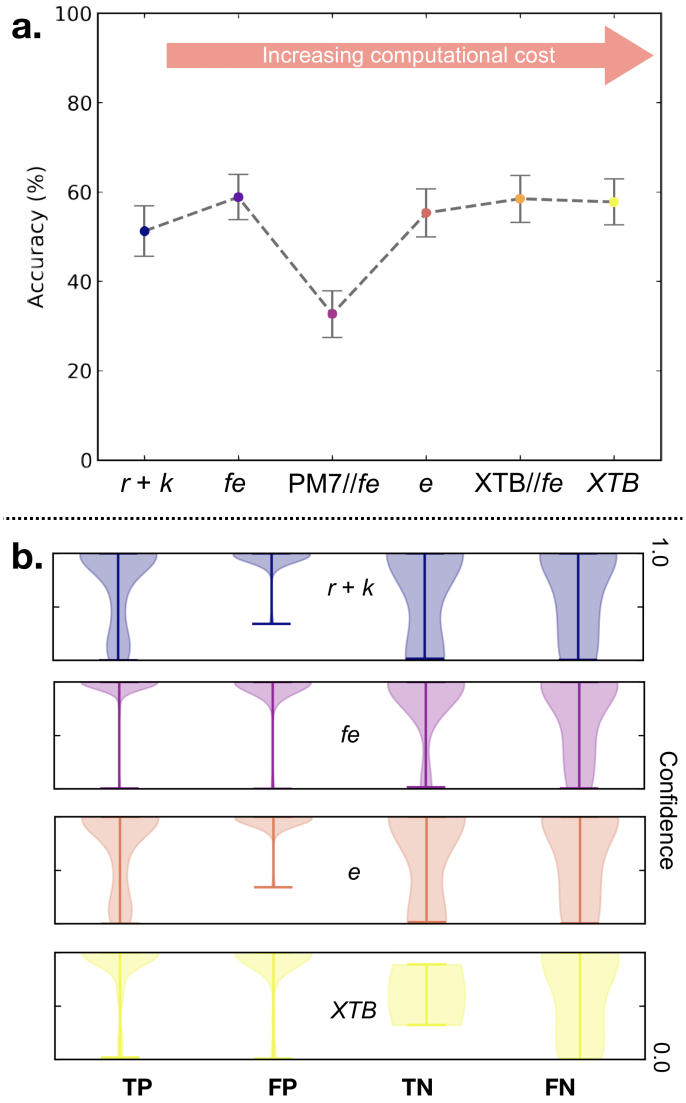


Figure 3.4.7. (a) Percentage accuracy in predicting a substrate binds or not considering 102 host-guest complexes reported in literature. Experimental data tabulated in full SI. All binding calculations were performed within the *cgbind* module interfaced through *autodE* to MOPAC and GFN-XTB. Error bars are plotted as a combined standard deviation, considering a small statistical error ($\sim$2 kcal mol^{-1}, Figure 3.5.6) with an assumed implicit normal error (σ = 2 kcal mol^{-1}) and bootstrapping with replacement (10,000 resamples) on the dataset. (b) Confidence for true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN) calculated as: $C = 1 - \exp(-(\Delta E/k)^2)$, where k = 5 kcal mol^{-1}.

Combinatorial Metallocage Library. To demonstrate the ability of *cgbind* to construct a library of cages from which subsequent property, screening and filtering calculations can be performed a library of Pd_2L_4 metallocages were generated (available for download as part of the full SI). Employing linkers comprised of three ends (E), 12 links (Z) and 29 centres (C) building blocks, joined linearly in a E-Z-C-Z'-E' pattern and Pd(II) as the metal centre, a theoretical total of 13,104 homoleptic cages are possible (building blocks shown in Figure 3.5.7). From these, using linkers initialized with 200 conformers, 5,639 metallocages were constructed with 'reasonable' geometries (defined to be the minimum pairwise distance above 0.8 Å, a higher proportion of cages would be constructed in a reasonable geometry were the linkers initialized with more conformers). Considering binding of benzoquinone as a guest, cages with cavity radii smaller than the centroid–O distance were removed to afford 1,353. This dataset was further reduced by eliminating those cages that were found to be too large to allow the formation of attractive NCIs between the substrate and the cage (cavity volume $> 2\times$ substrate), leading to 914 cages. Employing the algorithm `electrostactic_fast` to determine the optimal binding mode, 489 cages were removed based on $E_{\text{int}} > -15$ kcal mol^{-1}. Cages with formation energy [$\Delta E_f = E_{\text{C}} - (2E_{\text{Pd}} + 4E_{\text{L}})$ above 50 kcal mol^{-1} of an optimised reference Pd_2L_4 cage reported in ref. [224] calculated using XTB single point calculations were removed affording 275 cages. The most strongly binding hosts ($\Delta E_{\text{int}} < -12$ kcal mol^{-1}) were then taken forward. Finally, the three most synthetically accessible cages were selected, assuming the most accessible linker molecules lead to the most accessible metallocages, this was determined using the method from ref. [311] (Figure 3.4.8). From these core structures monovalent atom functionalisation e.g. H $\rightarrow$ Me, Et, Ph, CF_3 etc. may be performed using our Python module for molecular functionalisation (`github.com/duartegroup/molfunc`) to afford a vast library of heteroleptic cages.

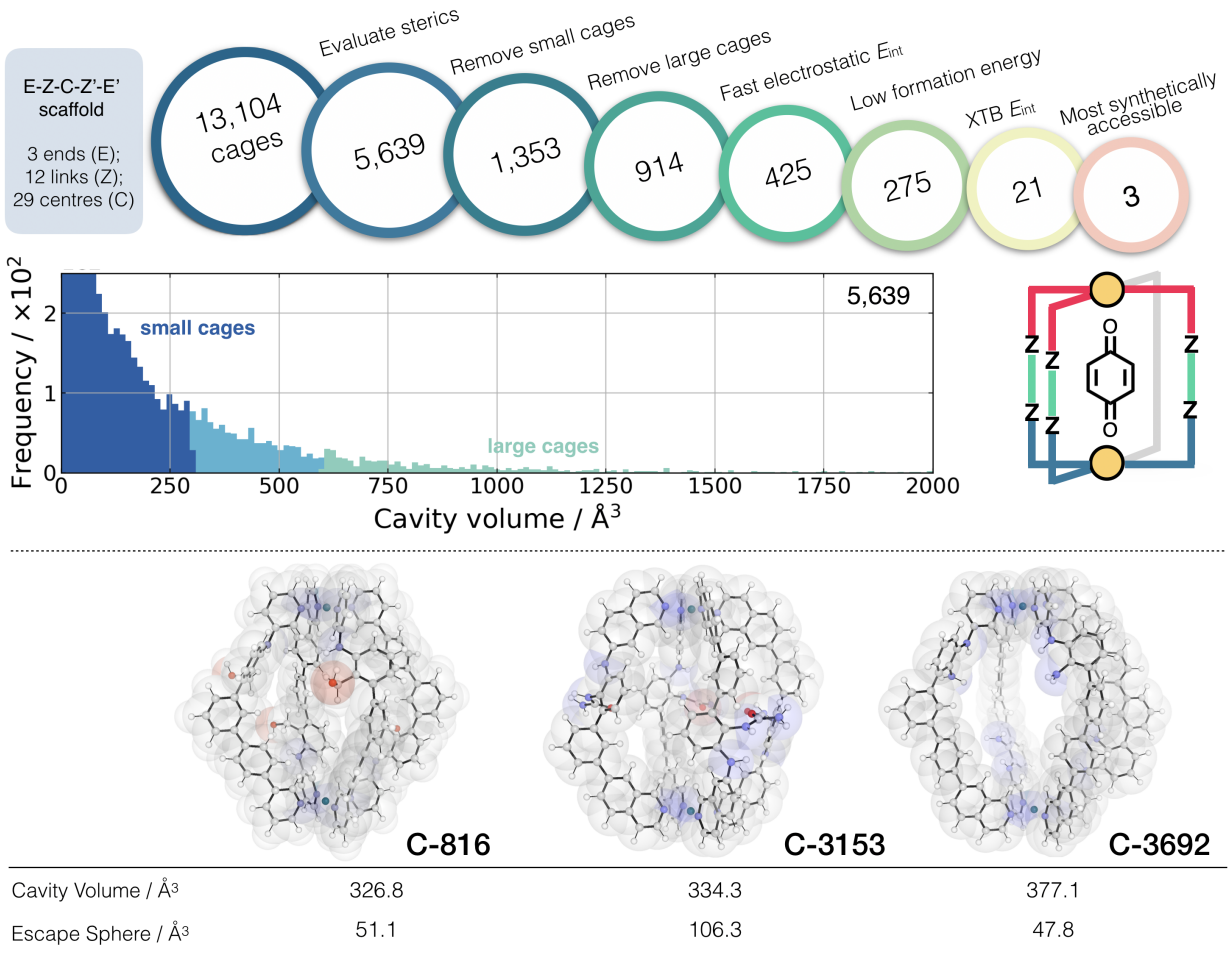


Figure 3.4.8. Filtering process used to find the three optimal hosts for benzoquinone based on simple geometric criteria, tight-binding DFT calculations and synthetic accessibility of the constituent linker molecules. Building blocks shown in Figure 3.5.7.

3.4.4 Conclusion

Here we have described a modular and extensible Python module *cgbind* for the automated construction and analysis of metallocage structure with arbitrary architecture (M_xL_y). Metallocages are constructed using a templating strategy and host-guest complexes generated by minimizing the energy of a simple forcefield developed here. Methods are available to analyse the resulting structures both qualitatively (electrostatic potential) and quantitatively (e.g. cavity size, window size) to analyse the resulting structures. We anticipate that *cgbind* and its web-app (`cgbind.chem.ox.ac.uk`) will facilitate the rational design of functional metallocage with applications in binding and catalysis. Further work towards the reverse design of metallocages is ongoing, where rather than modifying linkers to subsequently build metallocages and search for tight complexation, the cage is built around the substrate to maximize binding.

3.4.5 Computational Methods

Root mean squared deviations (RMSD) were calculated using the Kabsch algorithm implemented in the *rmsd* Python module,[312] reordering atoms and not considering hydrogens, all other parameters were set to their defaults. Metallocages had to be approximately aligned manually by fitting metal atoms prior to RMSD calculation, i.e. the reordering of atoms was not completely effective. All *cgbind* computations were performed with v. 1.0.0 beta. Tight binding (TB-)DFT calculations were performed using GFN2-XTB v. 6.2,[297] PM7 using MOPAC2016,[298] and resolution of identity DFT using ORCA v. 4.2.[246] Unless otherwise stated the latter using the PBE functional,[82] in combination with the D3BJ dispersion correction,[94, 313] def2-SVP basis set[247] (which employs effective core potentials for all metals used here)[314] and the default auxiliary basis sets.[248]

3.5 Selected Supporting Information II

Full Supporting Information including raw data can be found at:

`https://pubs.acs.org/doi/10.1021/acs.jcim.0c00519`

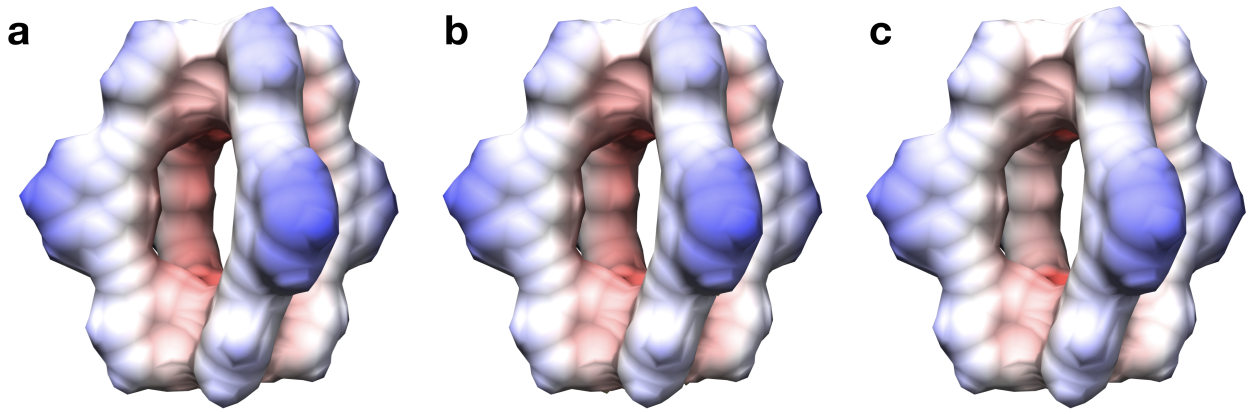


Figure 3.5.1. Electrostatic potential plotted on the van der Waals surface using Chimera for the metallocage reported in ref. [224], using a uniform grid over the whole structure evaluated using (a) partial atomic charges from GFN-XTB, (b) electron density from GFN-XTB with the in-built ESP generation and (c) Multiwfn calculated ESP from a PBE/def2-SVP electron density. All geometries are identical. (a) and (b) are plotted an identical colour gradient while for (c) the full electron calculation the ESP range is not identical.

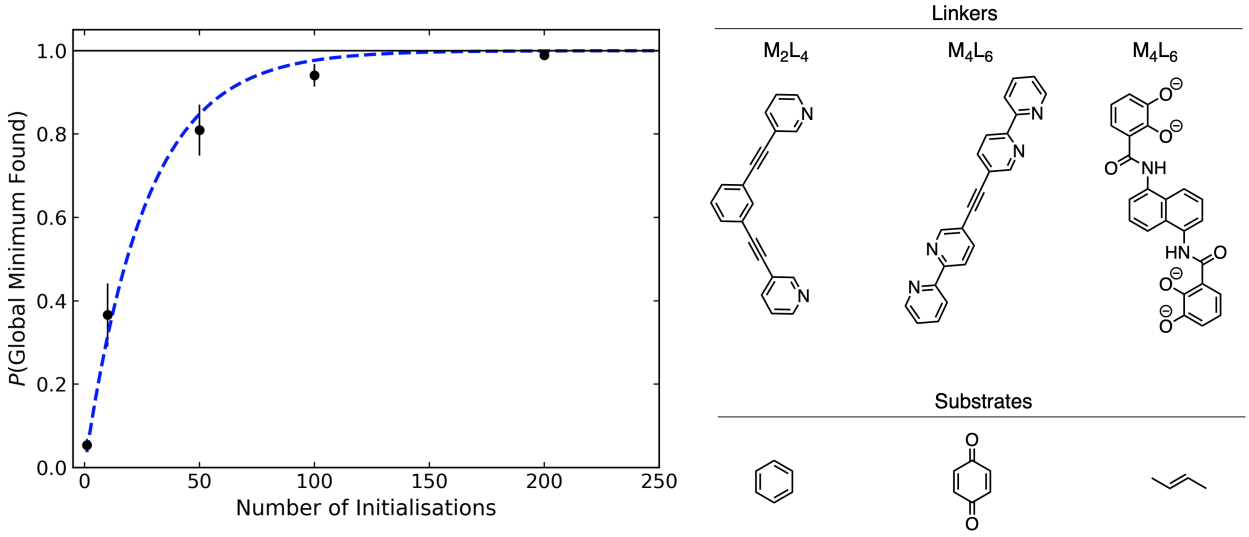


Figure 3.5.2. Probability of locating the global minimum on the rigid body cage-substrate PES as a function of the number of random initializations for the substrate rotation matrix. Error bars correspond to the standard error in the mean for nine cage-substrate complexes generated from the linkers and substrates shown. Probabilities are determined from 500 repeats. Energy threshold between the obtained binding energy and the true global minimum was set at 0.01 kcal mol^{-1}.

Linker	Cage	Metal ion	cgbind	GFN-XTB//cgbind	PBE-D3BJ/def2-SVP//cgbind
La	a[a,b]	Pd(II)	0.45559	0.41474	0.46321
Lb	b	Pd(II)	0.79541	0.91896	1.76651
Lc	c[b]	Pd(II)	0.62754	0.47084	0.47219
Ld	d[a,b]	Fe(III)	1.37837	2.01894	2.84652
Le	e	Mn(II)	0.82032	0.25010	0.67427
Lf	f	Fe(II)	1.09910	0.73566	1.17804
Lg	g[a]	Pd(II)	1.29529	1.73599	1.98475
Lh	h[b,c]	Pd(II)	1.21209	0.47521	0.68496
Li	i[b]	Cu(II)	1.40281	2.05635	1.76552

Table 3.5.1. Root mean squared displacements (Å) to crystal structures of *cgbind* generated geometries and subsequent optimisations with tight-binding and standard DFT. *a*. Crystal structure used to generate the template. *b*. Unmodified crystal structure contains encapsulated molecule(s). *c*. Conformer generated manually. n.c: Not calculated.

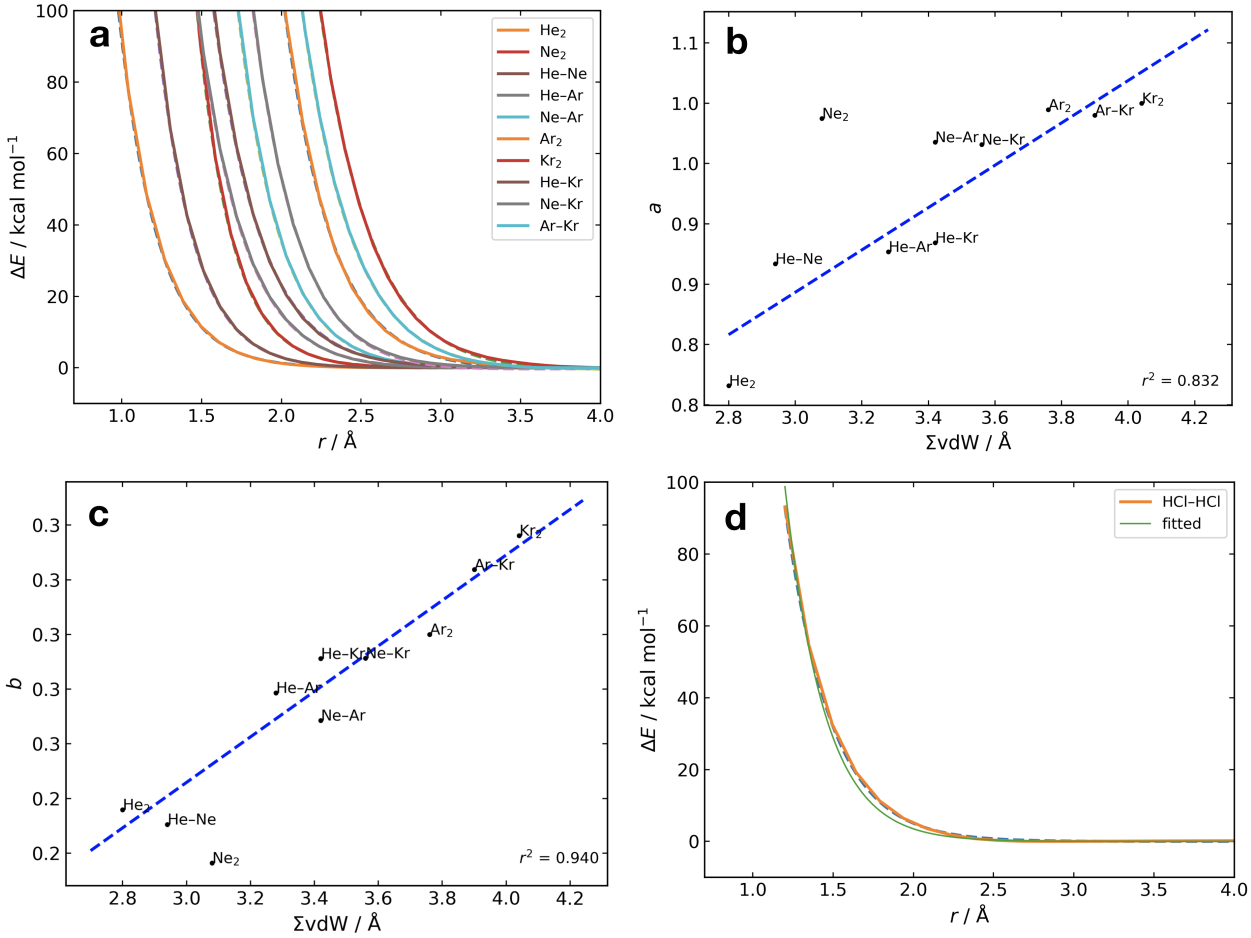


Figure 3.5.3. (a) Noble gas dimer potential energy surfaces calculated at DLPNO-CCSD(T)/ma-def2-QZVPP in ORCA (solid lines) and fit to $V(r) = \exp(a - r/b) - c/r^6$ with *scipy* (dashed lines). a, b as a function of the sum of van der Waals radii (b, c). Ne dimer excluded from the fits. a is normalized to its maximum value (Kr_2). Van der Waals radii taken from ref. [315]. (d) HCl–HCl head-to-tail PES calculated as (a), $r(H\text{–}Cl) = 1.27$ Å. Fitted potential (green) corresponds to a, b interpolated from the linear fits (b, c) where $\sum$vdW $= 2.95$ Å and $c = 32$ kcal mol^{-1}.

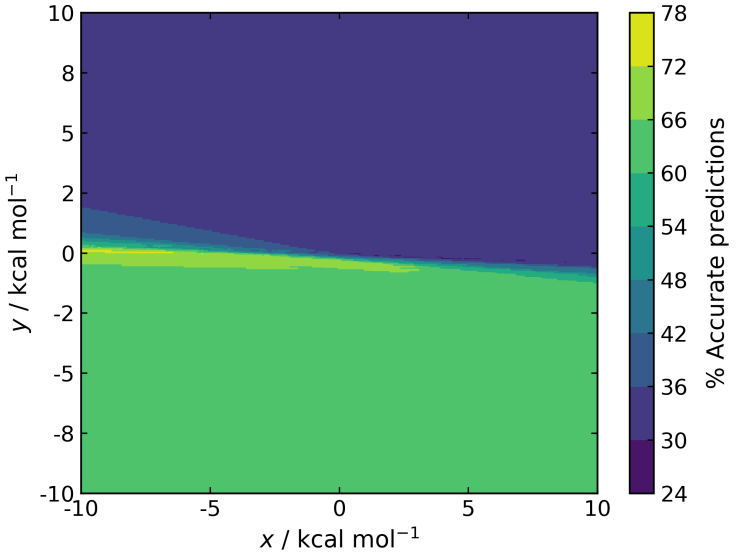


Figure 3.5.4. Accuracy on the prediction of 102 experimental binding affinities as a function of total binding affinity. $BA_t = BA_r + y \times N^{atoms}_{substrate} + x$. Repulsive binding affinity $BA_r = \sum_{i>j} c \times \exp(a(i,j) - r_{ij}/b(i,j))$ where $a(i,j) = 11.576415 \times (0.175541 \times (\gamma_i\gamma_j) + 0.316642)$, $b(i,j) = 0.083214 \times (\gamma_i\gamma_j)$ - 0.003768, as determined in Figure 3.5.2, $c = 1$ kcal mol^{-1}, γ_i is the van der Waals radius of atom i. An accurate prediction is made if a substrate expt. binds and $BA_t < 0$ or expt. does not bind and $BA > 0$. x and y are plotted on a grid of 500 × 500.

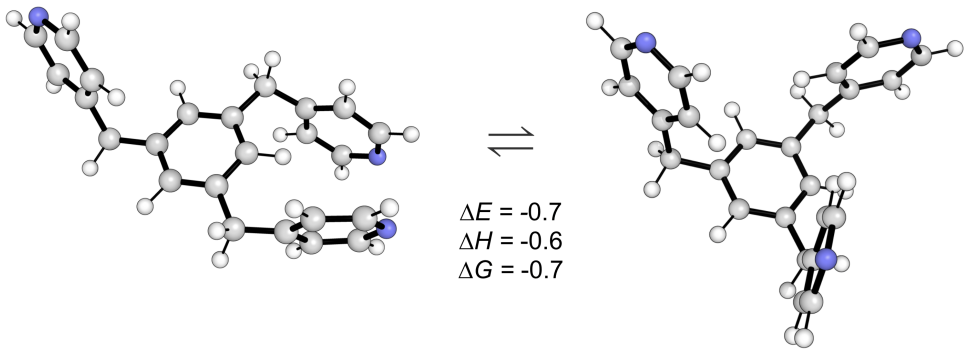


Figure 3.5.5. Selected conformers of L_h optimized at the PBE0-D3BJ/def2-SVP level of theory and single points calculated at DLPNO-CCSD(T)/def2-TZVPP. Frequencies were calculated to confirm that both correspond to minima (no imaginary frequencies). Energies quoted in kcal mol^{-1}.

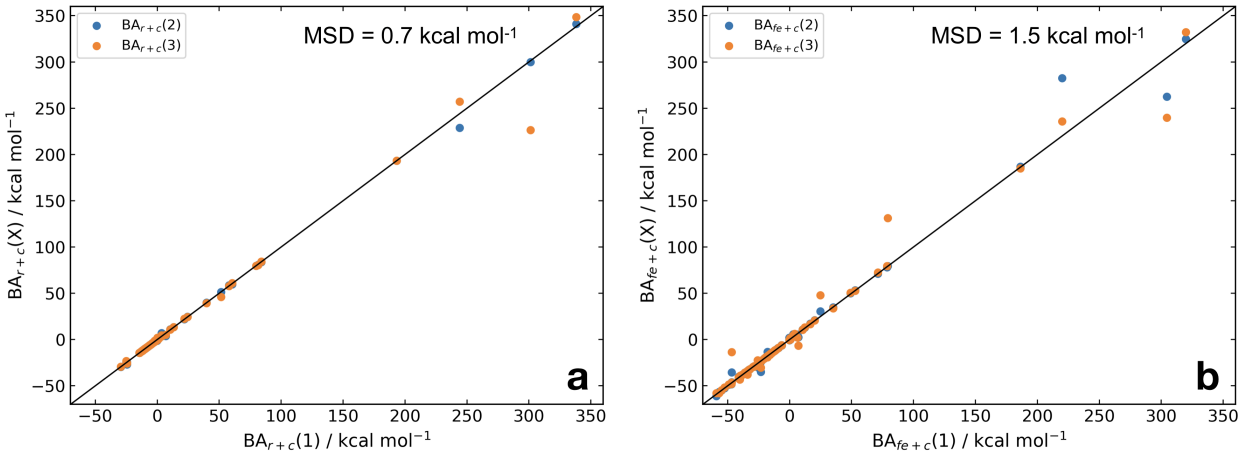


Figure 3.5.6. Statistical error for binding affinities obtained using 3 repeats with (a) repulsive including a constant attractive term E_{int}(r+k) and (b) fast electrostatic E_{int}(fe) methods for determining binding affinities. Mean signed deviations (MSD) are shown.

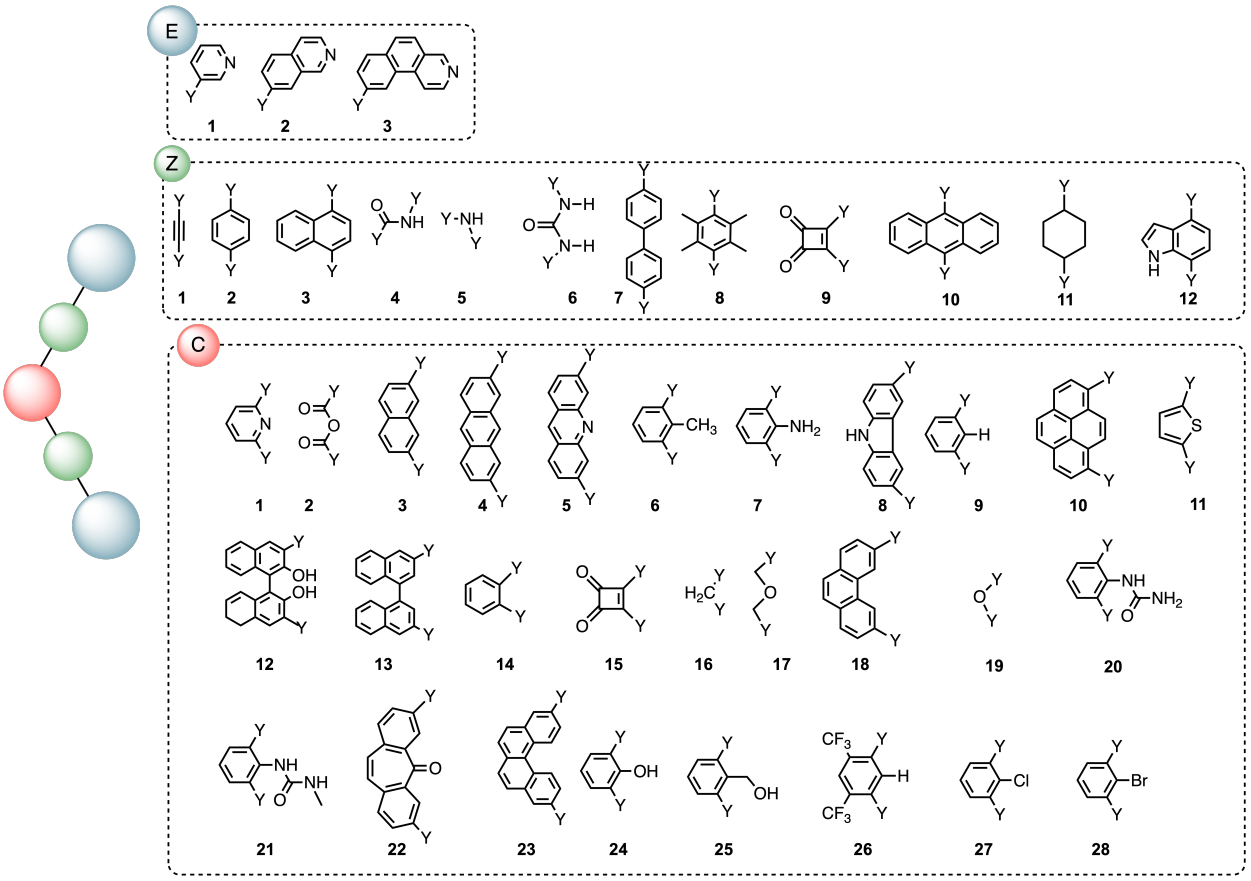


Figure 3.5.7. Building blocks used to generate the M_2L_4 metallocage library. Link atoms are represented by a bond to Y.

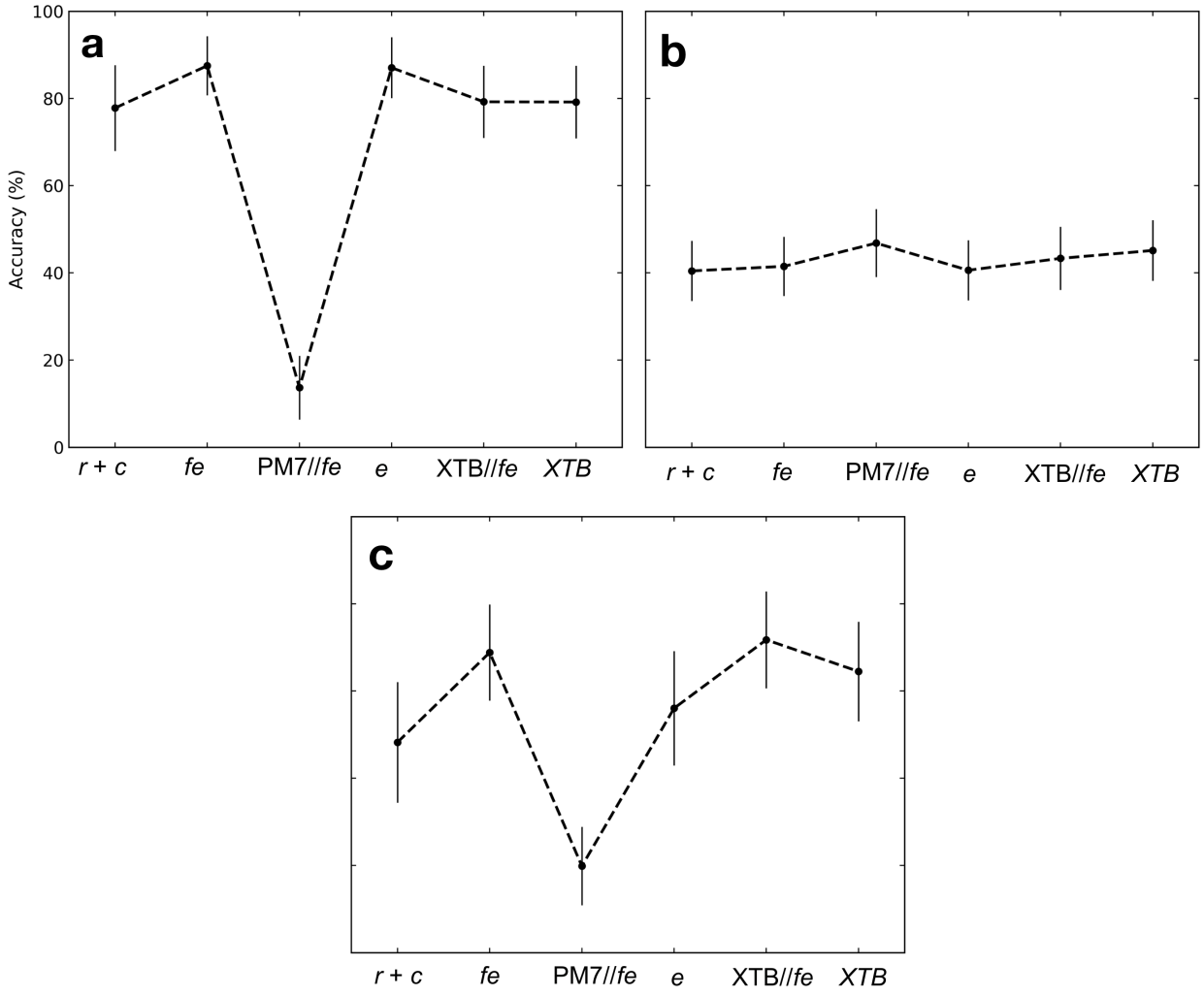


Figure 3.5.8. Binding affinity predictions for (a) 24 host-guest complexes with two different $Pd(II)_2L_4$ cages refs. [224] and [175]; (b) 54 hots-guest complexes with a series of aromatic-panelled $Fe(II)_4L_6$ tetrahedral cages ref. [309]; (c) 24 host-guest complexes with $Fe(II)_4L_6$ tetrahedral cages (ref. [310]).

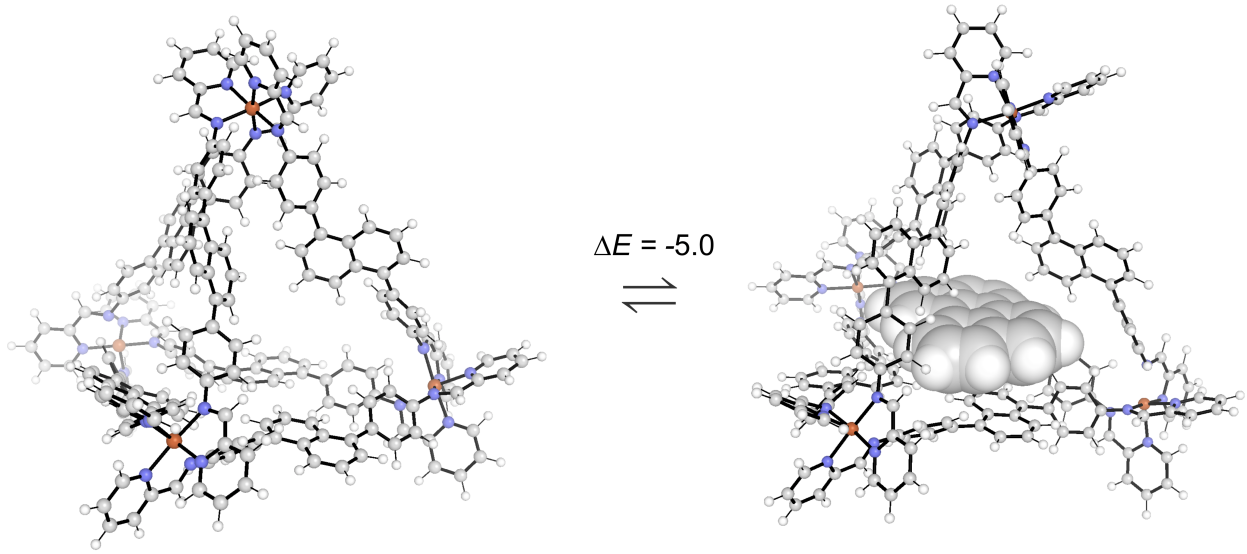


Figure 3.5.9. DFT optimized structures of cage (C-42) and cage-substrate complex (CS-42) and the binding affinity (ΔE in kcal mol^{-1}) is computed at the SMD(MeCN)-M06-2X/def2-TZVP//PBE-D3BJ/def2-SVP level of theory. Due to the large molecular size the geometry optimization of C-42 and CS-42 were converged with respect to binding affinity ($\Delta\Delta E < 1$ kcal mol^{-1}).

4 Solution Phase Entropy

4.1 Abstract

Predicting free energies using end-state methods from QM relies on calculating thermal corrections to the potential energy. The ideal gas model (IGM) provides a well established and accurate method for isolated molecules, but has limitations for large and condensed phase systems. One of the dominant terms and potential sources of error is the translational entropy, for which a new approximation is presented in this chapter. The potential for solute translation within a solvent cavity is spherical exponential, from which classical-quantum correspondence allows for rapid determination of the entropy. The resulting translational entropy is roughly half of that obtained from the IGM approximation, which has been previously suggested as an *ad hoc* correction. This method can correct for systematic entropy difference errors in bimolecular reactions.

4.2 Introduction

Accurate prediction of free energy differences is central to generate reliable estimates of binding and kinetics. Unfortunately, despite progress towards this goal, conventional methods

remain costly and/or inaccurate. Indeed, in Chapter 3 discrepancies of up to 10 kcal mol^{-1} were observed in binding and activation energies using conventional methods.

Although in some instances potential energy differences can be used to make predictions, computing absolute free energy differences at finite temperatures demand inclusion of an entropic component to the free energy ($F = E - TS$, where $E = H$ or $E = U$ in NPT or NVT ensembles respectively). The entropic component may be included implicitly in 'free energy methods' where sampling from one state to another affords a free energy difference (ΔG),[303, 316–318] or explicitly from end-states without any intermediate configurations. In the latter case, an estimation of the absolute entropy of a state (e.g. minimum/TS) is required.[319] Sufficient sampling to converge free energy methods with an accurate QM potential is generally not available for all but the systems of a few atoms.[149] Therefore, estimating absolute thermal (U and S) energy contributions will be the subject of this chapter (Figure 4.2.1).

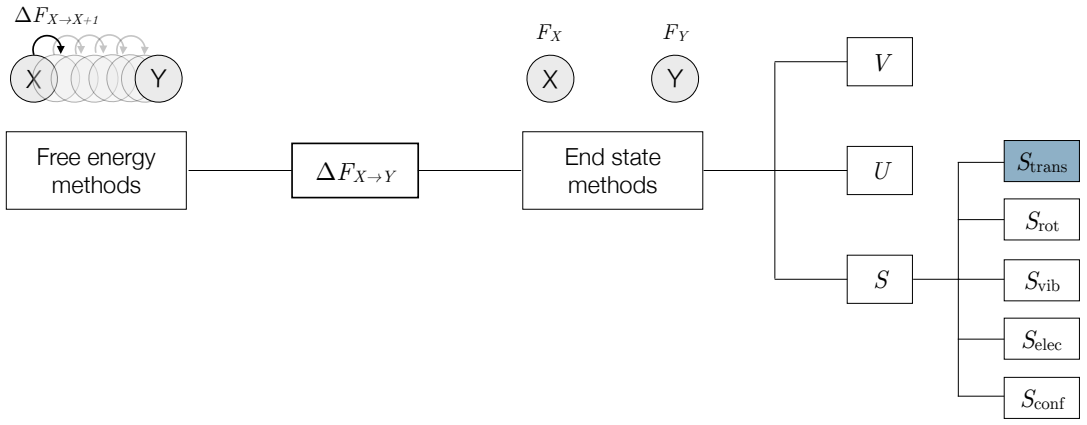


Figure 4.2.1. Schematic outline of methods to calculate free energy differences, with the focus of this chapter highlighted in blue.

Depending on the end points of the system and process, the entropic components may cancel and the entropy difference (approximately) vanish, e.g. in unimolecular transformations. However, in an intermolecular chemical process within an environment (e.g. solven-

t/supramolecular catalyst),

$$\mathrm{A} + \mathrm{B} \xrightarrow{\text{env.}} \mathrm{AB} \tag{4.1}$$

such a cancellation is not generally observed, as the minimum or transition state complex (AB) has lost translational and rotational degrees of freedom compared to the separated components.[319] An interesting case occurs when AB comprises a supramolecular structure, where one component (a guest, A) can translate within the host B (as alluded to in Chapter 3). Of particular importance are binding processes in proteins, which govern the affinity of a drug to its biological target.[320] While in the context of chemical reactivity, an accurate calculation of entropy is required as the rate is exponential in the difference.

4.3 Theory

Calculating F for a stationary point on a PES with QM is achieved by adding entropic and other contributions to the 0 K potential energy ($U_{0\,\mathrm{K}}$). As most chemical reactions take place at constant pressure, number of molecules and temperature, the appropriate ensemble is isothermal-isobaric (NPT). However, the volume change contribution ($P\Delta V$) to the Gibbs free energy is negligible for processes in the condensed phase, including small molecule[a] and enzymatic reactions.[322] It is therefore common to predict thermodynamic quantities in the isothermal-isochoric NVT canonical ensemble, where the associated free energy is the Helmholtz energy. From statistical mechanics A is given by,[323]

$$A = -k_B T \ln[Z(N, V, T)] \tag{4.2}$$

[a]For example, the expt. [321] volume of activation for the S_N2 reaction: allyl chloride + H_2O → allyl alcohol + HCl is $\Delta V^{\ddagger} = (-11.4 \pm 0.1)$ ml mol^{-1}, which at $T = 323.4$ K and P =1 atm translates to a contribution $P\Delta V = -0.00116$ kJ mol^{-1}.

where Z is the canonical partition function (PF) and k_B is the Boltzmann constant. Were the NPT ensemble to be used the PF would require summation over volume,

$$Z_P(N,P,T) = \sum_j Z(N,V_j,T)e^{\beta PV_j} \tag{4.3}$$

where $\beta = 1/k_BT$. To simplify, with minimal loss in accuracy, thermodynamic quantities are outlined in an NVT ensemble. With electronic structure methods affording $U_{0\,\mathrm{K}}$,[b] the problem broadly reduces to calculating S within the NVT ensemble.[c] From Eq. 4.2 then,

$$S = k_BT\Big(\frac{\partial \ln Z}{\partial T}\Big)_{N,V} + k_B\ln(Z) \tag{4.4}$$

which only requires the canonical PF for the system.

4.3.1 Classical Statistical Mechanics

Provided the system is sufficiently classical in nature (hot/massive), with negligible nuclear quantum effects and exclusively ground electronic states, the PF may be calculated classically. Classical PFs are integrals over all possible positions and momenta of the particles described by the canonical positions ($\boldsymbol{x}$) and momenta ($\boldsymbol{p}$),[323]

$$Z = \frac{1}{h^3}\int\int e^{-\beta\mathcal{H}(\boldsymbol{p},\boldsymbol{x})}\, d\boldsymbol{p}^3 d\boldsymbol{x}^3. \tag{4.5}$$

For a classical Hamiltonian of the form $\mathcal{H}(x,p) = p^2/2m + V(x)$ momentum may be inte-

[b]Where the approximate zero-point energy (ZPE) is generally obtained from a harmonic approximation of vibrations and the remaining thermal energy given by standard expressions involving the PF.

[c]Note the non-ZPE thermal contributions to the internal energy are small and largely cancel. For example, $\Delta E^{\ddagger}_{\mathrm{internal}} = 3.3$ kcal mol^{-1} for F^-+CH_3Cl.

grated out to give a purely configurational integral, which – with sufficient accuracy and sampling – provides access to all thermodynamic quantities of the system. In principle, calculating Z with classical MD is conceivable, as all possible particle positions and velocities are sampled as $t \to \infty$. However, in conventional NVT MD simulations the configurational space is already weighted by a Boltzmann factor, reducing the sampling of high-energy regions. Furthermore, regions of the potential energy landscape separated by large energy barriers will rarely be explored. Trying to numerically evaluate Eqn. (4.5) directly by integrating over $\boldsymbol{x}$ explicitly fails for all but the most simple systems ($N \ll 10$) due to the exponential scaling (3^N) of the configurational space for N particles.

4.3.2 Quantum Statistical Mechanics

The canonical quantum PF is given by,

$$Z = \mathrm{tr}[e^{-\beta \hat{H}}] = \sum_i g_i \, e^{-\beta E_i} \tag{4.6}$$

where g_i is the degeneracy of the i'th level with energy E_i. Unfortunately, for a macroscopically large chemical system, with $O(10^{23})$ particles, the problem is intractable. Even if all the eigenvalues of the system could be obtained, the sum is impractical to calculate. Indeed, even calculating Z for a particle in an infinite cubic well is challenging and usually approximated by an integral (Eqn. 4.10).

To progress, the complexity must be reduced from considering 10^{23} particles to just a few. Therefore, the system is assumed to be homogeneous and can be divided into n identical

and non-interacting copies to give,

$$Z = \frac{1}{n!}\Big(\sum_j g_j\, e^{-\beta E_j}\Big)^n \quad := \quad \frac{z^n}{n!} \tag{4.7}$$

where z is the canonical PF of the sub-system and the $n!$ factor corrects for the indistinguishability of the sub-systems. The sum over j now enumerates over all the eigenstates of the subsystem. Despite the system now containing 10^2–10^4 particles it still requires diagonalising a large system and is generally intractable.

A further simplification is therefore required; z is assumed to be factorisable into translational, rotational, vibrational, electronic and conformational contributions for each component in the subsystem i.e. a complete decoupling of molecular motion (Figure 4.3.1),

$$z = \prod_{\text{molecules}} z_{\text{trans}} \cdot z_{\text{rot}} \cdot z_{\text{vib}} \cdot z_{\text{elec}} \cdot z_{\text{conf}} \tag{4.8}$$

where approximations to each of these terms are outlined in the following sections.

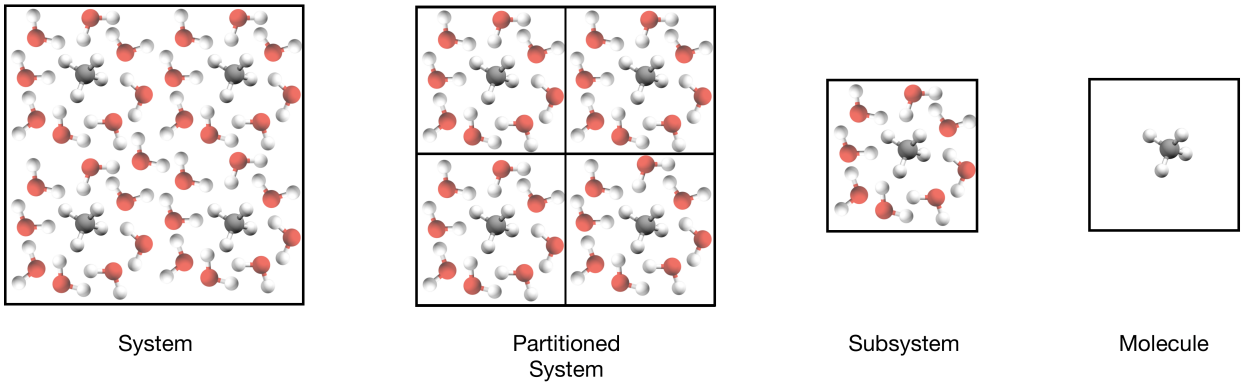


Figure 4.3.1. Example partitioning of a condensed phase molecular system.

4.3.3 Quantum: Gas Phase

To calculate the sub-system's PF comprised of a species with a 'gas phase' method the environment is completely neglected. Although this places the molecule in vacuum, it is commonly referred to as a 'gas phase' method due to its similarity to a low-pressure gas. In this section z [Eqn. (4.8)] will refer to a gas phase molecular PF.

If the gap between ground and excited electronic states of the system is large and only the ground state is accessible at room-temperature, thus the electronic component of the molecular PF (z_{elec}) is assumed to be unity. The conformational component ($z_{\text{conf.}}$) is often neglected due to a cancellation between states when calculating an energy or entropy difference. Only within the last few years accurate methods to estimate this component have been developed.[324, 325] This means only the translation, rotation and vibrational states for a single molecule must be approximated.

4.3.4 Ideal Gas Method

Within the ideal gas method (IGM), the molecular PF for a rigid molecule is approximated using analytic expressions for z_{trans}, z_{rot} and z_{vib}, derived from eigenvalues of a particle in a box (PIB), rigid rotor (RR) and harmonic oscillator (HO) respectively.[323]

4.3.4.1 Translation: Particle in a Box

For a cubic box with sides of length l and a particle of mass m the eigenvalues are,

$$E_n = \frac{\hbar^2\pi^2}{2ml^2}(n_x^2 + n_y^2 + n_z^2) \qquad : n_x, n_y, n_z \in \mathbb{N}_{>0} \tag{4.9}$$

from which the PF is then,

$$z_{\mathrm{PIB}} = \sum_{n_x,n_y,n_z}^{\infty} \exp\Big\{ -\beta \cdot \frac{\hbar^2\pi^2}{2ml^2}(n_x^2 + n_y^2 + n_z^2)\Big\} \approx \Big(\frac{2\pi m k_B T}{h^2}\Big)^{3/2} V \tag{4.10}$$

using an integral approximation to the sum.[d]

The effective volume (V) available is chosen based on the standard state at which the entropy, and so free energy, are required.[303] For gases this is usually taken to be the volume available to an ideal gas at 1 atm ($V_{1\ \mathrm{atm}} = 1/\beta p = 40605$ Å^3) and 1 mol dm^{-3} for solutions ($V_{1\ \mathrm{M}} = 1/C = 1661$ Å^3). For gas phase reactions this approximation is excellent – delivering quantitative accuracy to experiment in a selection of reaction rates.[326] However, in solution using an effective volume appropriate for an ideal gas is controversial, with 'free volume' theories proposed to correct for the potential overestimation of the effective volume and so entropy.[327] These strategies will be discussed in the context of a new method outlined below.

4.3.4.2 Rotation: Rigid Rotor

The standard IGM expression for the rotational PF is outlined below for completeness and is generally accurate for most molecules in the gas phase as it does not significantly contribute to the entropy.

Evaluating the rigid rotor PF requires the species' symmetry, as z_{rot} differs for symmetric, spherical and asymmetric tops (amongst others). Using a spherical top to simplify the

[d]The approximation is reasonable in $\ln(q)$, with a <1% error in $\ln(z_{\mathrm{PIB}}^{\mathrm{sum}})/\ln(z_{\mathrm{PIB}}^{\mathrm{integral}})$ for a water molecule at 298 K in a cubic box with side length >4 Å. Bound evaluated numerically by summation up to $n_{\mathrm{max}} = 200$, which is converged with respect to single precision in *numpy*.

calculation, with moment of inertia I, has eigenvalues,

$$E_J = \frac{\hbar^2}{2I} J(J+1) \qquad : J \in \mathbb{N}_{>0} \tag{4.11}$$

with degeneracy $g_J = (2J+1)^2$. The PF is evaluated at the continuum limit noting that large J will contribute most to the integrand (provided $\hbar^2/2I\beta \ll 1$) so $J+1 \sim J$,

$$z_{\text{rot}} = \sum_J^{\infty} (2J+1)^2 \exp\Big\{ -\beta \cdot \frac{\hbar^2}{2I} J(J+1) \Big\} \approx \sqrt{\pi} \Big(\frac{2Ik_bT}{\hbar^2} \Big)^{3/2} \tag{4.12}$$

Generalisation of Eqn. (4.12) to a molecule with unequal moments of inertia ($I_x \neq I_y \neq I_z$) is assumed to be possible and, when divided by a symmetry number σ_r, is given by,

$$z_{\text{rot}} = \frac{\sqrt{\pi}}{\sigma_r} \cdot \sqrt{\frac{(k_BT)^3}{B_xB_yB_z}} \tag{4.13}$$

where the rotational constants are of the form $B_k = \hbar^2/2I_k$ where I_k is the moment of inertia in the k axis.

4.3.4.3 Vibrational: Harmonic Oscillator

By diagonalising the mass-weighted matrix of second derivatives (the Hessian), the normal mode frequencies can be obtained from the eigenvalues, so that the vibrational PF can be obtained as a product over the individual uncoupled harmonic oscillators (HOs). Given for a quantum HO,

$$E_n = \hbar\omega \Big(n + \frac{1}{2} \Big) \qquad : n \in \mathbb{N}_{\geq 0} \tag{4.14}$$

then,

$$z_{\text{HO}} = \sum_n^{\infty} \exp\Big\{ -\beta \cdot \hbar\omega \Big(n + \frac{1}{2} \Big) \Big\} = \frac{e^{-h\nu/2k_BT}}{1 - e^{-h\nu/k_BT}} \tag{4.15}$$

where $\nu = \omega/2\pi$. The total molecular PF is then given by a product over normal modes as,

$$z_{\text{vib}} = \prod_{j}^{N} \frac{e^{-h\nu_j/2k_BT}}{1 - e^{-h\nu_j/k_BT}} \qquad : \qquad N = \begin{cases} 3n_{\text{atoms}} - 5 & \text{if linear} \\ 3n_{\text{atoms}} - 6 & \text{otherwise} \end{cases} \tag{4.16}$$

This simple expression is, however, limited in the low-frequency regime where $\nu \sim 10$ cm^{-1}, which are prevalent in molecules with more than a few atoms. These may correspond to dihedral rotations or translations and are not well approximated by a harmonic well.[e] Furthermore, these low frequency modes (LFMs) are subject to numerical noise and with a poorly converged geometry or an imprecise Hessian and can become imaginary.

To counter this deficiency Truhlar suggested that a *post hoc* shift should be applied to all harmonic frequencies below a certain threshold (ν_0, i.e. $\nu = \max(\nu, \nu_0)$) designated as LFMs.[259] However, the choice of cut-off is not uniquely defined and – as Harvey and co-workers found – can lead to ΔG differences in excess of 5 kcal mol^{-1}.[328]

Grimme and co-workers proposed an alternative LFM treatment that improved the agreement between theory and experiment for guest binding.[230] By treating LFMs as free rotors with average moment of inertia, then interpolating between the free rotor and HO regimes z_{vib} is scaled down. However, parameters must be set once again, this time governing the form of the weighting function.

4.3.5 Alternative Methods

While the IGM is commonplace in electronic structure theory codes for calculating the thermal (enthalpic and entropic) contributions to the absolute free energy, it is not without

[e] Indeed $\lim_{\nu \to 0^+}(z_{j,\text{vib}}) = \infty$ which affords an unphysical divergence in the entropy for a molecule.

limitations. Specifically, the translational component of the entropy is often considered to be overestimated for molecules in solution.[329] Indeed, various scaling factors have been proposed in order to improve computational accuracy with varying levels of justification. For example, neglecting both the translational and rotational components,[330] removing the translational component,[331] scaling by 0.5 and inverting the sign,[332] scaling by 0.5,[333] and scaling by 2/3.[334] In addition to empirical scalings, recent work by Garza[335] makes use of an effective volume within z_{PIB} given by,

$$V_{\mathrm{Garza}} = N_c v_c \tag{4.17}$$

where N_c is the number of available cavities for the solute and v_c the cavity volume within a solvent. The PIB approximation for condensed phase systems is, however, theoretically limited. Molecules are not hard spheres and have a free path in solution $O(1\ \text{Å})$;[336] indeed, most empirical forcefields make use of a Lennard-Jones $1/r^{12}$ repulsive term rather than a step function. This may result in a over/under-estimation of the translational entropy depending on the box size.

Other approaches have moved beyond the PIB treatment, to a harmonic oscillator approximation of the explicit translational motion. For example, the method by Henchmen and co-workers treats translational/rotational and vibrational components by diagonalising the force covariance matrix in a particular reference frame to afford frequencies that are used within a quantum harmonic oscillator expression for the entropy.[337, 338] A similar method based on QM calculations by Nakai and co-workers, referred to as the 'harmonic solvation model' (HSM), treats all normal modes of the solute as vibrations.[261] At least superficially this would seem reasonable as in solution a molecule has no free translation, or – to a lesser extent – rotation. This method recovers the equipartition limit for the internal energy for

translation and rotation as $\nu \to 0$ cm^{-1} namely,

$$E_{\text{trans/rot}} = \frac{k_B T}{2} \sum_{i}^{3} \lim_{\nu_i \to 0} \left[\beta \nu_i \left(\frac{1}{2} + \frac{e^{-\beta \nu_i}}{1 - e^{-\beta \nu_i}} \right) \right] = \frac{3}{2} k_B T. \tag{4.18}$$

Nakai suggested computing the frequencies using an implicit solvent model (CPCM); however, this must be performed numerically as the implicit solvent cavity must remain fixed i.e. not relaxed to accommodate the displaced geometry. Moreover, the numerical frequency calculation must be performed on a structure that has been optimised in the solution phase, thereby increasing the calculation time by hindering geometry convergence. For large systems this method therefore becomes prohibitively expensive.

The same authors reported a variant of this approach in 2018,[260] where the intramolecular vibrational modes are treated within the IGM (rigid-body HSM, RB-HSM). This model was tested on a small set of vaporisation (7) and combustion (5) entropies and performed similarly to the more expensive HSM. Translational and rotational motion are then calculated by diagonalising a 6×6 Hessian matrix consisting of the second derivatives with respect to translation and rotation in the Cartesian directions. For water this treatment results in harmonic frequencies > 70 cm^{-1}; however, for 1,2-dichloroethane (DCE) the lowest frequency mode treated harmonically is just 6.4 cm^{-1}. For such a small force constant, the harmonic approximation is a poor one, with the associated entropy contribution potentially overestimated.

It is worth noting that in the systems tested by Garza, Henchmen and Nakai the agreement to experiment is excellent in all cases, despite the disparity in methods. Therefore, the question of which method is physically based and so generally applicable remains.

4.3.6 Concentration Corrections

A meaningful comparison to experiment requires comparing free energies that have been obtained at the same (standard) state, which for molecules is usually taken to be 1 M. A correction is then needed to calculate a free energy at the standard state for an unequal simulation volume of,[339]

$$F^{\ominus} = F^{\mathrm{sim}} - k_B T \ln\left(\frac{V^{\ominus}}{V^{\mathrm{sim}}}\right) \tag{4.19}$$

where F is a free energy (G, A) and V^{sim} the volume used for the simulation.

In calculating the translational entropy within a defined cavity the solute concentration is far from the 1 M standard. While solvent dependent, it should be of the same order of magnitude as the solute volume. For example, considering a methane-in-water system, which is close to ideal spherical case, the molecular volume is ~ 30 Å^3[340] thus to correct into 1 M (an effective volume of 1661 Å^3) an absolute upper bound is 2.4 kcal mol^{-1}. Therefore, 2.5 kcal mol^{-1} provides a useful upper bound on the concentration correction the true 'simulation' volume from which the exponential well is obtained is larger than just the molecular volume. In fact, with a molecular volume more representative of those usually studied in solution phase reactions, with a radius ~ 5 Å ($\sim 500-700$ Å^3 is common in drug-like molecules[341]) the volume correction decreases to just 0.5—0.7 kcal mol^{-1} at room temperature. Thus, in all but the smallest solvated molecules, this contribution is small.

It is worth noting that this correction term is distinct from the entropy of mixing, as the maximum $TS_{\mathrm{mix}} \sim 0.4$ kcal mol^{-1},[f] and thus generally negligible at ambient temperature or below.

[f] Obtained from the standard statistical mechanics result $\Delta S_{\mathrm{mix}} = -k_B(x \ln x + (1-x)\ln(1-x))$ for a two component system one of which has mole fraction x.

4.4 Limitations of the IGM and Harmonic Approximation

4.4.1 Model Systems

While the quantum PF and thus entropy of a HO diverges as the frequency becomes small,[g] a classical treatment affords another an even greater deficiency. Specifically, the classical HO entropy is,

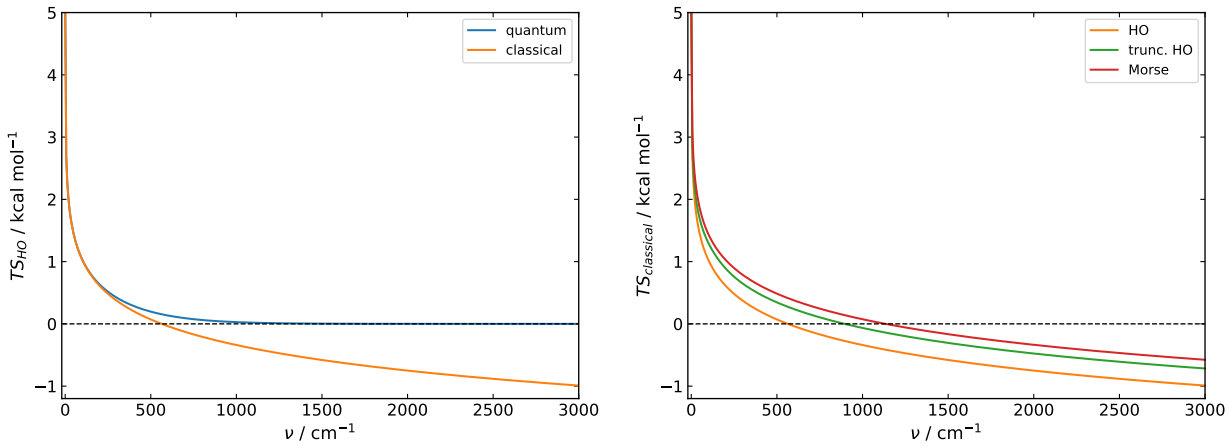
$$S_{\mathrm{HO}}^{\mathrm{classical}} = k_B \ln\left(\frac{k_B T \mathrm{e}}{h\nu}\right) \tag{4.20}$$

which becomes negative and unphysical as $k_B T \sim h\nu$ (Figure 4.4.1a). Plotting the classical and quantum entropies as a function of the oscillator frequency, the classical entropy becomes negative at a frequency of just 540 cm^{-1}. If the HO is used as an approximation for a bond (as for stretches in most force fields) the entropy should certainly be positive. Although this behaviour has been noted previously, and a correction proposed,[342] it is striking. Significant care must then be taken when estimating absolute or relative entropy differences with a classical method.

In calculating the PF for a classical HO the integral runs over an unphysical region beyond where the atoms pass through each other at $r = 0$, where r is the internuclear separation and r_e its equilibrium value. Two potentials with more realistic behaviour are that of a truncated HO,

$$V(r) = \begin{cases} \infty & r \leq 0 \\ k(r - r_e)^2 & r > 0 \end{cases} \tag{4.21}$$

[g] As it should with the Hamiltonian becoming that of a free particle, with $S_{\mathrm{HO}}^{\mathrm{quantum}} \sim k_B \ln(k_B T/h\nu)$: $\nu \ll 1$, but is clearly unphysical for a molecule.



(a) **(b)**

Figure 4.4.1. (a) Molar entropy (S) of non-interacting quantum and classical harmonic oscillators vs oscillator frequency at $T = 298.15$ K. (b) Molar entropy (S) of non-interacting oscillators vs oscillator frequency at $T = 298.15$ K. S_{Morse} was calculated by evaluating the partition function numerically over $[-\frac{1}{a}\ln(2), 5]$ Å, at 10^4 points with a trapezium integrator (converged quantities). D_e fixed at 400 kJ mol^{-1} and a modified according to the harmonic frequency. $\partial Z/\partial T$ calculated using a finite difference ($\delta T = 1 \times 10^{-8}$ K). $r_e = 1$ Å.

and a Morse oscillator,

$$V(r) = D_e(1 - \exp\left[-a(r - r_e)\right])^2. \tag{4.22}$$

The truncated HO captures the region where the internuclear separation becomes negative, but fails to describe the bond dissociation. The Morse oscillator has the opposite limitations describing the dissociation well but without $V \to \infty$ as $r \to 0$.[h]

Despite these improvements over the HO potential, the resulting entropy still becomes negative for an oscillator with a frequency above 1000 cm^{-1} (Figure 4.4.1). Derivations of truncated HO and Morse partition functions are presented in the Appendix 8.2.1.

These results imply that for systems with *high* frequency oscillators (> 500 cm^{-1}) any classical method for calculating thermodynamic quantities will be incorrect. This includes notable

[h]In fact, this region is largely captured correctly in calculation of the PF as the particle must be bound (to ensure convergence of the integral) thus $E < D_e$ and so $x > -\frac{1}{a}\ln(2)$ which is greater than $-r_e$ for $\omega > 630$ cm^{-1}

methods based on the covariance matrix from Karplus[343] and the interaction energy from Duan[344]. The latter experiencing a smaller deficiency, provided terms in the interaction energy do not include any *high* frequency oscillators.

This result implies that all molecular vibrations ($\gtrsim 200$ cm^{-1}) in a solute-solvent system must be treated quantum mechanically. This includes both bond stretches and bends; as for example, an aliphatic C–H stretch has $\nu \sim 2900$ cm^{-1} and bend $\nu \sim 1350$ cm^{-1}.[345] As such, the deficiency in computing the entropy classically cannot be relieved by simply fixing bond lengths using e.g. the LINCS algorithm.[346] Rather, all angles and bond lengths must be fixed or treated with QM to ensure only oscillators where $h\nu \ll k_B T$ contribute.

4.4.2 Water-in-Water

Limitations of the IGM become particularly apparent when considering cluster-continuum solvation of a solvent molecule. Taking water as a model system and considering the identity reaction,

$$(H_2O)_{n,(aq)} \rightarrow (H_2O)_{n-1,(aq)} + H_2O_{(aq)}$$

the free energy and entropy change is zero. However, when calculated using the IGM the free energy change is exergonic by up to 9 kcal mol^{-1}(Figure 4.4.2) for $n = 8$. Unfortunately, implicit solvent models do not decompose the solvation free energy into entropic and enthalpic components thus ΔS does not include $\Delta\Delta S_{\text{solv}}$. Nevertheless, irrespective of the approximation made within the IGM, and solvent model utilised, the free energy of dissociating a water octamer into a septamer and monomer is incorrect.[i] The dominant contribution to this energy is the $-T\Delta S$ term, which is approximately -10 kcal mol^{-1} (Figure 4.4.2, right). This

[i]Note that the electronic potential energies calculated at DLPNO-CCSD(T)//def2-TZVPP are not counterpoise corrected; however, recalculation with a more complete QZVPP basis afforded only a small difference ($\Delta E_{\text{QZ}} - \Delta E_{\text{TZ}} = -0.3$ kcal mol^{-1}).

effect is exacerbated when the cluster is fully dissociated, with the corresponding identity reaction exergonic by 60 kcal mol^{-1}.

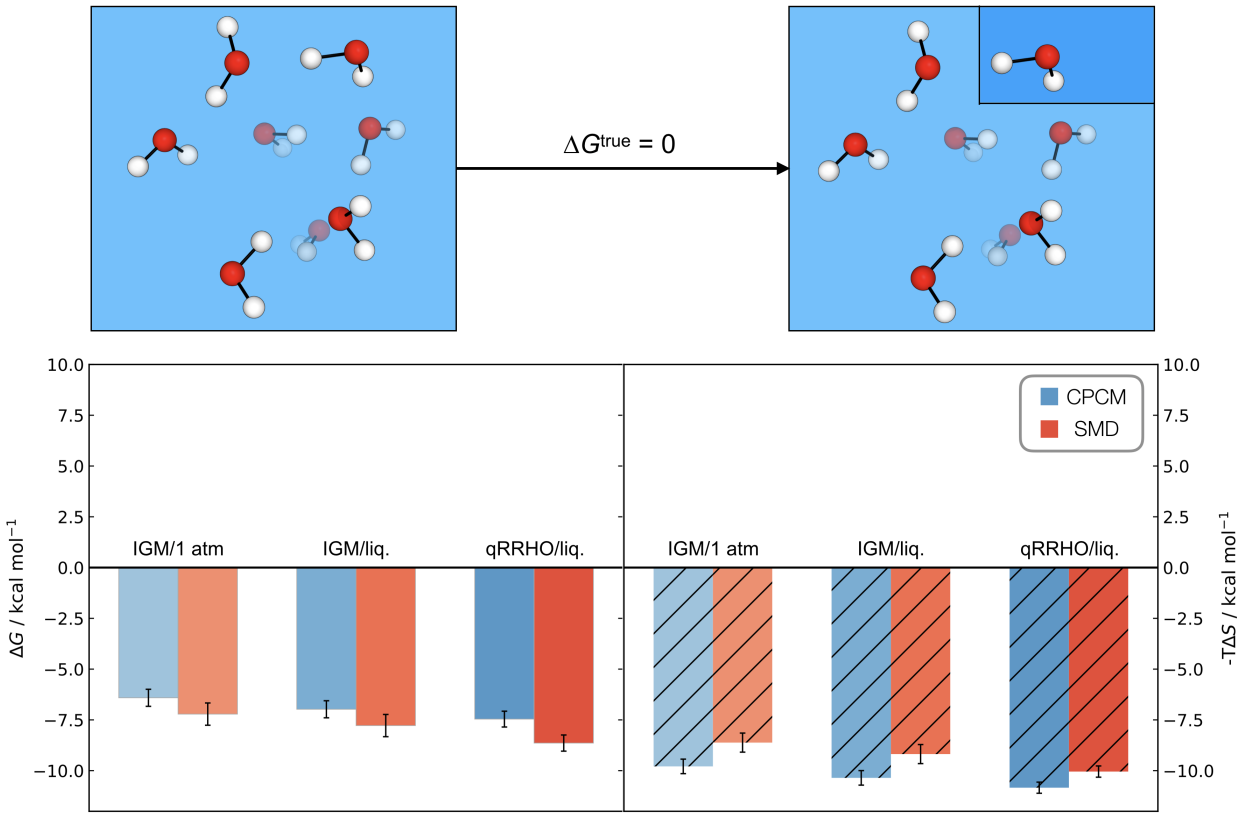


Figure 4.4.2. Free energy and entropy changes upon dissociation of a single water molecule from an octamer cluster immersed in implicit solvent modelled either with CPCM or SMD models (as implemented in ORCA). Potential energies calculated at DLPNO-CCSD(T)//def2-TZVPP//PBE0-D3BJ/def2-SVP and error bars quoted as standard error of the mean over the possible dissociation events. States appropriate for a 1 atm pressure (default in electronic structure codes) and the appropriate $55.5/n$ M molarity for a cluster of n molecules in water (liq.) are shown. Snapshot taken randomly from a revPBE0-D3 quality classical simulation of 64 water molecules at 300 K.

This is unsurprising when on the reactant side the translational entropy associated with a water molecule is treated as a vibration or rotation (within the qRRHO[230] approach) while on the right, the single water molecule has free translation within a PIB (and thus many more populated states). Likewise, with the rotational contribution, the motion on the left

is approximated as harmonic (or between a HO and free rotor within qRRHO) and on the right a free rotor. Clearly, this distinction is unphysical and there should be no difference in treatment for each molecule on the left and right hand sides of the reaction.

4.5 Translational Surface

To use a more realistic treatment of the translational and rotational motion for a solute molecule in solution, the corresponding energy wells are required ($V_{\text{trans}}(x, y, z)$, $V_{\text{rot}}(R_x, R_y, R_z)$, where R_k is a rotation around axis k). For a water-in-water system, the translational surface may be obtained by a time average over a simulation, which – by analogy to the ergodic hypothesis – should approach the average for many water molecules. Averaging the potential over frames from a revPBE0-D3 quality simulation of liquid water from ref. [347] and displacing the most central water molecule in the Cartesian x-direction affords the section function $V_{\text{trans.}}(x, 0, 0)$ (Figure 4.5.1).

Fitting these points to an infinite square well and harmonic potentials, with known QM eigenspectra, affords only moderate agreement to the averaged points on the potential.[j] In contrast, fitting an exponential well,

$$V_{\text{trans}}^{\text{exp}}(x) = k(e^{a|x|} - 1) \tag{4.23}$$

where k, a are parameters to be determined and k has units of energy, affords an excellent agreement between the fit and the true values, which are within error on all points (Figure 4.5.1). This agreement is not surprising considering short-range closed-shell (Pauli) repulsion

[j]Interestingly, Garza's method using the free-volume like approach to calculate the effective volume generates a similar estimate as the square well fit in Figure 4.5.1. $V \approx v_c$; $v_c = (V_{\text{water}}^{1/3} + V_{\text{free}}^{1/3})^3$; $V_{\text{free}} = M_{\text{w}}/N_A\rho - V_{\text{water}}$; $V_{\text{water}} = 27.971$ Å^3 using a spherical approximation of water and radii from ref. [348] , such that $v_c = 78.700$ Å^3 and the side length used in the PIB partition function $l = 4.285$ Å.

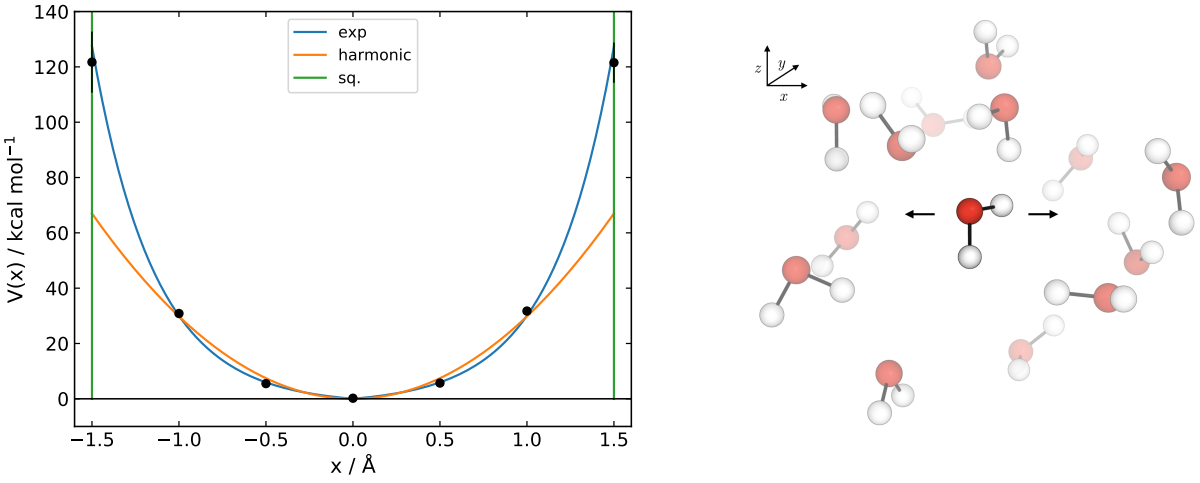


Figure 4.5.1. Translational potential energy surface for a water molecule in liquid water at 300 K (64 water molecules, cubic box, $l = 12.42$ Å), black points. Surface averaged over 500 frames separated by 60 fs at the PBE0-D3BJ/def2-SVP level of theory. Clusters generated by including neighbouring water molecules within a 5 Å spherical distance, shifting in the Cartesian x-direction and evaluating the single point energy. A sample water cluster containing 14 water molecules within a 5 Å sphere is shown. Fitted functions (lines) correspond to exponential (exp, $V \sim e^{|x|}$), harmonic ($V \sim x^2$) and infinite square wells (sq. $V = 0$ if $-1.5 < x < 1.5$, else ∞). Error bars shown as the standard error of the mean on all points.

has an exponential distance dependence.[197] Indeed, in the empirical Buckingham potential the pairwise repulsive term is exponential in distance.[349] With a model for V_{trans} in hand, the following sections explore faster methods to generate the averaged potential and its dependence on the solute and solvent.

4.5.1 Methodology Optimisation

To limit the computational expense of the calculations used to generate the averaged translational potential[k] the DFT calculations may be replaced by tight-binding DFT equivalents (Figure 4.5.2). The well is identical to within the average standard error across the points arising from incomplete sampling: MAD = 2 kcal mol^{-1} *cf.* avg. std. error = 3 kcal mol^{-1}.

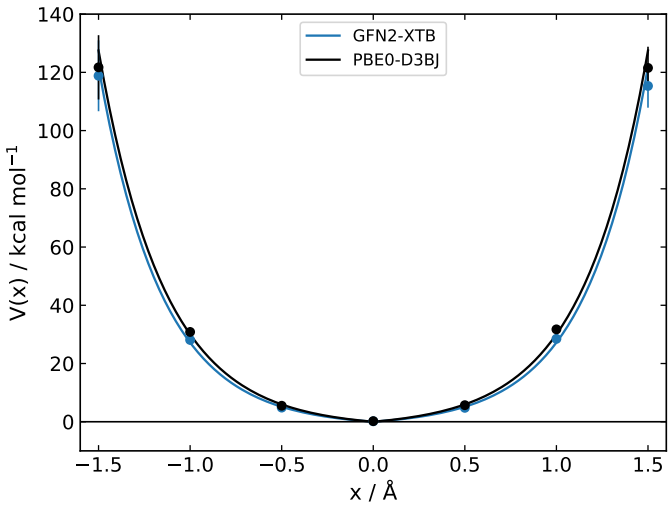


Figure 4.5.2. Comparison of GFN2-XTB and PBE0-D3BJ/def2-SVP surfaces for water-in-water. Trajectory frames taken as Figure 4.5.1.

Likewise, while high quality AIMD simulations provide an accurate representation of the configuration space of the condensed phase system, they are computationally expensive. As such, it is be advantageous to employ a faster method to generate frames over which a potential is averaged. Comparing the translational surfaces of bulk liquid water from DFTB and molecular mechanics (TIP4P) MD frames there is a quantitative change to the exponential fits, but no qualitative difference (Table 4.5.1, Figure 4.5.3).

Compared to the reference DFT-quality simulation, DFTB and MM under/overestimate the confinement of a water molecule within the solvent cage respectively, with the latter deviating

[k]Figure 4.5.1 required 7 points × 500 frames = 3500 DFT single-point calculations.

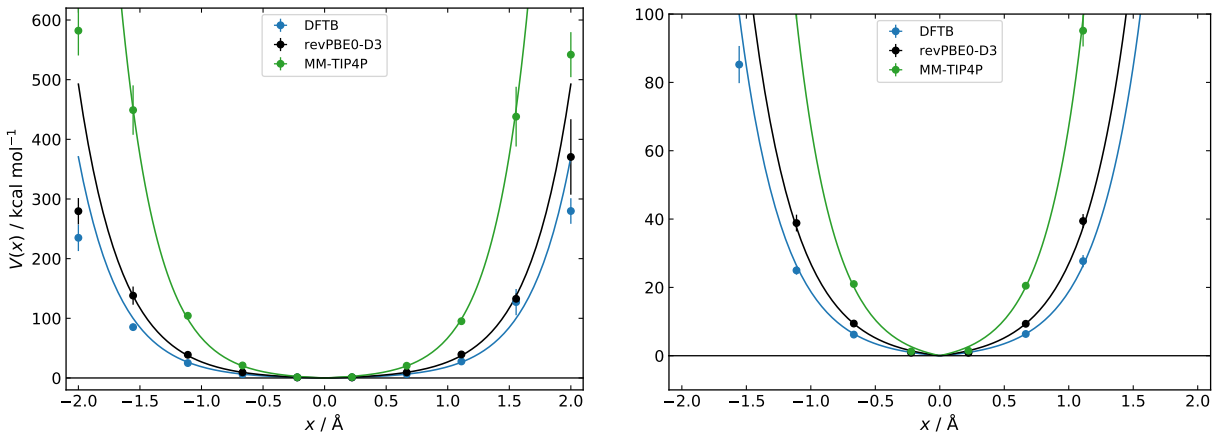


Figure 4.5.3. Comparison of methods to generate 500 frames over which the translational potential $V(x)$ is averaged for the water-in-water system. Frames generated from 300 K liquid water simulations of 30 ps using different reference methods to calculate energies and forces. DFTB used 3ob parameters while the molecular mechanics (MM) used TIP4P parameters, as implemented in GROMACS. Error bars are standard error of the mean.

Method	k / kcal mol^{-1}	a / Å^{-1}
DFTB	1.057	2.931
revPBE0-D3	1.650	2.851
MM-TIP4P	2.377	3.375

Table 4.5.1. Fitted parameters for exponential wells shown in Figure 4.5.3.

more significantly. Given molecular tumbling and solvent reorganisation is a reasonably fast process,[1] short time dynamics should be sufficient to sample all the relevant configurations for small molecules. Therefore, DFTB simulations provide a good compromise between accuracy and computational cost, where reference simulations are not available. With associated errors on k being ~ 50 % and on $a \sim 2$ % from only the methodology choice, any predictions must include that error.

[1]From the Stokes-Einstein relation for the rotational diffusion coefficient $\tau \sim 4\beta\eta r^3$, where r is the molecular radius and η the viscosity, which yields $\tau \sim O(10^{-12})$ for small molecules.

4.5.2 2D and 3D Surfaces

To ensure the exponential well is a good approximation beyond the $V_{\text{trans}}(x, 0, 0)$ section function and generalises to the other Cartesian directions, the corresponding 2D and 3D surfaces were generated for CO_2 in water. Fitting 2D exponential and harmonic wells to the averaged surface shows the former is a much better approximation to the true potential (Figure 4.5.4). This is also obtained over a wider energy window than shown for the water-in-water system, cementing its applicability. The corresponding 3D potential is more challenging to compare, requiring a fourth dimension in energy, but does convey the spherical symmetry. The equivalence of V_{trans} in all directions is unsurprising, as the potential is evaluated in the 'lab' frame and thus the choice of direction is arbitrary when averaged over solute and solvent rotations. Given its spherical symmetry, it is then puzzling why the translational partition function is most commonly evaluated from an infinite cubic well, when the eigenspectrum for the analogous spherical well is known.[350]

4.5.3 Solute and Solvent Dependence

To develop a widely applicable model that does not rely on explicitly solvated dynamics to generate parameters the solute and solvent dependence of the potential is required. To sample a range of interaction types in the solvation of CO_2, methane and alanine in water, acetonitrile and benzene were selected. In all cases an exponential well is an excellent approximation to the translational potential (Figure 4.5.5).

Correlating k and a for the solute-solvent systems reveals they are reasonably independent and, although some broad trends can be extracted, generating a model for *a priori* prediction of k and a may be challenging. Immediately, from Table 4.5.2 both parameters are tightly

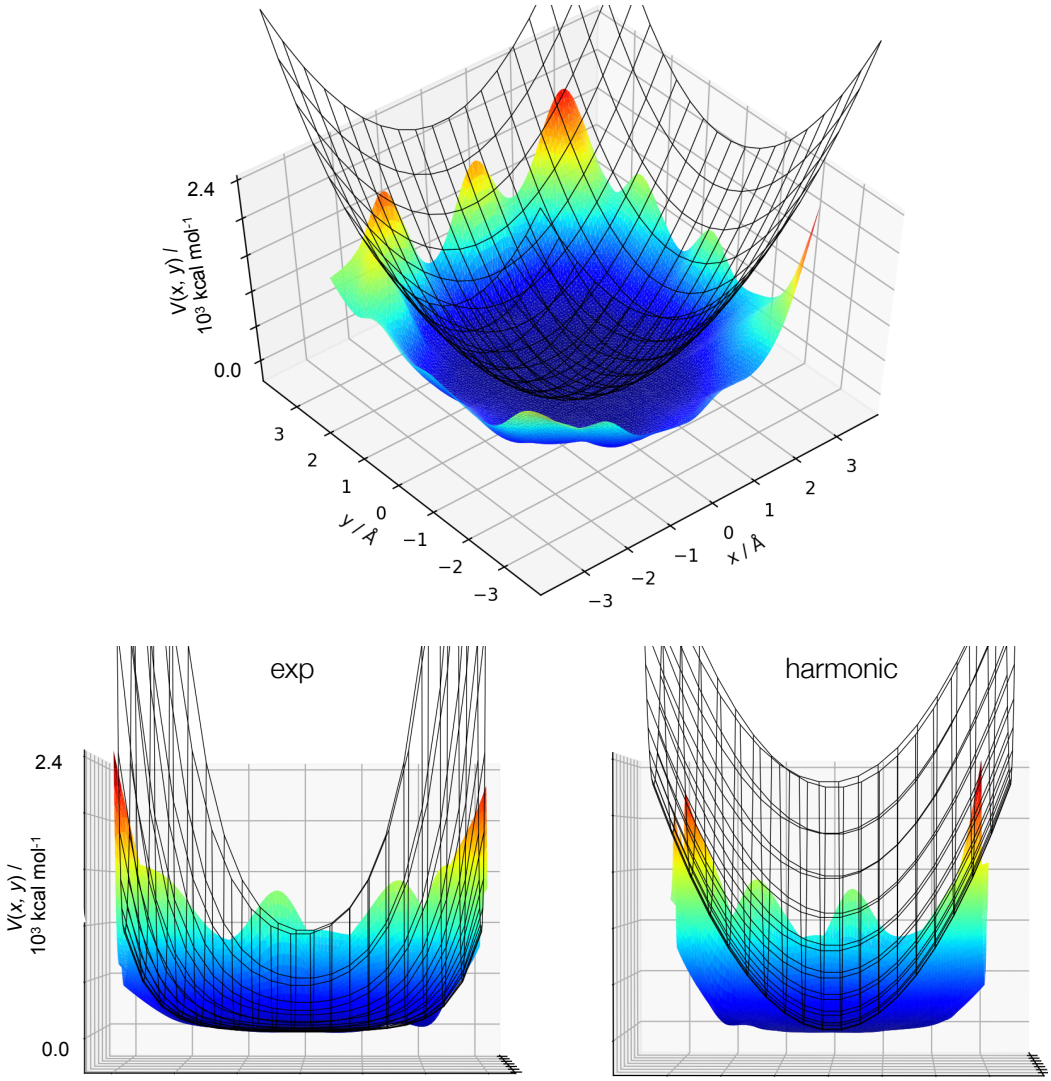


Figure 4.5.4. Translational potential energy surface for CO_2 in water evaluated by averaging GFN2-XTB single point energies over 100 frames separated by 1 ps obtained at the GAFF-TIP4P MM level using a 300 K NVT simulation.

clustered ($k \sim 1$ kcal mol^{-1}and $a \sim 2.6$ Å^{-1}) irrespective of the system. Furthermore, solvents with a higher number density, such as water are closer to the solute and thus generate a more enclosed cavity with a larger a. Unfortunately, this is not uniformly observed; CO_2 in benzene has amongst the largest a, which precludes a predictive model to avoid the time-consuming simulations. Nevertheless, with appropriate error bars on the derived quantities a semi-quantitative prediction may be made based on average parameters $(\bar{k}, \bar{a})$.

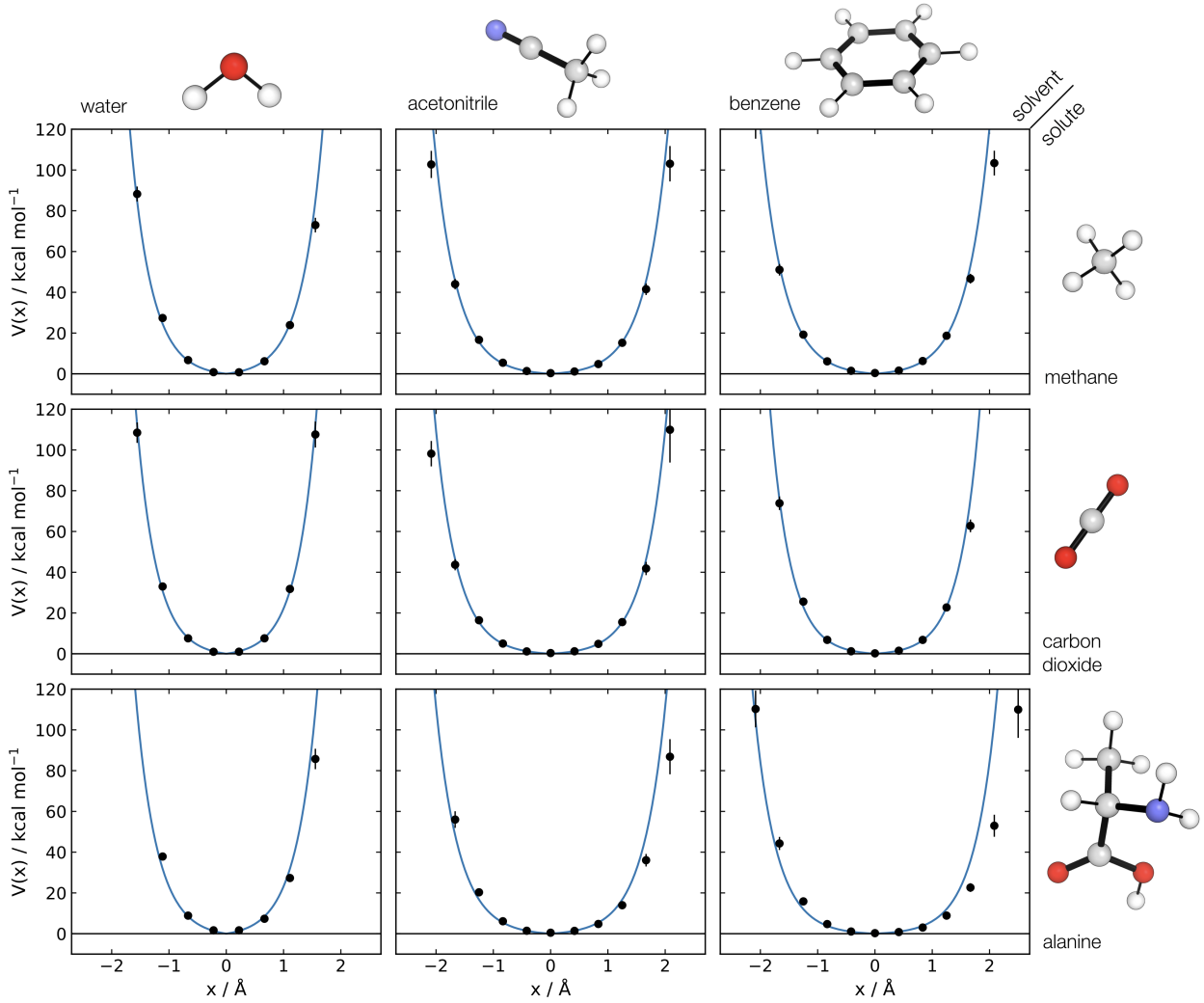


Figure 4.5.5. Translational potential energy surfaces for the displacement of solute molecules immersed in an explicit solvent. Points are averaged over 500 evenly spaced frames selected from 20 ps DFTB(3ob) 300 K Langevin dynamics in a cubic box (l = 12 Å) with a number of solvent molecule to reach the solvent density, excluding the solute volume. Potentials calculated using displacements of ≈ 0.35 Å in the x-direction. Error bars plotted as the statistical error of the mean over frames. Fully solvated clusters extracted from the periodic simulations by taking solvent molecules within 5 Å sphere around the solute. Overlaid exponential wells fit as Figure 4.5.1.

Solute	Solvent	k / kcal mol^{-1}	a / Å^{-1}
Methane	Water	1.278	2.705
Methane	Acetonitrile	0.718	2.500
Methane	Benzene	0.911	2.449
CO_2	Water	1.386	2.832
CO_2	Acetonitrile	0.655	2.562
CO_2	Benzene	0.707	2.809
Alanine	Water	1.692	2.688
Alanine	Acetonitrile	0.796	2.473
Alanine	Benzene	0.524	2.542
Average		1.0 ± 0.4	2.6 ± 0.1

Table 4.5.2. Fitted parameters for exponential wells shown in Figure 4.5.5 at the GFN2-XTB//DFTB(3ob) level of theory.

4.6 Translational Entropy from Spherical Wells

Given the average potential for solute translation is spherical, and well approximated by an exponential well (EW), the partition function must be obtained in order to calculate the translational entropy more accurately than the IGM approximation.

4.6.1 Infinite Spherical Well

First, considering the limiting infinite (step function) case of the exponential well, for an infinite spherical well the potential is given by,

$$V(\boldsymbol{r}) = \begin{cases} 0 & \text{if} \quad 0 \leq |\boldsymbol{r}| < c \\ \infty & \text{if} \quad c \geq |\boldsymbol{r}| \end{cases} \tag{4.24}$$

the Schrodinger equation is soluble,[350] and the eigenvalues are given by,

$$E_{l.n} = z_{l,n}^2 \frac{\hbar^2}{2mc^2} \tag{4.25}$$

where $z_{l,n}$ is the n'th zero of the l'th spherical Bessel function, and $E_{l.n}$ is $2l+1$ fold degenerate. For a light particle (1 amu) in a small cavity (radius, $c = 2$ au) the partition function converges rapidly with n, l which are evaluated on a rectangular grid converged in both n and l. The classical partition function for this potential is,

$$\begin{aligned} Z_c &= \frac{1}{h^3} \int\int e^{-\beta H(\boldsymbol{p},\boldsymbol{x})}\, d\boldsymbol{p}^3 d\boldsymbol{x}^3 \\ &= \frac{1}{h^3} \int_{-\infty}^{\infty} e^{-\beta \boldsymbol{p}^2/2m}\, d\boldsymbol{p}^3 \int_{-\infty}^{\infty} e^{-\beta V(x,y,z)}\, d\boldsymbol{x}^3 \\ &= \Big(\frac{2\pi m k_B T}{h^2}\Big)^{3/2} \int_0^c e^{-\beta \cdot 0} 4\pi r^2\, dr \\ &= \Big(\frac{2\pi m k_B T}{h^2}\Big)^{3/2} \frac{4\pi c^3}{3} \end{aligned} \tag{4.26}$$

and the classical entropy,

$$\begin{aligned} S_c &= k_B T \frac{1}{Z_c} \frac{\partial Z_c}{\partial T} + k_B \ln(Z_c) \\ &= k_B T \Big(\frac{h^2}{2\pi m k_B T}\Big)^{3/2} \frac{3}{4\pi c^3} \Big(\frac{2\pi m k_B}{h^2}\Big)^{3/2} \frac{4\pi c^3}{3} \cdot \frac{3}{2} T^{1/2} + k_B \ln(Z_c) \\ &= \frac{3k_B}{2} + k_B \ln(Z_c) \end{aligned} \tag{4.27}$$

The quantum and classical PFs for two systems are calculated as a function of the number of eigenvalues ($\bar{n}$) included in the summation (Figure 4.6.1). These are a reasonably classical system, with eigenvalues separated by far less than $k_B T$ ($c = 20$ au, $m = 48$ amu) and a more quantum one of a H_2 molecule in a cavity with radius 1 Å (Figure 4.6.1).

As expected for the classical system Z converges to its classical analogue (Figure 4.6.1a).

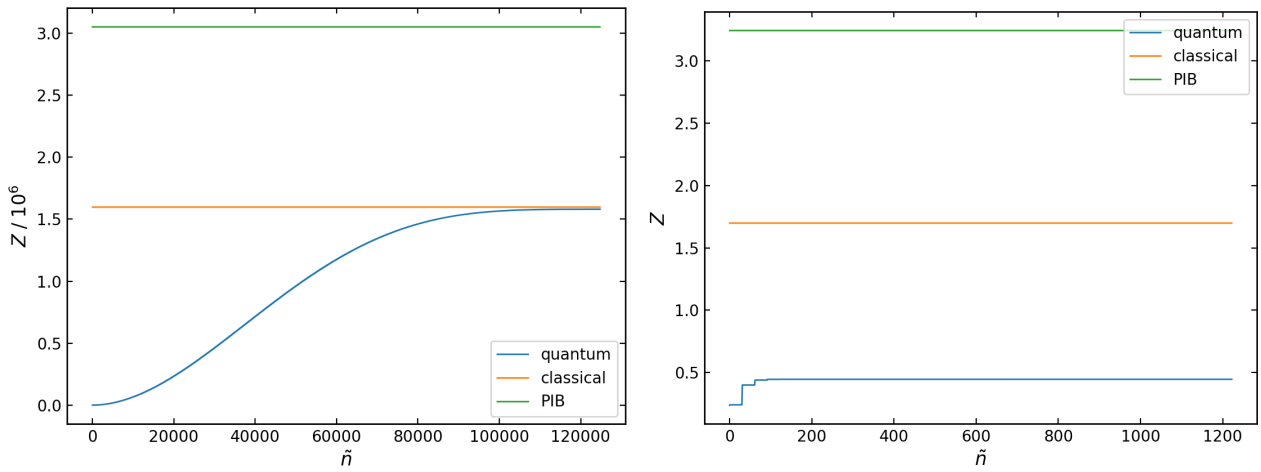


Figure 4.6.1. Z for a particle in a sphere (a) $c = 20$ au, $m = 48$ amu, $\delta x = 10^{-1}$, $\max(x) = 10^3$, $\max(l) = 10^3$ and (b) $c = 1$ au, $m = 2$ amu, $\delta x = 10^{-2}$, $\max(x) = 10^2$, $\max(l) = 10^2$. The partition function for a PIB (using the integral approximation) with $l = 2c$ is also plotted.

The PF for H_2 translation is, however, not well reproduced classically with an overestimation of a factor of 3 compared to the quantum analogue (Figure 4.6.1b). However, in the vast majority of cases the classical PF is likely to be a sufficient approximation, particularly as the entropy depends on the log of the PF.

For a system of mass 48 amu and $c = 2$ Å the PFs converge to within 5% and the resultant entropy to within 0.2 J K^{-1} mol^{-1} at $\tilde{n} = 10^5$, well below the acceptable error (Figure 4.6.2). Interestingly, a particle in a cubic box treatment with a side length equal to the diameter of the sphere is a reasonable approximation to the spherical analogue, with only a ~ 5 J K mol^{-1} difference in the corresponding entropy.

4.6.2 Exponential Well

For a spherical potential of the form $V(\boldsymbol{r}) = ke^{a|\boldsymbol{r}|}$ no exact analytic solution of the time-independent Schroedinger equation exists. As such, the eigenvalues must be either approxi-

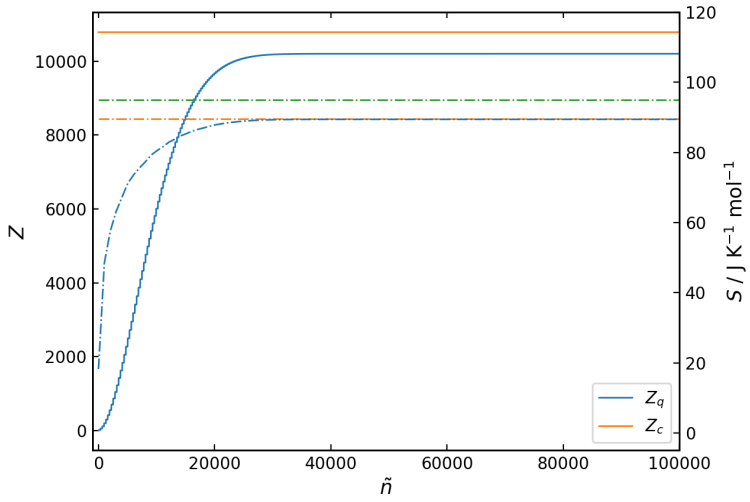


Figure 4.6.2. Classical and quantum PFs (solid lines) and entropy (dot-dashed lines) vs the number of eigenvalues included in the summation ($\bar{n}$). PIB entropy (green) and classical (orange) and quantum (blue) particle in a sphere. $m = 48$ amu and $c = 2$ Å, $T = 298.15$ K. $\partial Z/\partial T$ evaluated numerically $\delta T = 10^{-8}$ K.

mated or found numerically. For simplicity, the 1D case ($V(x) = ke^{a|x|}$) is considered initially, the eigenstates of which are solutions of,

$$\begin{aligned} \hat{H}\Psi(x) &= E\Psi(x) \\ \left\{ -\frac{1}{2m}\frac{d^2}{dx^2} + ke^{a|x|} \right\}\Psi(x) &= E\Psi(x) \\ \Psi''(x) + 2m(ke^{a|x|} - E)\Psi(x) &= 0 \qquad : x \in \mathbb{R} \end{aligned} \tag{4.28}$$

with boundary conditions $\Psi(-\infty) = \Psi(\infty) = 0$ ensuring the WF is normalisable. Seemingly this potential also does not have an analytic solution.[m] As such, the eigenstates and associated eigenfunctions are found numerically (Figure 4.6.3).

Henceforth the exponential potential refers to $V(x) = k(e^{a|x|} - 1)$, which takes the same zero of energy as a PIB/particle in a sphere. The first few eigenvalues of this potential are visualised in comparison with a 1D infinite square well (PIB) with a similar functional form

[m] An exponential potential on $x \in [0, \infty)$ with $\Psi(0) = \Psi(\infty) = 0$ has been solved.[351]

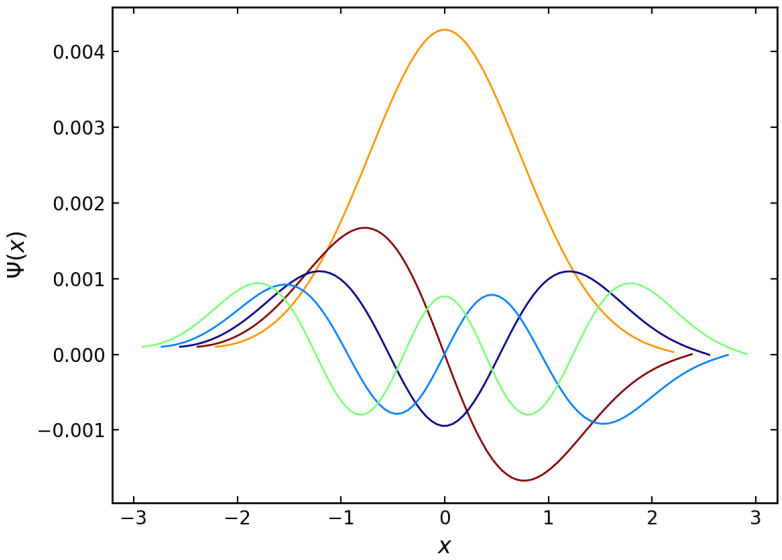
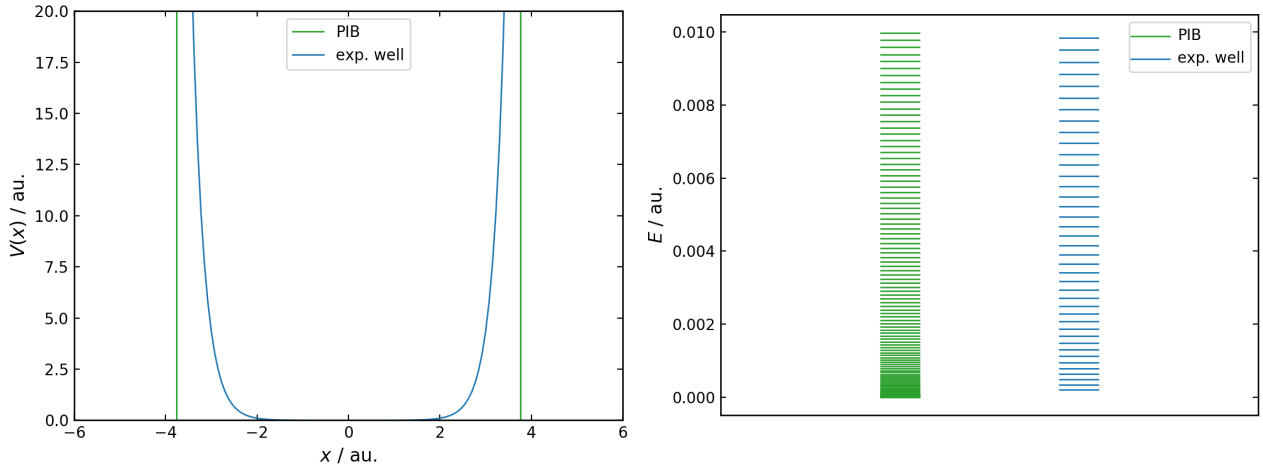


Figure 4.6.3. First few eigenstates of Eqn. (4.28) found numerically using a sweep over energies ($E \in [0.7, 10.0]$), $m = 1$, $a = 1$, $\hbar = 1$, $k = 1$, all in atomic units.

(Figure 4.6.4). The eigenvalues are significantly more separated than those of a similar PIB. Therefore, a classical approximation of the PF may not be suitable.



(a) **(b)**

Figure 4.6.4. (a) Potential energy and (b) eigenvalues ($E < 0.1$ au.) for an exponential (orange, $a = 3.0$) and infinite square well (blue, $l = 7.52$). $\hbar = 1$, $k = 0.057 k_B T$, all in atomic units. $m = 48$ amu.

To calculate the quantum PF only the low energy eigenstates are required. To do so requires a

rapid solution of Eqn. (4.28), which is achieved using *odeint* from the *boost* C++ library.[352] By sweeping energies and calculating the final value of the WF the eigenvalues may be obtained by searching for $\Psi(E, \max(x)) \sim 0$ i.e. a bound and normalisable solution to Eqn. (4.28). As $\Psi(E, \max(x))$ is oscillatory in E the task reduces to a root finding problem (e.g. Figure 4.6.5).

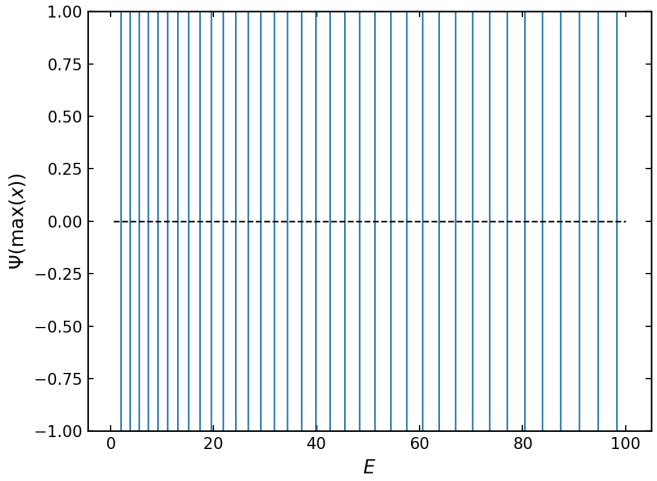


Figure 4.6.5. $\Psi(E, \max(x))$ vs E. $m = 1, a = 1$, $\hbar = 1$, $\max(x) = 7.5$, $x \in [-\max(x), \max(x)]$, $E \in [0.5, 100]$, $\delta E = 10^{-4}$, $\Psi(-\max(x)) = 10^{-14}$, $d\Psi(x)/dx|_{-\max(x)} = 10^{-14}$, all in atomic units.

The classical PF for the system is straightforward to derive from the classical Hamiltonian,

$$\mathcal{H}(p,x) = \frac{p^2}{2m} + k(e^{a|x|} - 1) \tag{4.29}$$

such that,

$$\begin{aligned} Z_c &= \frac{1}{h}\int\int e^{-\beta\mathcal{H}(p,x)}\, dpdx \\ &= \frac{1}{h}\int\int e^{-\beta\{p^2/2m + ke^{a|x|} - k\}}\, dpdx \\ &= \frac{e^{\beta k}}{h}\int_{-\infty}^{\infty} e^{-\beta p^2/2m}\, dp \int_{-\infty}^{\infty} e^{-\beta k e^{a|x|}}\, dx \\ &= \frac{2e^{\beta k}}{h}\cdot(2m\pi k_B T)^{1/2}\cdot\frac{1}{a}\int_{\beta k}^{\infty}\frac{e^{-u}}{u}\, du \\ &= -\left(\frac{8m\pi k_B T e^{2\beta k}}{h^2 a^2}\right)^{1/2}\mathrm{Ei}(-\beta k) \end{aligned} \tag{4.30}$$

where the substitution $u = \beta k e^{ax}$ has been utilised and $\mathrm{Ei}(x)$ is the exponential integral function.

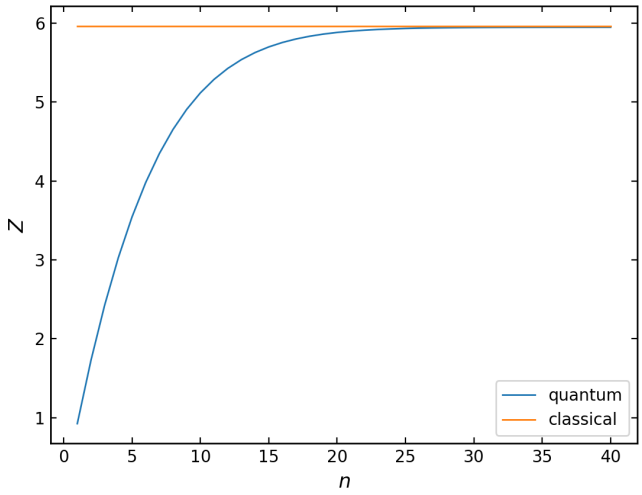


Figure 4.6.6. Quantum and classical PFs vs the number of eigenvalues included in the summation (n). $a = 3.0$ au$^{-1}, k = 0.057\ k_BT, T = 298.15$ K$, m = 8.75 \times 10^4$ au. Eigenvalues were found by root finding in $E = [10^{-5}, 10^{-3}]$ au with 10^4 intermediate values. $\max(x) = 5.0, \Psi(-\max(x)) = 10^{-20}$, $d\Psi(x)/dx|_{-\max(x)} = 10^{-20}$ all in au.

For a system of mass 48 amu and a potential given in Figure 4.6.4a (representative of a CO_2 molecule in a cavity surrounded by H_2O) the translational PF converges rapidly with the number of eigenvalues included in the summation (Figure 4.6.6). Despite the larger separation in eigenvalues compared to a PIB the classical approximation to the PF is still accurate ($\Delta Z_{q\to c} = 0.2\%$).

Proceeding to the 3D case the required eigenfunctions and eigenvalues are solutions of,[n]

$$\Big\{ -\frac{1}{2m}\nabla^2 + ke^{ar} \Big\}\Psi(r,\theta,\phi) = E\Psi(r,\theta,\phi) \tag{4.31}$$

where the wave function is most easily obtained in spherical coordinates for the spherical exponential well, with $r \in [0,\infty)$, $\theta \in [0,\pi]$, $\phi \in [0,2\pi]$. This bears resemblance to the hydrogen atom with a modified potential in r. Therefore, by analogy, the WF may be expressed as a product of known angular, and radial parts,

$$\Psi(r,\theta,\phi) = N\varphi_l(r)P_l^m(\cos\theta)\mathrm{e}^{im\phi} \tag{4.32}$$

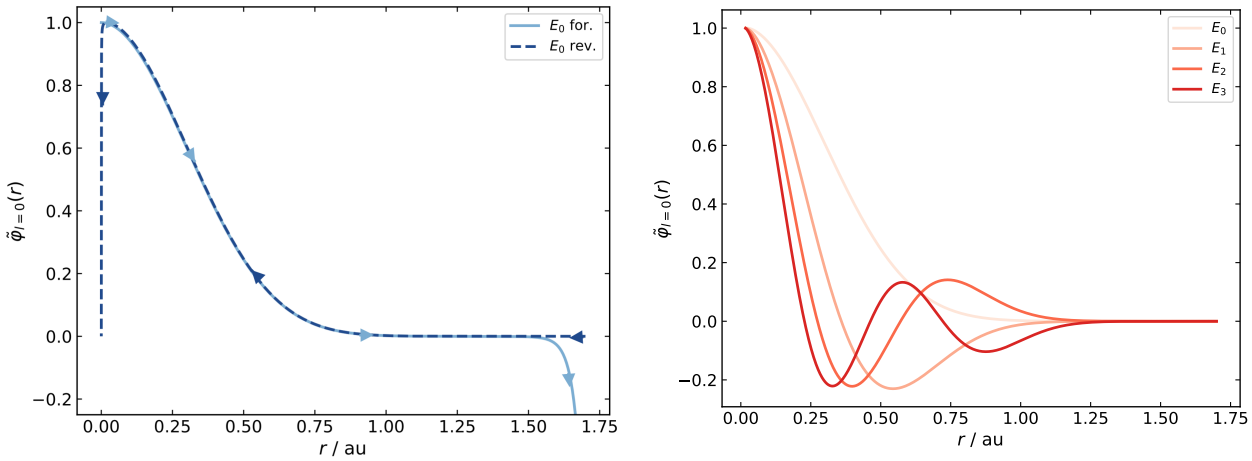
where the solutions are obtained from solving the radial equation with the exponential potential i.e.,

$$\frac{\mathrm{d}^2\varphi_l}{\mathrm{d}r^2} + \frac{2}{r}\frac{\mathrm{d}\varphi_l}{\mathrm{d}r} + \left(2m\left(E - k\mathrm{e}^{ar}\right) - \frac{l(l+1)}{r^2}\right)\varphi_l = 0 \tag{4.33}$$

where once again only a partial eigenspecturm is required so solving Eqn. (4.33) numerically over possible l ($l \in \mathbb{Z}_{\geq 0}$) at fixed a, k should be sufficient. In a similar vein to the 1D case Eqn. (4.33) may be solved numerically by ensuring $\varphi_r(r) \to 0$ as $r \to \infty$ as the potential diverges and $\varphi_r(0) \in \mathbb{R}$, so the WF is normalisable. Noting that the ODE diverges at $r = 0$ and will lead to numerical instabilities, the first few eigenfunctions are shown in Figure 4.6.7b,

[n]Where the constant shift $-k$ from setting the zero of energy is omitted as it just serves as a constant eigenvalue shift.

found by sweeping energies and selecting normalisable functions. Propagating away from the centre of the potential (forwards) again $\varphi_r(r = r_{\max})$ oscillates between $\pm\infty$ in E, thus the eigenvalue is taken between where $\varphi_r(r_{\max})$ changes sign. An almost identical eigenvalue ($\Delta E < 0.01$ cm^{-1}) is obtained when $\varphi(0) \nrightarrow \pm\infty$, with an eigenfunction that resembles a 1s orbital (Figure 4.6.7a). Obtaining higher lying eigenstates is then straightforward, with the first few shown in Figure 4.6.7b. In contrast to the hydrogen atom, the potential does not admit any higher symmetries to afford degeneracy between l levels, despite the eigenfunctions being qualitatively similar to the hydrogen radial wave functions. The $2l + 1$ degeneracy in the magnetic quantum number is retained due to the identical radial component.



(a) **(b)**

Figure 4.6.7. Numerical solution of Eqn. (4.33) for (a) the lowest energy eigenvalue E_0 located by propagating forwards and in reverse from $r = \delta : \delta \ll 1$ and (b) the lowest four eigenvalues where φ is normalised to 1 at δ for clarity. $m = 8.75 \times 10^4$,$k = 5.47 \times 10^{-5}$, $a = 3.0$, using $r_{\max} = 1.7$, $l = 0$. All quantities in atomic units.

The partial eigenspecturm is obtained as above and reasonably efficiently (execution time $\sim$ 1 min), from which summation over states converges reasonably to the classical result for CO_2 in water (Figure 4.6.8a), and is within 1 % for a limiting case (Figure 4.6.8b) at a

lower temperature.[o] This convergence is expected given as the dimensionality of a system increases, it becomes more classical in nature (*cf.* the quantum-classical correspondence in the limit of infinite spacial dimensions).[353]. Noting that the entropy depends on the log of the PF this suggests the classical approximation will be sufficient.

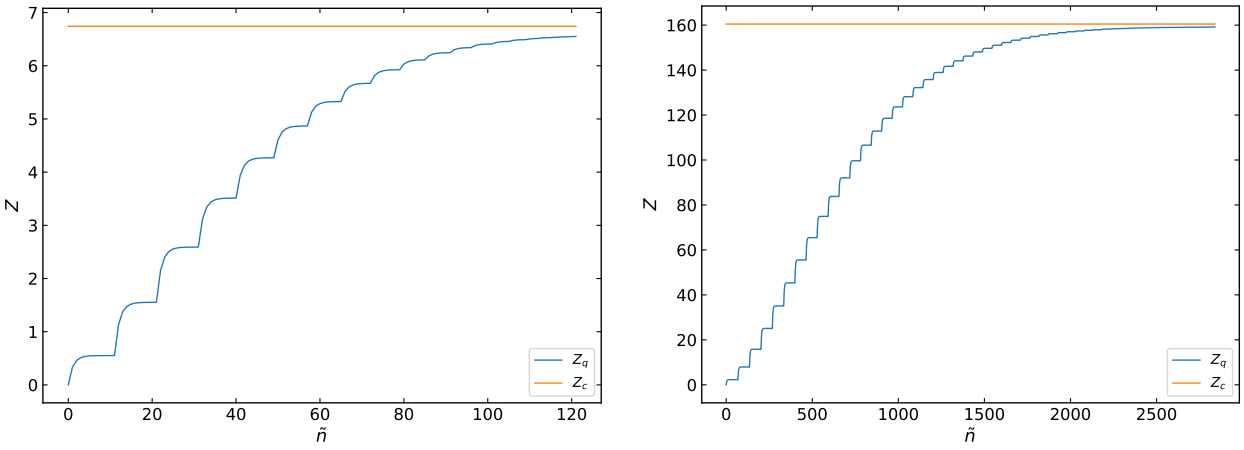


Figure 4.6.8. Quantum (Z_q) and classical (Z_c) partition functions for a 3D EW as a function of number of included eigenvalues ($\tilde{n}$) for (a) CO_2 in water ($a = 1.499$, $k = 0.00221$, $m = 8.75 \times 10^4$, $k_BT = 9.437 \times 10^{-4}$) and (b) a limiting case ($a = 1.499$, $k = 1 \times 10^{-4}$, $m = 8.75 \times 10^5$, $k_BT = 1.583 \times 10^{-4}$). Eigenvalues located using a numerical solution of Eqn. (4.33) by iterating through $E \in [1 \times 10^{-5}, 10k_BT]$ in 20,000 steps with integer $l \in [0, 50]$ from $r = 1 \times 10^{-3}$ to $r_{\max} = 1.7$. All quantities in atomic units.

The classical PF is straightforwardly,

$$\begin{aligned} Z_c^{\exp} &= \frac{1}{h^3} \int \int e^{-\beta \mathcal{H}(\boldsymbol{p},\boldsymbol{r})}\, d\boldsymbol{p} d\boldsymbol{r} \\ &= \frac{1}{h^3} (2m\pi k_B T)^{3/2} \int_{-\infty}^{\infty} e^{\beta k} e^{-\beta k e^{a|\boldsymbol{r}|}}\, d\boldsymbol{r} \\ &= e^{\beta k} \left(\frac{2m\pi k_B T}{h^2} \right)^{3/2} \int_0^{\infty} e^{-\beta k e^{ar}} 4\pi r^2\, dr \\ &= 4\pi e^{\beta k} \left(\frac{2m\pi k_B T}{h^2} \right)^{3/2} \int_0^{\infty} r^2 e^{-\beta k e^{ar}}\, dr \end{aligned} \tag{4.34}$$

[o]A reduced temperature ($T = 50$ K) is used in this system to limit the number of eigenvalues required to converge the quantum PF. The quantum–classical correspondence is expected to be even more exact at larger higher temperatures.

where the integral is non-analytic, and thus must be evaluated numerically, but is smooth and well localised around 0 (Figure 4.6.9a), making a numerical solution possible. However, from Eqn. (4.34) the integrand is exponential in k and super-exponential in a making the Z_c dependence significant (logarithmic, Figure 4.6.9b). Here it is worth noting that the dominant contribution to the integral is for $V < 10k_BT$, making fits up to a 100 kcal mol^{-1} maximum non-optimal when the integrand is zero above $V \sim 5$ kcal mol^{-1}. Instead, the fit should be focused in the low-energy regions for optimal accuracy.

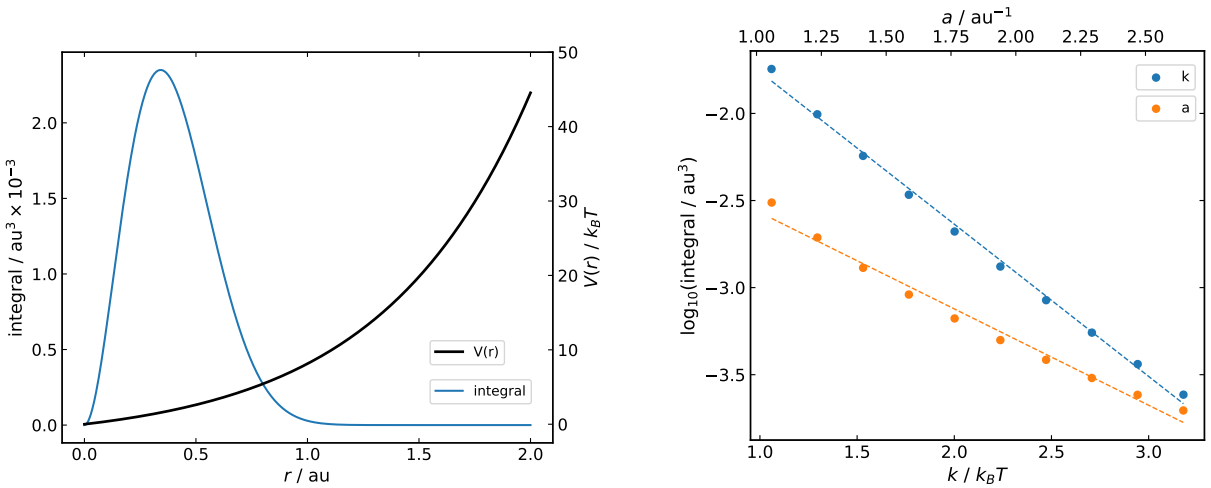


Figure 4.6.9. (a) Integrand from Eqn. (4.34) and potential $V(r) = k(e^{ar} - 1)$, and (b) dependence of the integrand on k and a. Parameters as Figure 4.6.8a with all quantities in atomic units.

The corresponding entropy for a molecule confined to an EW can be derived from the PF Eqn. (4.34) as,

$$S_{\text{trans}}^{\text{exp}} = k_B\left(\frac{3}{2} - k\beta + \ln(Z_c^{\text{exp}}) + \frac{4\pi k\beta\Lambda}{Z_c^{\text{exp}}}\int_0^\infty r^2 e^{ar} e^{-\beta k(e^{ar}-1)}\,\mathrm{d}r\right) \tag{4.35}$$

where $\Lambda = (2m\pi/\beta h^2)^{3/2}$ and the temperature derivative of the PF is shown in Appendix

8.2.1. The associated internal energy is then,

$$\begin{aligned} E_{\text{trans}}^{\text{exp}} &= k_B T^2 \left(\frac{\partial \ln Z_c^{\text{exp}}}{\partial T} \right)_{N,V} \\ &= \frac{k_B T^2}{Z_c^{\text{exp}}} \left(\frac{\partial Z_c^{\text{exp}}}{\partial T} \right)_{N,V} \\ &= \frac{3k_B T}{2} - k + \frac{4\pi k \Lambda}{Z_c^{\text{exp}}} \int_0^\infty r^2 e^{ar} e^{-\beta k (e^{ar}-1)} \, \mathrm{d}r \end{aligned} \tag{4.36}$$

which reproduces the correct equipartition limit as $k \to 0$, and likewise as a vanishes, again $E_{\text{trans}}^{\text{exp}} \to 3k_B T/2$.

Regenerating surfaces over a narrower energy window generates more accurate parameters and translational entropies that are around one half (43 ± 2 %) of their PIB analogues (at a 1 M effective volume) and provide a theoretical basis for sub-unity scalings proposed in refs. [331–334] (Table 4.6.1). As expected, and in contrast to the IGM, the EW recovers a solvent dependence on the translational entropy (and energy, from Eqn. (4.36)) which should be present. For higher number density solvents e.g. water that are more tightly packed around the solute, the translational entropy is uniformly smaller than for acetonitrile or benzene, as would be expected.

4.6.3 Test Cases

With a new method to calculate the translational entropy experimental validation is required. Two benchmark sets that comprise accurate activation and reaction entropies were selected (Appendix 8.2.5). Calculating IGM entropy differences reveals they are subject to a reasonably systematic shift using a 1 M effective volume (Figure 4.6.10). Both the entropies of reaction (blue, Figure 4.6.10) and activation (red, Figure 4.6.10) are bimolecular and therefore do not have a vanishing translational entropy contribution. Introducing an

Solute	Solvent	k / kcal mol^{-1}	a / Å^{-1}	$TS_{\text{trans}}^{\text{PIB, 1 atm}}$	$TS_{\text{trans}}^{\text{PIB, 1 M}}$	$TS_{\text{trans}}^{\text{exp}}$
Methane	Water	1.048	2.918	9.623	7.729	2.5
Methane	Acetonitrile	0.529	2.793	9.623	7.729	3.162
Methane	Benzene	0.679	2.736	9.623	7.729	2.997
CO_2	Water	0.545	4.075	10.52	8.626	3.364
CO_2	Acetonitrile	0.446	2.930	10.52	8.626	4.104
CO_2	Benzene	0.415	3.431	10.52	8.626	3.877
Alanine	Water	0.53	4.083	11.147	9.252	4.009
Alanine	Acetonitrile	1.005	2.127	11.147	9.252	4.625
Alanine	Benzene	0.368	2.878	11.147	9.252	4.903

Table 4.6.1. Fitted parameters for exponential wells for different solute-solvent systems at the DFTB//XTB level of theory Fits performed over the [0, 10] kcal mol^{-1} energy window on ~ 10 points for average potentials (500 points) with $x \in [-1, 1]$ Å. $T = 298.15$ K.

EW treatment of the translational entropy reduces the error to experimental values by some 3 kcal mol^{-1}, while also including an error that almost always encompasses the correct value (at 1σ). This error arises from an average model generated using $\bar{k}, \bar{a}$ parameters (from Table 4.6.1), their associated standard deviations and propagating the assumed uncorrelated errors through to ΔS.

Once again, the solvent reorganisation entropy is not included in the calculated values as it is parametrised into the implicit solvent models. Although the overall solvent reorganisation free energy for a reaction is thought to cancel exactly, the environmental $\Delta H_{\text{env}}, T\Delta S_{\text{env}}$ terms may be significant (up to ~ 10 kcal mol^{-1}, with opposing signs).[354] However, a recent literature review of protein-ligand binding found little entropy-enthalpy compensation suggesting a system dependence.[355] Here, it would seem surprising if the solvent reorganisation entropy were to cause such a systematic shift given the diversity of solvents (methanol, dimethyl-formamide, acetonitrile, dimethyl-sulfoxide) suggesting that the origin lies in the

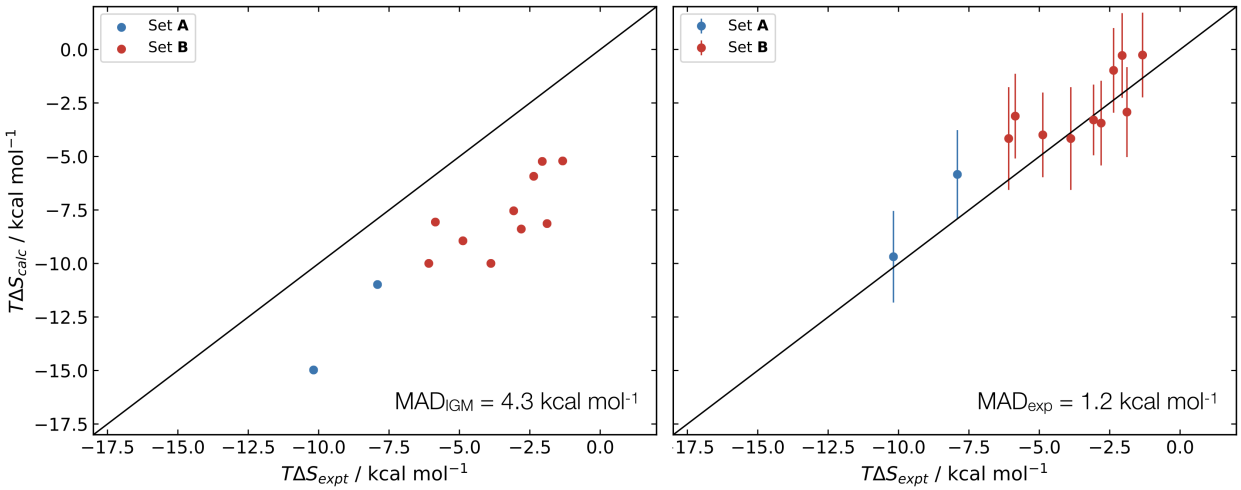


Figure 4.6.10. Correlation between calculated and experimental entropy contributions to the free energy for test sets **A, B**, shown in Appendix 8.2.5. Left are ideal gas values (1 M) and right EW generated from an average model and 1σ errors plotted from standard deviations in k, a.

intrinsic translational contribution. Using the EW approximation to the translational entropy here increases the absolute accuracy over the standard IGM treatment and provides a quantification of the likely error.

Moving to the striking case of water-in-water presented above, which illustrated the limitation of the IGM, an EW corrects almost all of the IGM deficiencies and affords a free energy difference close to zero (Figure 4.6.11). The agreement is somewhat surprising given (1) the SMD solvent model is parametrised to the IGM, (2) the rotational contribution remains at the rigid rotor level and (3) an average EW model is used. Explanations for some of these points are given in the following.

Free energy differences for the $(H_2O)_8 \longrightarrow (H_2O)_7 + H_2O$ identity reaction using an EW may be calculated in two ways (blues, Figure 4.6.11); either treating the translation of one molecule within the octamer and the separated components [$(H_2O)_8$, $(H_2O)_7$, H_2O] using an EW, and deleting three low frequency modes approximately corresponding to translation

(exp_1), or just modifying (PIB→EW) the translational entropy and internal energy of each whole component (exp_2). Interestingly, these two methods lead to an almost identical free energy change ~ 0. This is due to the vibrational entropy (treated as mostly rotation within the qRRHO with frequencies < 50 cm^{-1}) matching the EW translational entropy (~ 3 kcal mol^{-1}, at this temperature).

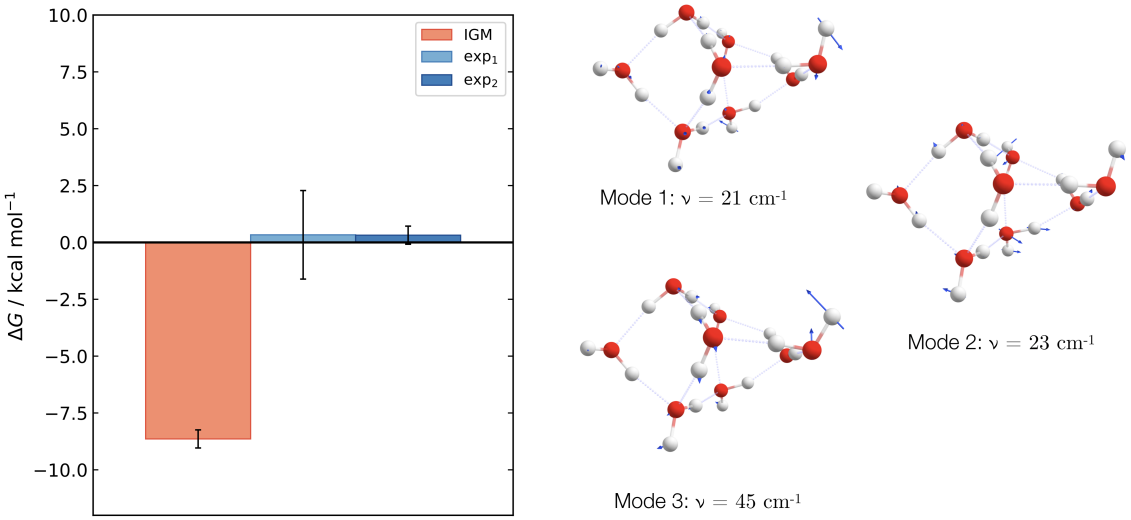


Figure 4.6.11. Comparison of ideal gas (qRRHO) and ideal gas with a EW treatment of the translational entropy free energy changes for the $(H_2O)_8 \longrightarrow (H_2O)_7 + H_2O$ identity reaction in implicit (SMD) water. The three lowest energy modes corresponding to overall translational modes treated with an EW rather than a harmonic oscillator approximation for the water $(H_2O)_8$ are also shown.

While the SMD implicit solvent model has been parametrised to IGM solvation free energies, water clusters are absent from the test set,[131] making the large discrepancy less surprising. Furthermore, while the treatment of the rotational contribution to the entropy and internal energy changes when dissociating a water molecule from a cluster the absolute magnitude is small ($TS_{\text{rot}}^{H_2O} = 3.2$ kcal mol^{-1}) and is compensated on the left-hand side of the reaction by the vibrational modes that are mostly treated as rotations. Finally, using an average model leads to a larger error for exp_1 value, but encompasses the correct value; while, in exp_2 the translational component broadly cancels.

Finally, host-guest binding affinity calculations performed in Chapter 3 may now be re-evaluated with this new methodology. A set of binding entropy changes are available in the literature,[356] the systems comprise more than 300 atoms, making accurate DFT calculations (hybrid, with Hessians) prohibitively expensive. Therefore, the $[Pd_2L_4]^{2+}$ systems ($\sim$ 140 atoms) investigated in ref. [280] will form a test set to calculate binding free energies using an EW approximation. Scaling the translational entropy down with an average EW approximation leads to excellent agreement to experiment where low frequency modes are treated as rotors (qRRHO, Figure 4.6.12), in contrast to the standard IGM. This suggests that, again, the EW provides improved accuracy while being more theoretically justified.[p]

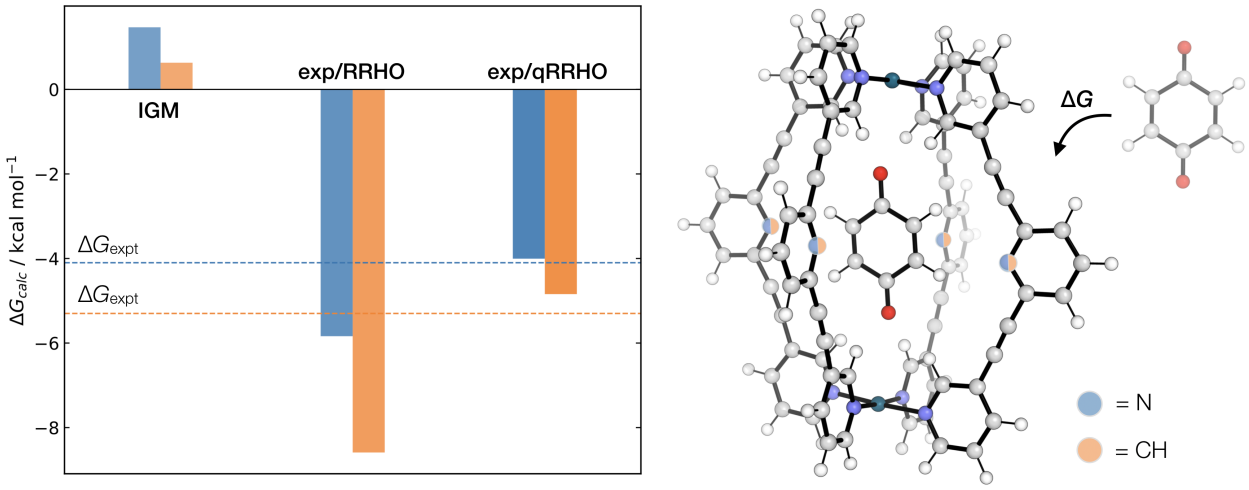


Figure 4.6.12. Comparison of free energy contributions added to SMD(DCM)/PBE0-D3BJ/gcp-def2-SVP binding energies for canonical benzoquinone binding within two $[Pd_2L_4]^{2+}$ metallocages. IGM uses a PIB treatment of the translational entropy while exp uses an exponential well.

[p]There, however, remains a significant functional dependence on $\Delta E_{\text{bind, calc}}$ of $\sim$ 5 kcal mol^{-1}, and higher level coupled cluster calculations are not routinely possible for a system with $\sim$ 150 atoms.

4.7 Conclusion

The ideal gas model (IGM) has been implemented and used in quantum chemistry programs for the best part of 30 years with almost no development and, while successful for gas phase reactions, it remains suboptimal for the solution phase. The translational contribution to the entropy, which dominates the free energy cost of associating two molecules, uses a particle in an infinite cubic well approximation to calculate the partition function. Here, we have shown an exponential well (EW) can be used to evaluate the translational entropy more accurately, including the correct functional form and symmetry of the problem. The classical partition function and resulting entropy are easily obtainable from numerical integrals and, in general, converge to their quantum counterparts.

However, with this treatment some limitations remain. Namely, (1) high-quality dynamics simulations are required to generate frames over which the EW is fit. This is partially alleviated by using an average model but introduces a $\sim$ 2 kcal mol^{-1} error. (2) The overestimation of the entropy lost by the IGM is often cancelled directly by the basis set superposition error (or dispersion overestimation). For example, in the water-in-water test case using uncorrected PBE0-D3BJ/def2-SVP free energies affords agreement largely within 1 kcal mol^{-1} to the true free energy difference. This observation hints to why the prototypical B3LYP/6-31G(d) method affords such good agreement to experimental free energies. (3) Perhaps most importantly, implicit solvent models that are generally (excluding non-polar solutes in non-polar solvents) important for accurate solution phase energetics implicitly include corrections for using the IGM in solution by their parametrisation to experiment.

In summary, while an EW approximation provides a more sound theoretical basis for translational motion in solution, it is unlikely to provide significant improvements over the IGM. A more effective approach to calculate accurate free energy differences is perhaps to use

'free energy' methods in combination with fast, accurate and reactive potentials. Methods towards this latter strategy are developed in Chapter 6.

4.8 Methods

Density functional theory (DFT) and wave function (WF) calculations were performed in ORCA v. 4.2.1[246] with Ahlrichs def2 basis sets[247]. DFT calculations used PBE0[82, 86] and M06-2X[89] density functional approximations with and without D3BJ[94, 313] damping respectively. Both DFT and WF (DLPNO-CCSD(T)[118]) calculations made use of default auxiliary basis sets. Approximate counterpoise corrections performed using the gCP scheme from ref. [357] with parameters appropriate for the calculation. Fitted functions used *scipy.optimise.curve_fit*, numerical ordinary differential equation (ODE) propagation from *scipy.integrate.odeint* for toy models.[292] Converged partition functions obtained by summing partial eigenspecta located by root finding on numerically propagated ODEs using 4th order Runge Kutta[358] propagation implemented in *odeint* from the *Boost* C++ library.[359] In all cases, numerical integration used *scipy.integrate.quad* over a region that converges the integral with respect to the upper/lower bound (where infinite). Ideal gas and quasi-rigid rotor harmonic oscillator calculations performed in *otherm*. Implicit solvent calculations made use of the CPCM[360] and SMD[131] models implemented in ORCA while explicitly solvated calculations used MD simulations. Classical molecular mechanics MD performed in GROMACS v. 2019.2 using a minimise-equilibrate-production sequence.[244] Unless otherwise specified, all simulations were performed at 298.15 K using velocity rescale or Langevin thermostats. Periodic DFTB calculations performed with DFTB+[361] using 3ob[362] parameters.

5 autodE

5.1 Overview

Mechanistic elucidation has played a central role in the development of chemistry, including discovering and optimising chemical reactions and the design of catalysts. For example, new designs are suggested to reduce an energetic barrier to enhance activity, or increase the energy difference between two transition states and, perhaps, enhance selectivity. While there are systems where computation is challenging (e.g. the Morita Baylis–Hillman reaction[363]), computational investigations of reaction mechanisms remain a useful tool.

Introducing automation may significantly accelerate reaction mechanism elucidation and enable the faster development of novel catalysts. Towards this goal, *autodE* was developed. For a known set of reactants, intermediates and products, *autodE* targets the construction of reaction (free) energy profiles, enabling the simultaneous exploration of different systems and chemical pathways. The following sections describe the implementation and applications of this tool to a wide range of reaction mechanisms. It also outlines recent developments to accelerate the workflows of computational chemists beyond the examples described in this work. Indeed, several collaborative projects used *autodE* to find transition states faster than would otherwise be possible. These include an unusual ring contraction from a 7-membered

heterocycle,[364] diphosphane metathesis via a radical chain process[365] and locating TSs within the final manuscript presented in this thesis.[347]

5.2 Manuscript III

autodE: Automated Calculation of Reaction Energy Profiles – Application to Organic and Organometallic Reactions

5.2.1 Abstract

Calculating reaction energy profiles to aid in mechanistic elucidation has long been the domain of the expert computational chemist. Here, we introduce *autodE* (`https://github.com/duartegroup/autodE`), an open-source Python package capable of locating transition states and minima and delivering a full reaction energy profile from 1D or 2D chemical representations. *autodE* is broadly applicable to study organic and organometallic reaction classes, including addition, substitution, elimination, migratory insertion, oxidative addition and reductive elimination; it accounts for conformational sampling of both minima and TSs, and is compatible with many electronic structure packages. The general applicability of *autodE* is demonstrated in complex multi-step reactions, including metal-catalysed cobalt- and rhodium-catalysed hydroformylation, and an Ireland-Claisen rearrangement.

5.2.2 Introduction

Automating the search for new chemical reactions is seen as one of the 'grand challenges' in computational chemistry.[46, 366] Discovering such thermodynamically and kinetically accessible reactions requires knowledge of the underlying potential energy surface (PES), which – to a first approximation – involves the characterization of reactants, products, intermediates, and transition states (TSs).[51, 367, 368] While this has been routine for computational chemists for decades, locating TSs remains time consuming and a highly non-systematic en-

deavour.[369] To this end, many automated TS-search algorithms have been developed which may be broadly classified into single and double-ended methods.[370, 371] The former requires a single chemical structure from which TSs, intermediates, and products are located with no *a priori* information. Representative examples include the ab initio nanoreactor (AINR),[26] which employs high temperature and pressure molecular dynamics (MD) simulations to identify chemical transformations; the single-ended growing-string method (SE-GSM),[372] which iteratively generates structures along the PES and optimises them until a TS and minima at each side of the string are found; the artificial force induced reaction (AFIR) method,[373] which adds artificial external forces to the original PES of the target system, and the transition state search method which uses high-energy chemical dynamics simulations (CDS) and a geometry-based algorithm to identify reactive pathways.[374] While they allow for free exploration of unknown pathways, these methods remain costly and limited to small systems ($\sim$ 10 reactant atoms) if left to explore all transformations.

On the other hand, double-ended methods use knowledge of both reactants and products to generate TSs. Depending on the specific method either single or multiple steps can be studied. The latter includes reaction network enumeration algorithms, which allow for a systematic search of the reaction space, but hit an exponential wall of complexity if arbitrarily many intermediates are permitted.[13, 16] Single-step methods, although seemingly limited in scope, remain a powerful and affordable approach to explore many chemical processes determined by elementary steps. This is the case, for example, when the goal is to elucidate the origin of regio- or enantioselectivity or compare different synthetically-plausible pathways for a given transformation. Several methods have been developed following this philosophy, from linear and quasi-synchronous transit[375, 376] and related approaches[377, 378] to more recent developments based on growing[379] and freezing string methods[380,

381] (GSM/FSM, Figure 5.2.1). Finally, heuristic approaches guided by either graph-based chemical rules or electronic-structure theory have been employed to explore a wide range of chemical reactions; including the Reaction Mechanism Generator,[382] which follows approaches pioneered by NetGen,[383] and other reaction network approaches.[384, 385]

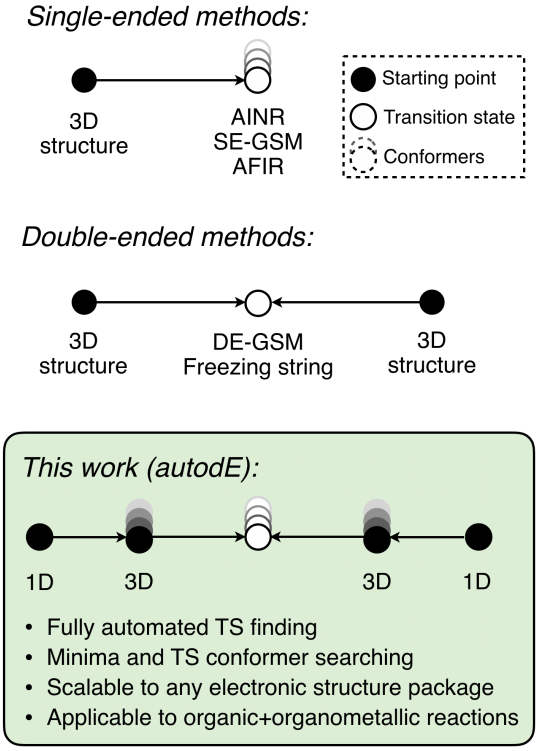


Figure 5.2.1. Comparison of approaches to generate reaction profiles.

Here, we introduce *autodE*, which combines elements of previously reported double-ended methods with the development of a general, easy to use, and freely available package to automate the characterization of reaction pathways. The *autodE* algorithm takes inspiration from related methods, such as AutoTS[386] developed by Schrödinger™ and the Reaction Mechanism Generator developed by Green et al.,[382] and aims to overcome current limitations in applicability, conformational sampling and accessibility; specifically (1) it provides a broadly applicable framework to study organic and organometallic reaction classes; (2) it accounts for conformational sampling of both minima and transition states, which is essential

particularly when exploring flexible systems; (3) it is freely available (open-source Python package distributed under an MIT license) and requires minimal user input and expertise. Moreover, it is compatible with several electronic structure theory packages and extensible to others. This work describes the algorithm and implementation of *autodE* and demonstrates its capabilities in a range of representative organic and organometallic reactions. We demonstrate that *autodE* is capable of locating the different TSs and minima along the PES and delivering a full reaction energy profile using only SMILES representations as inputs, available from most 2D chemical sketching software. To illustrate the functionality and general applicability of *autodE*, we apply it to a range of reaction classes, including complex organic and metal-catalysed reactions.

5.2.3 Methodology

In general, human-guided characterization of TSs and reaction pathways requires: (1) locating reactants and products; (2) locating the TS, usually starting from a guess TS structure generated by chemical intuition; (3*) once reactants, TSs, and products have been characterized, perform a conformational search to identify the lowest energy conformer in each case; and (4*) performing single point energy evaluations to refine energies. While the starred steps are not always performed, they are usually necessary to achieve meaningful conclusion about the reactivity and selectivity of a given reaction step. Our method follows this workflow, which is described in detail in the following sections, along with representative examples (Figure 5.2.2).

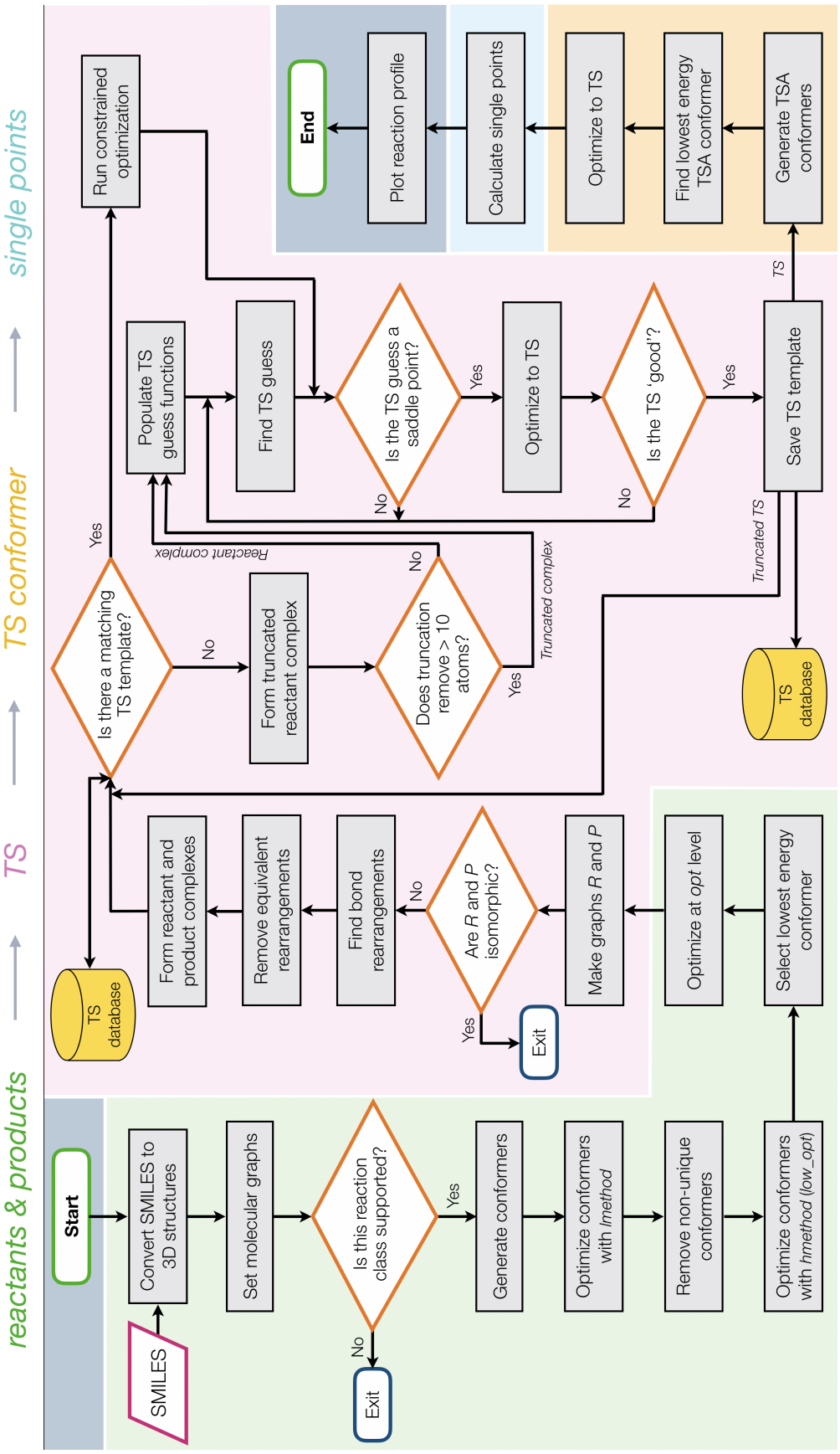


Figure 5.2.2. Diagrammatic workflow used by *autodE* to generate a reaction profile from a set of reactant(s) and product(s) SMILES strings. TSA = transition state analogue, where active bonds are constrained at distances found at the TS found in previous the step.

With a view to generating reaction profiles efficiently using currently available quantum mechanics methods, our implementation makes use of two levels of theory: a lower-level (*lmethod*) and higher-level method (*hmethod*). Both methods have default settings (Table 5.2.1) selected based on their transferability, which can be modified by the user depending on the methods available in the respective software. For example, DFT methods can be used as *lmethod* for optimizations, and wavefunction (WF) methods for *hmethod* single point energy evaluations.

Software Package	Calculation	Default Method
ORCA, Gaussian09, NWChem[a]	low_opt	PBE-D3BJ/def2-SVP
	opt, optts, hess	PBE0-D3BJ/def2-SVP
	sp	PBE0-D3BJ/def2-TZVP
MOPAC	low_opt, opt, sp	PM7
XTB	low_opt, opt, sp	GFN2-XTB

Table 5.2.1. Default methods used in *autodE* calculations. sp = single point energy, opt = optimization, low_opt = low level optimization, optts = transition state optimization, hess = Hessian. *a*. Uses D3 rather than D3BJ dispersion.

autodE is readily extensible to many electronic structure theory codes and force fields in addition to those with currently implemented wrappers (Figure 5.3.1). In the following paragraphs we describe the implementation and key features of *autodE*.

Finding Lowest Energy Conformers for Reactant and Products. Locating the global minimum, or a structure close to it, on a PES is challenging but essential in estimating the feasibility of a reaction.[387] To characterize the lowest energy conformer at a given level of theory requires transforming the 1D SMILES representation into a 3D geometry and searching the conformational space. Several open-source tools are available including

RDKit[294] and CONFAB[306] implemented in OpenBabel, but are generally limited to organic species. *autodE* uses the ETKDGv2[295] algorithm implemented in RDKit, which is known to perform well for organic species, along with its SMILES parser.[307] To determine optimal parameters for the number of conformers generated and the root mean squared displacement (RMSD) threshold used to exclude identical conformers, a set of 20 amino acids were considered. Using CREST[388] as the benchmark method, the generation of 600 conformers and RMSD threshold = 0.1 Å; was sufficient to obtain reliable minima, rendering a conformer within 1 kcal mol^{-1} of the most stable CREST conformer in 90% of the cases (18/20, Figure 5.3.3). Therefore, this inexpensive conformer search was kept as the default for organic molecules and allows for *autodE* to remain method agnostic (i.e. *lmethod*/*hmethod* can be chosen by the user).

For metal complexes, where OpenBabel and RDKit fail to interpret the SMILES string and/or generate a sensible 3D geometry, we utilize our own (Open)SMILES parser and a simple repulsion+bonded (RB) forcefield (Eqn. (5.1)) by randomizing atoms then minimizing the function (Figure 5.3.4–5.3.6):

$$U_{\mathrm{RB}}(\boldsymbol{x}) = \sum_{ij}^{\mathrm{bonds}} k_1(r_{ij} - r_{ij}^{\mathrm{avg}})^2 + \sum_{i>j} \frac{k_2}{r_{ij}^n} \tag{5.1}$$

For both organic molecules and metal complexes, the lowest energy conformer is found by optimizing each conformer at the *lmethod* and excluding non-unique structures based on an RMSD threshold (default 0.3 Å), then an energy threshold (default 0.2 kcal mol^{-1}). The remaining structures are finally optimised at the desired low_opt level with the *hmethod*, and the lowest energy is kept (Figure 5.3.8).

TS Location. Within *autodE* , each species has an associated molecular graph, in which

the atoms correspond to vertices (V) and 'bonds' to edges (E). For a discussion of how bonds are defined, see the SI (Figure 5.3.10). From the set of reactant graphs $\{R_i\}$, the reactant graph (R) is formed as the graph union, and likewise with products $\{P_i\}$ to generate (P). This is represented in Figure 5.2.3 for a Claisen rearrangement, where the graphs R and P are formed from a single reactant and product molecule. If R and P are isomorphic, as in an identity reaction, then the chemical transformation of interest is not obvious from the graph transformation and an exception is raised. Once this has been checked, to find an atom mapping (bijection) from R to P, we generate a set of transformations of $R, \{R'\}$, obtained by adding and/or removing edges. This leads to a set of functions $\{g\}$ that represents all the possible bond rearrangements $[g : E(R) \rightarrow E(R')]$.

The atom mapping(s) $[f : V(R') \rightarrow V(P)]$ are then found where R' and P are isomorphic using NetworkX.[389] A brute force search for $\{g\}$, obtained by enumerating over all possible edge additions/deletions becomes intractable quickly, as the number of operations grows as 2^N for a system with N atoms. However, chemical reactions usually involve a limited number of bonds that are formed and broken in a single elementary step, thereby substantially reducing the search space. We therefore limit the transformation to a maximum of four 'active' bonds where two edges are added and two removed, denoted (2, 2). For a graph X with b bonds ($b_X = |E(X)|$) the principle of microscopic reversibility is used to switch reactants and products if $b_R < b_P$.

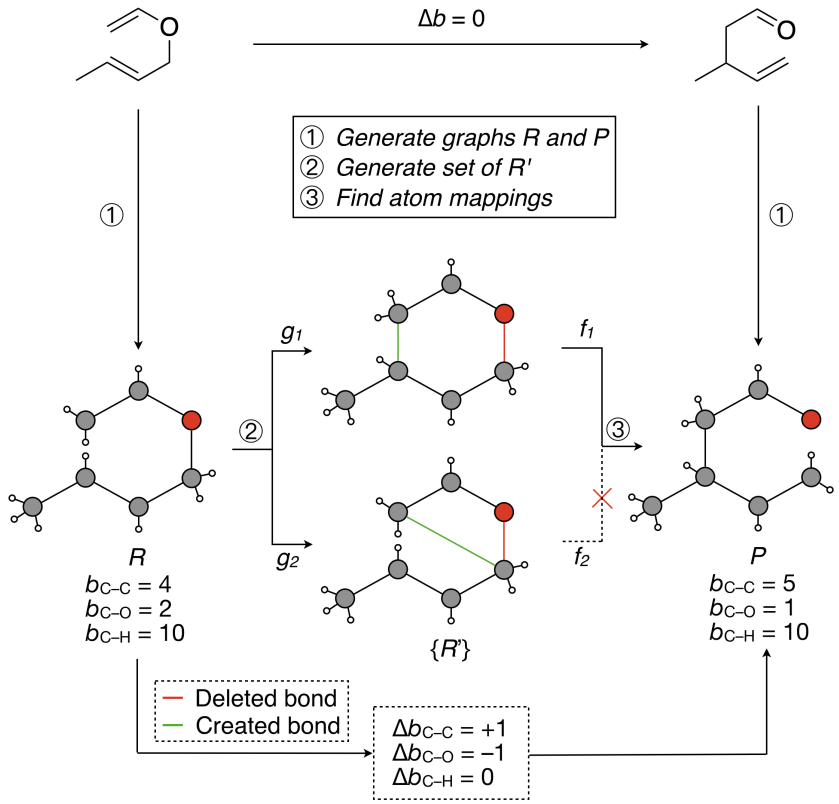


Figure 5.2.3. Atom mappings between reactants and products for a Claisen rearrangement.

From these constraints, five scenarios exist:

I. (1, 1) – substitution reactions e.g., S_N2.

II. (2, 2) – substitution reactions e.g., epoxidation with peroxide.

III. (0, 1) – dissociations e.g., E1 eliminations.

IV. (0, 2) – dissociations e.g., reverse of Diels-Alder cycloaddition.

V. (1, 2) – eliminations e.g., E2 eliminations.

Defining the change in the number of bonds as $\Delta b = b_R – b_P$ as in refs. [386, 390] only when $\Delta b = 0$ does the full set of deleting and adding two edges need to be enumerated. Further acceleration is achieved by calculating $\{\Delta b_k\}$ where k is a bond type (e.g. C–C)

and ensuring any atom (a) does not exceed a maximal valence (e.g., $d(a) \leq 4$ for a carbon atom). For example, in a Claisen rearrangement $\Delta b_{\text{C–C}} = 1$ and $\Delta b_{\text{C–O}} = -1$ such that the enumeration over (1, 1) transformations is targeted to only C–C and C–O bonds (Figure 5.2.3). Once the set of valid $\{g\}$ is found, a TS is located for each of them (if it exists), and following conformational sampling, the lowest energy TS is taken for the calculation of a reaction profile. For the Claisen reaction shown in Figure 5.2.3, only one bond rearrangement is obtained while for the all-carbon analogue (O $\rightarrow$ CH_2) two rearrangements are found ($g = \{g_1, g_2\}$). While reasonably exhaustive, we are yet to find a reaction where the process of generating/checking R' and P isomorphisms and finding the mapping is comparable or more demanding than a DFT calculation.

To generate a saddle point, while ensuring compatibility with multiple electronic structure codes, we favour finding a TS guess by performing a set of constrained geometry optimizations along a linear path between reactants and products. In general, *autodE* attempts 1D constrained searches over forming/breaking bonds except when $\Delta b = 2$ (e.g., Diels-Alder). If this procedure fails and the number of active bonds is more than one, a 2D PES exploration is performed by constrained optimizations and fitting a polynomial up to order 3.[292] From a 2D PES, the lowest energy saddle point is found by locating where the first derivative vanishes, and then using Dijkstra's algorithm[391] implemented in NetworkX.[389] A constrained optimization at the saddle point is then performed to generate the TS guess. Once the TS guess is obtained, it is optimised with the TS optimisers implemented in the electronic structure theory package selected by the user. 'Goodness' of the TS is defined by a significant imaginary mode ($|\nu_{\text{imag}}| > 45$ cm^{-1}) and a significant contribution from the active bonds, or forward/backward displacement along the mode affording graphs isomorphic to reactants/products (quick reaction coordinate calculation,[392] see description of `true_ts()` function in §5.3.6 for further details). In cases where the reactant or product(s)

are not isomorphic to R or P (e.g. barrierless proton transfers), rigid-body conformers of the reactant- and product- complex are generated (see SI §5.3.5) and isomorphisms checked for the forwards/backwards displaced structures to all optimised conformers. We envisage the implementation of nudged elastic band (NEB) approaches to accelerate the computationally intensive 2D PES scans using an initial a linear path from the reactant(complex) by making/breaking bonds to drive towards a product state.

Truncation. Performing tens of constrained optimizations along one or two coordinates in a large system is currently computationally demanding. To accelerate the generation of a TS guess in such systems, the system is truncated. Our implementation truncates over C–X bonds where the carbon atom is fully saturated and is at least two bonds away from the active atoms, and the truncated group is replaced by a hydrogen atom (Figure 5.3.11). The TS is then located using this truncated complex and saved in the template library from which the full non-truncated TS may be found using the templating method described below. For truncation to be utilised it must remove > 10 atoms.

Finding Lowest Energy TS Conformers. TS conformers are located in a similar manner to the protocol described for metal complexes in §5.3.3 (n=8 in Eqn. (5.1) and randomization using a displacement vector of 3 Å). However, here the 'active' bonds are constrained at distances found at the TS using a harmonic potential with a force constant $k_1' = 10k_1$ (Eqn. (5.2)).

$$U_{\mathrm{RB'}}(\boldsymbol{x}) = \sum_{ij \in \mathrm{bonds}} k_1 (r_{ij} - r_{ij}^{\mathrm{avg}})^2 + \sum_{ij \in \mathrm{const.}} k_1' (r_{ij} - r_{ij}^{\mathrm{avg}})^2 + \sum_{i>j} \frac{k_2}{r_{ij}^8} \tag{5.2}$$

An identical method is then used to find the lowest energy transition state analogue (TSA) – as a model for the TS – by performing optimizations with the active bonds constrained to their length in the TS. From the TSA a TS is found by rerunning the transition state optimization and checking that it is both a 'good' TS and is lower in energy than the

previously found TS.

Transition State Templates. To accelerate the location of TSs, if available, a template is employed to generate a TS guess structure. Templates are saved into a local library following the successful location of a TS and contain a graph, solvent, charge and multiplicity. For a template to be used, the reactant (complex) must have the same charge, multiplicity, and solvent parameters as the one used to obtain the template. In addition, the active bonds must match (same number and atoms) along with their nearest neighbours, based on their connectivity and atom type. If a matching template is found in the library, then a TS guess is generated by constrained optimization, where the active bonds are fixed at values found in the saved TS, from which TS optimization is performed. For example, a TS found for EtCl + F^- → EtF + Cl^- enables a 10× faster characterization of the TS for the propyl analogue.

Reactive Reactant Complexes. For bimolecular reactions (substitution and elimination, not addition as $b_P > b_R$) an initial association complex on the PES must be found that favours a linear arrangement of the nucleophile and the leaving group, while also reducing the required number of optimizations. For this reason, the energy function U_{RR} (Eqn. (5.3)) is minimized with respect to rigid-body rotation and translation of one reactive component with empirical parameters $c_1 = c_2 = c_3 = 1, c_4 = 10$ and $c_5 = 2.5(1.5)$ for charged (uncharged) complexes.

$$U_{\mathrm{RR}}(\boldsymbol{x}_i) = \sum_{\{\mathrm{ac}\}} c_1(r_{\mathrm{ac}} - c_5 r_{\mathrm{ac}}^{\mathrm{avg}})^2 + \sum_{\{\mathrm{acx}\}} c_2(1-\cos\theta) + \sum_{\{\mathrm{acx}\}} c_3(1-\cos\phi) + \sum_{\{\mathrm{ij}\}} \frac{c_4}{r_{\mathrm{ij}}^4} \qquad (5.3)$$

This energy function has been designed to maximize the collinearity of vectors $\boldsymbol{v}_{an}$, $\boldsymbol{v}_{ac}$ and $\boldsymbol{v}_{cx}$ while maintaining both a low steric repulsion and distance ($r_{ac} = |\boldsymbol{v}_{ac}|$) close to that desired (Figure 5.2.4). This choice is guided by chemical intuition, and there are rare cases

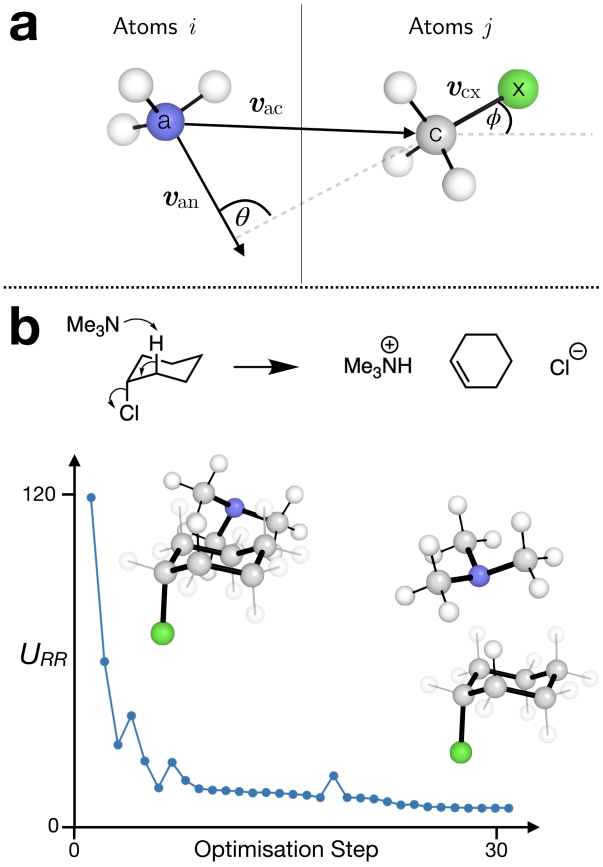


Figure 5.2.4. (a) Illustration of the geometric parameters used in Eqn. (5.2) the S_N2 reaction NH_3 + CH_3Cl as an example: van is the average of vectors $\{\boldsymbol{v}_{am}\}$ where m is a nearest neighbour to a. (b) Minimization of U_{RR} for an E2 reaction.

where this is not adhered to (e.g., front-side S_N2,[393] which is found upon TS conformer searching). For example, adequate initial geometries are obtained for S_N2 substitution, E2 elimination and S_NAr reactions (Figure 5.2.4, 5.3.12). Note that there may be several minima on this surface; thus, multiple initial random rotations and translations ($\sim$ 10) are performed to have a better chance of locating the global minimum while remaining computationally inexpensive (execution time $\sim$ seconds). Non-reactive association complex conformers are also available from optimization of random molecular orientations (SI §5.3.5).

5.2.4 Results and Discussion

To demonstrate the applicability of *autodE* in multiple reaction classes, we explored some textbook organic reactions (e.g., S_N2, E2, etc.) alongside industrially relevant organic and metal-catalysed reactions involving more than 50 atoms. By way of example, we demonstrate that *autodE* is broadly applicable, efficient in exploring conformational space, and straightforward to use.

Even with a small amount of programming experience, using *autodE* is as simple as declaring reactants and products involved in the reaction from their respective SMILES strings, then calling a method. An input for calculating an S_N2 reaction profile immersed in an implicit aqueous environment is shown in Figure 5. Here, by calling the method `calculate_reaction_profile` with complexes turned on, the association complexes are also calculated along the energy profile. Alternatively, one can decide only to obtain the TS, with the function `locate_transition_state`.

Without specifying which electronic structure codes to use, *autodE* employs the first available *hmethod* and *lmethod*, (sequentially ORCA, Gaussian09, NWChem and XTB, MOPAC respectively), with the default DFT methodologies listed in Table 5.2.1. Using this setup, the reaction energy profile shown in Figure 5.2.5 is obtained in about 30 minutes. While it would be ideal to obtain reaction free energies, the cost and limitations associated with the calculation of entropy and thermal contributions using currently available models (ideal gas) mean that the addition of such corrections should be considered with care. For this reason, potential energies are set as the default in *autodE* , but free energies are available with the `free_energy=True` flag. Further algorithmic details are outlined in the SI (§5.3.6).

TS Conformers. A thorough exploration of the conformational space in order to find the lowest energy conformer for reactants and TSs is essential in characterising the kinetic

```python
from autode import Reactant, Product, Config, Reaction

Config.n_cores = 4

F = Reactant(name='F-', smiles='[F-]')
MeCl = Reactant(name='CH3Cl', smiles='ClC')
Cl = Product(name='Cl-', smiles='[Cl-]')
MeF = Product(name='CH3F', smiles='CF')

reaction = Reaction(F, MeCl, Cl, MeF, solvent_name='water')
reaction.calculate_reaction_profile(with_complexes=True)
```

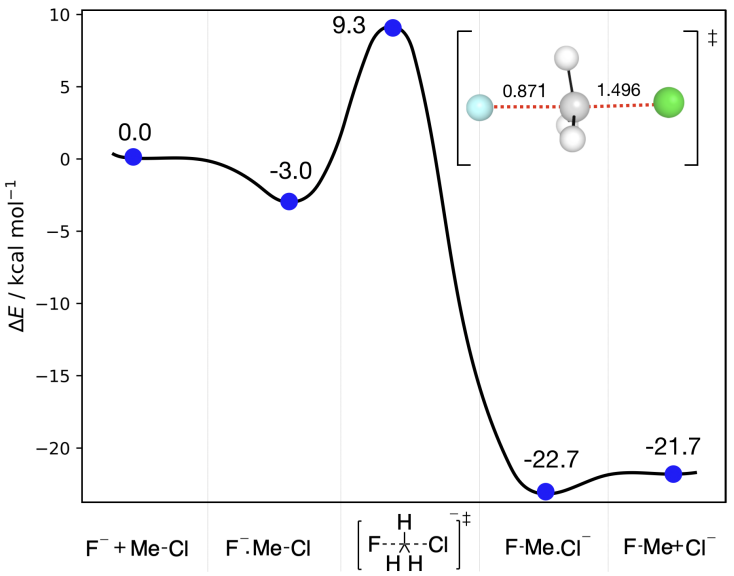


Figure 5.2.5. Reaction energy profile for an S_N2 reaction calculated using *autodE* with GFN2-XTB and DFT level *lmethod*/*hmethod*in XTB/ORCA. Final energies quoted at the CPCM(H_2O)-PBE0-D3BJ/def2-TZVP//CPCM(H2O)-PBE0-D3BJ/def2-SVP level of theory and distances quoted in Å.

and thermodynamic parameters of a reaction. *autodE* provides two routes to locating TS conformers. The first uses reactant/product complex conformers, from which different TSs may be traversed, and the second locates TS conformers directly. If using templates, both approaches are efficient, generally requiring only one constrained optimisation and one TS optimisation per conformer once a TS has been found. Direct TS conformer searching is, however, faster as only the TSAs are optimised and the lowest energy found whereas full enumeration from reactant/product complexes can require rescanning over the PES.

For example, for an E2 elimination reaction, the initial search may locate a transition state

in which the deprotonation occurs with the proton and the leaving group in a *syn* conformation, rather than the favoured *anti* conformer. The latter TS is automatically found in *autodE* by randomizing and minimizing Eqn. (5.2) (Figure 5.2.6a). Similarly, the *autodE* algorithm correctly locates the lowest energy H-migration TS for the radical decomposition of 1-propanol, with an RMSD < 0.1 Å compared to the human-generated TS from ref. [394] (Figure 5.3.16). The importance of this unbiased TS conformer generation is highlighted in the Michael addition of nitromethyl acetate and methyl vinyl ketone, where several rotatable bonds exist (Figure 5.2.6b). For this reaction, an exhaustive search from product conformers generated 21 TS conformers, which upon optimization, led to a range in activation energies of more than 5 kcal mol^{-1} and reaction energies that differ by up to 17 kcal mol^{-1}. The weak correlation between $\Delta E^{\ddagger}$ and ΔE_r from the global reactant conformer to each TS-product conformer pair highlights the importance of a systematic conformer search at both minima and transition states.

Organic Reaction. For *autodE* to become routinely used in mechanistic investigations, it must be applicable to synthetically relevant, and usually more complex, reactions. Here, we considered the Ireland–Claisen rearrangement explored computationally via trial-and-error and the AFIR method.[395] With known intermediates, processed as SMILES strings from Chemdraw™, *autodE* was tested and compared to the human-guided approach, which originally involved the testing of various conceivable mechanisms. *autodE* delivered an essentially identical reaction profile and reduced the time from the year of reported human and compute effort (that included searching other possible pathways) to a few minutes of setup time and ∼600 CPUh (one day of computer time on 24 cores, Figure 5.2.7). Interestingly, *autodE* deviates < 2 kcal mol^{-1} from the trial and error-based search (blue), with INT3 showing a larger difference of 5 kcal mol^{-1}, due to *autodE* finding a more stable intermediate than the one located previously.

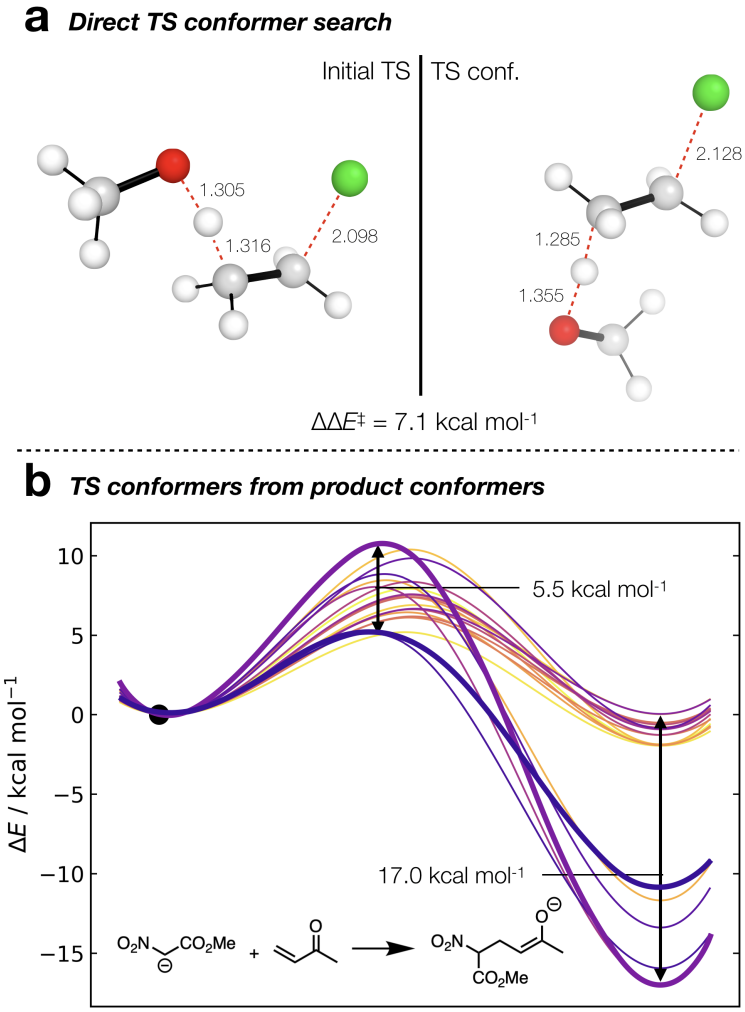


Figure 5.2.6. (a) Lowest energy transition state conformers found with *autodE* for methoxide + chloroethane E2 elimination. (b) TS and product conformer distributions for the addition reaction between nitromethyl acetate and methyl vinyl ketone. Calculations performed at the CPCM(H_2O)-PBE0-D3BJ/ma-def2-SVP level of theory and distances quoted in Å.

This example demonstrates that a combination of chemical knowledge, required to hypothesize reasonable mechanisms, and the use of *autodE*, can substantially speed up joint experimental and computational efforts to elucidate complex reaction mechanisms, and in this way advance the optimization of synthetic routes and the design of novel catalysts.

Organometallic Reaction. The conversion of alkenes into aldehydes via addition of CO and H_2 is an exceptionally important industrial process, making catalyst optimization the

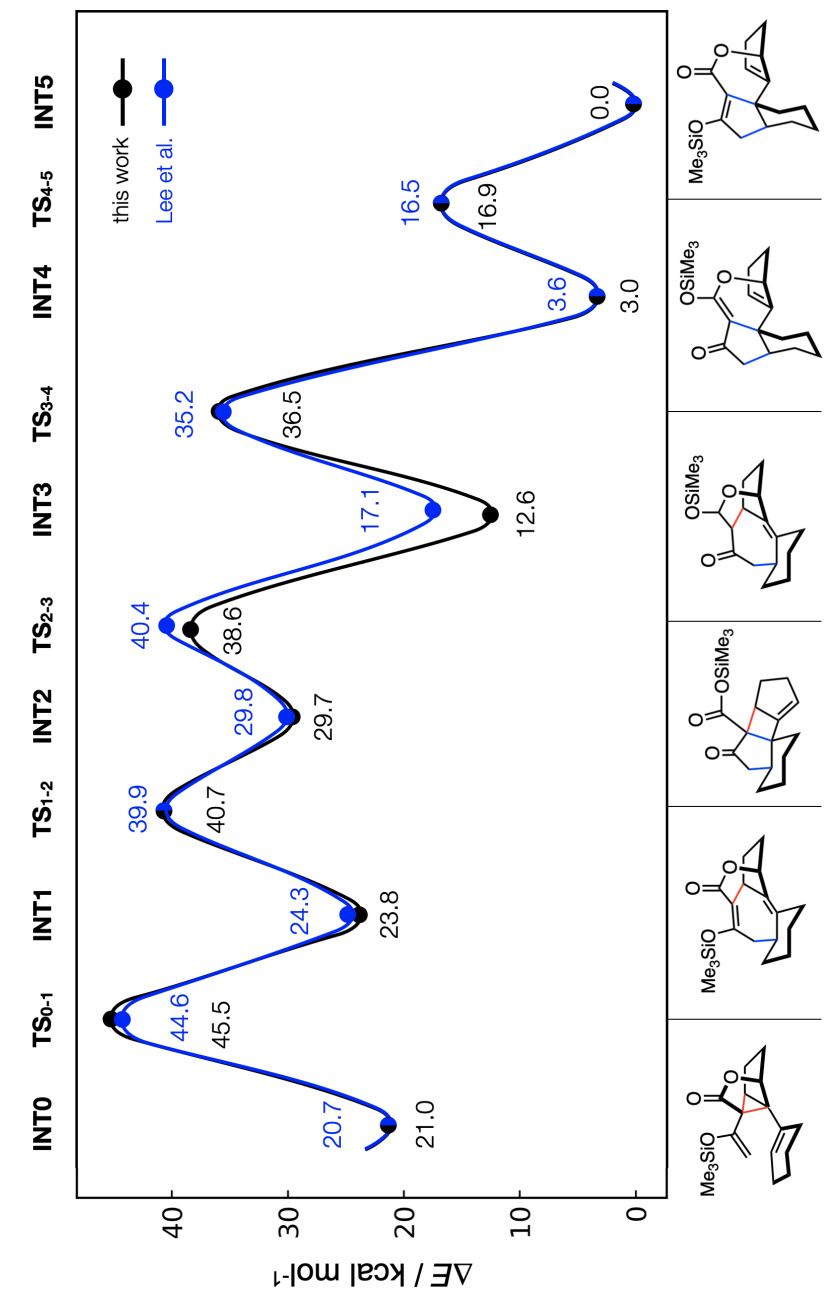


Figure 5.2.7. Ireland–Claisen rearrangement calculated using *autodE* (black) and described in ref. [395] (blue) calculated at B3LYP-D3BJ/6-311++G(2d,2p)//B3LYP-D3BJ/6-31G(d). CPCM(hexane) solvent model used in ORCA (this work) and IEF-PCM(hexane) in Gaussian09 (from ref. [395]). 3D Structures of the most stable intermediates and transition states are shown in Figure 5.3.17.

subject of considerable study.[396] The general mechanism was discovered by Heck and Breslow[397] and has been the subject of numerous computational studies (for an overview of these studies see ref. [398] and references cited therein). Those works have provided significant insights into the mechanism of the reaction, and highlighted the challenges associated with it. In fact, only finding the intermediates and transition states along the PES is already a laborious process, requiring extensive sampling, due to the presence of several conformers and isomeric forms (Figure 5.2.8a).

Applying *autodE* to study this catalytic cycle, using only the SMILES strings of the catalyst, ethylene, intermediates and products, the full reaction profile was obtained in less than 26 hours (16 CPU cores, excluding electronically barrierless ligand association steps, Figure 5.2.8b). *autodE* correctly identifies the most stable isomer of all intermediates; for example, the axial isomer for strong σ-donors (H: INT1, alkyl: INT3, acyl: INT5). TSs for all steps in the cycle are located successfully, i.e. they contain a single imaginary frequency and have geometries very similar to those found for analogous phosphine-based catalysts (Figure 5.3.18).[399]

The same process was applied to the analogous Co-catalysed hydroformylation (Figure 5.3.19), which has been used as a representative example in other automated reaction generation algorithms.[13, 16, 400] Once again, all TSs were successfully located with minutes of setup time.

With a view to expand the use of *autodE* for catalyst design campaigns, and following previous efforts in this field,[33] we have also developed the *molfunc* tool (`https://github.com/duartegroup/molfunc`). Combining *autodE* with *molfunc* enables a rapid screening of different catalysts by performing functionalisation of hydrogenic or other monovalent atomic positions. For example, for the H-insertion step, *molfunc* facilitates the

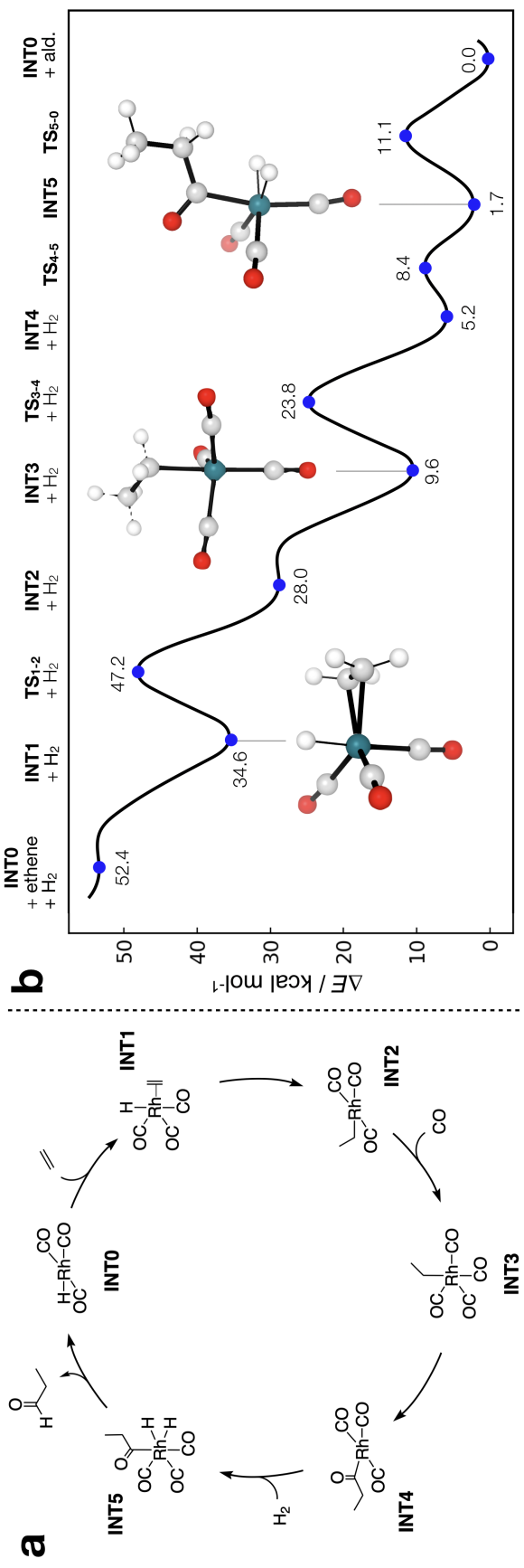


Figure 5.2.8. (a) Heck-Breslow mechanism of hydroformylation in the absence of a phosphine catalyst. (b) Reaction profile calculated at PBE0-D3BJ/def2-TZVP//PBE0-D3BJ/def2-SVP in autodE.

automated exploration of different groups at the para position of triphenyl phosphine on the barrier (Figure 5.3.20). The obtained positive correlation between the enhanced electron withdrawing ability of the catalyst and rate agrees with the experimental observation.[398]

Further Examples. To demonstrate the generality of our method, we tested *autodE* on a set of 15 simple organic reactions of different classes. *autodE* finds all TSs in all cases and generates reactions profiles in a matter of hours (Figure 5.3.21). Further examples include metal-catalysed carbonylation of methanol to generate acetic acid via the Monsanto and Cativa processes (see full SI),[401] alkaline ester hydrolysis (Figure 5.3.12), a Houk-List TS (Figure 5.3.25),[402] and 57 carbene insertion reactions reported in ref. [403] (see full SI). In this latter case *autodE* correctly locates 55/57 insertion TSs, with two failures occurring when a QRC+graph isomorphism check fails to confirm the TS as linking reactants and products, despite the TS being correct. Moreover, TSs for synthetically relevant reactions including a key Diels-Alder step in a total synthesis of Brevianamide A investigated by Domingo et al.[404] and a diastereoselective epoxidation[405] are presented in the full SI.[a]

Limitations. The graph-based approach affords two immediate limitations of the method: The need to define bonds, and checking stereochemistry. While bonds are generally well-defined in organic systems, they are not a rigid concept. This is partially alleviated by using not regenerating the molecular graphs from the 3D structure, rather than a molecular graph representation, and retaining the SMILES-defined connectivity but may afford an incorrect assignment of a 'good' TS, which actually does or does not lead to products/reactants. Furthermore, the isomorphism condition on reactants/products does not currently include a stereochemistry check – the algorithm simply finds the lowest energy TS. While this is often defined by the reaction type e.g., stereochemical inversion in S_N2 is generated by traversing the lowest energy TS, there are reactions where this is not adhered to. Furthermore, by

[a]Examples presented in the full SI were performed by Joseph J. Silcock.

enforcing a maximum connectivity change of four (up to 2 bonds breaking and forming) means that, for example, the TS in a synchronous Fritsch–Buttenberg–Wiechell rearrangement mechanism cannot be found. In this case, human intervention is necessary to prevent the very slow enumeration of $\sim$ 10 2D PES scans. Finally, we note a limitation in the current *lmethod* electronic structure theory methods used to generate conformers of reactant and product complexes. Reactions involving anions are particularly problematic, as at the GFN2-XTB level they are generally not sufficiently stable (even in implicit solvent) such that the reaction may be barrierless at the *lmethod* level. We hope that as semi-empirical/TB electronic structure methods become more accurate and/or DFT and WF methods become faster, this limitation will be mitigated. Because *autodE* does not rely on one specific method implementing novel methods should be straightforward.

Further Development. There are several opportunities to increase the accuracy and efficiency of *autodE*, including (1) addition of explicit solvation; (2) inclusion of an online and open-source transition state library; (3) treatment of dynamical effects and (4) enhanced error correction. Initial progress has been made on explicit solvation using a simple QM/MM embedding scheme, but for the time being is too computationally demanding to routinely replace implicit solvation models. Future developments to the code will focus on these aspects to approach a fully automated protocol to predict reaction mechanisms with chemical accuracy.

5.2.5 Conclusion

Converting a 2D representation of a reaction to a reaction profile has long been the domain of expert computational chemists. Here we have shown that our method, *autodE*, can largely automate this process and present a selection of both simple and synthetically interesting

organic and organometallic reactions. By building a flexible framework in a way that does not rely on a particular electronic structure theory method, future methods for calculating points on the potential energy surface are easy to include. Indeed, the dominant source of failure of the methodology is due to inaccuracies in electronic structure methods rather than issues associated with *autodE*. Crucially, *autodE* is open source, making the development faster, more collaborative and more robust. We believe *autodE* will facilitate faster computational investigation for both newcomers and experts alike.

5.3 Selected Supporting Information III

Full Supporting Information including raw data can be found at:

`https://onlinelibrary.wiley.com/doi/10.1002/anie.202011941`

5.3.1 Computational Methods

All *autodE* calculations were performed with a development version of the code (1.0.0a0). Unless otherwise stated the *lmethod* used was GFN2-XTB v. 6.2[297] and the *hmethod* ORCA v. 4.2[246] All ORCA optimizations employed resolution of identity DFT (RI or RIJCOSX for hybrid functionals)[406, 407] using the PBE[82] or PBE0[86] functional, in combination with the D3BJ dispersion correction,[94, 313] def2-SVP or def2-TZVP basis set[247] (with effective core potentials for Rh) and the default auxiliary basis set.[248] Gaussian09[242] calculations employed identical methods without density fitting unless otherwise specified. NWChem v. 6.6[299] calculations used no dispersion correction as D3BJ is unavailable with PBE0. Default integration grids were used in all DFT calculations. Conformers generated with RDKit[294] version 2020.03.1 and OpenBabel version 2.4.1.

5.3.2 Compatibility of autodE with Different Electronic Structure Theory Codes

Figure 5.3.1 shows the energy profile for the endo/exo Diels Alder cyclization between methyl acrylate and cyclopentadiene obtained when combining *autodE* with each of the software for which wrappers are currently implemented and show invariability of the results to the chosen method.

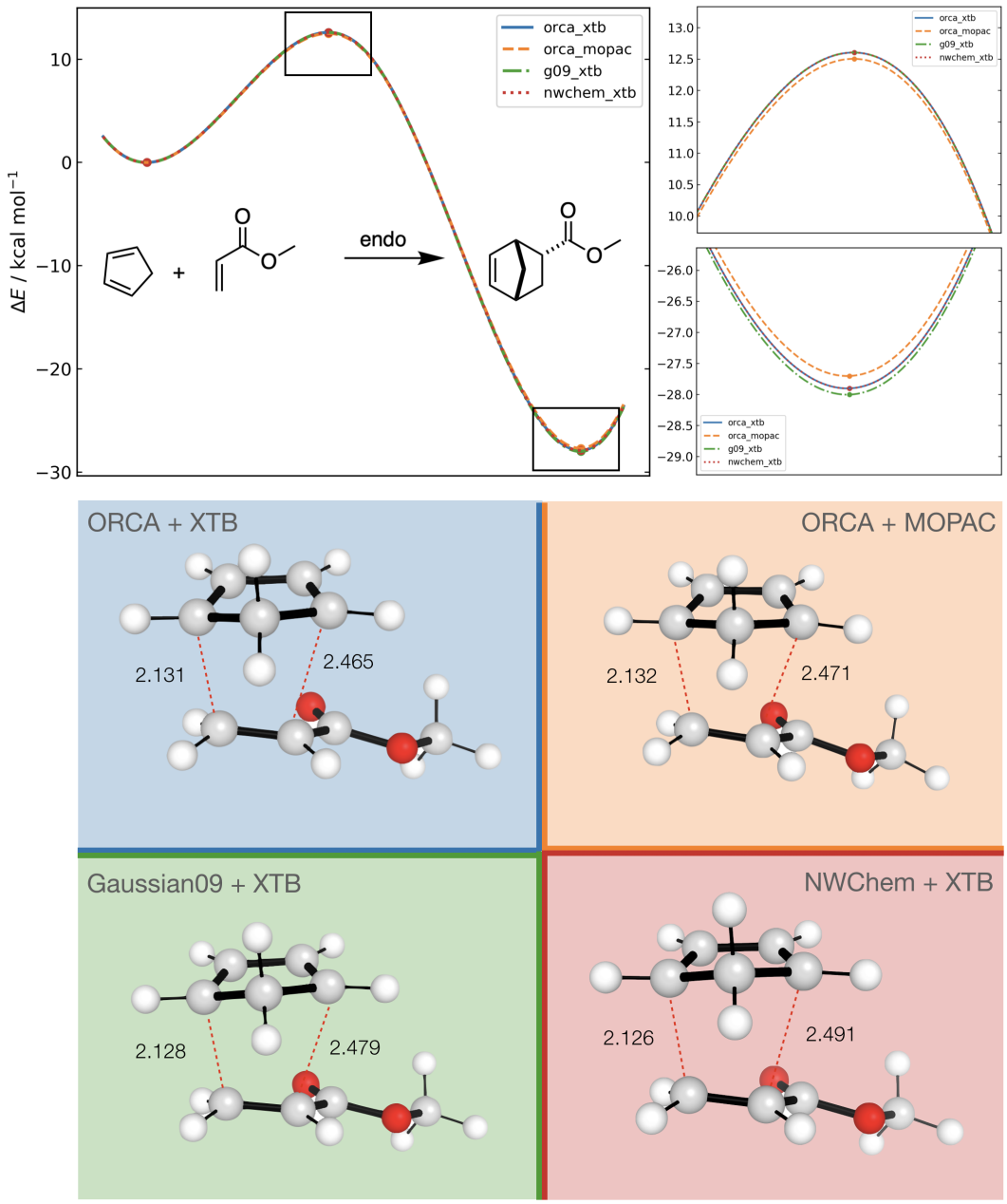


Figure 5.3.1. Comparison of different codes used to perform optimizations and single point energy evaluations in the reaction between *endo* Diels–Alder cyclisation between methyl acrylate and cyclopentadiene. All profiles are calculated at the PBE0/def2-TZVP//PBE0/def2-SVP level of theory and the low-level method is tight-binding DFT (XTB) or semi-empirical PM7 (MOPAC). Forming bond distances are quoted in Å.

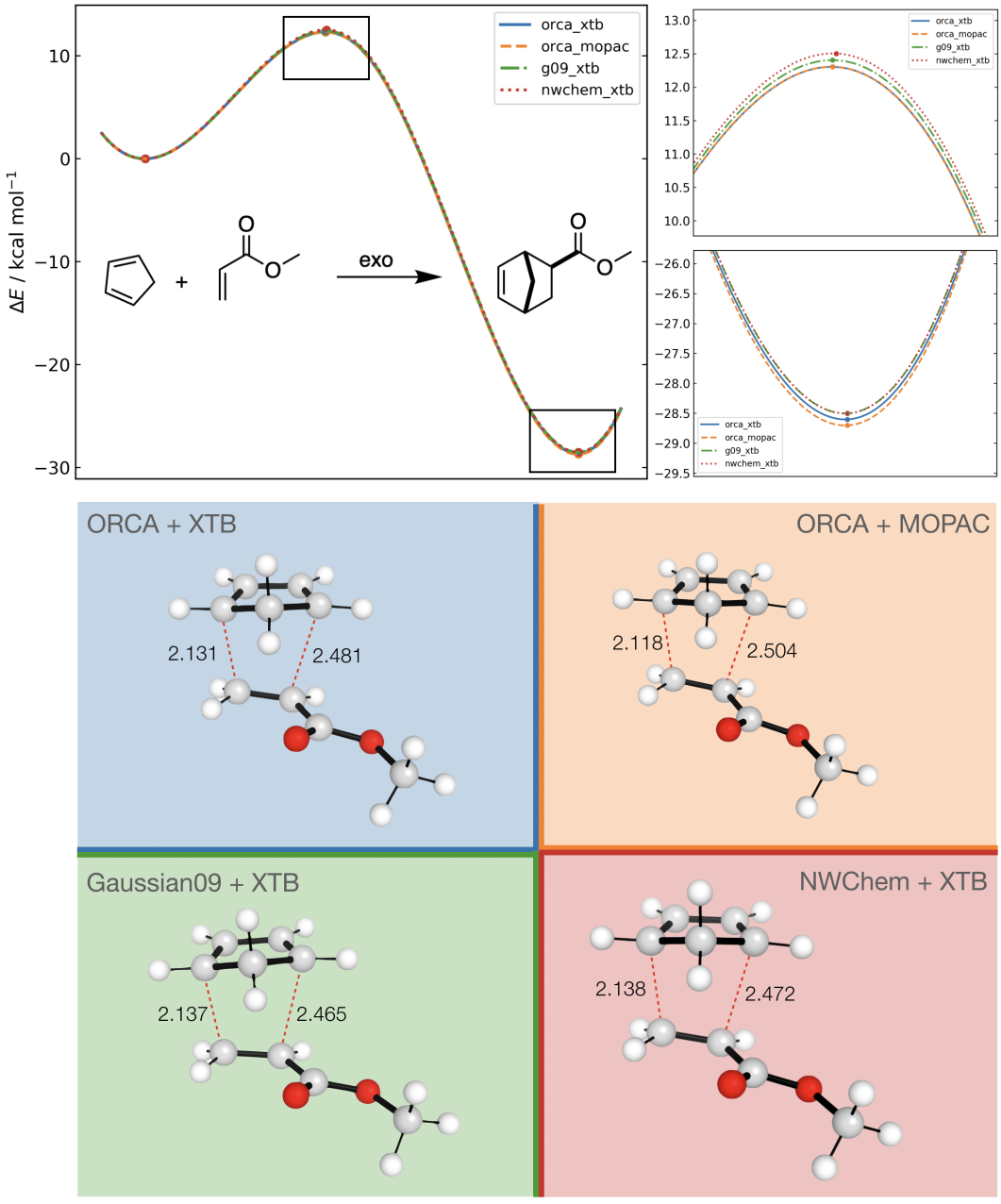


Figure 5.3.2. As Figure 5.3.1 for *exo* products.

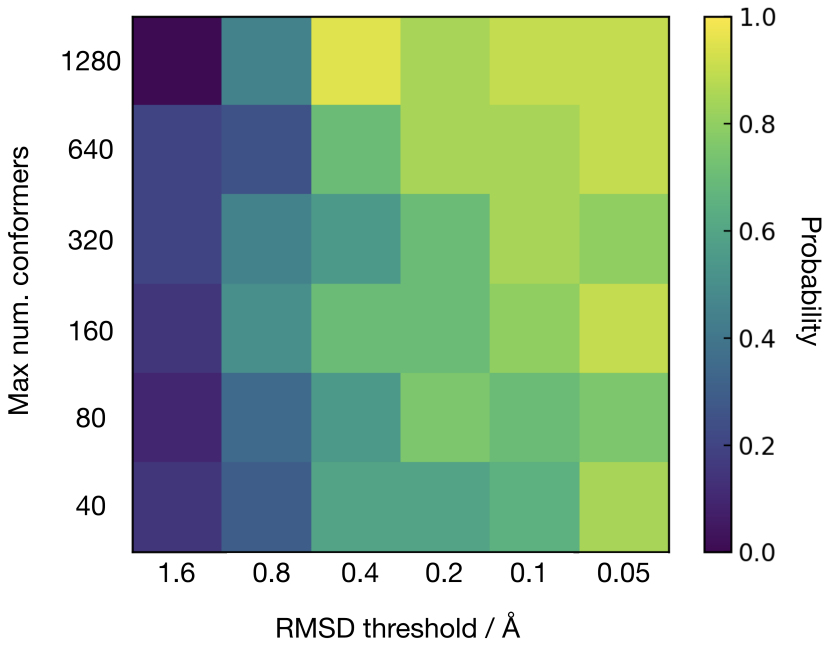


Figure 5.3.3. Probability that a conformer will be found – using RDKit + XTB optimization – within 1 kcal mol^{-1} of the most stable conformer generated by CREST given a specific root mean squared deviation (RMSD) and a number of conformers threshold in the EKTDGv2 algorithm. Molecule set contains 20 L-amino acids. All optimizations were performed at the GFN2-XTB level. Crest v. 2.9 & XTB v. 6.2.3 to find the lowest energy conformer, as it generated a more diverse set than the exhaustive algorithm in CONFAB. For example, CREST generated 90 conformers for serine for which 85 were found in ref. [408] while CONFAB did not identify the correct rotatable bonds and afforded only 36 conformers.

5.3.3 Metal-complex Conformers

autodE generates metal complex conformers using Eqn. (5.4),

$$U_{\mathrm{RB}}(\boldsymbol{x}) = \sum_{ij \in \mathrm{bonds}} k_1 (r_{ij} - r_{ij}^{\mathrm{avg}})^2 + \sum_{i>j} \frac{k_2}{r_{ij}^n} \tag{5.4}$$

where the first harmonic term describes bonds between atoms, and the second term introduces a non-physical repulsion that enforces a notion of steric repulsion. Parameters $k_1 = 1$ and $k_2 = 0.01$ were selected based on empirical experience and comparison to optimised structures (Figure 5.3.4) while ideal bond lengths (r^{avg}) are obtained as averages from the Cambridge Structural Database.

To generate reasonable 3D structures for organometallic complexes, each atom is added sequentially and randomly in a 10 Å cubic box, and the function U_{RB} is minimized with respect to all coordinates after each addition. Using a smooth potential with few local minima ($n = 2$ in Eqn. (5.4)) is required to obtain stable structures for large complexes (Figure 5.3.5). For a test set of 20 metal-complexes with up to 100 atoms, our approach delivers a stable conformer in all cases, while RDKit successfully generates only 15 geometries and CONFAB failed in all cases (Figure 5.3.6 and Table 5.3.1). With the analytic derivative of U_{RB} and a conjugate gradient algorithm implemented in *scipy*,[292] an initial structure is available in a few seconds for complexes with more than 100 atoms. Stereo-defined metal complexes are currently not supported, as random initialization does not respect any chirality.

An alternative and slightly faster approach is used to generate metal complex conformers; a random displacement vector (length $\sim$ 3 Å) is applied to all atoms and U_{RB} minimized with $n = 8$. The steeper repulsive term is used as it generates a more realistic PES; for example, generating the expected three minima in the butane dihedral PES, Figure 5.3.5.

While simple, this strategy affords more conformers than both RDKit and CONFAB for the metal complex test set (Figure 5.3.6 and Table 5.3.1). It is also worth noting that we found no advantage in using the EMBED algorithm[409] to generate initial coordinates for organic systems (Figure 5.3.8). Moreover, conformers can be generated using arbitrary distance constraints specified by the user (e.g., to retain a square planar geometry given an 3D initial structure).

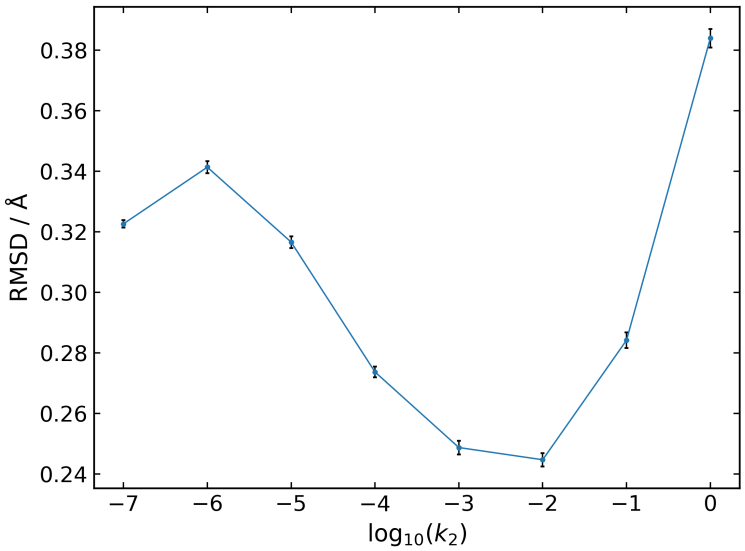


Figure 5.3.4. Using the molecules butane, proline and [RhH(ethene)$(CO)_3$], the calculated RMSD between the geometries optimised with the repulsion + bonded forcefield [(5.4) $n = 8$] and XTB for different k_2 values (Eqn. (5.4)). Initial conformer geometries are obtained by random displacements from XTB-optimised geometries (normally distributed, σ=0.5 Å, μ=0.0 Å, on each direction). This approach provides a more realistic starting structure than optimizing at the local XTB minimum i.e. not overly favouring a very small repulsion where atoms do not move from their optimised positions. Error bars are quoted as the standard error in the mean of 20 repeats.

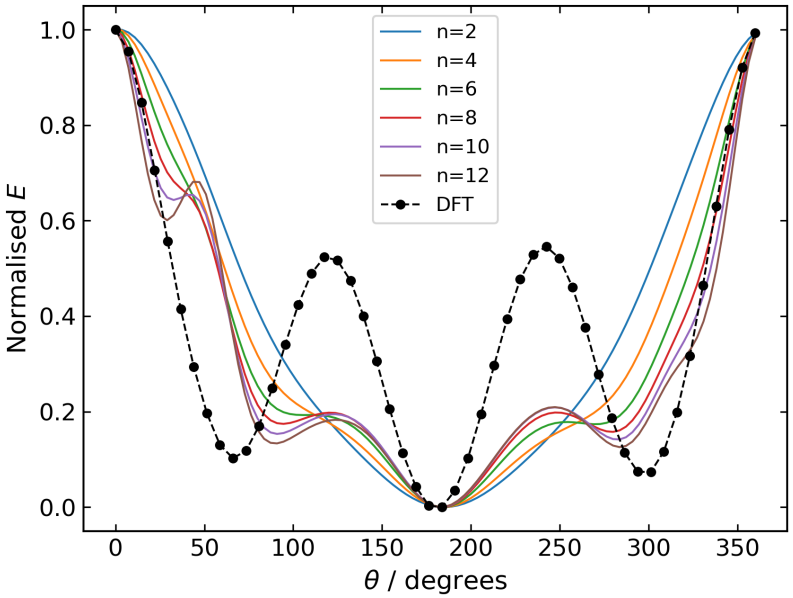


Figure 5.3.5. Rotational barrier of butane using a simple repulsion + bonded potential Eqn. 5.4, with different n values, compared to a DFT reference (PBE0-D3BJ/def2-SVP). Relative energies normalized to the eclipsed conformation ($\theta = 0$ °). Initial structure not symmetrized.

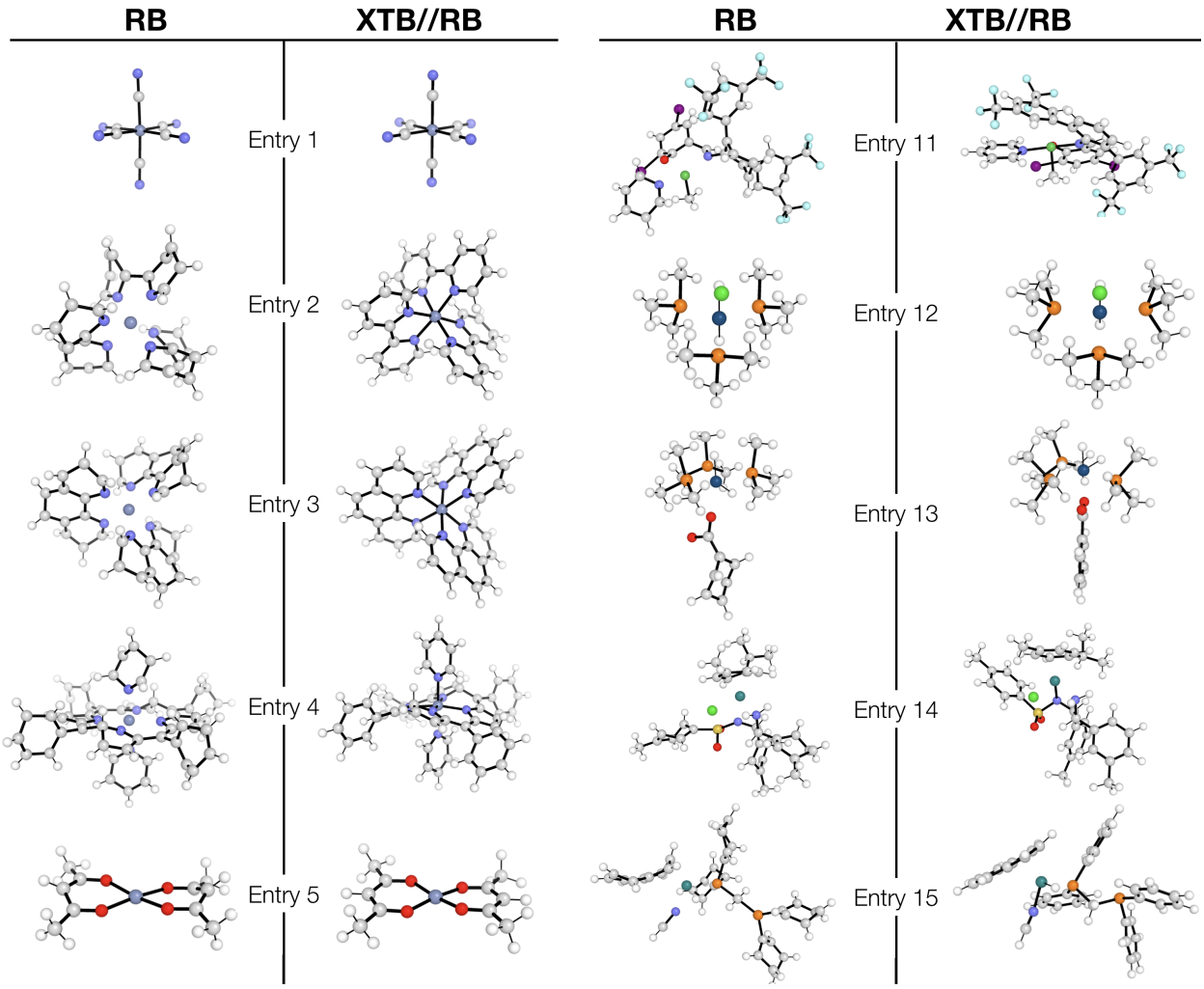


Figure 5.3.6. Metal complexes itemized in Table 5.3.1 generated by minimizing a simple repulsion + bonded potential Eqn. (5.4) and their subsequent XTB optimised geometries. Structures initialized by adding atoms sequentially from a random position within a 10 Å cube and minimizing $U - \mathrm{RB}$ with $n = 2$ repeatedly until all atoms were added, then performing a final minimization with $n = 8$ ($k = 1$, $c = 0.01$).

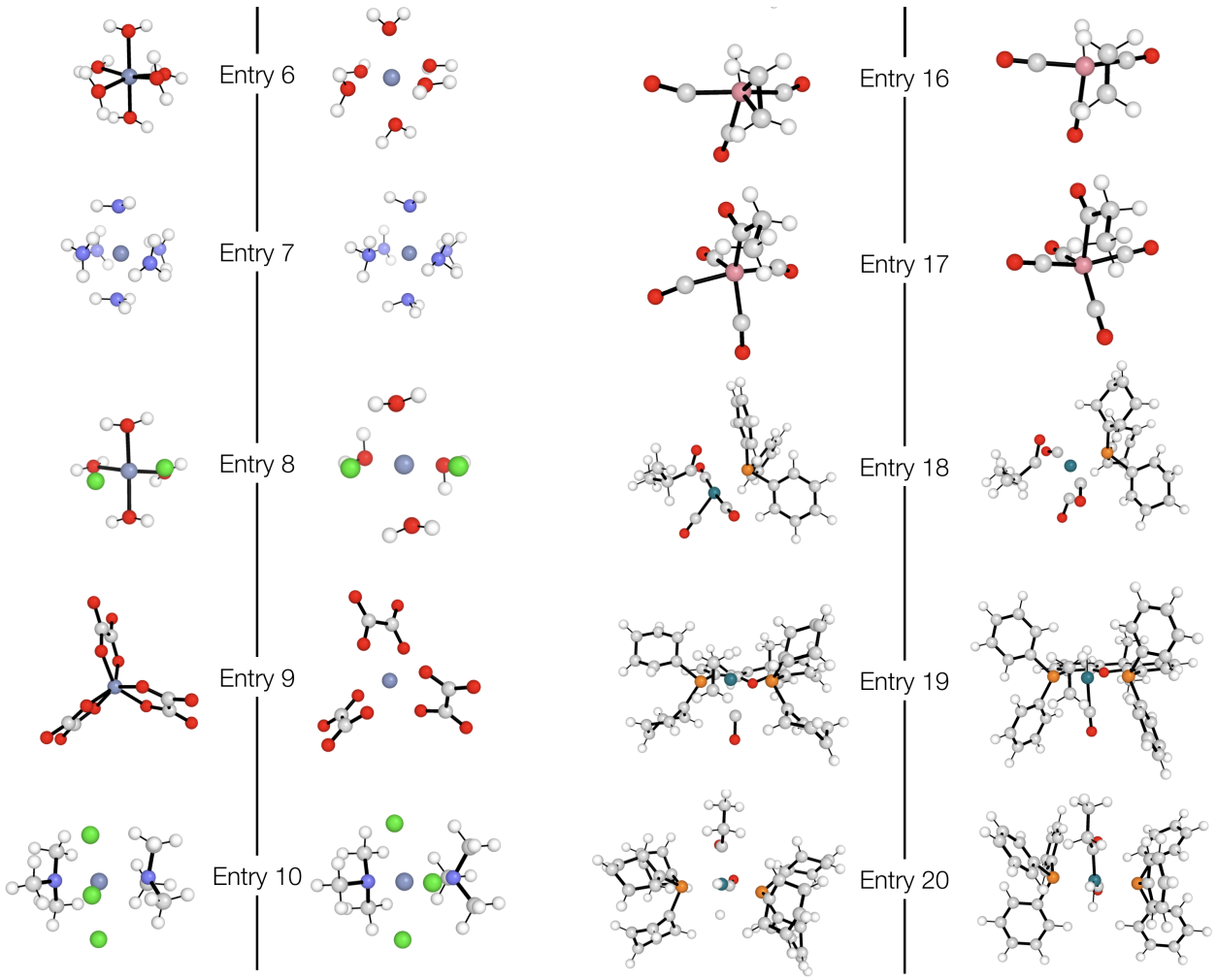


Figure 5.3.7. Figure 5.3.6 continued.

Entry	RDKit		CONFAB		RB	
	SG	N	SG	N	SG	N
1	✗	0	✗	0	✓	1
2	✓	1	✗	0	✓	2
3	✓	1	✗	0	✓	5
4	✓	15	✗	0	✓	17
5	✓	1	✗	0	✓	8
6	✗	0	✗	0	✓	10
7	✗	0	✗	0	✓	1
8	✗	0	✗	0	✓	10
9	✓	2	✗	0	✓	1
10	✓	2	✗	0	✓	5
11	✓	6	✗	0	✓	23
12	✓	8	✗	0	✓	16
13	✓	9	✗	0	✓	18
14	✗	0	✗	0	✓	20
15	✓	22	✗	0	✓	24
16	✓	1	✗	0	✓	5
17	✓	3	✗	0	✓	5
18	✓	13	✗	0	✓	18
19	✓	14	✗	0	✓	19
20	✓	28	✗	0	✓	30

Table 5.3.1. Metal complex test set. The set comprises the first 10 complexes used to benchmark MolSimplify (ref. [410]., all Cr(II)), 5 large complexes from ref. [411] and 5 intermediates in hydroformylation catalysis. N is the number of conformers/isomers generated using RDKit (v. 2019.09.3), CONFAB (OpenBabel v. 2.4.1), and the repulsion + bonded (RB, Eqn. (5.3)) algorithm introduced in this work. Unique conformers are found by discarding those with energies within 0.24 kcal mol^{-1} of others. RDKit and RB requested 50 conformers and CONFAB the default. 3D structure generated (SG) successfully is indicated with a tick. See full supporting information for SMILES strings and Figure 5.3.6, 5.3.7 for geometries.

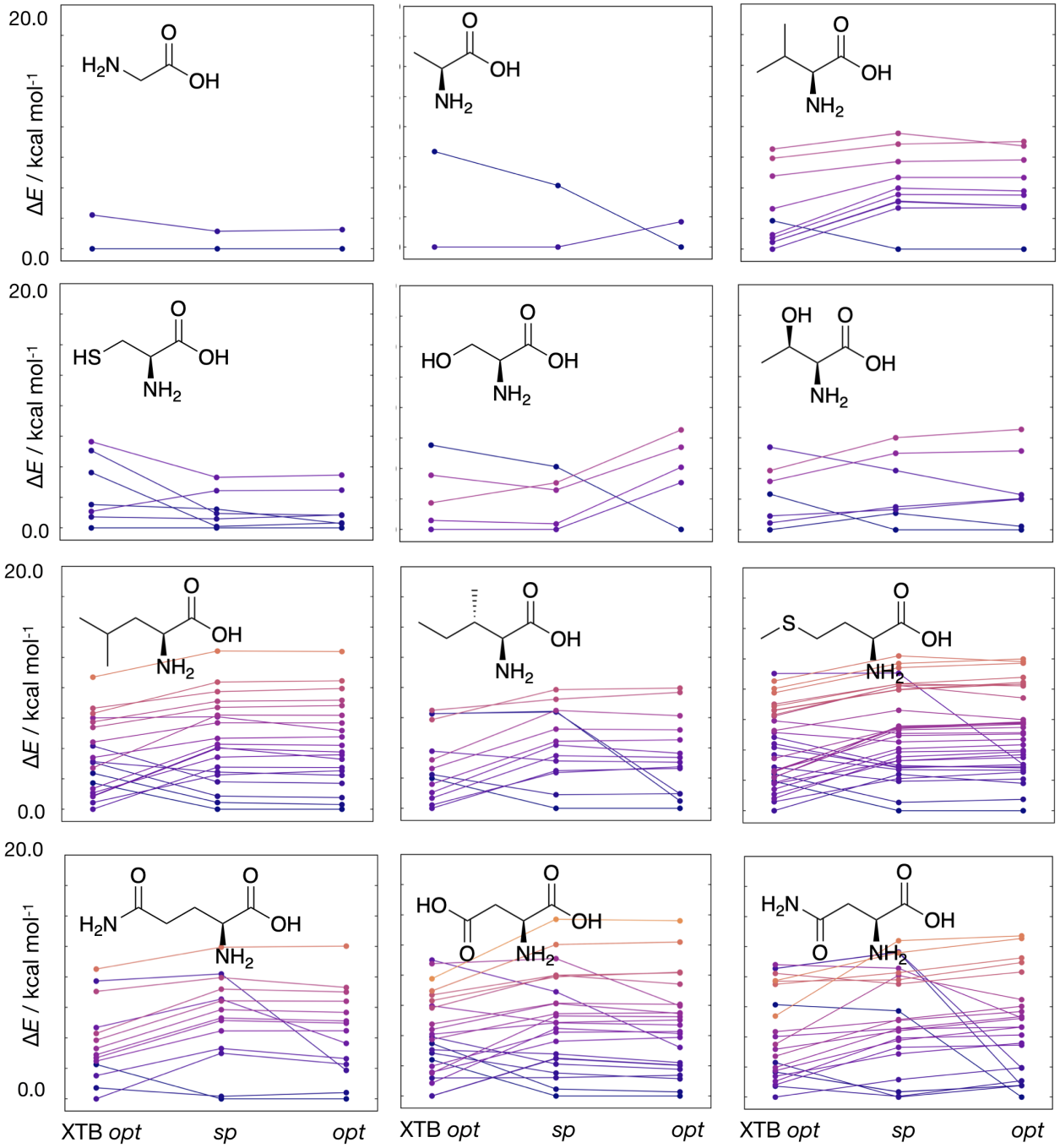


Figure 5.3.8. Conformer ranking for 20 amino acids in their neutral form using energies from XTB optimization (XTB opt), DFT single points (sp, XTB optimised geometries) and DFT loose optimization (opt). DFT calculations performed at the PBE-D3BJ/def2-SVP level of theory. Conformers generated using RDKit with a maximum of 1000 requested and an RMSD threshold of 0.05 Å.

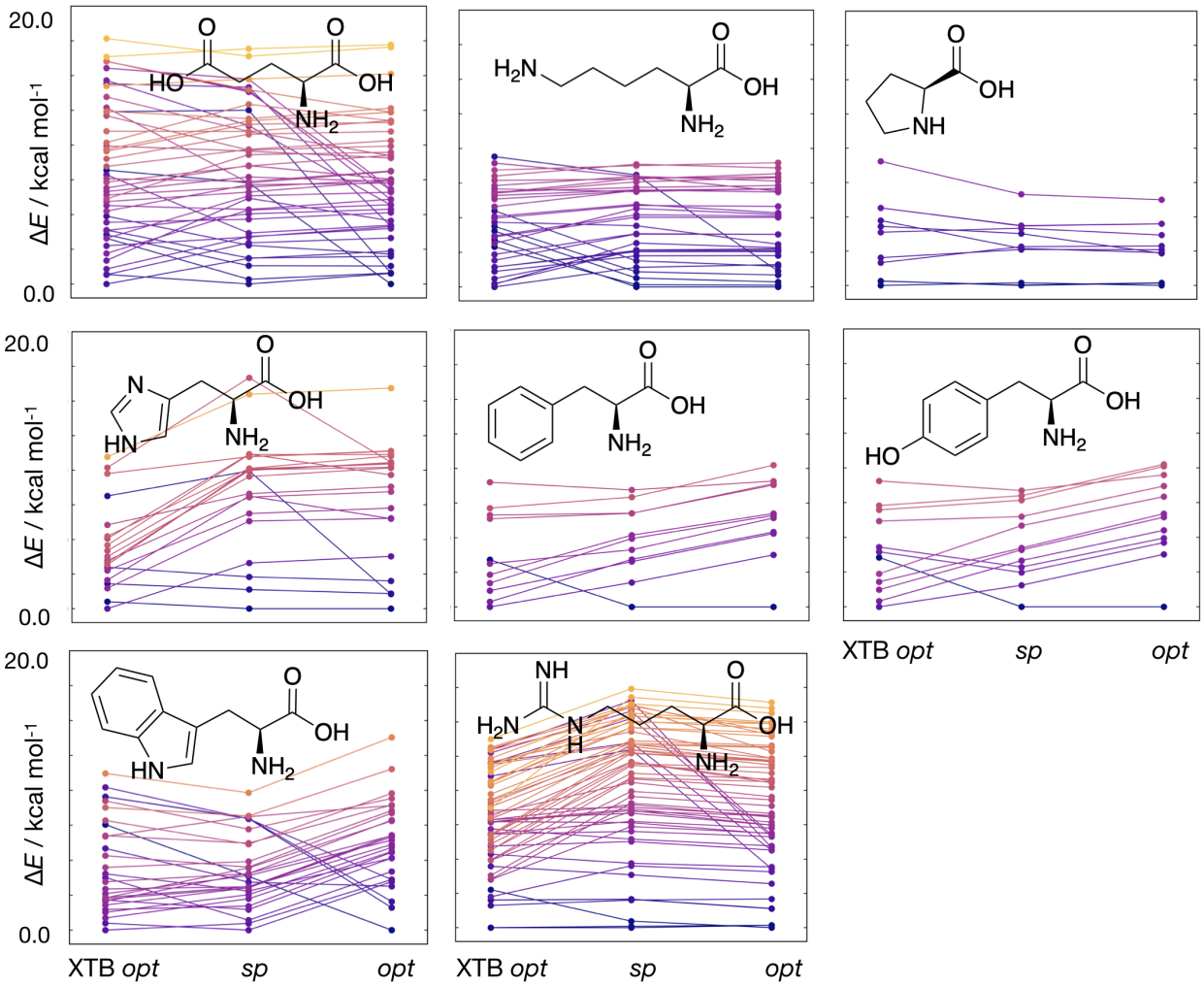


Figure 5.3.9. Figure 5.3.8 continued.

5.3.4 Constructing Molecular Graphs

The precise nature and properties of a chemical bond are still not uniquely defined,[412] this is despite a rather inclusive IUPAC definition of 'When forces acting between two atoms or groups of atoms lead to the formation of a stable independent molecular entity, a chemical bond is considered to exist...'.[413] However, to form a molecular graph the bonds must be rigidly defined. Given a SMILES string the molecular graph is constructed by adding edges defined explicitly (e.g. C=C) and implicitly (C–H) in the string. If a molecular graph is constructed from a 3D geometry an edge between two atoms (a, b) is added if $r_{ab} < 1.25 \times r_{ab}^{\text{avg}}$, where r_{ab}^{avg} is the CCDC average for that atom pair (i.e. $r_{\text{C–C}}^{\text{avg}} = 1.427$ Å). Edges are added by iterating through atoms sorted by their atomic weight, then for each atom index i enumerating atoms $j > i$ sorted by the distance closest to atom i. If an atom exceeds its maximal valance e.g. hydrogen: 1, carbon: 4 then the longest edge(s) is(are) removed. Some simple examples are shown in Figure 5.3.10.

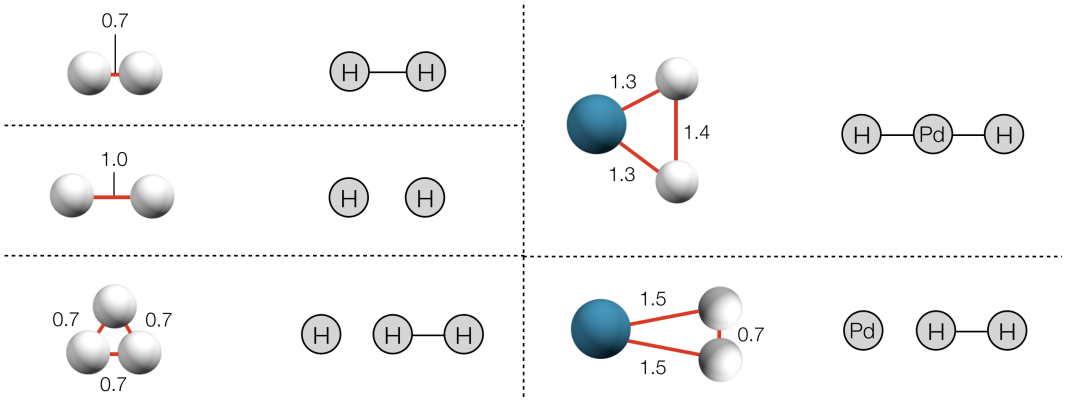


Figure 5.3.10. Mapping of 3D structures to molecular graphs. Distances quoted in Å.

In the molecular graphs used here there is no concept of multiple bonding with the exception of allowing nodes to be 'pi atoms', so that suitable distance constraints can be added to ensure no rotation about π-bonds in RB conformer generation and a molecule is not truncated over

a π-bond. A node may also be a 'stereo atom' which again is to prevent inversion of chiral centres in RB conformer generation.

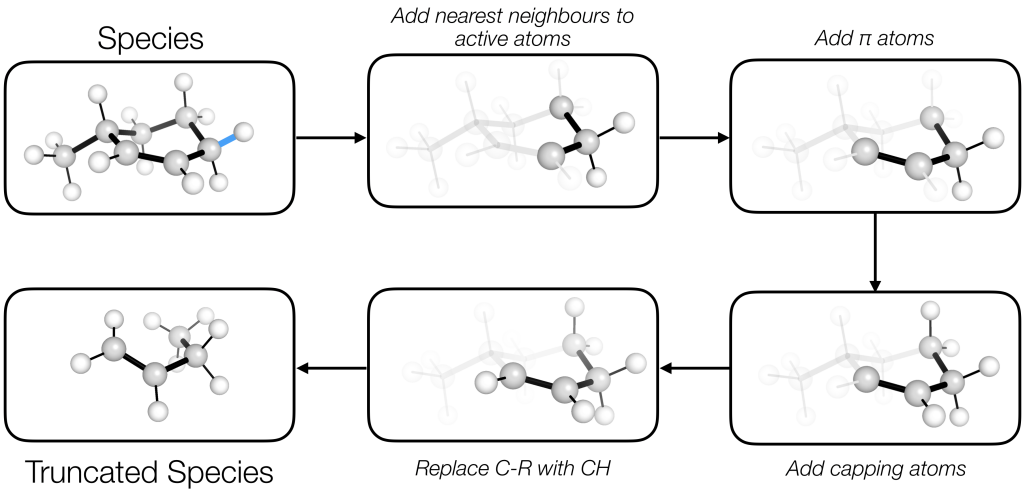


Figure 5.3.11. Schematic process of truncating 3-methylcyclohex-1-ene. The active bond is highlighted in blue and the active atoms are those that form the active bond (C, H).

5.3.5 Non-covalent and Reactive Reactant Complexes

In many chemical processes, the formation of a reactant/product complex precedes/follows the chemical step of interest, and is therefore fundamental to determine the kinetics of the process. In some cases, the molecular graph of the reactant and product complexes may not be isomorphic to the separated reactants and products. For example, in the base-catalysed hydrolysis of methyl acetate, loss of MeO^- from the tetrahedral intermediate proceeds with concurrent deprotonation of the acid (Figure 5.3.13). Thus, to successfully characterize the TS as 'good', reactant and product complex conformers are required from which isomorphisms can be checked to the forward/reverse displaced species from the TS. This approach allows the lowest energy conformer of a non-covalent interaction (NCI) complex to be located systematically. For a complex formed by molecule A and B, conformers of the A.B complex are constructed by adding molecule B at points equally spaced on the surface of a sphere in a random orientation around A, then energy minimizing with an *lmethod* (Figure 5.3.14). For a complex with N_m molecules using N_s points on a sphere and N_r random rotations, this approach generates $(N_s \times N_r)^{N_m-1}$ conformers. To maintain efficiency, a maximum threshold number of conformers is set (1000 by default). This approach also provides additional functionality, facilitating the generation of hydrogen bond complexes of relevance in anion-recognition without prior knowledge (Figure 5.3.15).

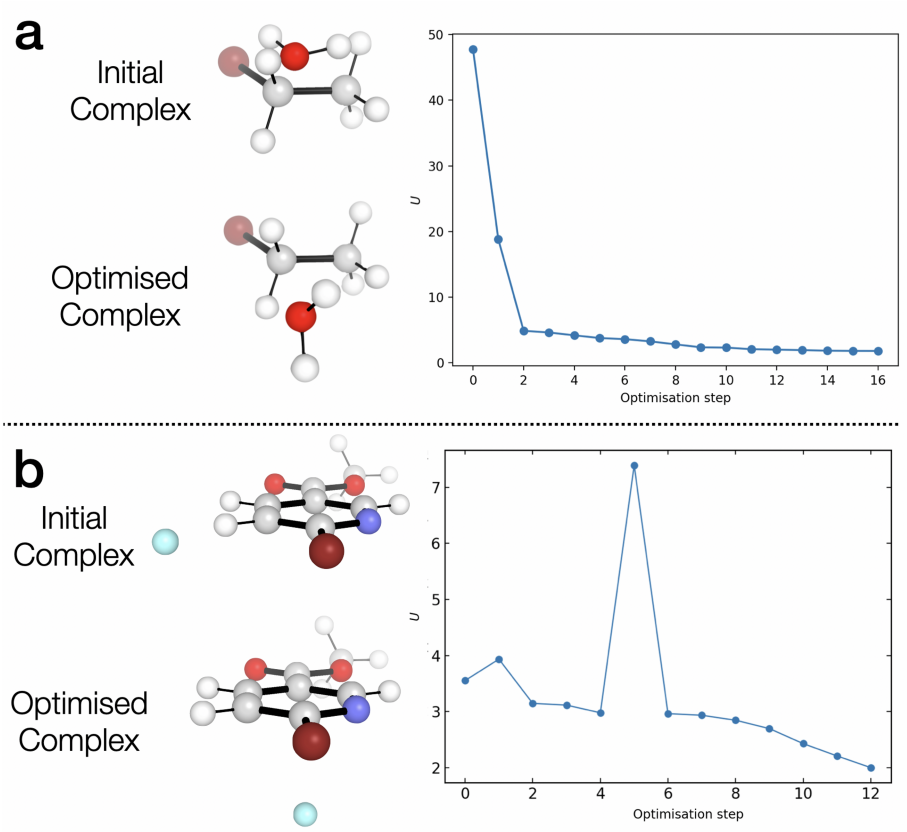


Figure 5.3.12. Representative optimization of the reactant complexes for (a) ethylbromide + water and (b) methyl 6-bromonicotinate + fluoride (concerted S_NAr from ref. [414]) under equation (5.3). Energies (U) are in arbitrary units.

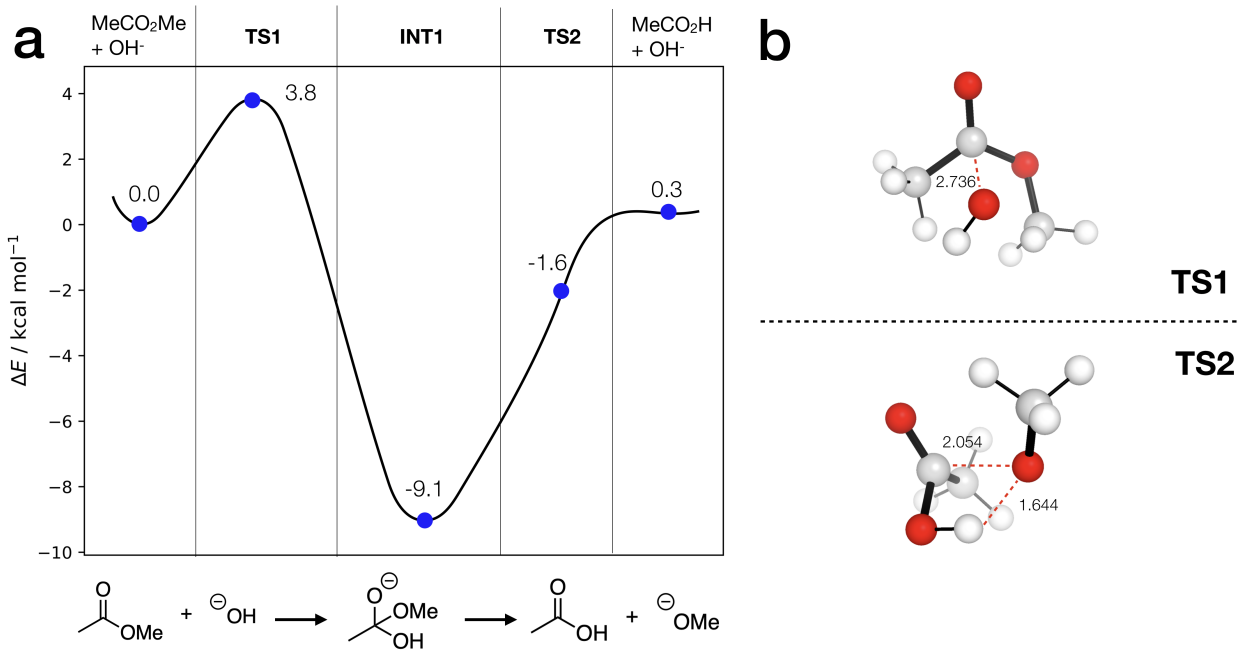
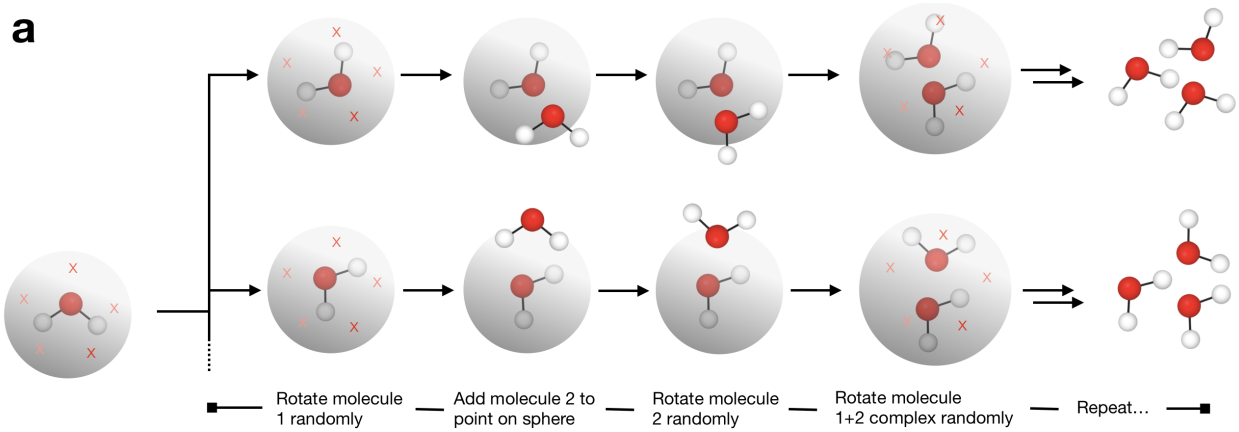


Figure 5.3.13. (a) Reaction profile for alkaline ester hydrolysis generated in *autodE* (ORCA/XTB, CPCM(water)-PBE0-D3BJ/def2-TZVP//CPCM(water)-PBE0-D3BJ/ma-def2-SVP). TS for methoxide loss is more stable than separated acetic acid + methoxide species at the chosen level of theory. (b) TS structures calculated at CPCM(water)-PBE0-D3BJ/ma-def2-SVP. Key distances are quoted in Å.




```python
from autode import Molecule, Config, NCIComplex
from autode.wrappers.XTB import xtb

Config.num_complex_sphere_points = 5
Config.num_complex_random_rotations = 3

# Make a water molecule and optimise at the XTB level
water = Molecule name='water', smiles='O')
water.optimise method=xtb)

# Make the NCI complex and find the lowest energy structure
trimer = NCIComplex(water, water, water, name='water_trimer')
trimer.find_lowest_energy_conformer low_level_method=xtb)
trimer.print_xyz_file( )
```

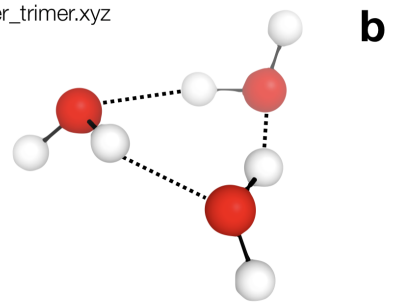


Figure 5.3.14. NCI complex conformer (a) generation methodology and (b) application to the water trimer. $(5 \times 3)^2 = 225$ conformers are generated and optimised.

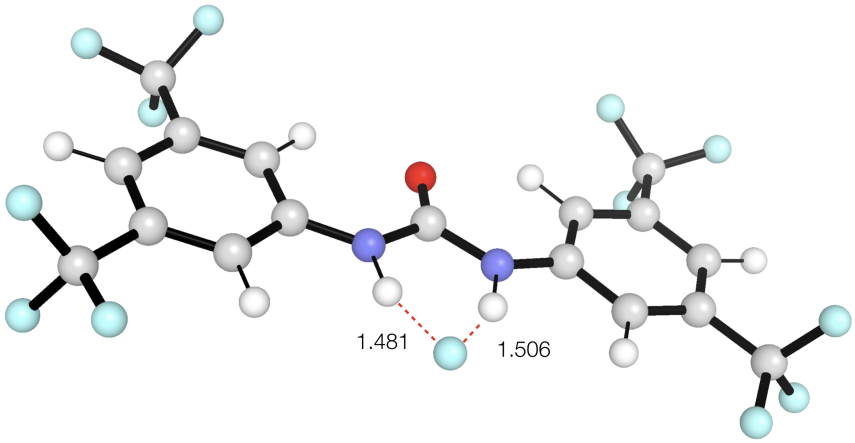


Figure 5.3.15. Most stable NCI complex conformer for a urea-fluoride hydrogen-bonding complex located using *autodE* at the XTB level. Distances shown in Å. See full supporting information for the associated input script.

5.3.6 Algorithm Details

To outline the algorithm in more detail than shown in the high-level workflow (Figure 5.2.2) for an S_N2 reaction (Figure 5.2.5, with the input reproduced below) the partial trace of function calls is shown and commented where the function is not self-explanatory, or there are multiple possibilities exist depending on the system. Note that this trace will likely change in development but is accurate for the 1.0.0a0 release. Only a single molecule initialization (MeCl), reaction initialization and calculating reaction profile functions calls are shown. Functions without a prefix (e.g. RDKit.) are *autodE* functions.

`MeCl = Reactant(name='CH3Cl', smiles='ClC')`

---

1. `init_smiles('ClC')`.

    a) `any(metal in smiles for metal in metals)` → False.

    If true then use the RB algorithm to generate 3D structure.

    b) `init_organic_smiles('ClC')`

        i. `rdkit.Chem.MolFromSmiles('ClC')`

        ii. `rdkit.Chem.AddHs(..)`

        iii. `rdkit.AllChem.EmbedMultipleConfs(..)`

        iv. `make_graph(..)` Add nodes and edges from the atoms and RDKit bonds defined by the SMILES connectivity, set stereocenters inc. Z/E alkenes.

        v. `are_coords_reasonable` → True If the coordinates aren't reasonable then revert to RB algorithm to generate a 3D structure.

        vi. `check_bonds(..)` If the connectivity based on the 3D geometry defined by

distances is different to the SMILES connectivity display an error.

`reaction = Reaction(Flouride, MeCl, Chloride, MeF, name='sn2', solvent_name='water')`

1. `classify(..)` Classify the reaction based on the number of reactants and products. If there are no reactants/ products throw an exception.
2. `get_solvent(solvent_name)`
3. `_check_solvent()` If the reactant and product molecules that comprise this reaction have an associated solvent this function checks they match. Also permissible for all molecule.solvent = None for a reaction in the gas phase.
4. `_check_balance()` Ensure that the number of atoms and the total charge are conserved reactants → products, otherwise throw an exception.

`calculate_reaction_profile()`

1. `find_lowest_energy_conformers()`
   a) `for mol in reacs + prods: mol.find_lowest_energy_conformer()`
      i. `_generate_conformers()` If RDKit conformers are fine then use the ETKDGv2() algorithm to generate conformers with a specified RMSD and max number cutoff. Otherwise use the RB algorithm with the same cutoffs. Both are parallelized.
      ii. `for conf in mol.conformers: conf.optimise(lmethod)` Initially optimise all conformers at the low-level method with keywords.opt to generate structures (hopefully) closer to the minima on the high-level method surface
      iii. `for conf in mol.conformers: conf.optimise(hmethod)` Failures in Gaussian internal coordinate bends are corrected by running a few optimisation steps in Cartesian coordinates then switching back to internals .

iv. `_set_lowest_energy_conformer()` Enumerate all the conformers in this molecule and set mol.atoms, mol.energy as the conformer with the lowest energy that is also graph isomorphic to the molecule's graph to remove any structures with different connectivity i.e. not conformers.

2. `optimise_reacs_prods()`

   a) `for mol in reacs + prods: mol.optimise(hmethod)` Optimise the geometry with the high-level method but don't reset the molecular graph

3. `find_complexes()`

   a) `ReactantComplex(reacs, ..)`

      i. `charge = sum(reacs.charge)` $\rightarrow -1$

      ii. `mult = sum(reacs.mult) - (n-1)` $\rightarrow 1$ Spin multiplicity is the sum of the two where n is the number of reactants by default. atoms = reacs.atoms

      iii. `solvent = reacs[0].solvent` $\rightarrow$ water

      iv. `graph = union(reacs.graph)`

   b) `ProductComplex(prods, ..)` As for reactant complex.

4. `locate_transition_state()`

   a) $b_R > b_P \rightarrow$ False.

   If True then call `switch_reacs_prods()`, then switch back after locating TSs.

   b) `find_tss()` Locate all possible transition states for this R $\rightarrow$ P rearrangement.

      i. `species_are_isomorphic(reac, prod)` $\rightarrow$ False.

      If the graphs of reactant and product (R, P) are isomorphic then the bond

rearrangement of interest is not obvious and this returns true and `find_tss` returns None.

ii. `get_bond_rearrs(reac, prod, ..)` Get all possible bond rearrangements that generate an R' that is isomorphic to P.

iii. `bond_rearrs is None` → False.

If a set of bond rearrangements cannot be found then `find_tss` returns None.

iv. `for g in bond_rearrs:` Attempt to locate a TS for all possible bond rearrangements (g).

v. `translate_rotate_reactant()`

vi. `reoder_product_complex(..)` Reorder the atoms in the product complex using the atom mapping from R' to P so R and P have atoms that map.

vii. `is_worth_truncating(..)` → False.

Form the truncated reactant complex by swapping fragments for hydrogen if they are far enough away from the active atoms, then calculation the difference in number of atoms between the full complex and the truncated, if > 10 then use the truncated complex then revert.

viii. `for ts_func in funcs()`: Enumerate possible TS guess functions, here a 2D low-level scan, 1D breaking-bond scan and 1D high-level forming bond scan in sequence until either there are no more functions or a TS has been found.

ix. `ts_guess = ts_func(reaction, ..)` Use the function to generate a transition state guess geometry as, in the first instance, the saddle point in the 1D/2D surface.

x. `could_have_correct_mode(ts_guess)` → True.

The 'lowest' mode in the Hessian must be negative (imaginary), be $|\nu| > 45$ cm^{-1}, and has a contribution from the active atoms. The latter is defined by the mass weighted $|\nu_i|$ for an active atom i being above a threshold.

xi. `ts.optimise()` Use the TS optimiser in the high-level method to optimise to a first order saddle point. If there is > 1 imaginary frequency following the successful termination displace along the spurious mode forwards and backwards to try and remove it.

xii. `is_true_ts()` Check that the forward/reverse displacement leads to the expected change in active bond lengths (all $\Delta r > 0.3$ Å) or that displacement forwards and backwards along the largest magnitude imaginary mode leads to a structure that has a graph which is isomorphic to either the reactant complex or an optimised conformer of the reactant complex.

xiii. `ts.save_template()` If `Config.make_ts_template` then save a TS template containing the atoms, distances, charge, multiplicity and solvent for subsequent use.

xiv. `len(tss) == 0` → False.

If no transition states can be found then `find_tss` returns None.

c) `find_lowest_energy_ts()` If there is more than one TS that leads to products then choose the lowest energy.

5. `find_lowest_energy_ts_conformer()` Use the repulsion + bonded force field to randomise coordinates then minimize using the low-level method with fixed active bond lengths, removing the similar based on RMSD and energy tolerance then reoptimising.

at the high-level method from which a TS optimisation is run from the lowest energy. If this is both a 'good' TS and is lower energy than previously found then use this TS.

6. `calculate_single_points()`

   a) `for mol in reacs + prods:  mol.single_point()`

7. `plot_reaction_profile(..)` Spline between the points in the profile and optimise the points in the spline to generate stationary points at the minima/maxima for a smooth profile.

In checking if the largest magnitude imaginary mode afforded by a Hessian calculation on the transition state guess geometry could be the 'correct' transition state mode that separates reactants and products there has to be a 'significant contribution' from the active atoms before a transition state optimization is performed. The significance is quantified by $d > 0.1$ based on empirical experience. Further *autodE* developments will introduce a more robust and interpretable measure where,

$$d = \frac{\sum_{i \in \text{active}} w_i}{\sum_{i \in \text{atoms}} w_i} \quad , \quad w_i = \frac{M_i}{c_M} + 10\frac{|\boldsymbol{v}_{\text{imag}}(i)|}{c_d} \tag{5.5}$$

M_i is the mass of atom i, $c_M = 1$ amu, $v_{\text{imag}}(i)$ the normal mode displacement vector and $c_d = 1$ Å.

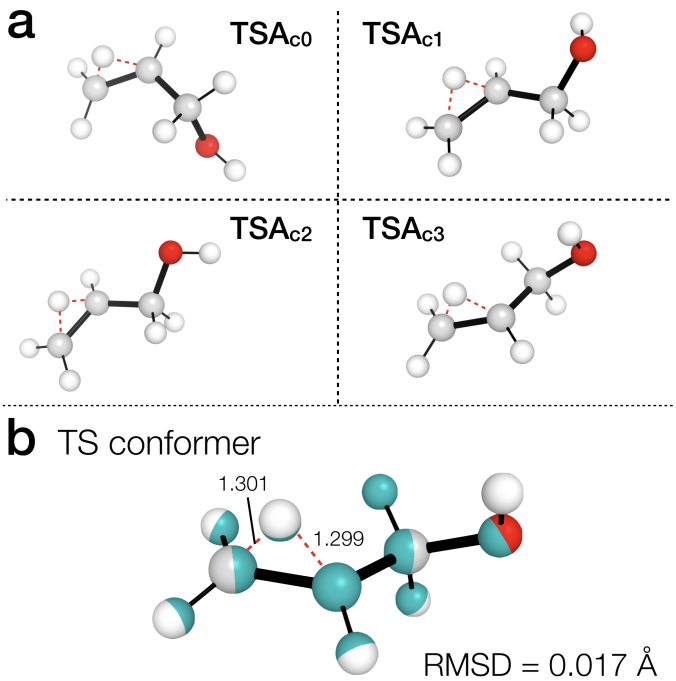


Figure 5.3.16. H-migration reaction involved in the radical decomposition of 1-propanol. (a) Transition state analogue (TSA) conformers located with *autodE* and ranked in increasing energy (c0-c3)(b) optimised TS, starting from TSAC1 conformer, at PBE0-D3BJ/def2-SVP compared to the lowest energy TS conformer reported in ref. [394] (cyan). RMSD calculation include all atoms. Key distances quoted in Å.

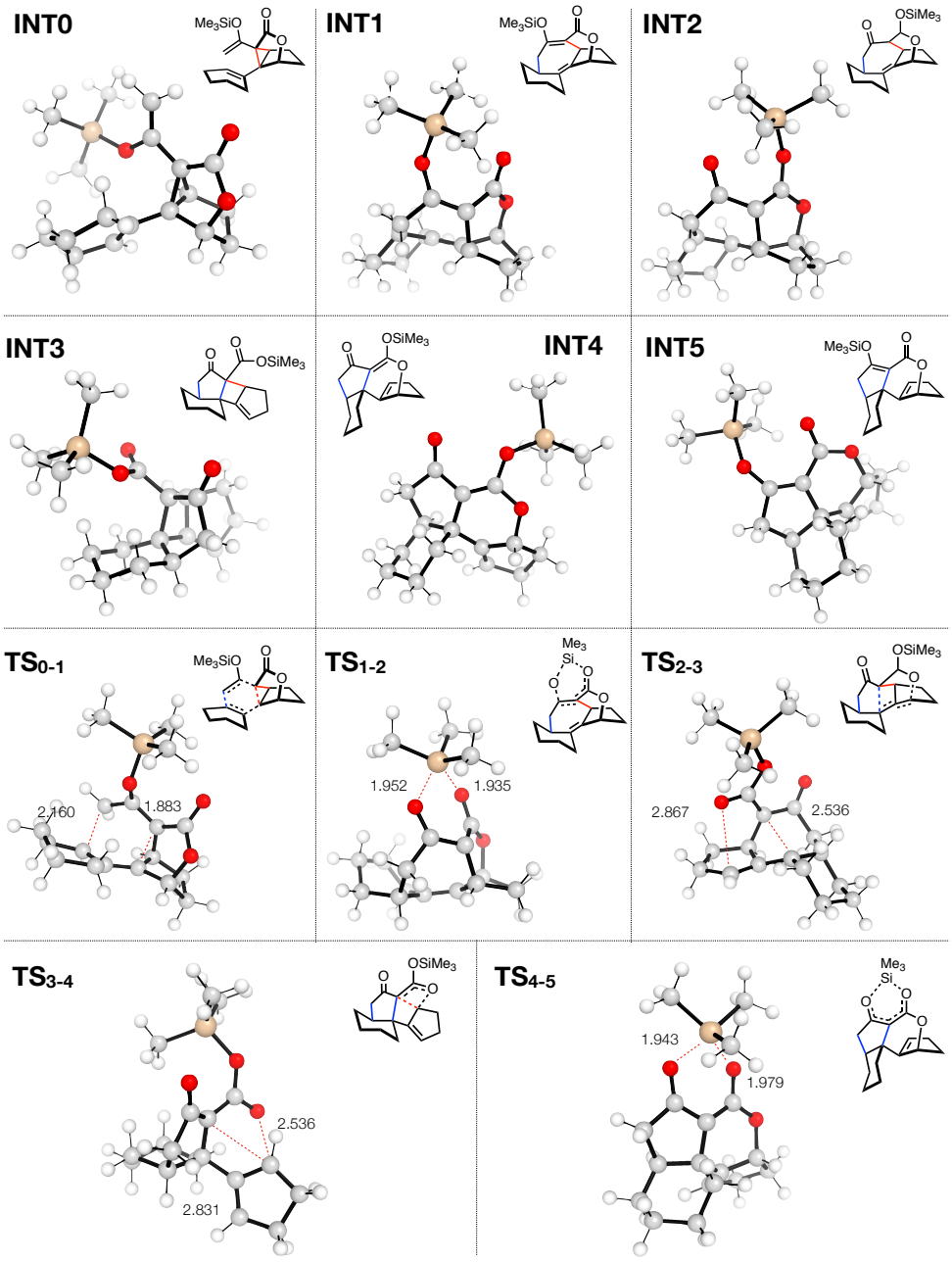


Figure 5.3.17. Most stable intermediates and transition states for the Ireland-Claisen reaction profile shown in Figure 5.2.7 located by *autodE* at the B3LYP-D3BJ/6-31G(d) level of theory using ORCA/XTB.

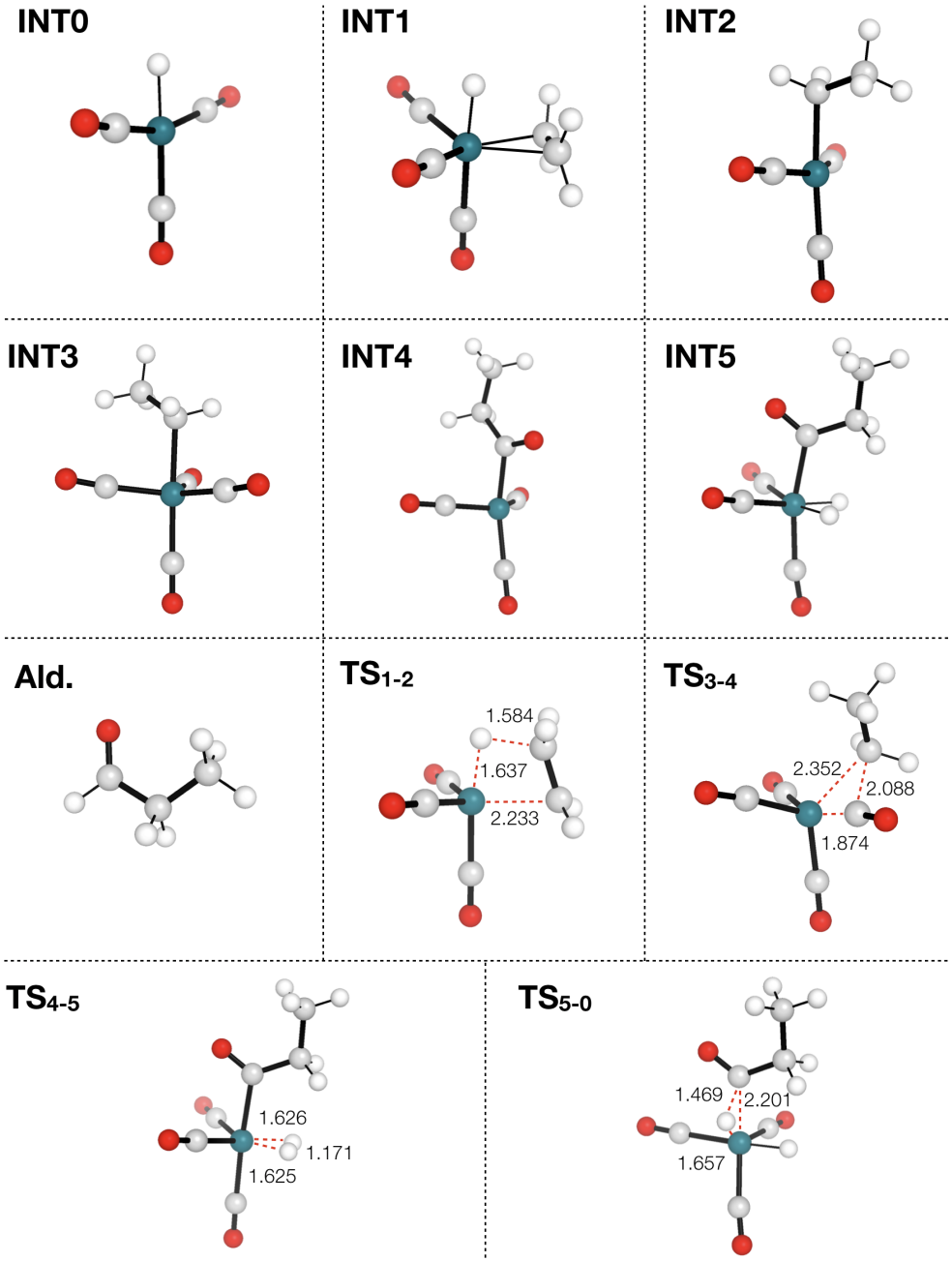


Figure 5.3.18. Most stable intermediates and transition states for the Heck-Breslow mechanism of hydroformylation shown in Figure 5.2.8 located by *autodE* at the PBE-D3BJ/def2-SVP level of theory.

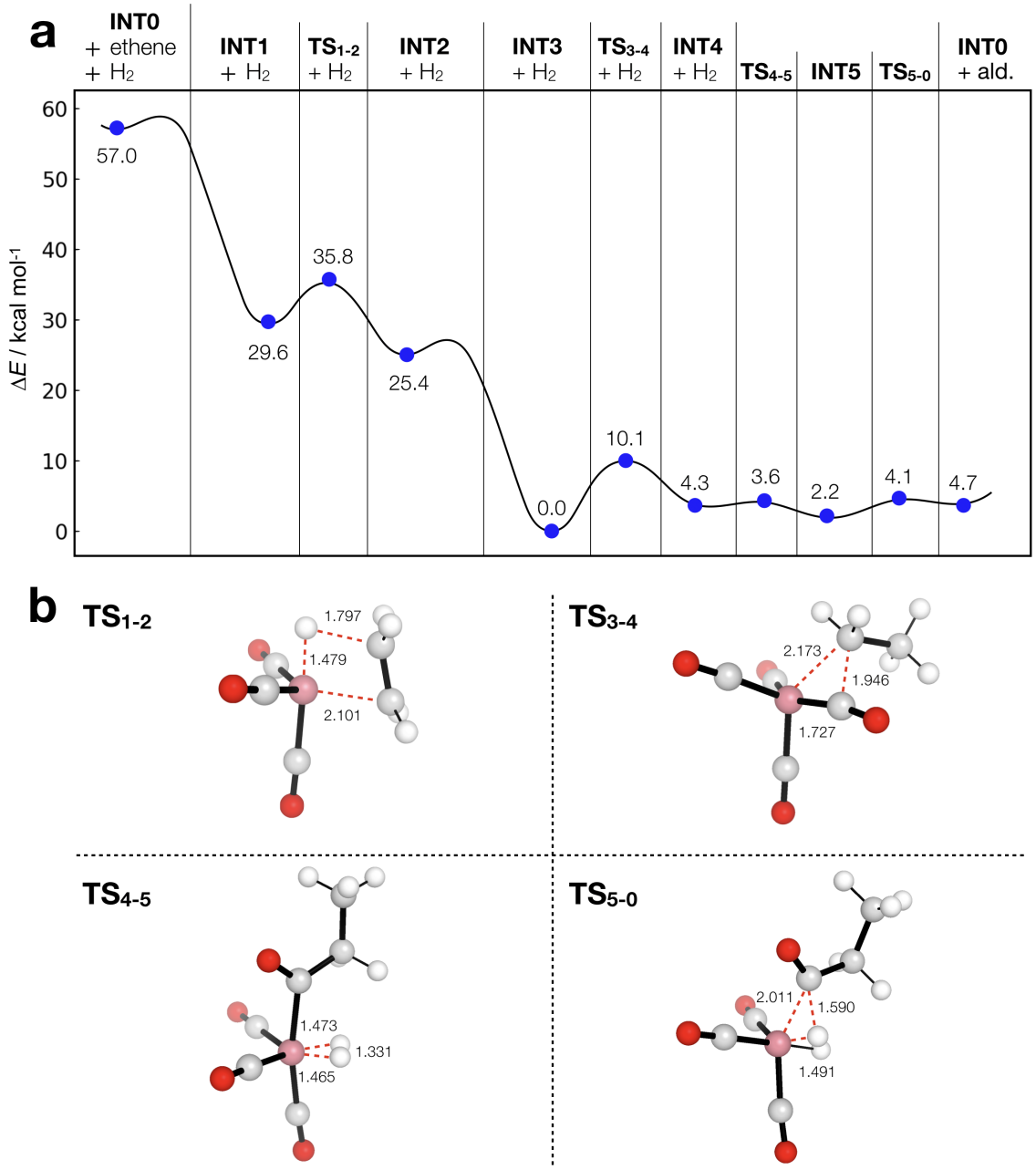


Figure 5.3.19. Co-catalyzed hydroformylation (a) reaction profile and (b) transition state geometries generated in *autodE* (ORCA/XTB, PBE0-D3BJ/def2-TZVP//PBE0-D3BJ/def2-SVP). ald. = propionaldehyde. TS_{5-0} is not a local maximum due to basis set superposition error making the TS more stable than the separated aldehyde and INT0. Electronically barrierless ligand association TSs are not calculated.

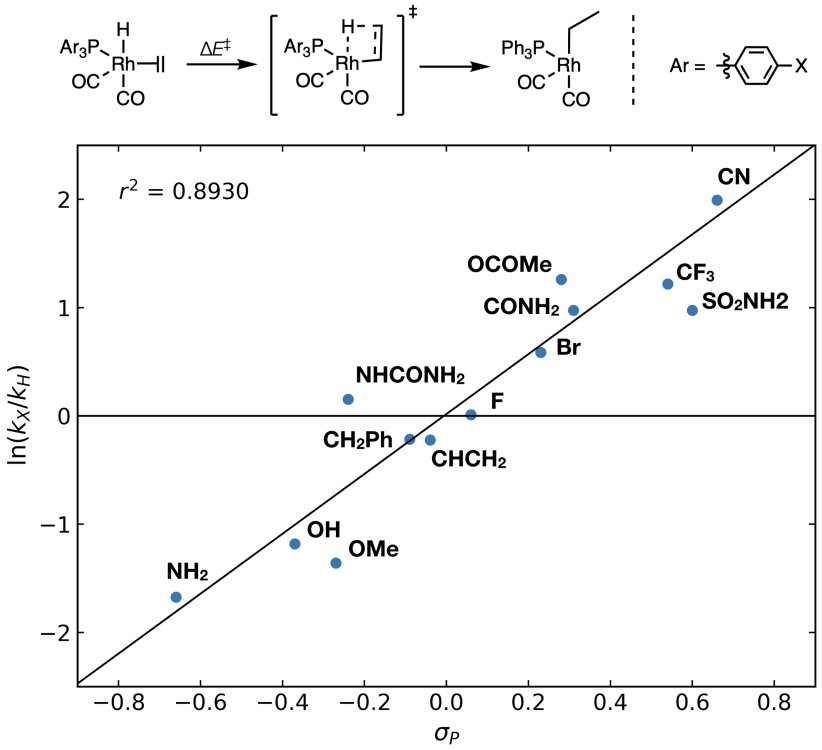


Figure 5.3.20. Relative rate ($\ln(k_X/k_H) \approx -\Delta\Delta E_A^{\ddagger}/RT$, T = 373.15 K) for the hydrogen migration catalysed by [RhH(CO)$_2$(PAr$_3$)(ethene)]. $\Delta\Delta E_A^{\ddagger}$ is the energy difference between the TS analogue and the canonical TS with X = H. The TS analogues are generated by keeping bonds fixed and only optimizing the added functional group). Energies calculated at PBE0-D3BJ/def2-SVP where Ar = *p*-C$_6$H$_5$X and X = {H, F, Br, CF$_3$, CH=CH$_2$, CONH$_2$, NO$_2$, NH$_2$}.

5.3.7 Reaction Classes

To demonstrate the applicability of *autodE* to multiple reaction classes we have generated reaction profiles for a selection of organic reaction types choosing a minimal example in each case (Figure 5.3.21). Several reactions (Figure 5.3.21 g, i, j, k) which have a TS below either products or reactants due to less stable separated anionic components at the DFT method used. Of particular note is acylium addition to benzene (Figure 5.3.20), where our TS conformational search algorithm finds a true TS (that links reactants and products) that is more stable (0.7 kcal mol^{-1}) than the Wheland intermediate conformer obtained with RDKit (2.3 kcal mol^{-1}). These differences are within the error of the level of theory used and suggest that this reaction is electronically barrierless.

All of the reaction profiles were generated successfully with the exception of the stepwise conversion of acetic acid to acetyl chloride by thionyl chloride, for which only the chloride addition TS was located with *autodE* (Step II, Scheme 1). This resulted from the substitution reaction being barrierless at the chosen *lmethod* (GFN2-XTB) and saddle point found on the 2D surface at the `low_opt` (PBE-D3BJ/ma-def2-SVP) level of theory not being a saddle point on the opt level (PBE0-D3BJ/ma-def2-SVP) surface. Note that *autodE* by default does not perform a 2D scan at the opt level due to computational expense. Both of these limitations are not intrinsic to the method, but rather of the accuracy electronic structure theory packages currently available.

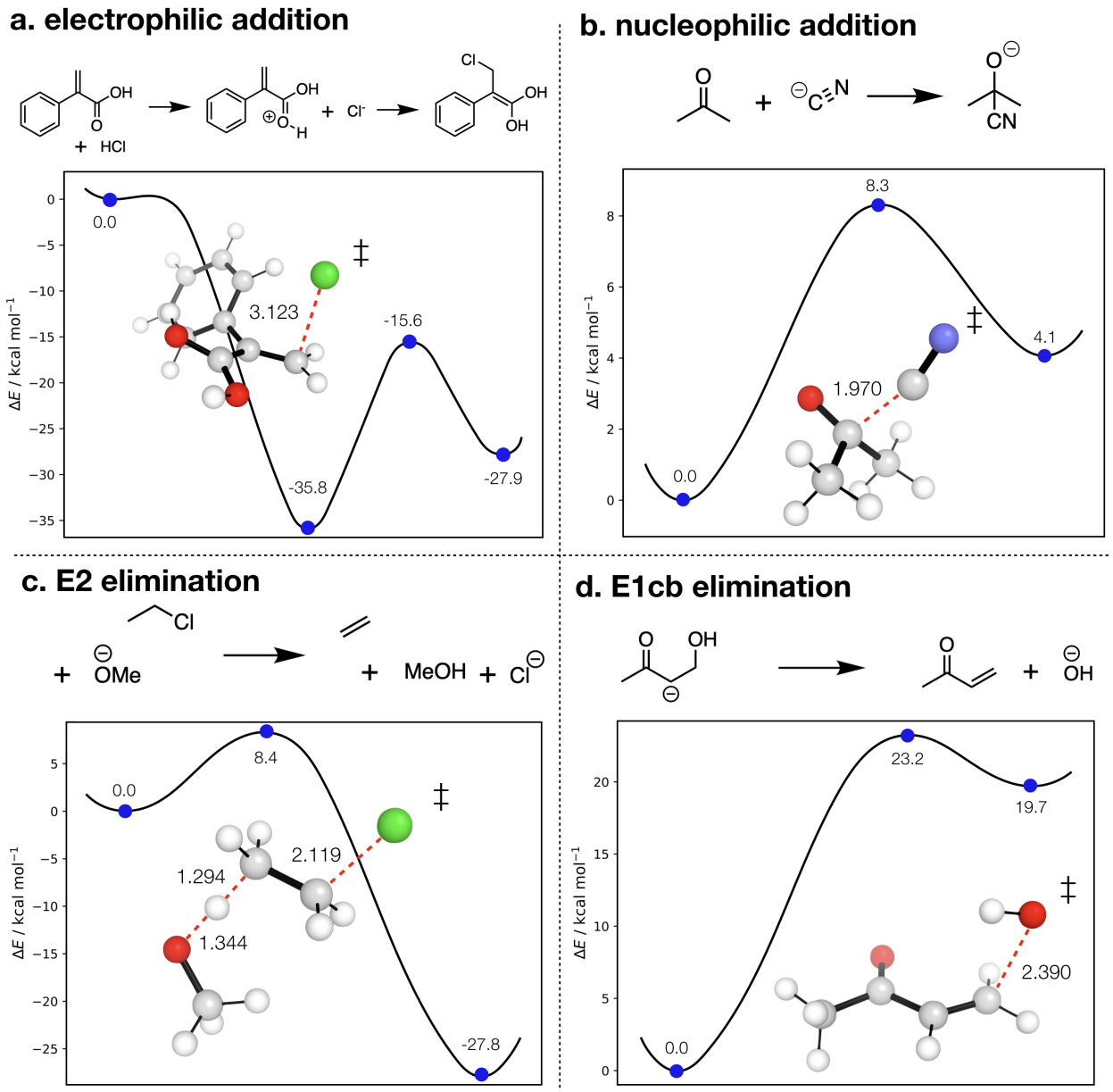


Figure 5.3.21. Reaction profiles for a range of organic reaction classes calculated at PBE0-D3BJ/def2-TZVP//PBE0-D3BJ/def2-SVP for neutral and CPCM(water)-PBE0-D3BJ/ma-def2-TZVP//CPCM(water)-PBE0-D3BJ/ma-def2-SVP for reactions with an anionic component with ORCA/XTB. (c) TS contained a second spurious imaginary frequency $< 50i$ cm^{-1} that could not be removed via displacement along the imaginary vector. b. TS contained two spurious imaginary frequency $< 50i$ cm^{-1}. Distances quoted in Å.

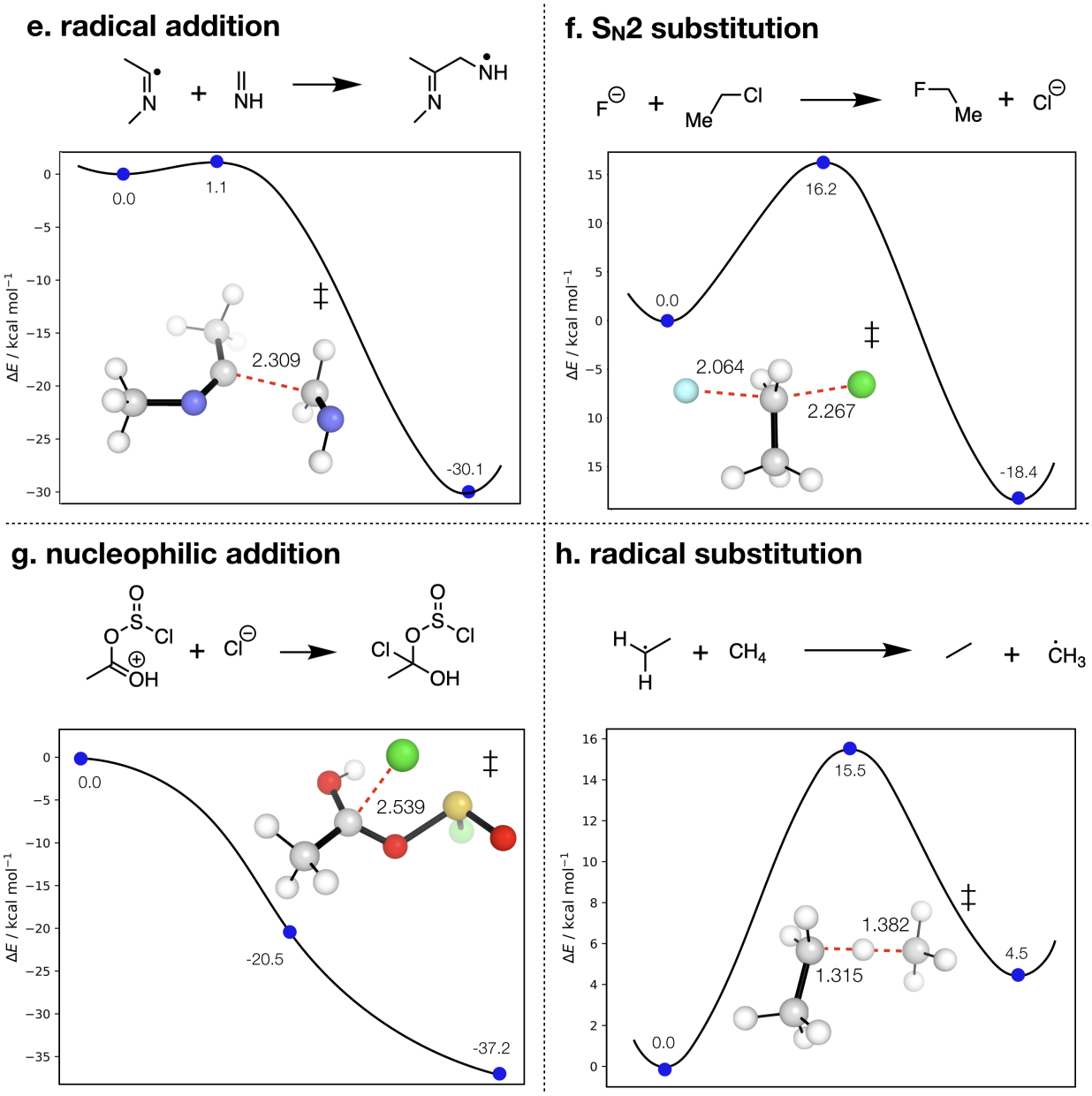


Figure 5.3.22. Figure 5.3.21 continued.

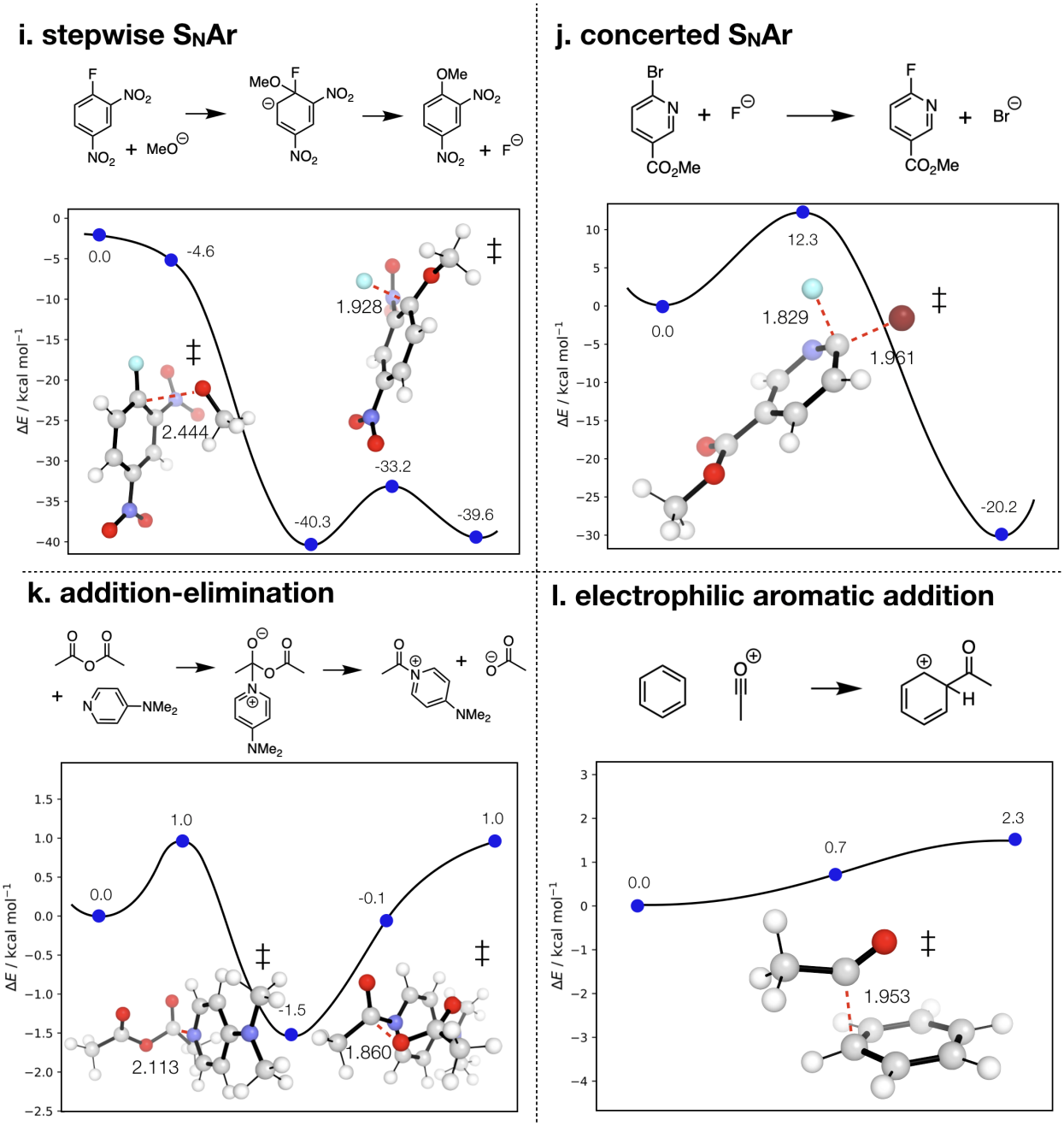


Figure 5.3.23. Figure 5.3.21 continued.

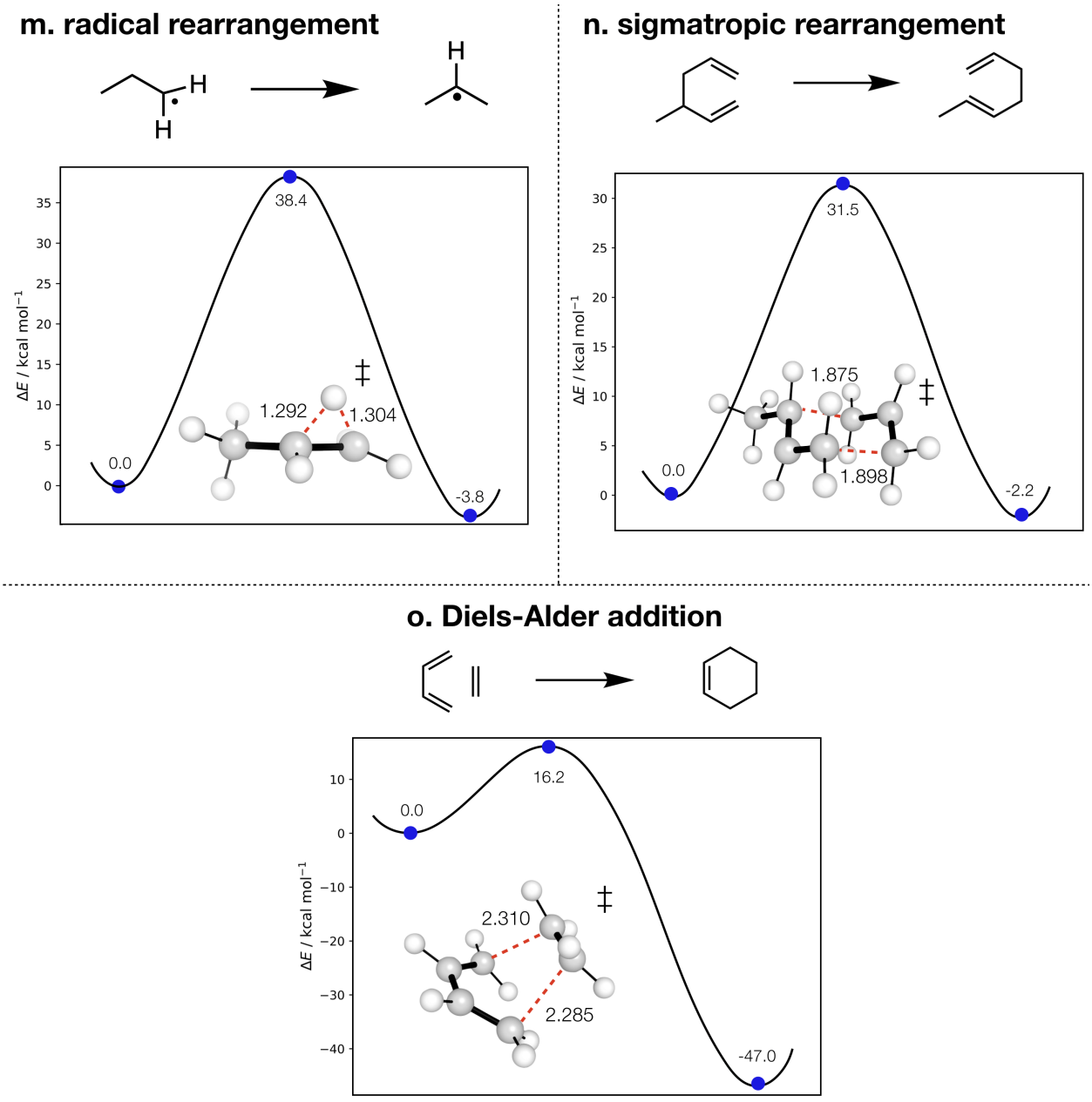


Figure 5.3.24. Figure 5.3.21 continued.

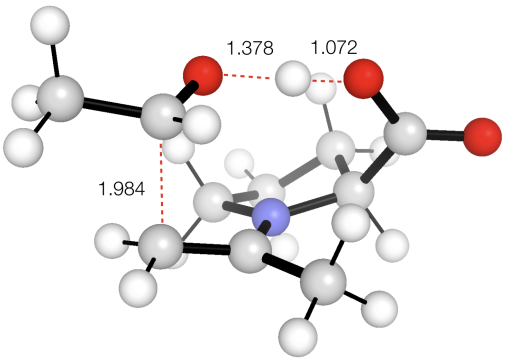


Figure 5.3.25. Houk-List TS for an asymmetric aldol reaction found with *autodE* (ORCA/XTB, PBE0-D3BJ/def2-TZVP//PBE0-D3BJ/def2-SVP). See full supporting information for input. Distances quoted in Å.

5.4 Further Developments

Since its publication in late 2020, a number of developments have been introduced into *autodE* to enhance functionality and reduce computational expense. These are outlined in the following sections, with in-progress developments itemised on the online repository.[a]

5.4.1 Transition State Guesses

The initial *autodE* implementation to generate TS guess geometries made use of relaxed 1D and 2D relaxed PES scans over the breaking/forming coordinates to search for a saddle point. This approach was employed based on the conventional human-guided method of calculating reaction profiles, but has some limitations: (1) a uniform step is taken, meaning the computational effort cannot be focused around the saddle point; (2) for high energy unphysical reactions involving two active bonds 1D low-level, 1D high-level, 2D low-level and 2D high-level searches were employed in sequence only to perhaps not find a TS; (3) 2D high-level surfaces are expensive to compute, thus for e.g. a Diels-Alder reaction where a low GFN2-XTB method is insufficiently accurate, 64 DFT optimisations were required. Therefore, a nudged elastic band (NEB)[415] along with its climbing image variant (CI-NEB)[416] were implemented as follows. The energy of a number of structures ('images') along a path are minimised under a force $\boldsymbol{F}$. Specifically, the implementation follows that

[a] See `https://github.com/duartegroup/autodE/projects` for current developments.

of Henkelman[417] where, for an image i in the nudged elastic band,

$$\boldsymbol{\tau}_i = \begin{cases} \boldsymbol{\tau}_i^+ & \text{if} \quad V_{i-1} < V_i < V_{i+1} \\ \boldsymbol{\tau}_i^- & \text{if} \quad V_{i+1} < V_i < V_{i-1} \\ \boldsymbol{\tau}_i^+ \Delta V_i^{max} + \boldsymbol{\tau}_i^- \Delta V_i^{min} & \text{if} \quad V_{i-1} < V_{i+1} \\ \boldsymbol{\tau}_i^+ \Delta V_i^{min} + \boldsymbol{\tau}_i^- \Delta V_i^{max} & \text{if} \quad V_{i+1} < V_{i-1} \end{cases} \tag{5.6}$$

where,

$$\begin{aligned} \boldsymbol{\tau}_i^+ &= \boldsymbol{R}_{i+1} - \boldsymbol{R}_i \\ \boldsymbol{\tau}_i^- &= \boldsymbol{R}_i - \boldsymbol{R}_{i-1} \end{aligned} \tag{5.7}$$

and,

$$\begin{aligned} \Delta V_i^{max} &= \max(|V_{i+1} - V_i|, |V_{i-1} - V_i|) \\ \Delta V_i^{min} &= \min(|V_{i+1} - V_i|, |V_{i-1} - V_i|) \end{aligned} \tag{5.8}$$

and $\boldsymbol{R}_i$ are the coordinates of image i. The spring force is,

$$\boldsymbol{F}_i^s|_{\parallel} = (k_i|\boldsymbol{R}_{i+1} - \boldsymbol{R}_i| - k_{i-1}|\boldsymbol{R}_i - \boldsymbol{R}_{i-1}|)\hat{\boldsymbol{\tau}}_i \tag{5.9}$$

and the total force on the image,

$$\boldsymbol{F}_i = \boldsymbol{F}_i^s|_{\parallel} - \nabla V(\boldsymbol{R}_i)|_{\perp} \tag{5.10}$$

where,

$$\nabla V(\boldsymbol{R}_i)|_{\perp} = \nabla V(\boldsymbol{R}_i) - \nabla V(\boldsymbol{R}_i) \cdot \hat{\boldsymbol{\tau}}_i \hat{\boldsymbol{\tau}}_i \tag{5.11}$$

and finally $\hat{\boldsymbol{\tau}} = \boldsymbol{\tau}_i / |\boldsymbol{\tau}_i|$.

The climbing image (CI) NEB implementation follows that in ref. [416], where after a few

iterations the force on the maximum energy image (m) is given by,

$$\boldsymbol{F}_m = -\nabla V(\boldsymbol{R}_m) + 2\nabla V(\boldsymbol{R}_m) \cdot \hat{\boldsymbol{\tau}}_i \hat{\boldsymbol{\tau}}_i \tag{5.12}$$

which is the force due to the potential along the band being inverted.

Employing a NEB relaxation of a path and taking the highest lying point on the approximate minimum energy pathway (MEP) as a transition state guess, however, relies on an initial reasonable path to relax along, and the NEB not corner cutting (i.e. the images relaxing into the reactant/product basins). To develop an initial path reactant and product complexes were overlapped and interpolation in Cartesian coordinates was employed. It is, however, challenging to align and rotate dihedrals to minimise the difference. Instead, a linear path from reactants to products over the changing bond lengths (i.e. sequential constrained minimisations) was an improvement but can lead to unexpected dissociations.[b] In an attempt to reduce corner cutting an adaptive force constant, where k is larger at images with higher energy, was introduced and is, generally efficient.[417]

$$k_m = k_{\max} - \Delta k \left(\frac{V_{\max} - V_m}{V_{\max} - V_{\text{ref}}} \right) \quad \text{if} \quad V_m > V_{\text{ref}} \qquad \text{else} \quad k_m = k_{\min} \tag{5.13}$$

where $V_{\text{ref}} = \max(V_n, V_0)$ and $\Delta k = k_{\max} - k_{\min}$.

For example, for a 1,2-shift in a terminal propyl radical using a fixed force constant and relaxing the NEB leads to two images very close to the fixed end points while the two higher lying images cut the barrier peak and miss a close TS geometry (Figure 5.4.1). Although improved using an adaptive force constant, this effect is one reason why NEB relaxation is not employed by default in *autodE*.

[b] For example, the E2 elimination using methoxide+propyl chloride on a linear path using GFN2-XTB generates an unphysical $[H_3CCHCH_3]+$ cation.

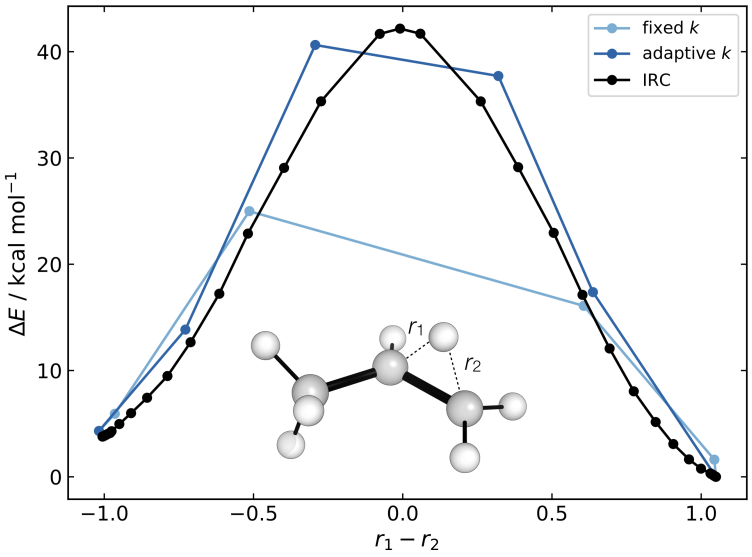


Figure 5.4.1. Comparison of approximate minimum energy pathways for the H-shift in a terminal propyl radical at the PBE0-D3BJ/def2-SVP level of theory. NEB paths in blue $k = 0.1$ Ha Å^{-1}, fixed or adaptive according to Eq. (5.13) and the reference IRC in black.

While searching for an improved method to generate reaction paths upon which NEB relaxation may be performed, an algorithm was developed for traversing a reasonable initial pathway. Motivated by steepest-decent minimisation where the direction is defined by the gradient of the surface an 'adaptive' path may be defined using gradient dependent constraints. The initial constraints for the first point are,

$$r_b^{(1)} = r_b^{(0)} + \mathrm{sgn}(r_b^{\mathrm{final}} - r_b^{(0)})\Delta r_{\mathrm{init}} \tag{5.14}$$

for an active bond b, where the superscript denotes the current step. Δr_{init} is an initial step size, e.g. 0.2 Å. Constraints for subsequent steps are then given by,

$$r_b^{(k)} = r_b^{(k-1)} + \mathrm{sgn}(r_b^{\mathrm{final}} - r_b^{(0)})\Delta r_b^{(k-1)} \tag{5.15}$$

$$\Delta r_b^{(k)} = \begin{cases} \Delta r_{\max} & \text{if } \operatorname{sgn}(r_b^{\text{final}} - r_b^{(0)}) \nabla E_j \cdot \boldsymbol{r}_{ij} > 0 \\ \Delta r_{\text{m}} \exp\left[-\left(\nabla E_j^{(k)} \cdot \boldsymbol{r}_{ij} / g \right)^2 \right] + \Delta r_{\min} & \text{otherwise} \end{cases} \tag{5.16}$$

where $\Delta r_{\text{m}} = \Delta r_{\max} - \Delta r_{\min}$, E the total potential energy (in the absence of any harmonic constraints) and g a parameter to control the interpolation between $\Delta r_{\max}$ and $\Delta r_{\min}$ e.g. 0.05 Ha Å^{-1}. Atom indices i, j form part of the bond indexed by b with j being an atom not being substituted. In the case that neither i nor j are being substituted the gradient is taken as an average over i and j.

This approach generates a parameter dependent but remarkably accurate way of traversing close to a MEP with no *a priori* knowledge. For the standard model London–Eyring–Polanyi –Sato (LEPS) analytic 2D PES (see ref. [418] for a review and definition) a judicious choice of parameters can afford an excellent path (Figure 5.4.2).

Motivated by this accuracy a more complex, true chemical reaction path was traversed using the same method with parameters $\Delta r_{\max} = 0.3$ Å and $g = 0.05$ Ha Å^{-1}. Specifically, for the fluoride + methyl chloride reaction employing a small minimum step size, a point within 0.1 Å of the true TS is traversed, which is a good TS guess (i.e. transition state optimisation is successful in just a few geometry optimisation steps).

Further testing of this method suggests a significant performance benefit, for a set of small ($<$ 20 atoms) organic reactions adapted from ref. [419] (Figure 5.4.4) and small organometallic reactions[296] (Figure 5.4.5) a guess TS is generated in $\sim$ 10 minutes on 8 CPU cores in each case.

In particular, for Diels-Alder reactions the performance increase over previously employed

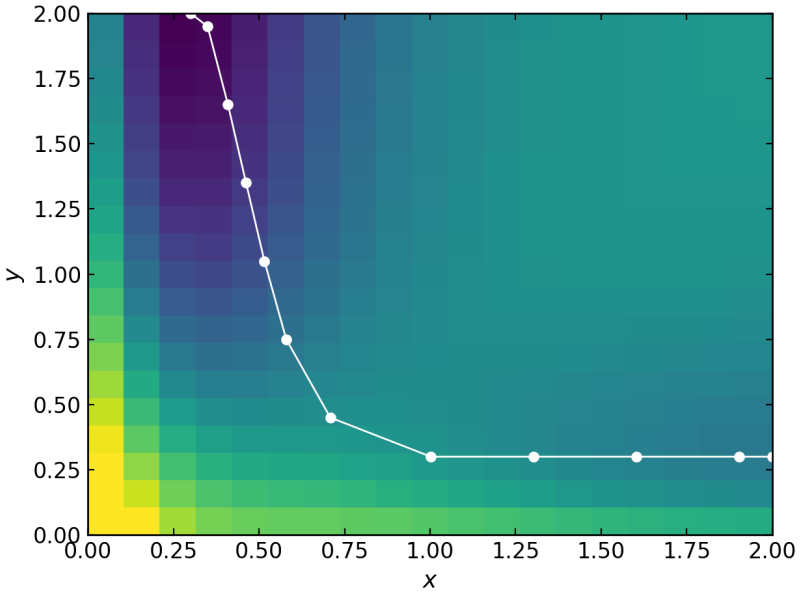


Figure 5.4.2. LEPS potential ($a = 0.05; b = 0.8; c = 0.0; r_0 = 0.742; \alpha = 1.942$) overlaid with an adaptive path, with parameters $\Delta r_{\text{min}} = 0.05, \Delta r_{\text{max}} = 0.3, g = 50$.

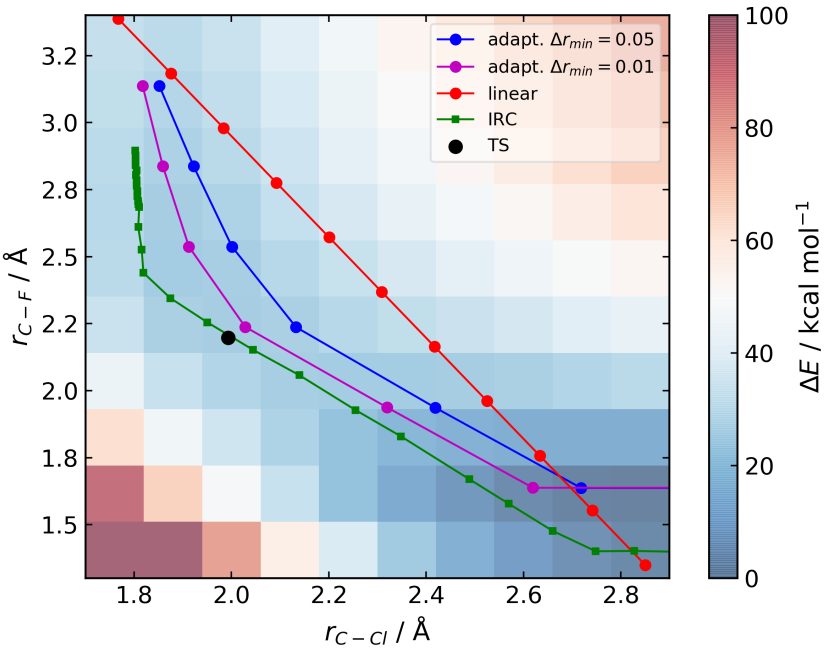


Figure 5.4.3. Relaxed 2D PES calculated at PBE0-D3BJ/ma-def2-SVP/CPCM(water) overlaid with adaptive paths the IRC and the true (optimised) TS highlighted.

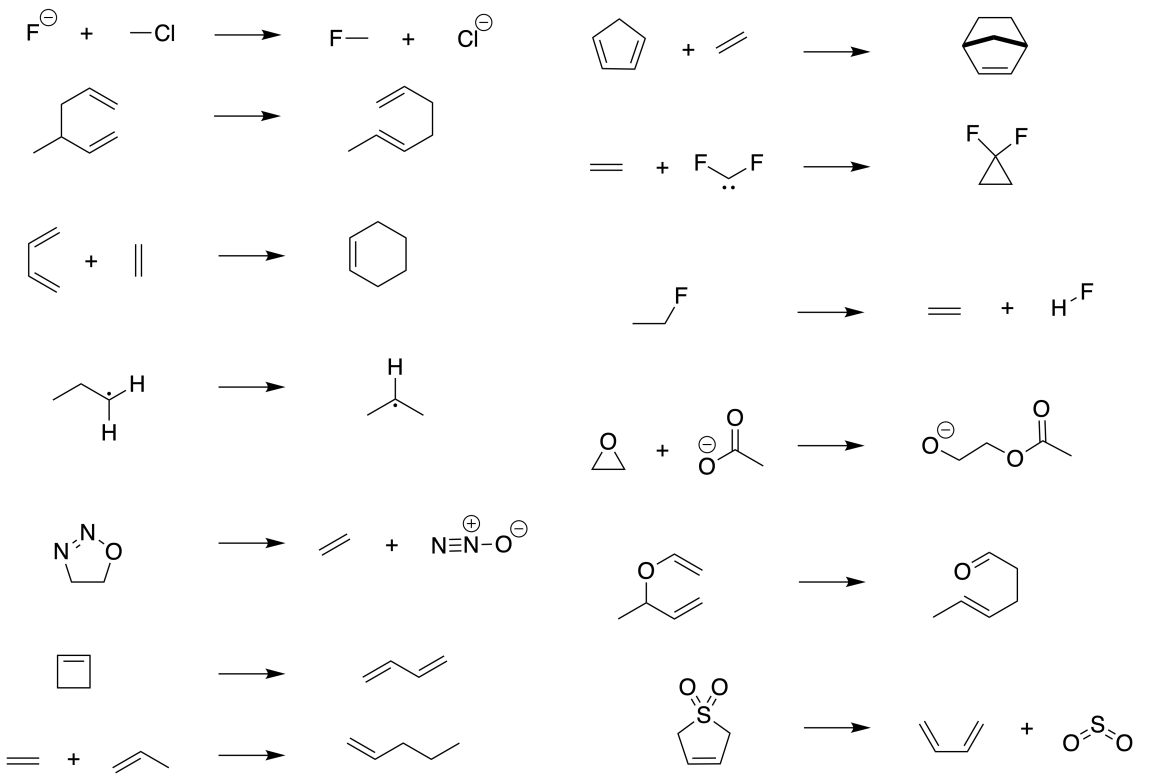


Figure 5.4.4. Small organic reactions.

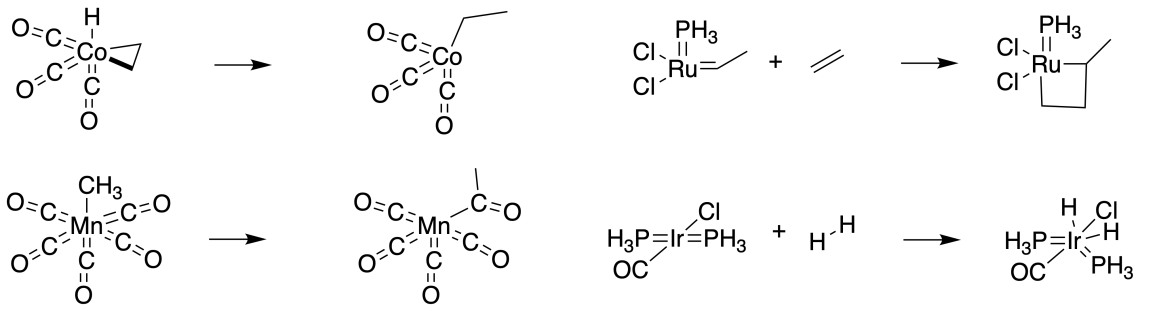


Figure 5.4.5. Small organometallic reactions. Explicitly defined valencies are required for accurate SMILES strings

2D relaxed PES scans is significant, as only a single coordinate is traversed resulting in at least a 10× speed-up. For this reason the adaptive path algorithm is now the default in *autodE* for all (x, y) transformations (where x bonds form and y bonds break).

5.4.2 Molecule Building

Converting a SMILES string into a 3D geometry was achieved within *autodE* using RDKit[294] for organic molecules and using a SMILES parser and a randomise+relax algorithm on all atoms for metal complexes. This approach, however, comes with two drawbacks: (1) cyclic organic molecules containing *trans* (E) double bonds are (as of RDKit v. 2020) generated as their *cis* (Z) analogues and (2) randomising all atoms does not respect any stereochemistry defined in the SMILES string (point tetrahedral or E/Z).

To rectify these limitations a completely new (Open)SMILES parser was developed that is fast, well tested and extensible. From this, 3D geometries are built by:

1. Converting all implicit hydrogen atoms defined in the SMILES string to explicit atoms (e.g. 'CC' refers to a ethane molecule).

2. Setting the atom types with templated vacant sites, for example a tri-coordinate atom is assigned a trigonal template unless it is a group 15 element in which case a trigonal-pyramidal geometry is defined.

3. For each atom in turn:

    a) Translate the atom to the origin if it's the first to be added, or a vacant site on an already-translated atom.

    b) Rotate the empty sites on the atom such that one is coincident with the neighbouring atom.

While simple, this approach is remarkably efficient at generating linear molecules, with an execution time generally < 100 milliseconds. There are, of course, special cases that require some care. For example, double bonds with a defined E/Z stereochemistry require the

dihedral angle across the double bond to be adjusted appropriately. Rings and particularly fused rings are challenging to build using this approach, as added atoms can no longer be chosen to minimise the repulsion with the rest of the structure. Once ring atoms have been added sequentially the dihedral angles are adjusted to minimise,

$$V = k(r - r_{\text{avg}})^2 + \sum_{i>j} c_{ij}/r_{ij}^n \tag{5.17}$$

where n is a positive integer and average distances r_{avg} are taken from the CCDC. This 'closes' the ring while maintaining a small steric repulsion.

For a flat aromatic ring (e.g. phenyl) then the three dihedrals must be simply be zeroed. In a saturated or semi-saturated ring, however, many dihedrals must be minimised on as to close the ring. This is achieved using a stochastic global steepest-decent minimisation where random points in the space are selected.[c] Further complications arise when small rings ($<$ 5-membered) or metallocycles are closed, as rotating dihedrals is no longer sufficient to adequately close a ring. In these contexts, either the bond angles within the ring are adjusted, or the whole ring minimised with a repulsion+bonded (RB) potential, as previously used to generate the whole structure.

5.4.3 Future Development

Currently, *autodE* provides access to reaction profiles for a wide range of reactions. However, some limitations remain that encourage further developments. Currently a maximum of four bonds can be been broken and formed in a a single step and the performance limited by the inaccuracy of 'low-level' electronic structure methods. Addressing these limitations will be

[c]Note an improved semi-stochastic algorithm by maximising the distance between the starting points in the n-dimensional space for n dihedrals is under development.

the subject of future development, with the current focus on the following aspects:

- Improved load balancing in conformer optimisation. Currently, each conformer optimisation or single-point energy evaluation is performed in serial. While this approach is close to optimal for calculations run on $\lesssim 8$ CPU cores, for which most electronic structure codes parallelise almost linearly up to, it is not for *autodE* calculations performed on more cores. A context-aware and dynamic multiprocessing approach therefore would be beneficial to reduce the total wall time.

- Consistent thermochemical calculations. Currently, *autodE* uses thermochemical data calculated by the electronic structure codes and is therefore inconsistent. Hessian extraction and diagonalisation to calculate frequencies thus enthalpy/entropy/free energies would provide a consistent framework.

- Implementation of minima and transition state optimisers. Currently, optimisations are performed using electronic structure codes, which means (1) Atomic Simulation Environment (ASE[420])-wrapped codes cannot be used for efficient optimisations, as ASE optimisation are in Cartesian coordinates; (2) convergence criteria cannot be defined explicitly and (3) optimisations can crash irrevocably in both ORCA and Gaussian.[d] Due to (2), an *autodE*-based optimisation could afford faster path searches by employing loose convergence criteria for regions not close to the TS. Implementing minimisers would also enable greater functionality with newly emerging machine learned potentials (MLPs).

- Implementation of wrappers for other electronic structure packages e.g. TurboMole and PSI4. Adding these methods will enhance the flexibility and usability of *autodE*.

[d]For example, when an angle approaches 180° the calculation fails in Gaussian09, which is handled in *autodE* but requires a recursive fix, which may fail after a few attempts. On the other hand ORCA does not stop when the internal→Cartesian back transformation fails, but enters a high-energy region of phase space.

- Improved conformer enumeration using an exhaustive algorithm. Currently, conformer generation is performed with RDKit for organic molecules and adding random displacements then minimising for organometallic species. However, for small and/or rigid molecules complete sampling of the conformational space may be possible and will provide the global minimum.

I expect most or all of these to be included with version 1.1, which will be released at some point in 2021, further enhancing the accuracy and speed of *autodE*.

6 Machine Learned Potentials

6.1 Overview

Although DFT remains steadfast as the computational method of choice to study reactions of more than a few atoms,[421] there remain significant limitations in speed and accuracy.[76] For example, while a DFT estimate of the catalytic activity of a metallocage may be obtained in a day[280] (faster than the corresponding experiment), efficient methods are required for high-throughput virtual screening. Moreover, absolute barrier heights are rarely within 5 kcal mol^{-1} of the true value (e.g. ref. [422]). Therefore, a predictive computational approach to catalyst design using enumerative *ab initio* calculations requires faster access to accurate absolute energies.

Accurate inclusion of solvent effects is also a long-standing challenge, with solvation free energy errors in the 5 kcal mol^{-1} range for neutral and >5 kcal mol^{-1} for charged species.[130, 423] Inclusion of solvent effects capturing directional interactions missing from implicit models require either cluster-continuum models or QM/MM. The former often lacks sufficient sampling and/or convergence with respect to the number of explicit solvents,[424] while the latter also suffers from how to couple the regions, and forcefield accuracy.[425] Methods that could calculate total periodic system energies quickly would thus alleviate these problems

and enhance accuracy where solvent effects are appreciable.

Reaction dynamics can also be important in (catalytic) reactions,[426, 427] but generally require computationally expensive AIMD. Therefore, the accuracy is generally limited to the GGA DFT-level in the condensed phase.[428] Methods to rapidly generate energies and forces that are reactive and accurate would increase the understanding of these processes.

Manuscript IV outlines a route to developing bespoke machine learned potentials capable of delivering energies and forces in O(ms) for a wide range of systems. Using a Gaussian Approximation Potential framework,[429] hyperparameters and a method of generating reference configurations were developed so an ML potential can be trained for a variety of systems in a day or so. This method enabled CCSD(T)-quality dynamics for an S_N2 reaction, accurate bulk (periodic) water simulations using hybrid DFT and reactive dynamics of a condensed phase reaction. The training method is also applicable to reaction discovery; by initialising velocities sufficient to make/break bonds, an exothermic S_N2 reaction was sampled from an association complex of reactants.

The method and freely available code (`https://github.com/t-young31/gap-train`) will enable more routine development of ML potentials where, for example, possible bifurcations occur and barrier heights are insufficient to predict selectivity.[66] Solvent effects can be interrogated using this method and a GGA DFT-quality potential generated within a day. It is also applicable to finding intermediates and transition states of new reactions. However, the generated potentials are intrinsically bespoke and thus limited in transferability. Therefore, more data-intensive approaches are likely to prevail in this domain.[160, 430]

6.2 Manuscript IV

A Transferable Active-Learning Strategy for Reactive Molecular Force Fields

6.2.1 Abstract

Predictive molecular simulations require fast, accurate and reactive interatomic potentials. Machine learning offers a promising approach to construct such potentials by fitting energies and forces to high-level quantum-mechanical data, but doing so typically requires considerable human intervention and data volume. Here we show that, by leveraging hierarchical and active learning, accurate Gaussian Approximation Potential (GAP) models can be developed for diverse chemical systems in an autonomous manner, requiring only hundreds to a few thousand energy and gradient evaluations on a reference potential-energy surface. The approach uses separate intra- and inter-molecular fits and active learning to maximise a prospective error metric used to quantify accuracy. We demonstrate applications to a range of molecular systems: from bulk solvents, a solvated metal ion and a metallocage onwards to chemical reactivity, including a bifurcating Diels–Alder reaction in the gas phase and non-equilibrium dynamics (S_N2 reaction) in explicit solvent. The method provides a route to routinely generating machine-learned force fields for reactive molecular systems.

6.2.2 Introduction

Molecular simulations are a cornerstone in computational chemistry, providing dynamical insights beyond experimental resolution.[431] Realistic simulations of (bio)chemical reactions require the inclusion of the chemical environment where they occur (e.g. solvent and/or enzyme) and often extended timescales. Therefore, the generation of accurate and efficient

approaches has been central to the development of this field.

Empirical interatomic potentials (force fields), in combination with molecular dynamics (MD) or Monte Carlo (MC) simulations, have been widely used to sample the potential-energy surface (PES). However, they are limited in accuracy and transferability.[432] Moreover, most of these potentials are parameterised for isolated entities with fixed connectivity and thus unable to describe bond breaking/forming processes. In contrast, *ab initio* methods provide an accurate description of the PES, which is particularly critical for reactions in solution. However, because of their high computational cost and unfavourable scaling behaviour, they are limited to a few hundred atoms and simulation times of picoseconds in *ab initio* molecular dynamics (AIMD) at the DFT level, and practically impossible at the computational 'gold-standard' [CCSD(T)].[149]

Machine learning (ML) approaches have the potential to revolutionise force-field based simulations, aiming to provide the best of both worlds,[169, 433, 434] and have indeed begun to provide new insights into a range of challenging research problems.[435–444] The development of an ML potential applicable to the whole periodic table mapping nuclear coordinates to total energies and forces is, however, precluded by the curse of dimensionality. Within small chemical subspaces, models can be achieved using neural networks (NNs),[158, 160, 434, 445–447] kernel-based methods such as the Gaussian Approximation Potential (GAP) framework[171, 429] or gradient-domain machine learning (GDML),[448] and linear fitting with properly chosen basis functions,[164, 449] each with different data requirements and transferability.[450] GAPs have been used to study a range of elemental,[451–453] multicomponent inorganic,[454, 455] gas-phase organic molecular,[441, 456] and more recently condensed-phase systems, such as methane[457] and phosphorus.[458] These potentials, while accurate, have required considerable computational effort and human oversight. Indeed, condensed-

phase NN[459, 460] and GAP fitting approaches typically require several thousand reference ('ground truth') evaluations.

Active learning (AL), where new training data is added based on the current state of the potential, has been used for generating databases and accelerating the fitting process.[454, 461–465] Notable examples in materials modelling include an early demonstration of a "query-by-committee" approach in fitting a high-dimensional NN potential for elemental copper,[462] the fitting of Moment Tensor Potential[449] models[466] to predict elemental crystal structures[461] and multicomponent alloys,[463] and the deep potential generator (DP-GEN)[467, 468] that provides an interface to deep NN potential models for materials.[469] AL schemes have also been combined with GP based force fields including GAP,[470] and included within a first-principles MD implementation such that it allows the "on the fly" fitting of force fields for a specific simulation system.[471, 472]

Efficient approaches to generate reactive ML potentials become even more important when exploring chemical reactions in molecular systems, which often require a description at a computational level beyond DFT, and therefore require reference data at the same level. Very recently, AL approaches have started to be adopted for fitting reactive potentials for organic molecules based on single point evaluations at quantum-chemical levels of theory. Notable examples include the modelling of gas-phase pericyclic reactions,[440] the exploration of reactivity during methane combustion,[473] and the decomposition of urea in water.[464]

In the present work – with a view to developing potentials to simulate solution phase reactions – we consider bulk water as a test case and develop a strategy which requires just hundreds of *total* ground truth evaluations and no *a priori* knowledge of the system, apart from the molecular composition. We show how this methodology can be directly transferred to chemical systems in the gas phase as well as in implicit and explicit solvent, focusing on the

applicability to a range of scenarios that are relevant in computational chemistry.

6.2.3 Results and Discussion

Despite GAP fitting being increasingly used for inorganic systems, we found that the same fitting strategies did not easily transfer to the description of complex molecular environments. Even with a high correlation and low error on energies in unseen test data, some potentials were not stable for more than a few femtoseconds. In the following section, we therefore outline a training strategy along with a prospective error metric to develop robust models for gas-phase and condensed-phase molecular systems.

6.2.3.1 A Prospective Error Metric

The initial step in validating supervised machine learning (ML) tends to follow the splitting of a dataset into training and test sets, training the model, then evaluating its performance on the test set with a squared error (RMSE/MSE) or a correlation (R^2) metric. As with model overfitting, this 'retrospective' validation strategy ultimately limits the applicability of these models.[474–480] In an ML potential, the minimum required domain of applicability is the region of configuration space likely to be sampled during a simulation with the potential. However, this region is not known *a priori*, making the choice of test data problematic if not impossible for use in a standard train/test data split approach. In addition, one would also like to ensure high accuracy in regions sampled on the ground truth surface (especially for early versions of an evolving potential), but being able to quantify this accuracy requires dynamics at the ground truth method level in the first place, which is much more expensive than sampling with an efficient potential.

Using a train/test set split with high structural similarity between the two sets can lead

to highly misleading accuracy whenever the potential is to be taken outside the training region in computational practice. For example, splitting an AIMD trajectory of water into a training and test set with an odd/even frame split (50:50) and training a simple GAP model yields an energy error on the order of 1 kcal mol^{-1} (Figure 6.3.1a). However, simulations with this potential in the same configuration space sample unphysical configurations within 10 fs (Figure 6.3.1b), making an RMSE over *a priori* test data an insufficient metric in quantifying the quality of a potential.

Considering that single-point reference energy evaluations are reasonably cheap, a 'prospective' validation scheme is possible, where the error metric operates in the configuration space sampled in a simulation. With this in mind, we propose a temporal cumulative error metric (τ_{acc}, Eqn. (6.1)), defined as the time required for the cumulative error (absolute difference between true (E^0) and predicted (E^{GAP})) to exceed a given threshold (E_T); the larger τ_{acc}, the more robust the potential. Note that the time for which a potential is stable in MD can far exceed τ_{acc}, as shown in the following. Here only errors above a lower-bound threshold value (E_l) contribute to the cumulative error. The lower threshold is required to account for the residual error that is due to the finite radial cut-off of the model. In the following we take E_T to be 10 times E_l, but it may be adjusted depending on the simulation context.

$$\tau_{\text{acc}} = \text{time} \quad : \quad E_T < \sum_{i \in \text{frames}} \max(|E_i^0 - E_i^{\text{GAP}}| - E_l, 0) \tag{6.1}$$

This metric has several advantages in that (a) it ensures that a potential with high accuracy will result in stable dynamics; (b) it allows the user to specify the level of accepted error according to the quality of the training method, thus not penalising where the error is within the difference between the ground truth and the true PES (i.e. a larger threshold may be suitable for a less accurate reference method); (c) it penalises large errors, even

if they only occur for single configurations, which is important as such errors may lead to instabilities in the ML-driven MD trajectory and (d) it enables a quoted accuracy to include regions that may not be accessible to direct evaluation at the ground-truth level (e.g. long-time behaviour). Overall, this metric depends on the lower bound and total error, interval between evaluations, and the simulation on which it is evaluated; so while not unique, it is – crucially – prospective. We found this metric to be essential in developing an efficient training strategy and accurate potentials for bulk water (Figure 6.2.1).

6.2.3.2 Water Models

For bespoke ML potentials to be routinely developed for molecular systems, one would hope to complete the data generation, model training, and know the accuracy of the resulting potential within a matter of hours to days. With this in mind, here we train GAP models to simulate bulk water, aiming to minimise the number of required ground truth evaluations as well as the required human intervention, while maximising stability (measured by τ_{acc}). A selection of training strategies is discussed in the following paragraphs and their results are outlined in Figure 6.2.1.

We initially employed training strategies found to work well in elemental materials by, for example, fitting a combined potential with two- and three-body GAPs. However, this approach was found to be detrimental to the potential's stability. This can be understood considering a water dimer ($\mathrm{HO–H}_c \cdots \mathrm{OH}_2$); here, a two-body description that treats the two O–H_c interactions on the same footing is a poor approximation, in view of the different order of magnitude between the interactions at their respective minima (Figure 6.3.3). Therefore, we decided to proceed employing a smooth overlap of atomic positions[161] (SOAP) descriptor for an exclusively many–body description of atomic environments (with the exception of

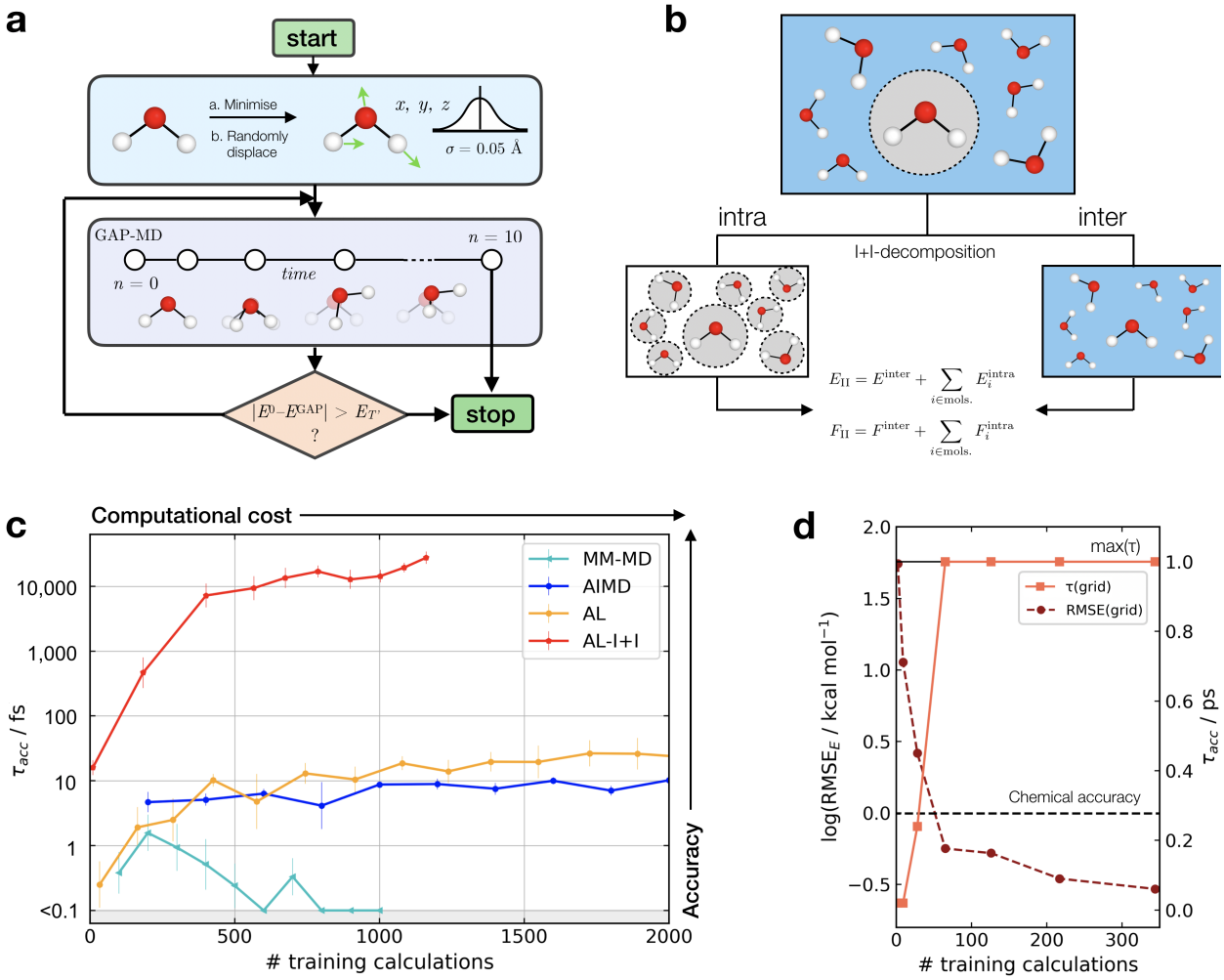


Figure 6.2.1. Active learning of machine-learning potentials for liquid water. (a) Schematic of the active learning loop implemented for fitting GAP models, where the GAP-MD exploration is run for n^3+2 femtoseconds, where n (the number of evaluations) is incremented after each time the error is evaluated. (b) Schematic illustrating the separation into inter- and intra-molecular terms (I+I) for a bulk water system; these are described by separate GAP models and then added to give the combined prediction for energies, E, and forces, F. (c) Learning curves for a bulk water GAP model using different training strategies. τ_{acc} with $E_l = 0.1$ eV, $E_T = 1$ eV, 10 fs interval, 300 K, from the same random minimised configuration of 10 waters in a 7 Å cubic box. Error bars quoted as the standard errors in the mean from 5 independent repeats. The horizontal axis denotes the number of evaluations in training data generation. See Tables 6.3.1–6.3.2 for detailed methods. DFTB(3ob) ground truth. Minimum τ_{acc} is shown as 0.1 fs to enable plotting on a log scale. (d) Water monomer model training performance as characterised by τ_{acc} and RMSE over the full 3D PES; see Figure 6.3.2 for details.

AL-I+I, which uses 2+3 body for the intramolecular component as discussed later).

We also explored different approaches to generate the database and their influence on the generated potential. An emerging approach to generate training data for elemental GAPs is to initialise the database with randomised configurations (with reasonable constraints, as in *ab initio* random structure searching[481]), and to gradually explore configuration space with evolving versions of the potential (see, e.g., ref. [482]). However, randomly placing water molecules does not in itself afford a stable potential. A similar result is observed when the most diverse configurations are selected using the CUR algorithm[482, 483] (Figure 6.3.4) or when applying intramolecular displacements, following minimisation (Figure 6.3.4). Selecting frames from classical MD simulations at temperatures of 100–1000 K was also found to be an ineffective strategy (Figure 6.3.5), reaching τ_{acc} of only a few fs (MM-MD, Figure 6.2.1c). This is in line with the results reported in ref. [441]. Note that this is not because the GAP cannot fit reference energies and forces from MM configurations (Figure 6.3.6), but because of a poor configuration space overlap with the ground truth PES (Figure 6.3.7). Selecting configurations from an AIMD simulation at 300 K (AIMD, Figure 6.2.1c) was an improvement over training on random and MM-generated configurations, with $\tau_{\text{acc}} \sim 10$ fs. However, by adding additional AIMD configurations the increase in accuracy saturates quickly, even if those are obtained at higher temperatures (Figure 6.3.9). Using AIMD configurations can also involve a significant cost (requiring thousands of evaluations). Finally, active learning from only a few randomly generated configurations provides a modest uplift in accuracy (AL, Figure 6.2.1c), with accuracy on-par with GAP trained on AIMD configurations at a third of the required reference data.

Only when the relevant length and energy scales of the system are decomposed by treating intra- and inter-molecular components separately (Figure 6.2.1b) a potential that is stable

for picoseconds is obtained (AL-I+I, Figure 6.2.1c). We note that this approach is related to the hierarchical fitting of GAPs[161, 457] and related ML models[456, 460, 484, 485] using different levels of computational approaches, and the decomposition strategy by Wengert and co-workers.[486] In the present work, we employ the same ground-truth method throughout rather than combining different levels of theory for the input data, but as in prior studies we describe the stronger (e.g., covalent) and weaker intermolecular terms with separate fits that are afterwards combined to give the final model. The intramolecular GAP for water contains only 2- and 3-body terms and the training data are chosen using an evenly spaced grid over the full 3-atom PES ($8 \times 8 \times 8$ grid points in r_{OH} and r_{HH}, $\sim$0.1 Å spacing, Figure 6.3.10). Energy and force evaluations of this potential are a simple sum of intra- and inter-molecular terms, but require the former to be evaluated in an expanded simulation box to ensure no non-bonded hydrogen atoms are present within the cut-off radius of the 2- and 3-body descriptors on oxygen (Figure 6.3.11). Here the intramolecular PES is fairly low-dimensional, so a full and reasonably dense grid is available, which in turn allows us to define an error measure over the whole PES, where we find the error to be inversely correlated with τ_{acc} (Figure 6.2.1d). Using an acceptable error of 0.2 kcal mol^{-1} per H_2O molecule for a description of bulk water, which is similar to that achieved in a recent NN fit of water,[459] we find that this potential affords $\tau_{\mathrm{acc}} > 10$ ps with just a few hundred ground truth evaluations (AL-I+I, Figure 6.2.1c). To put this value in context, we measured τ_{acc} for the fully reactive water NN of Cheng et al.,[459] which was trained on $\sim$7000 reference configurations (DFTB energy/forces) and has shown to provide a highly accurate water model over multiple states. For this state-of-the-art ML potential, a τ_{acc} value of 7.6$\pm$0.7 ps for liquid at 300 K is obtained, comparable to the one obtained for the new AL-based potentials of the present work ($>$ 10 ps). Of course, direct comparison requires caution because the GAP and NN potentials are different in scope: the former employs a large reference database to develop

a general water model, whereas the present study targets robust potentials for liquid water with minimal computational effort, in turn allowing the user to apply similar approaches to other chemical systems (as will be shown below).

The model fitted using our approach (AL-I+I) yields radial distribution functions (RDFs) in good agreement with the ground-truth method, initially chosen to be DFTB, both considering the location and intensities of the peaks corresponding to the first and second coordination shells (Figure 6.2.2a-c). This is despite the relatively short-range atomic cut-offs (3 Å, O only) used. Only in the O–O pair RDF there is a slight deviation from the DFTB ground truth, precisely where the potential is zero outside the 3 Å cut-off radius of the SOAP descriptor. Interestingly, for a DFT-quality GAP simply re-evaluating energies and forces on DFTB-derived active-learnt configurations is insufficient, with the DFTB configurations being high in energy at the DFT level ($\sim$ 5 eV, Figure 6.3.12). However, applying an active learning strategy with a PBE reference method and a slightly larger 3.5 Å cut-off generates excellent agreement with the AIMD simulation from ref. [487], in only a few hours of total training time (Figure 6.2.2d-f). At this level of theory, the local structure of liquid water is predicted largely correctly, with two distinct peaks in the O–H RDF, corresponding to first and second solvation shells with the largest deviation from the ground truth again in the O–O pair around the descriptor cut-off. The real significance, of course, is in moving to more accurate ground-truth methods, for which a full MD would not be straightforward: indeed, using the same method, a hybrid DFT-quality water model can be generated within a few days, which would be inaccessible with other methods (the generation of the GAP model required $\sim$ 5 days on 20 CPU cores, Figure 6.2.3g-i). These results suggest that the training strategy (and hyperparameter selection) presented here is suitable independent of the reference method.

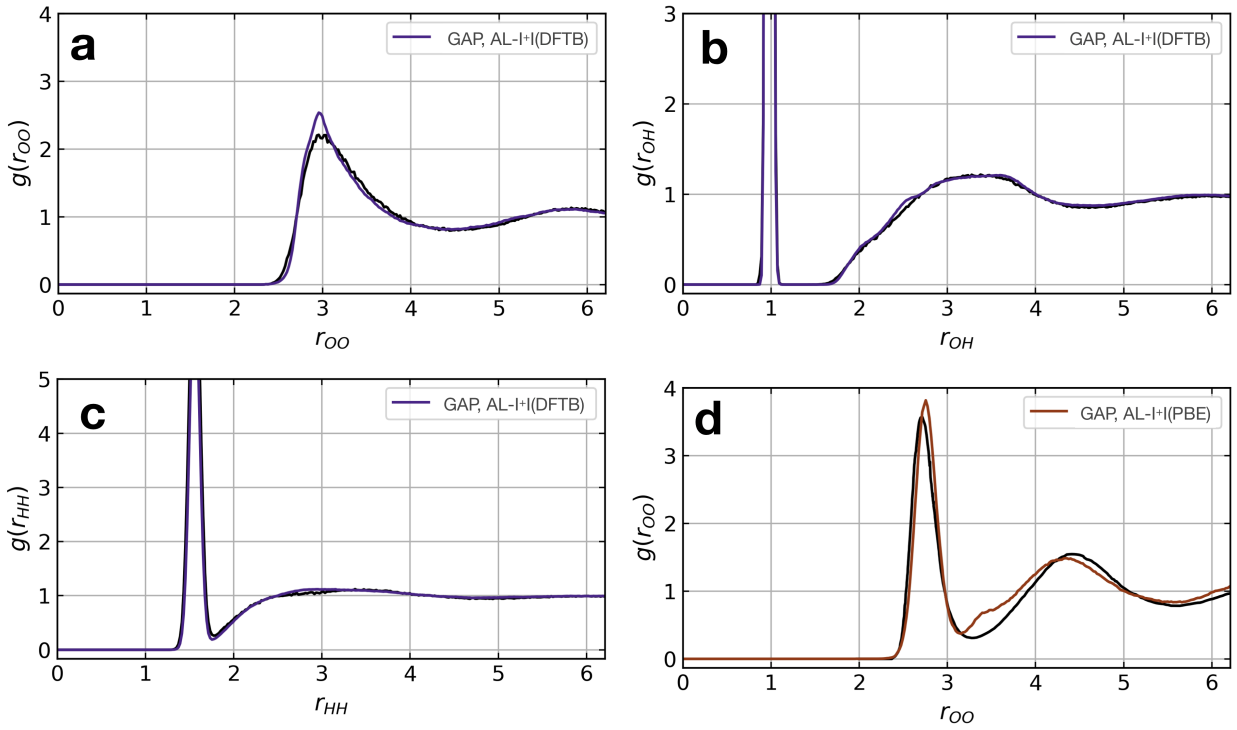


Figure 6.2.2. Liquid water simulations. Active learning of bulk water models at various levels of theory. Shown here are oxygen–oxygen, oxygen–hydrogen, and hydrogen–hydrogen radial distribution functions from NVT MD simulations of 64 water molecules in a 12.42 Å cubic box, with ground truth (black) and GAP (purple / red) simulations. (a–c) DFTB(3ob params) ground truth, 100 ps, 300 K, $r_c^{\mathrm{SOAP}}(O) = 3$ Å.

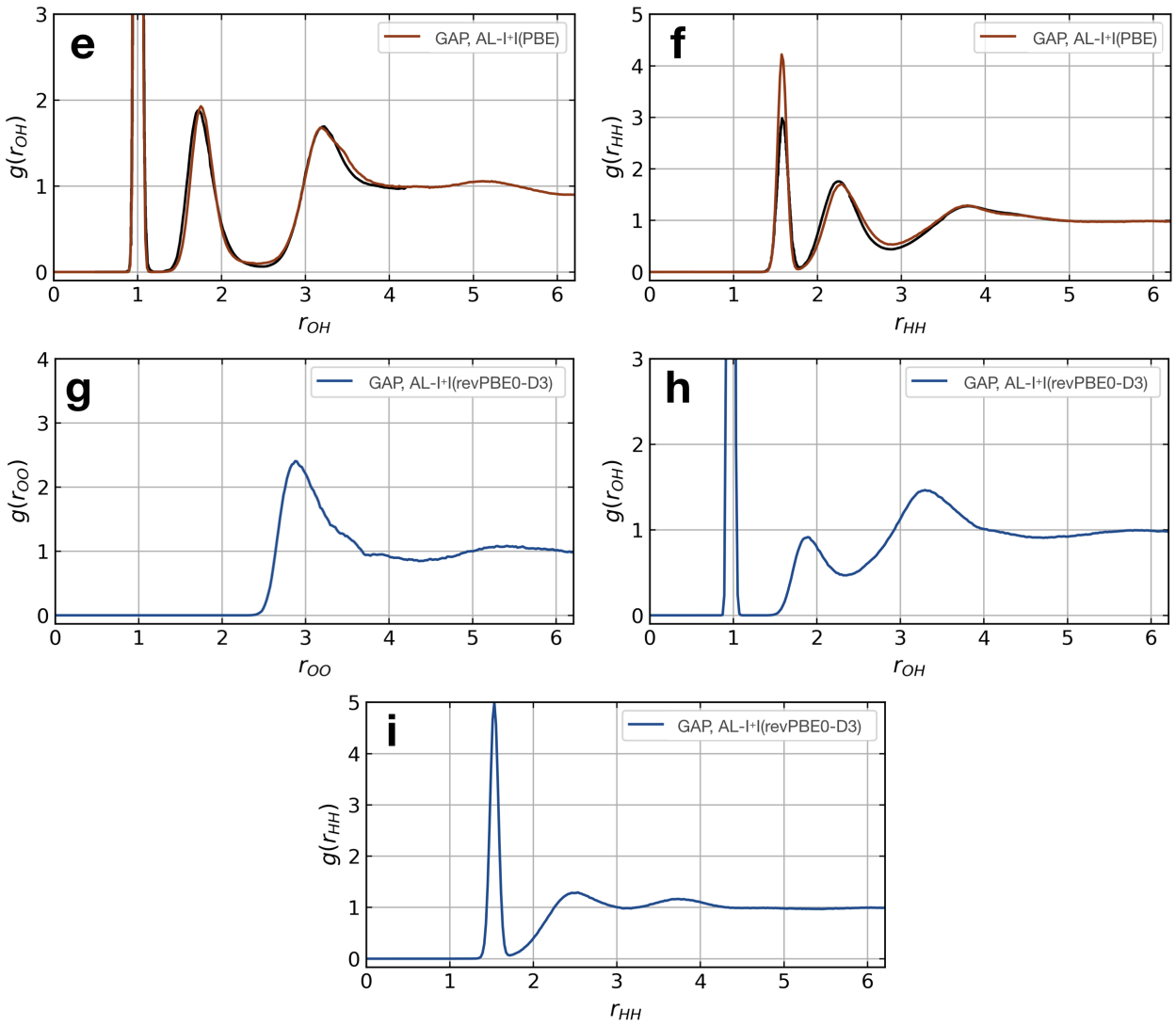


Figure 6.2.3. Liquid water simulations cont. (e–f) DFT(PBE) reference RDF data extracted from ref. [487], 30 ps, 330 K, $r_c^{\mathrm{SOAP}}(O) = 3.5$ Å. (g-i) DFT(revPBE0-D3) GAP, 30 ps, 330 K, $r_c^{\mathrm{SOAP}}(O) = 4.0$ Å.

6.2.3.3 Other Solvent Systems

Organic reactions often take place in solvents other than water. Using an identical training strategy to the one described for water, we trained GAPs for a selection of organic solvents with various types of intermolecular interactions. To quickly generate the reference simulation data for this proof-of-concept, a DFTB ground truth is employed; chlorinated solvents were not selected due to a large discrepancy between the DFTB-generated and experimental C–Cl bond dissociation energy (Figure 6.3.13). A uniform grid over the intramolecular PES is now no longer possible; thus, AL is used to develop an initial intramolecular potential trained using GAP-MD at 1600 K (Figure 6.3.14). This temperature is used to sample higher-energy configurations more efficiently. In all cases, only hundreds of ground truth evaluations were necessary to generate GAPs affording stable dynamics, with τ_{acc} values on the order of picoseconds (Figure 6.3.15, Table 6.2.1).

Solvent	SOAP on	N_{intra}	N_{inter}
Acetonitrile	C, N	269±12	120±60
Methanol	C, O	221±13	292±49
Acetone	C, O	566±80	359±29
Pyridine	C, N	249±36	243±11
Ammonia	N	38±40	109±24

Table 6.2.1. Average number (N) of total ground truth evaluations (over 5 repeats quoted with a standard error in the mean) required to obtain a potential with $\tau_{\mathrm{acc}} > 3$ ps, where $E_T = 1$ eV, $E_l = 0.1$ eV, 300 K. All SOAP descriptors used 3.0 Å cut–offs; they are centred on the stated atomic species, and include all atoms within the neighbourhood of those atoms (including hydrogen).

For a representative example, the computed RDFs for acetonitrile compare well with the ground truth (Figure 6.3.16). As with the water models above, to quantitatively evaluate bulk properties training an accurate reference method and the inclusion of nuclear quantum effects would be necessary.[457, 459] Nevertheless, this example demonstrates that the

training method is applicable to a range of chemical systems beyond water.

6.2.3.4 Aqueous Zn(II)

Modelling metal ions in solution remains one of the main challenges for general-purpose force fields.[488] Historically, metal centres have been developed by fitting van der Waals parameters to reproduce RDFs and hydration free energies of aquo complexes, which are expected to be transferable to more chemically complex environments. However, while simple, these models have often led to unstable simulations or poorly describe structural properties.[488] Considering these challenges and their relevance in biomolecular modelling, we decided to use our strategy to generate a GAP for aqueous Zn(II) ion as a representative system. Here the system was decomposed into a $[Zn(H_2O)_6]^{2+}$ cluster and the remaining water molecules. A strategy identical to the one described for water was used, training the intermolecular interactions separately with a 4.0 Å intermolecular cut-off for the oxygen atoms. Using this potential, MD simulations were propagated at 300 K reproducing the experimental[489] coordination number (CN=6), and Zn–O distances of both the first (2.08 Å) and second hydration shells without further optimisation (Figure 6.2.4).

The accuracy of the local structure compared to experiment in the first and second solvation shell indicates that this partitioning is effective at capturing both strong dative M–OH_2 interactions and weaker hydrogen bonding effects. From random points in the configuration space of $[Zn(H_2O)_6]^{2+}$ and 20 water molecules (intermolecular distances > 1.7 Å, 10 Å cubic box), τ_{acc} reached 0.5 ps ($E_l = 0.8$ kcal mol^{-1} per H_2O, 20 fs interval). Note this value is far short of the 100 ps simulations performed to generate the RDF and illustrates that a potential may be 'stable' and not sample any high energy regions for $t \gg \tau_{\text{acc}}$. Here, the potential for the Zn-water cluster was trained on almost 1000 configurations, suggesting that

tens of atoms per component may be the upper limit in dimensionality for which a model can be trained within a day.

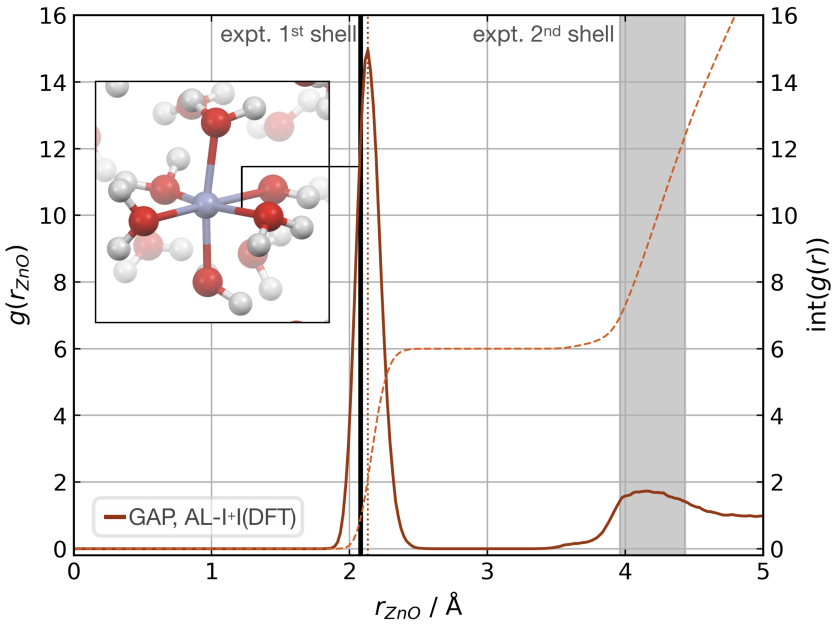


Figure 6.2.4. Zn(aq) simulation. Zn–O radial distribution function averaged from 1 ns of cumulative (10 × 100 ps) NVT MD simulations of Zn(II) in aqueous solution at 300 K, with the experimental modal Zn–O distance shown in black. Experimental (X-ray diffraction) Zn–O distances from ref. [489], octahedral first hydration shell. Shaded area denotes the range of experimental second hydration shell (ref. [489] and cited within). GAP trained as those in Table 1 using a PBE/400 eV ground truth, intra-$Zn(H_2O)_6$ used an O SOAP $r_c = 3.0$ Å and inter $r_c = 4.0$ Å.

6.2.3.5 Metallocage Dynamics

With a method capable of generating high-quality potentials for modestly sized chemical systems, we next demonstrate the applicability of the strategy to investigate a supramolecular metallocage consisting of >100 atoms including metal ions. As a representative example, we selected the $[Pd_2L_4]^{4+}$ metallocage architecture (L = organic pyridine-based ligand), which occupies a prominent place in supramolecular chemistry. Previously, we have studied this system due to its catalytic proficiency in Diels-Alder reactions employing both classical and

DFT modelling.[280] The different flexibility of two similar cage architectures was found to be key in explaining their contrasting catalytic activity. Taking advantage of the symmetry in the system, a representative fragment containing one full ligand and three pyridine molecules coordinated to a Pd(II) metal ion (68 atoms) was used to fit a GAP for the entire cage (138 atoms) in the gas phase. This potential was trained in a few days (∼1400 CPUh). We used the resulting GAP to perform nanosecond MD simulations on the whole metallocage at 300 K in the gas phase. This simulation took one day and ∼100 CPUh to complete. For comparison, an equivalent AIMD simulation would take around 50 years with the reference level of theory employed here. The flexibility of the system was monitored and compared to the one obtained using classical MD simulations in dichloromethane solvent.[280] Compared to classical MD simulations, using helicity as a measure of flexibility, our potential describes the cage as being more rigid; this suggests that the classical potential overestimates the dynamic flexibility (Figure 6.2.5). This difference is expected as the classical potential has no C-C≡C-C dihedral barrier, which is presumably correctly captured in the GAP. This example illustrates the general applicability of the approach to increasingly complex systems, where the training of a simpler but representative fragment is sufficient to capture the relevant features of the full system.

6.2.3.6 Reaction Dynamics in Gas and Solvent Phase

The high dimensionality and ensuing flexibility of ML potentials make them highly suitable to study reaction dynamics – the latter usually require many costly electronic structure calculations to obtain atomic-level descriptions of reaction mechanisms, solvent effects or post-transition state (TS) dynamics.[490, 491] In the following section, we show that our data-efficient strategy enables accurate reactive potentials ($\tau_{\mathrm{acc}} > 100$ fs) with only a few hundreds of DFT evaluations for a set of prototypical organic reactions.

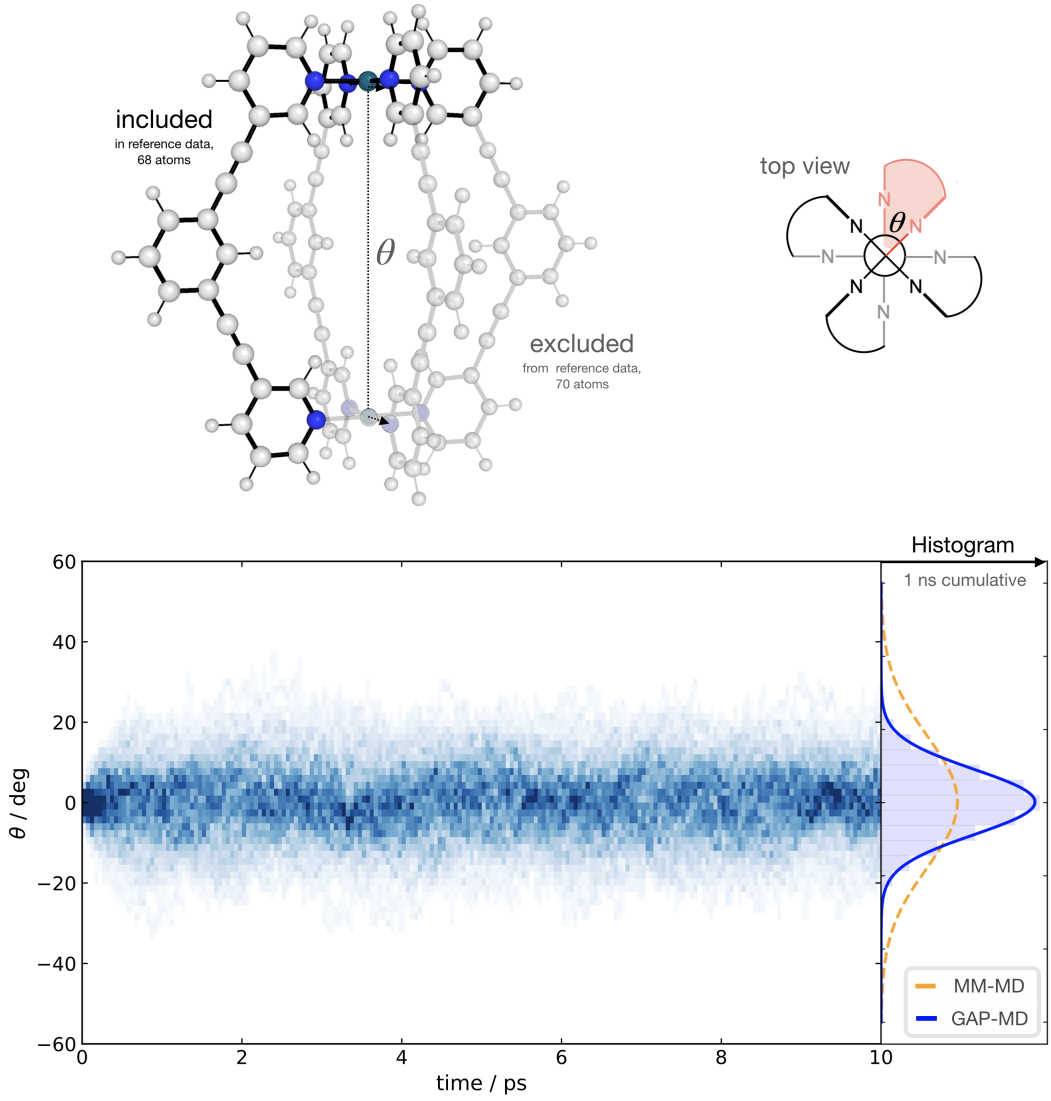


Figure 6.2.5. Temporal twist angle θ for an $[Pd_2L_4]^{4+}$ metallocage (top) obtained from 100 independent GAP-MD trajectories (each run for 10 ps at 300K), GAP trained on a 68-atom $[PdL(py)_3]^{2+}$ system (py = pyridine), representative of the whole metallocage as shown, at 500 K with a 0.2 eV error threshold for the active-learning protocol. Time-dependent histogram generated from 50 fs time chunks over the whole 10 ps time period. PBE0-D3BJ/def2-SV(P) ground truth surface.

Gas phase bimolecular nucleophilic substitution. The S_N2 nucleophilic substitution reaction is fundamental in organic chemistry and has been extensively studied using AIMD and analytically fit PES.[490, 492] However, even with efficient approaches to fitting PES, AIMD methods still require tens of thousands of energy evaluations.[493] Here, we generated

a reactive GAP to study the reaction between chloride and methyl chloride as a prototypical case often employed to validate QM/MM reactions and for which extensive literature exists (see ref. [494] and references therein). By initialising active learning from the transition state (TS), the true intrinsic reaction coordinate is reproduced to within 1 kcal mol^{-1} (Figure 6.3.17).

Interestingly, and unlike our previous attempt to generate a DFT-quality GAP by evaluating energies and forces on DFTB active-learnt configurations, here, uplifting a DFT-level GAP to an accurate wavefunction-level GAP is possible. This method allows coupled cluster-quality energy profile (Figure 6.2.6a) and dynamics to be propagated from the TS with just 55 energy and (numerical) force evaluations at the CCSD(T) level (Figure 6.2.6b/6.3.17). The resultant GAP is considerably more accurate than the underlying DFT energy profile (dashed, Figure 6.2.6a). Active learning can also be initialised from an association complex and the IRC learned without prior knowledge of the TS. For the exothermic reaction between cyanide and methyl chloride, training from reactants and initialising velocities such that 11 kcal mol^{-1} (0.5 eV) was present in the breaking bond, the reaction is sampled in the training (Figure 6.3.17c). Relaxing a nudged elastic band (NEB) using the trained GAP over an interpolated path between reactants and products affords an IRC within chemical accuracy of the true profile (RMSE = 0.9 kcal mol^{-1}, Figure 6.2.6c). In this case, adding a NEB refinement step to the training is essential to adequately sample the product region and reach chemical accuracy in the reaction energy (orange vs. blue lines Figure 6.2.6c). Here, uplifting the GAP to CCSD(T) affords chemical accuracy in the minima and TS regions, with a more limited accuracy in the region in between (Figure 6.3.17d); this is likely due to the large differences between the PBE and CCSD(T) surface in that region. Despite this, the uplifted profile is considerably more accurate than the underlying DFT (Figure 6.3.17d).

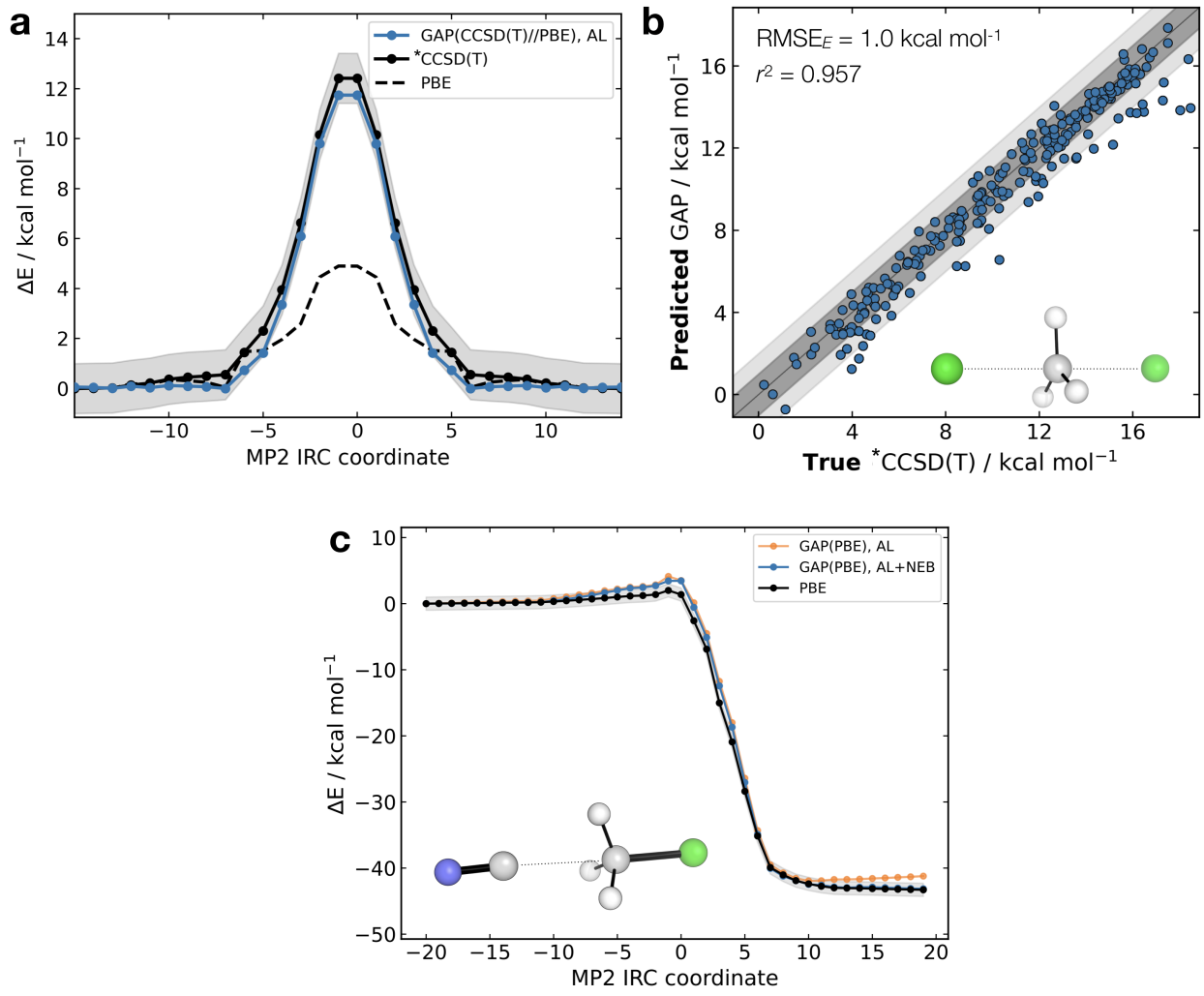


Figure 6.2.6. Gas-phase S_N2 reactive dynamics. Energetics of a mode $Cl^- + CH_3Cl \rightarrow Cl^- + CH_3Cl$ l S_N2 reaction in the gas phase. (a) Predictions (GAP) and ground-truth (*CCSD(T) ≡ DLPNO-CCSD(T)/ma-def2-TZVPP) energy values on the MP2/ma-def2-TZVPP intrinsic reaction coordinate (IRC). Shaded region bounds the 'chemically accurate' (±1 kcal mol^{-1}) region. (b) Parity plots of between GAP predictions and true energies from ten 100 fs GAP-MD trajectories initialised from the TS (300 K). Dark and light grey area bound the ±1 kcal mol^{-1} and ± 2 kcal mol^{-1} error regions, respectively. (c) IRC for $CN^- + CH_3Cl \rightarrow Cl^- + CH_3CN$ but trained using uphill active learning, with nudged elastic band refinement; see Figure 6.3.17 for additional details.

Post-TS bifurcating pathway in a Diels-Alder reaction. GAPs for more complex reactions involving reactions that proceed on a bifurcating PES can also be trained. These reactions typically require AIMD simulations, where selectivity is determined from the av-

erage behaviour of many trajectories leading to either product. Other approaches have also been developed.[495] We explored the dimerisation between cyclopentadiene, for which endo selectivity has been rationalised on the basis of bifurcating reaction pathways.[496]

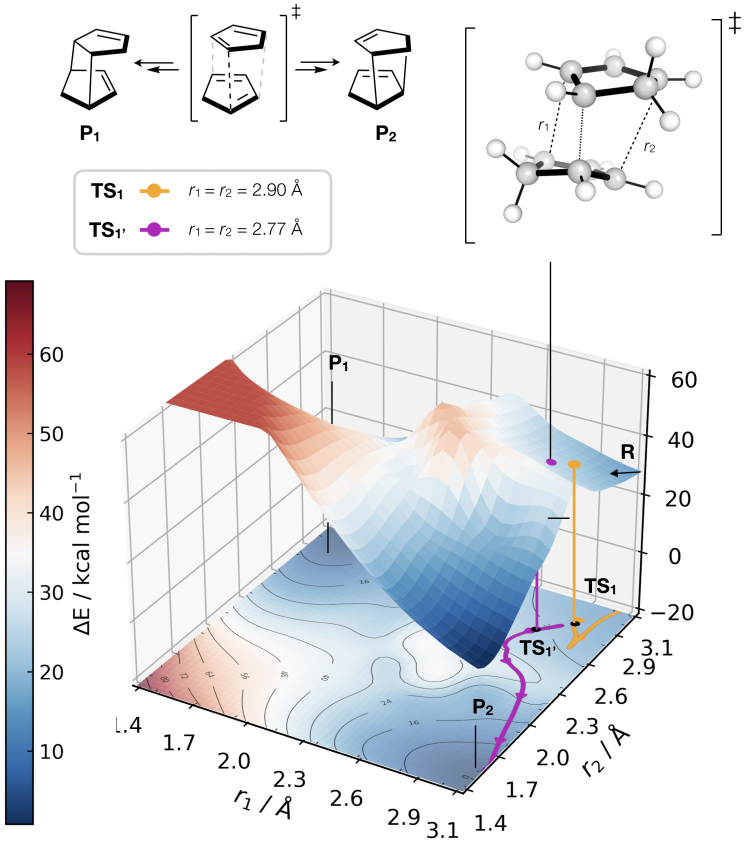


Figure 6.2.7. GAP dynamics on a bifurcating surface. 2D PES (B3LYP/def2-SVP) along the forming bond distances (r_1, r_2) in the dimerisation of cyclopentadiene. An example of GAP-propagated reactive dynamics (300 K) is shown from TS_1 (7_N in ref. [496]), which leads to reactants, (representative trajectory in orange), and from TS_1' which leads to products (a representative trajectory is shown in purple). 3D projection is truncated at 2.5 eV above the minimum for plotting. Interpolated surface used a cubic spline using *scipy.interp2d* with the raw surfaces shown in Figure 6.3.21. All trajectories shown in Figure 6.3.19 and 6.3.20.

Once again, initiating active learning from the literature TS (TS_1, Figure 6.2.7) and using a DFT method analogous to the one used in the original work by Caramella and co-workers we obtain a reactive potential from which 500 fs trajectories were propagated. Interestingly, we found that propagating the system from this TS did not afford any products (P_1 or P_2,

Figure 6.2.7), with all trajectories leading to the reactant state (Figure 6.3.19).

Further investigation and generation of the relaxed 2D potential energy surface over the two possible forming C–C bonds (r_1, r_2) leading to products provided a rather different surface to the one suggested in ref. [496], with a flat portion then an incline as r_1, r_2 shorten below 2.9 Å, with a steeply exergonic reverse reaction (intrinsic reaction coordinate, IRC, see full Supporting Information). As noted by Caramella, following the IRC forwards from TS_1 the reaction proceeds to another TS_1' which is similar in energy ($\Delta E = 2$ kcal mol^{-1}). By training a GAP at 500 K and propagating GAP-MD from TS_1' and sampling the area of the PES around a valley-ridge inflection point (VRI), trajectories lead to the expected two products (e.g., purple line, Figure 6.2.7 and Figure 6.3.20). This example demonstrates that with no *a priori* knowledge, apart from the structure of TS_1', the topology of the bifurcating surface can be revealed efficiently using GAP dynamics. This strategy is completely automated, requiring training of a few hours to days, thus providing a promising approach to routinely examine reaction dynamics in organic molecules.

Solution phase bimolecular nucleophilic substitution. The ability to accurately describe bond-breaking/forming paths in the condensed-phase is crucial if this strategy is to be applied to increasingly complex processes, such as enzymatic reactions. Towards this goal, and having generated potentials for condensed-phase molecular systems and gas-phase reactions, we decided to extend our active learning strategy to explicitly solvated reactions; once again, using the S_N2 reaction between chloride and methyl chloride as a test case. S_N2 reactions have been used as a test case for a recent ML potential, but those studies have been limited to the gas phase and used thousands of training points.[447] Literature examples of ML potentials to study reactions in explicit solvent are limited. Recently, Parrinello and co-workers reported an NN potential to study urea decomposition in water.[464] Training

a model for the implicitly solvated reaction proceeds in a similar way as for the gas phase analogue and affords a surface close to the ground truth (Figure 6.2.8a).

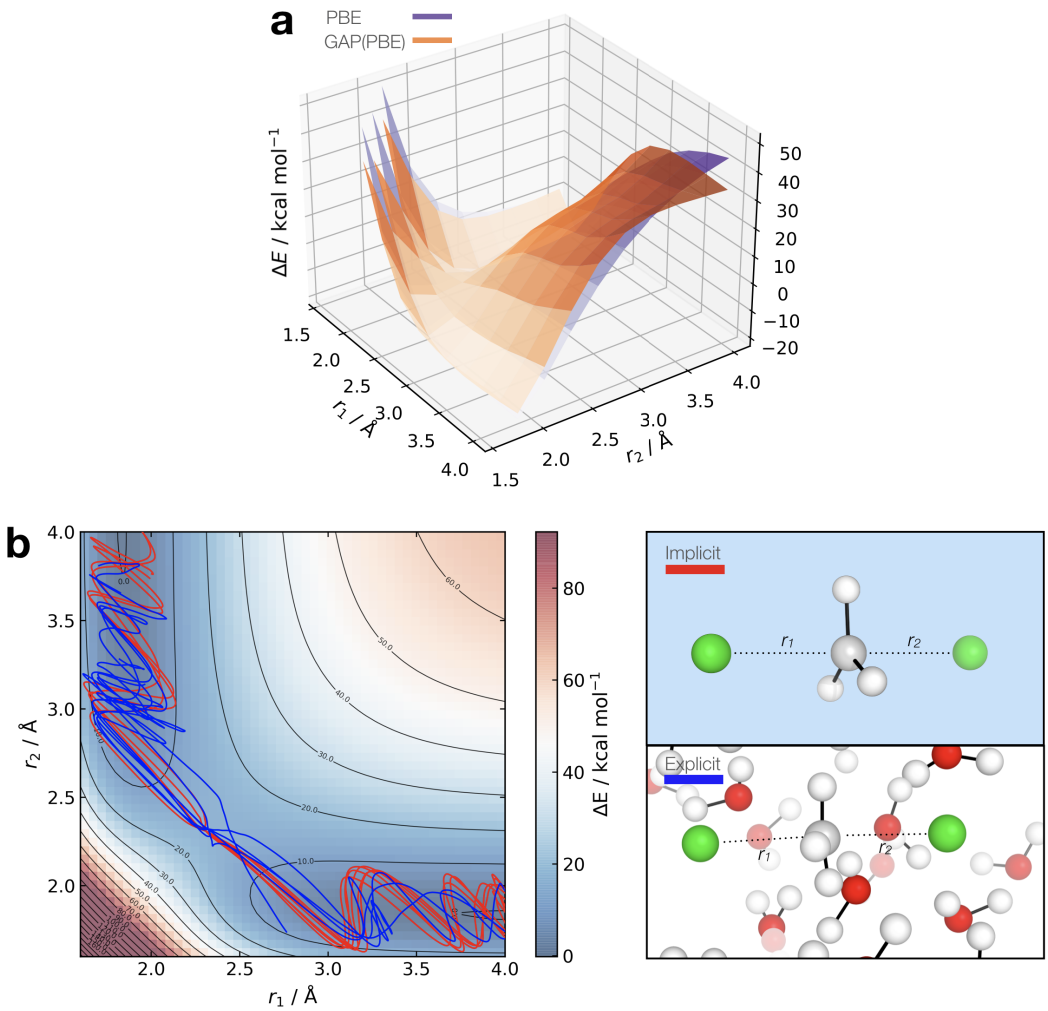


Figure 6.2.8. Solution-phase S_N2 reactive dynamics. (a) True (CPCM(Water)-PBE/def2-SVP, purple) and GAP predicted (orange) relaxed 2D PESs in implicit solvent, zeroed to the transition state energy. (b) Reactive GAP MD trajectories (lines) propagated from the TS, trained on implicitly solvated configurations (CPCM(Water)-PBE/def2-SVP, red) and explicitly solvated configurations (PBE/400 eV, blue). Ground truth implicitly solvated surface (cubic interpolated, 5:1) underneath, with the error to the GAP prediction in Figure 6.3.23. Intramolecular GAP used C-only (r_c^{SOAP} = 6 Å) descriptor with the intermolecular explicit solvent using a SOAP on O (r_c^{SOAP} = 3.5 Å) and Cl (r_c^{SOAP} = 4.5 Å, Figure 6.3.24).

Adopting an identical strategy to the one employed for condensed phase systems, the intra- and inter-molecular PES dynamics can be propagated from the TS and the effect of explicit

solvation interrogated (Figure 6.2.8b). This only requires knowing *a priori* the gas-phase TS for the training to be complete in explicit water. Interestingly, the behaviour in explicit water (blue, Figure 6.2.8b) differs from the implicit counterpart (red, Figure 6.2.8b). This can be understood considering that in implicit solvent reorganisation is instantaneous, which results in oscillations in the C-Cl bond characteristic of a gas phase reaction. In contrast, the dynamics are more complex in explicit solvent, with a slower transition from the product channel. Additionally, one of the 10 trajectories re-crosses the barrier after 170 fs of simulation (Figure 6.3.22), where the solvent has not reorganised to accommodate the anionic chloride yet, making the path to products shallower in energy.

The component-wise separation of the system also leads to the possibility of training to a more accurate *ab initio* surface for the gas phase reaction, in a similar way to QM/MM, but here a ML(A)/ML(B) partition is available where A and B are two different ground truth methods. Application of this kind of hierarchical ML potential fitting will be the subject of further work.

Limitations. In its current implementation, this method is well suited for studying systems of up to 50 atoms if training is expected to be completed in a day, using a single node of 20 CPUs. For larger systems, the speed will depend on the complexity of the PES, where even using inexpensive methods thousands of training configurations and several iterations may be needed to learn the different atomic environments. In contrast, training can be achieved efficiently for large systems with higher symmetry, and effectively lower dimensional PES. This is the case, for example, for the supramolecular cage shown above, which contains 138 atoms and can be modelled in full with GAP only using 68 atoms for training (Figure 6.2.5). Furthermore, the current intra+intermolecular decomposition remains fixed throughout the simulation making the water potential generated initially incapable of auto-ionisation. Re-

determining the connectivity (and therefore re-assigning the 'intra' components) every few steps during the simulation might help to address this limitation. However, provided the model has been trained in a region where a chemical change has occurred, the molecular components in that region do not need to retain their connectivity. This is the case, for example, in the S_N2 reaction shown in Figure 6.2.8, where connectivity changes between the two molecular units that constitute the solute.

6.2.4 Conclusion

Studying dynamic processes and the effect of explicit solvation on chemical reactions demands a rapid method to develop bespoke force-field models with high accuracy. Here, we demonstrated that within the Gaussian Approximation Potential (GAP) machine learning framework, accurate and robust models can be developed efficiently for gas-phase and condensed-phase molecular reactions. Our strategy starts from a small number of randomly selected points in the configuration space, from which active learning training of intra- and inter-molecular components of the energy and forces is carried out. The developed method is publicly available (`https://github.com/duartegroup/gap-train`). We also define a prospective error metric, which is found to be crucial in developing robust active-learning-based potentials, whereas correlation on a predefined test set is insufficient to assess the quality of such a potential. We illustrate the generality of this approach by modelling bulk water, Zn(II) in aqueous solution, and chemical reactions in the gas phase and explicit solvent, including post-TS cyclisation and S_N2 reactions. The diversity of the examples presented here demonstrates the general applicability of the strategy and encourages applying this approach in the modelling of more complex reactions in homogeneous and heterogeneous environments.

6.2.5 Methods

All Gaussian Approximation Potentials (GAPs) were trained using the GAP and QUIP codes (Singularity distribution, commit # 66c553f) and a Smooth Overlap of Atomic Positions (SOAP)[161] kernel with hyperparameters as defined in Table 6.3.1. For a single component condensed phase system such as water, two GAPs are fitted for the intra and intermolecular components, respectively, while for the solute-solvent systems, such as the S_N2 reaction between chloride and methyl chloride, three GAPs were fitted: one for the gas-phase solute, a second for gas-phase solvent and a third one for the intermolecular interactions; see Table 6.3.2, entry 6 for details. An example of the input script required to train a bulk water model is shown in Figure 6.2.9. In all systems the intramolecular water potential was trained at the reference level on an evenly spaced grid (512 points, $r_{OH} \in [0.8, 1.5]$ Å, $r_{HH} \in [1.0, 2.5]$ Å). Other intramolecular GAPs employed SOAP descriptors with cut-offs shown outlined in the figure captions. Intermolecular GAPs were trained by subtracting intramolecular energies and forces of all the defined components from the reference total energy. All pure condensed phase systems were trained and simulated at or close to their experimental liquid densities using 10 solvent molecules (e.g. 10 H_2O in a 343 $Å^3$ cubic box). Aqueous Zn and explicitly solvated S_N2 reactions used 20 water molecules in a cubic box with side length of 10 Å. All files required to reproduce the examples shown here can be found in `https://github.com/duartegroup/gap-train/tree/master/examples/paper_examples`.

GAP-MD simulations were performed with ASE[497] interfaced to QUIP with the quippy wrapper using the Langevin integrator with 0.5 fs timesteps at 300 K unless otherwise specified. Condensed phase MD simulations were performed with three-dimensional periodic boundary conditions following minimisation and equilibration for at least 20 ps. Initial configurations, CUR[483] selection and all learning curves were generated with the *gap-train*

module, which was used to run the automated fitting.[498] All active learning was performed using a 'diff' strategy, where a configuration is added to the training set if $|E^0 – E^{\mathrm{GAP}}|$ is larger than a threshold. With system-dependent hyperparameter optimisation, using a threshold on the maximum atomic variance predicted by the Gaussian process ('gp_var') can result in accelerated learning (Figure 6.3.21). CUR selection used SOAPs averaged over atoms in a configuration using the Dscribe[499] package ('inner' averaging, over entries of the expansion coefficient vector). Intra+Inter (I+I) energy and force evaluations used an expansion factor of 10 to ensure no intermolecular atoms were within the intra GAP cut-off. The NumPy[291] implementation introduces a negligible computational overhead for expanding the box (~ 0.1 ms step^{-1} real time) but requires two GAP calculations on the inter and intra components, currently carried out in serial. All generated potentials, with the exception of the revPBE0-D3 water potential and metallocage, were trained in less than a day on 10 CPU cores. The revPBE0-D3 water potential was constructed without any prior data in 5 days (1 intra + 4 inter) and used 20 CPU cores, while the metallocage fragment was trained for 3 days also on 20 CPU cores. Explicit S_N2 reaction dynamics was performed using intra components for H_2O and $[Cl\cdots CH_3Cl]^-$, where the latter, due to the finite atomic cut-off employed, has the correct dissociation behaviour when Cl^- and CH_3Cl are distant.

Periodic DFTB calculations performed with DFTB+[361] using 3ob[500] parameters, and molecular equivalents using GFN2-XTB[297] in XTB v. 6.2.3. Periodic pure DFT calculations were performed with GPAW[501, 502] v. 19.8.1 with the PBE[82] functional and a 400 eV plane-wave cut-off from a dzp LCAO initial guess at the gamma point. Hybrid periodic DFT calculations with the revPBE0[86, 503] functional combined with the D3[94] dispersion correction were performed with CP2K.[504]

Molecular DFT, MP2 and coupled cluster [DLPNO-CCSD(T)] calculations performed with

ORCA[246, 505] v. 4.2.1 wrapped with *autodE*[296] using PBE[82] and PBE0[86] functionals, (ma)-def2-SVP, def2-TZVP, ma-def2-TZVPP basis sets[247] AIMD calculations at the DFTB level performed with DFTB+ with 3ob parameters[500] and MM simulations performed with GROMACS[244, 506] 2019.2 with TIP3P parameters.[257] Behler NN[158] trained on 7258 RPBE configurations from ref. [507], re-evaluated at DFTB(3ob) with n2p2[508] using the default parameters and symmetry functions.

```python
import gaptrain as gt
gt.GTConfig.n_cores = 4

h2o = gt.System(box_size=[10, 10, 10])
h2o.add_solvent('h2o', n=20)

data, gap = gt.active.train_ii(h2o, method_name='dftb')
```

Figure 6.2.9. Example Python input script required to train a bulk water model from scratch at the DFTB level using four CPU cores.

6.3 Selected Supporting Information IV

Full Supporting Information can be found at: `https://doi.org/10.26434/chemrxiv.13856123`

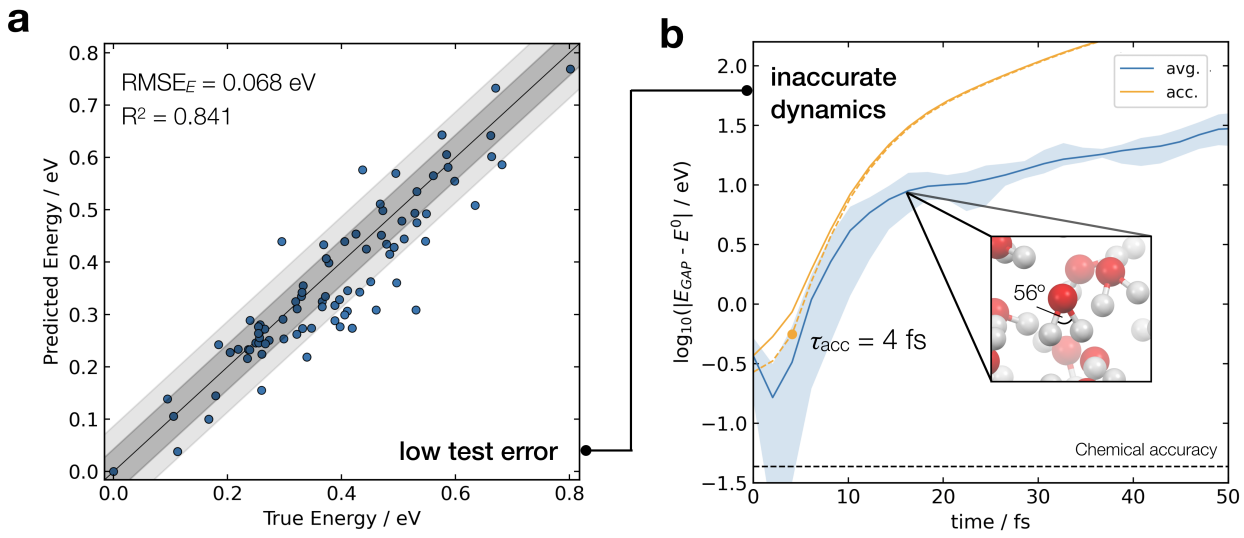


Figure 6.3.1. (a) A GAP model for molecular water (fitted using the P1 parameter set, Table 6.3.1) as characterized by its accuracy on external validation data not encountered in training; dark and light shaded area bound the 'chemically accurate' ± 1 kcal mol^{-1} and ± 2 kcal mol^{-1} regions, respectively. (b) the resulting difference between predicted and true single point energies on MD frames at 300 K ($\delta t = 0.5$ fs). Dynamics with 30 water molecules in a cubic box (l = 10 Å, $\rho \sim 1$ g cm^{-3}). Error range (min–max, shaded) and average of five simulations using the same trained GAP from different initial randomly placed then minimized points. DFTB(3ob) ground truth. Orange lines depict the total cumulative error (solid) and cumulative error above 0.1 eV (dashed).

Set	Type	Parameter	Value
P1	GAP	σ_E	0.316 meV
		σ_F	0.1 eV Å^{-1}
		ζ	4
		r_c	5.5 Å
		n_{sparse}	30
	2b descriptors:	δ_{2b}	1.0 eV
	O–H, H–H, O–O	θ	1.0 Å
		sparse method	Uniform
	SOAP descriptor: O	r_c	3.0 Å
		n_{sparse}	100
		δ_{SOAP}	0.1 eV
		$\sigma_{\text{at}}^{\text{SOAP}}$	0.5 Å
		$n_{\text{max}}, l_{\text{max}}$	6
		sparse method	CUR points
P2	GAP	σ_E	0.316 meV
		σ_F	0.1 eV Å^{-1}
		ζ	4
		r_c	various
		n_{sparse}	500
	SOAP descriptors	$\sigma_{\text{at}}^{\text{SOAP}}$	0.5 Å
		$n_{\text{max}}, l_{\text{max}}$	6
		sparse method	CUR points

Table 6.3.1. Parameter sets for GAPs, SOAPs and 2/3b descriptors.

Strategy	Notes	τ_{acc} / ps	N
MM-MD	Classical molecular mechanics MD simulations were performed at 100, 300, 500 and 1000 K using GROMACS v. 2019.2 with a timestep of 1 fs and TIP3P water. Following a 1 ns NVT equilibration of a random configuration of water 10 ns of NVT dynamics were performed taking 1000 evenly spaced frames from the simulation.	0	1000
rand.	Configurations were generated by adding GFN2–XTB optimised water molecules into the box in a random position and orientation, ensuring that there are no intermolecular distances < 1.5 Å. Random displacements are/were added to each Cartesian coordinate sampled from a random normal distribution with $\sigma = 0.05$ Å, which samples over intramolecular bond stretches and bends.	0	1000
rand. min.	As rand. where each random configuration is minimized to $\lvert F_i \rvert < 1$ eV Å^{-1}, where F_i is the force on atom i. Subsequently, random normal displacements were added to each atom to ensure some sampling of the intramolecular modes.	0.0008 ± 0.0007	7490 ±20
AIMD	*Ab-initio* MD simulations were performed at the ground truth level (DFTB, 3ob parameters) for 1 ps at 300 K with a 0.5 fs timestep. Frames are randomly selected from the trajectory with at least a 2 fs interval.	0.011 ± 0.003	2570 ± 10

Table 6.3.2. Outline of training strategies used to train a GAP for bulk water (Figure 6.2.1). N total ground truth evaluations used to train the final potential. Errors are quoted as standard errors in the mean from 5 independent samples where appropriate to one significant figure. All training used 10 water molecules in a cubic box with side length 7 Å.

Strategy	Notes	τ_{acc} / ps	N
AL	Active learning is initiated from a GAP trained on 10 random configurations with a 1.5 Å minimum distance between water molecules. MD simulations at 300 K was then propagated using this potential for $n^3 + 2$ fs, where n is the number of iterations of the MD trajectory. The error between the ground truth (E0) and predicted energy (E^{GAP}) is evaluated for the final frame ($\lvert E^0 – E^{\text{GAP}} \rvert$) and, if above 0.1 eV the configuration is added to the training data. If above 10× the threshold then the error is backtracked in intervals of 2 fs until a suitable configuration is obtained, as to not add any very high energy configurations. If the error is 100× the threshold, then the first frame of the trajectory is returned, as the backtracking is likely to be too slow. If the error is below the threshold, then a further MD trajectory is propagated, and n incremented by one. If n > 10 then no configuration is added from this set of trajectories.	0.07 ± 0.02	3200 ± 200
AL-I+I	Intra- and intermolecular interactions are trained independently, and separate GAPs trained for each term. The intramolecular GAP is trained on a cubic grid of configurations (see Methods) with 2- and 3-body descriptors with 3 Å cut-offs and 30 sparse points. The intermolecular GAP is trained using the active learning method outlined above. Energies and forces are calculated as a sum of terms and the intramolecular GAP prediction evaluated in a box expanded by a factor of 10 while maintaining the fractional coordinate of the centre of mass of each water molecule fixed, as to ensure no intermolecular hydrogens are in the radius of the intramolecular descriptors.	31 ± 7	1160 ± 90

Table 6.3.3. Table 6.3.2 continued.

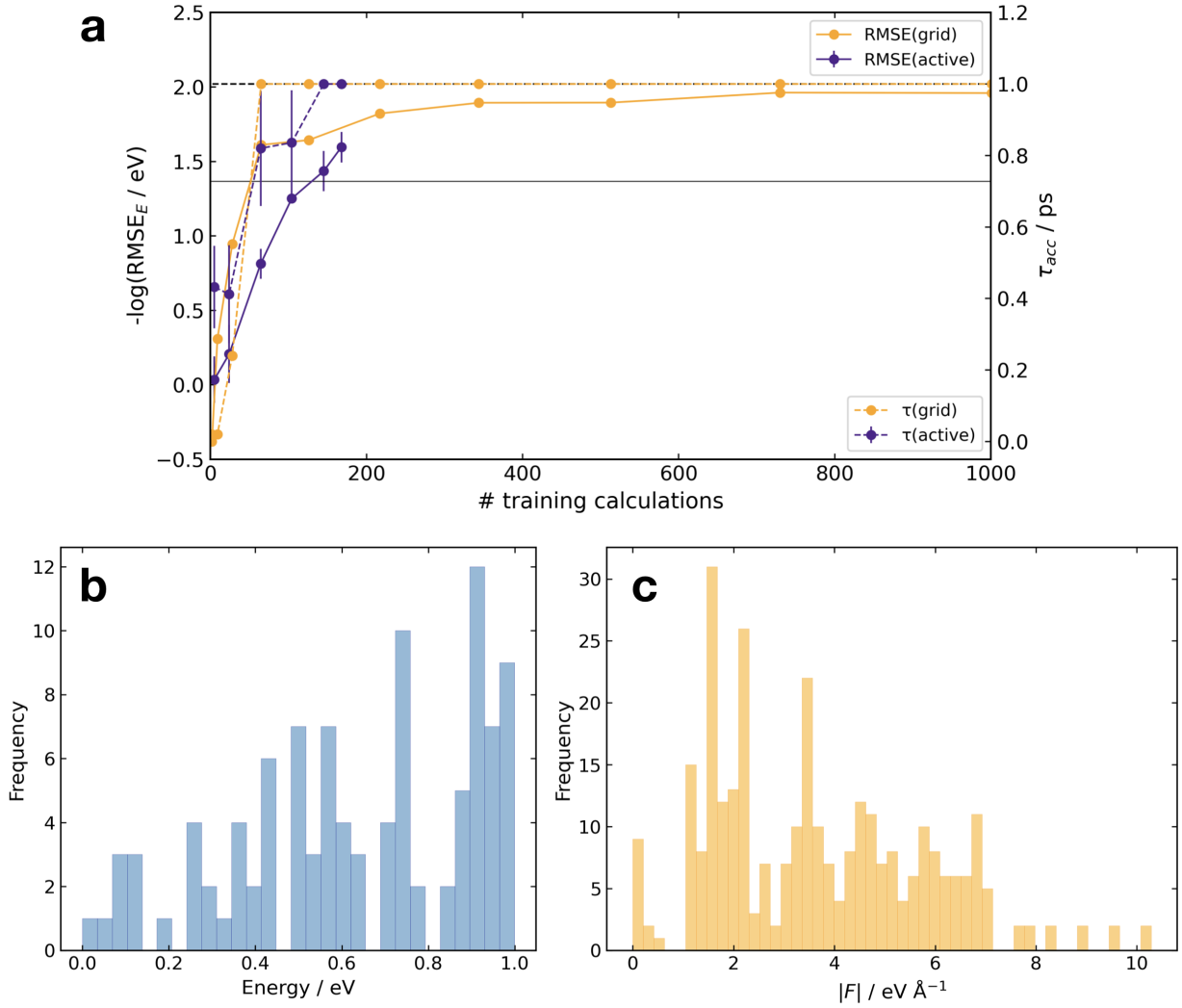


Figure 6.3.2. (a) Comparison of active learning and a grid-based approach for training a water monomer. 2b+3b GAP with $r_c = 3.0$ Å without a SOAP all other parameters as P2 (Table 6.3.2). $\max(\tau_{\text{acc}}) = 1$ ps calculated in a 2500 K simulation for a ∼1% probability of accessing a configuration 1 eV above the minimum, $E_l = 0.043$ eV, $E_T = 0.43$ eV. (b, c) Energy and force distribution of the test data used to calculate a root mean squared error (RMSE) generated on a grid over $r_{\text{OH}_a}, r_{\text{OH}_b} \in [1.0, 1.3]$ Å and $r_{\text{HH}} \in [1.0, 2.5]$ Å and truncated above 1 eV of the minimum to 103 data points that are not coincident with any training data.

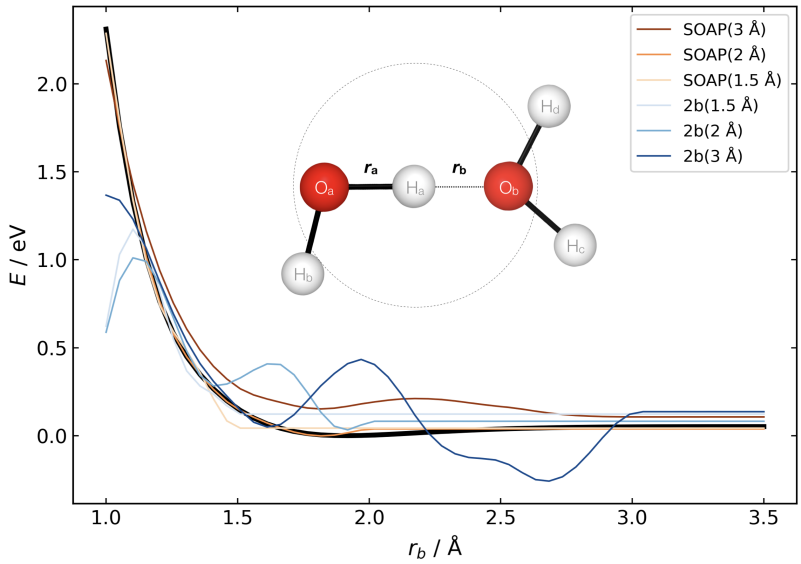


Figure 6.3.3. Water dimer PES predicted using SOAP and 2b GAPs with the ground truth (DFTB(3ob)) in black. Trained on then evaluated on the PES points. Intramolecular component subtracted using the same intra-GAP as Figure 6.2.1 (2b+3b, r_c=3.0 Å) evaluated in separate boxes.

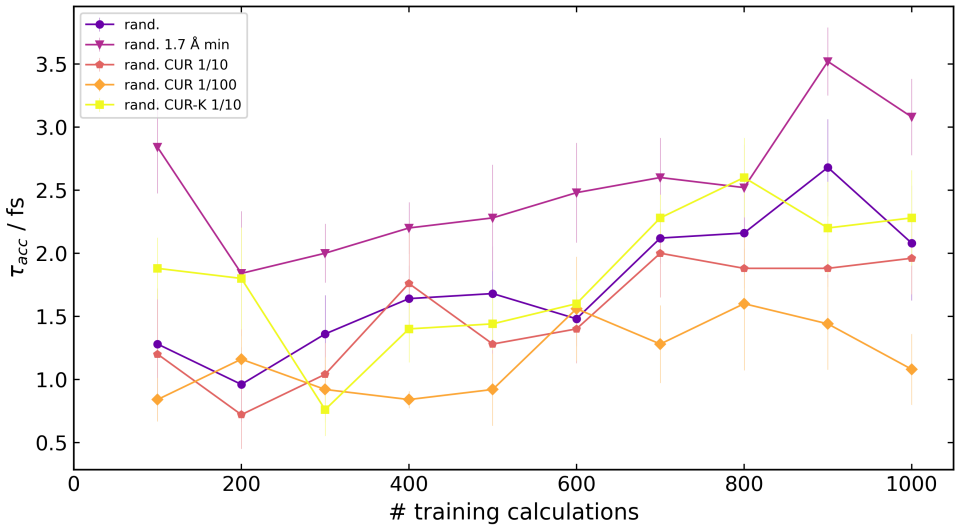


Figure 6.3.4. Learning curves for a bulk water GAP trained on random configurations, with or without selection strategies. Water molecules randomized in the box by applying a random rotation and translation to each water molecule ensuring no intermolecular distance < 1.5 Å, apart from rand. 1.7 Å min (purple triangles) where the minimum distance is 1.7 Å. τ_{acc} calculated with a 1 fs interval, $E_l = 0.1$ eV, $E_T = 1$ eV averaged over 5 initial random configurations. Error bars are standard error in the mean over 5 independent iterations. CUR-K 1/10 is a CUR selection of the square kernel matrix between SOAP descriptors calculated using Dscribe,[499] keeping 1 in 10 rows, CUR is selection on the SOAP matrix averaged over atoms in the box.

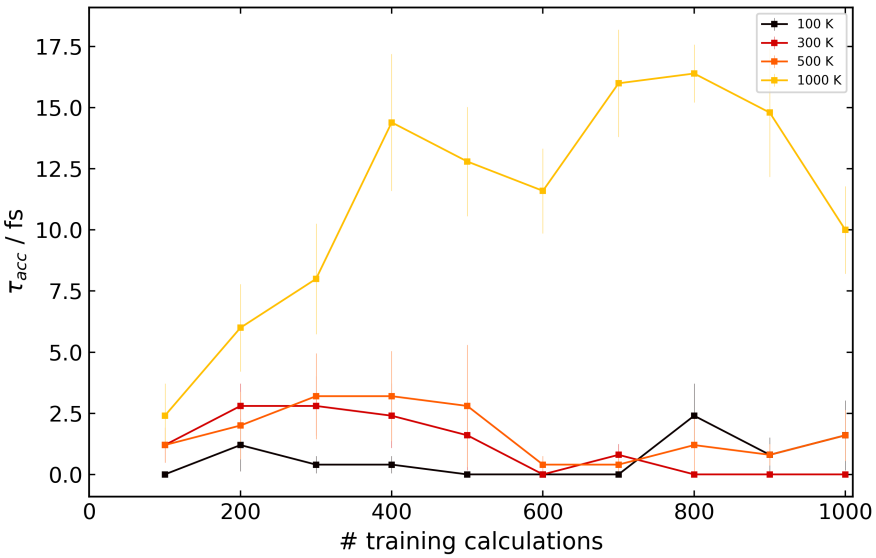


Figure 6.3.5. Learning curves for a bulk water GAP trained on classical molecular mechanics configurations at different temperatures (performed by Tristan Johnston-Wood). Initial random configuration minimized then equilibrated for 1 ns, TIP3P parameters, flexible water. Configurations taken evenly spaced from a total of 10 ns of simulation time. τ_{acc} calculated with a 10 fs interval, $E_l = 0.1$ eV, $E_T = 1$ eV averaged over 5 initial random configurations. Error bars are standard error in the mean over 5 independent iterations.

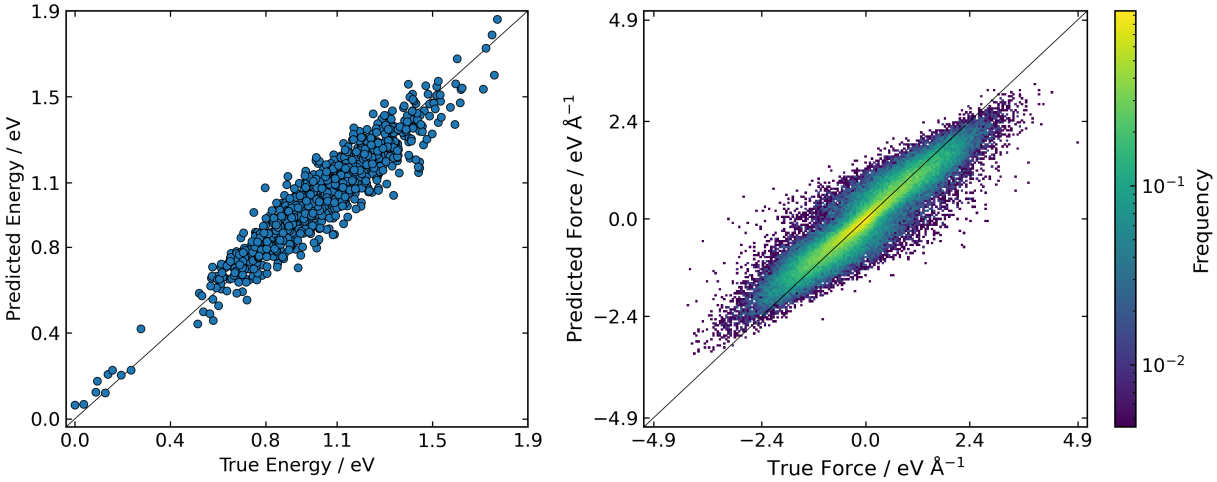


Figure 6.3.6. Correlation plot between predicted and 'true' (DFTB) energies and forces on MM training data (300 K) for bulk water (TIP3P parameters).

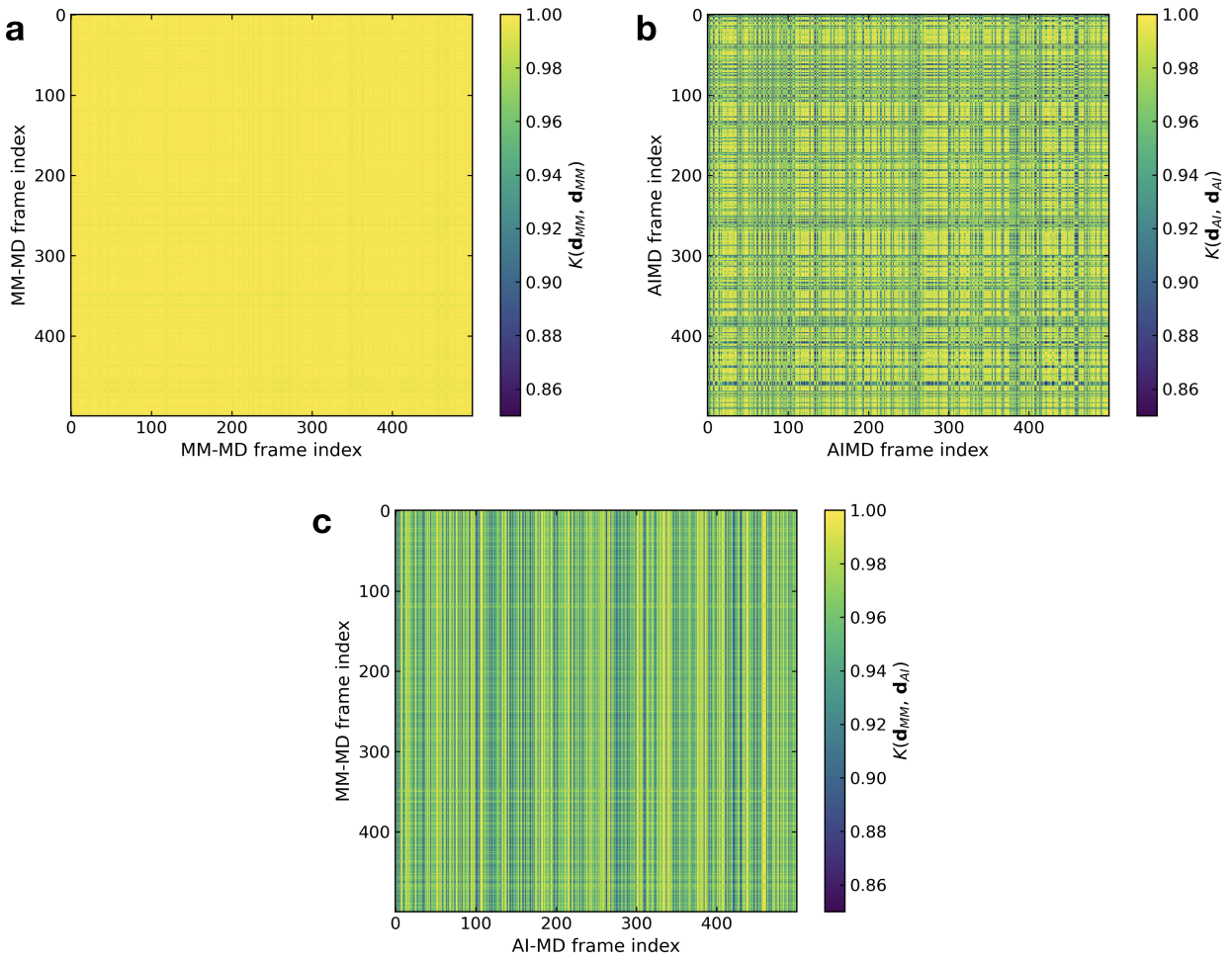


Figure 6.3.7. Kernel matrices on normalised SOAP vectors between MM-MD and AIMD frames simulated at 300 K calculated using Dscribe,[499] with 'inner' averaging (average coefficients over sites before summation over angular projection, m) over unique elements (H, O), $r_c = 5$ Å, $n_{\max} = l_{\max} = 6$, raised to the 4th power (i.e. $\zeta = 4$). Values 0–1 correspond to minimal–maximal similarity in configurations.

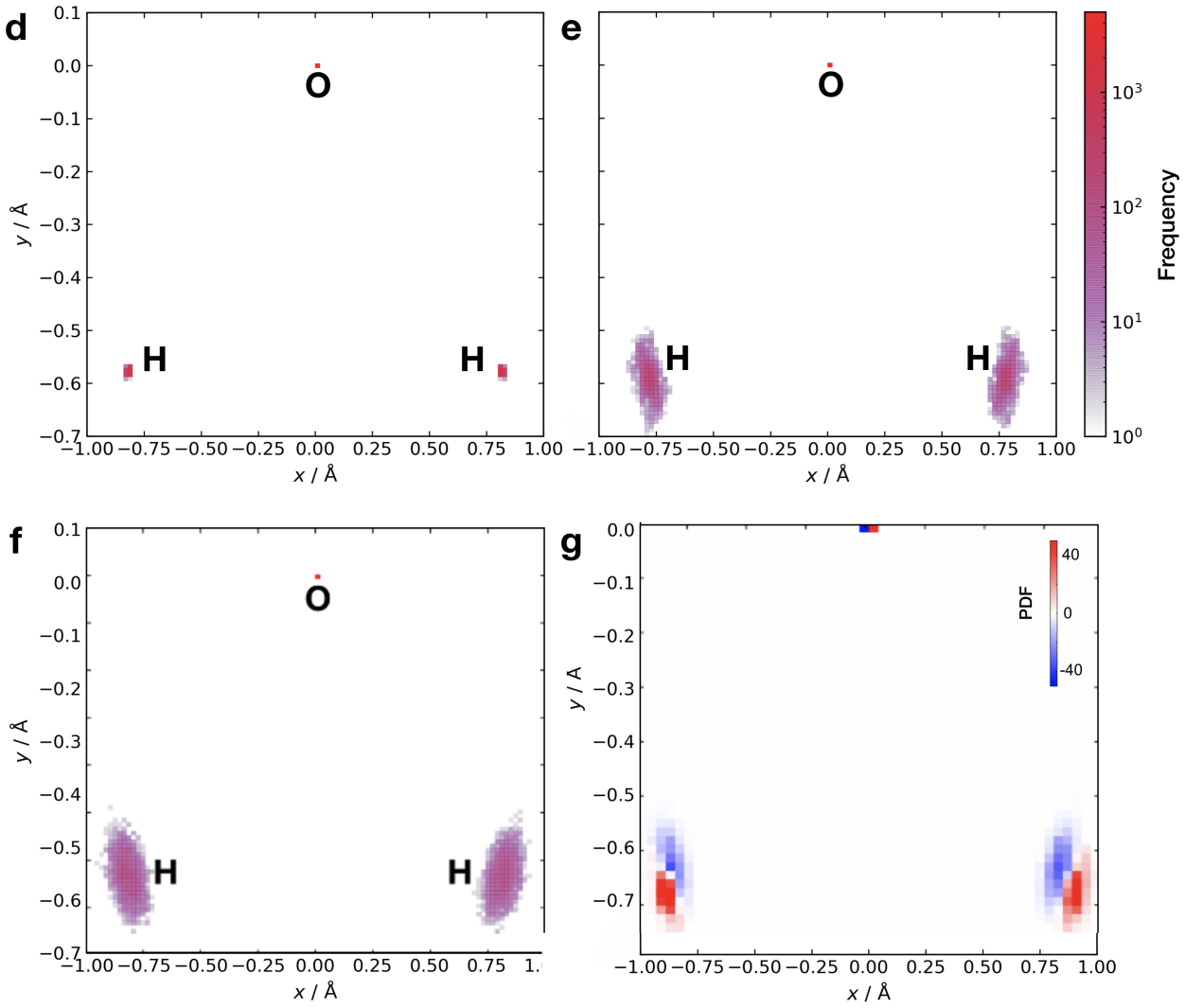


Figure 6.3.8. Figure 6.3.7 continued. Histogram of atomic positions in (d) MM(TIP3P) with LINCS H-constraints and without (f), (e) AIMD(DFTB) configurations generated over a 300 K trajectory and the difference between MM(ITIP3P, no LINCS) and AIMD (g). All water molecules fitted to an initial reference in the xy plane with the Kabash algorithm, oxygen centred at (0, 0). Only molecules wholly within the box included.

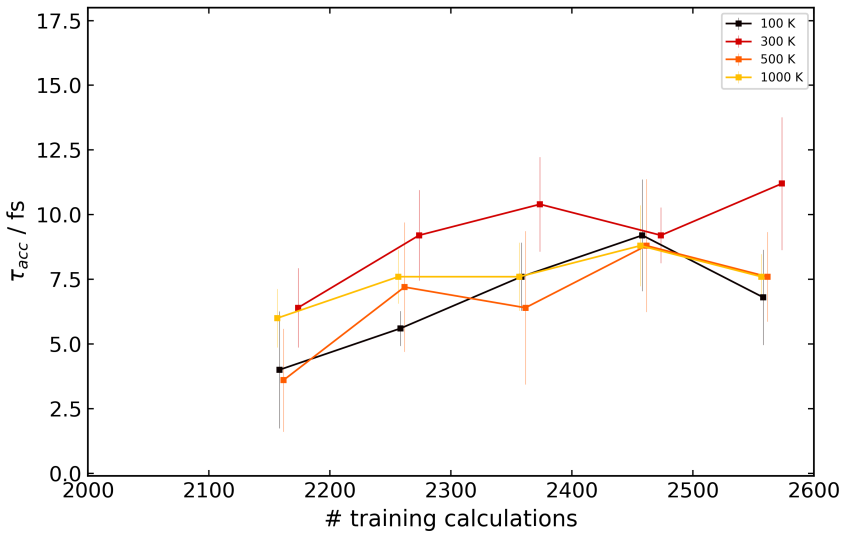


Figure 6.3.9. Learning curves for a bulk water GAP trained on classical 'ab-initio' [AIMD, DFTB(3ob)] configurations at different temperatures. Initial random configuration minimized at DFTB. Configurations taken evenly spaced from a total of 1 ps of simulation time. τ_{acc} calculated with a 10 fs interval, $E_l = 0.1$ eV, $E_T = 1$ eV averaged over 5 initial random configurations. Error bars are standard error in the mean over 5 independent iterations.

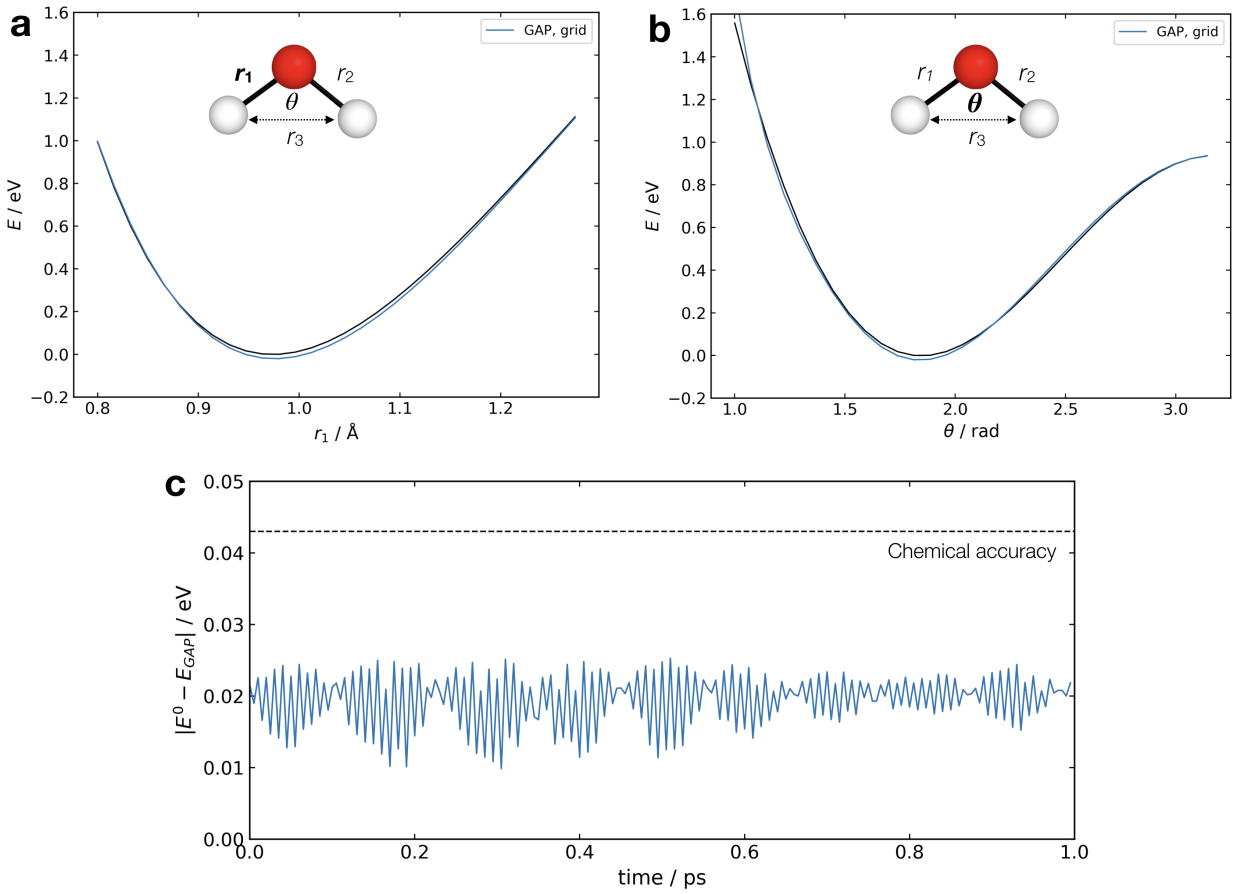


Figure 6.3.10. Unrelaxed potential energy surfaces for the O-H stretch (a) and HOH bend (b) and (c) error in gas phase monomer dynamics generated by GAPs trained on 65 data points in a grid over $r_1, r_2 \in [0.8, 1.5]$ Å, $r_3 \in [1.0, 2.5]$ Å including the minimum at the DFTB(3ob parameters) level of theory. Two-body and three body GAP trained with 3.0 Å cut-offs.

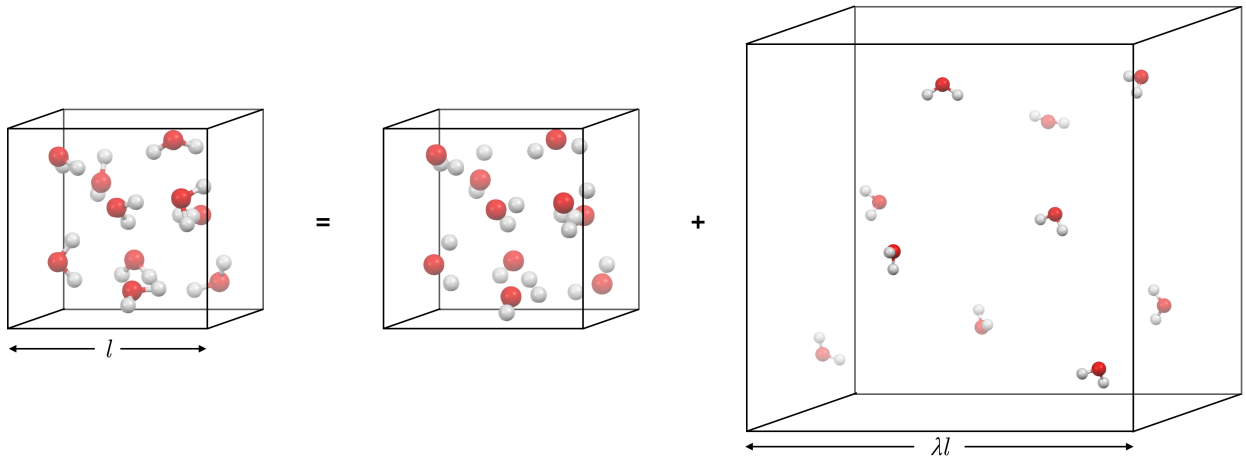


Figure 6.3.11. Schematic representation of an intra+inter (II) energy evaluation using two GAPs. Total energy is a sum of the intermolecular interactions in a box, plus the intramolecular energy calculated by expanding the box by a factor λ, keeping the fractional centre of masses of each molecule fixed.

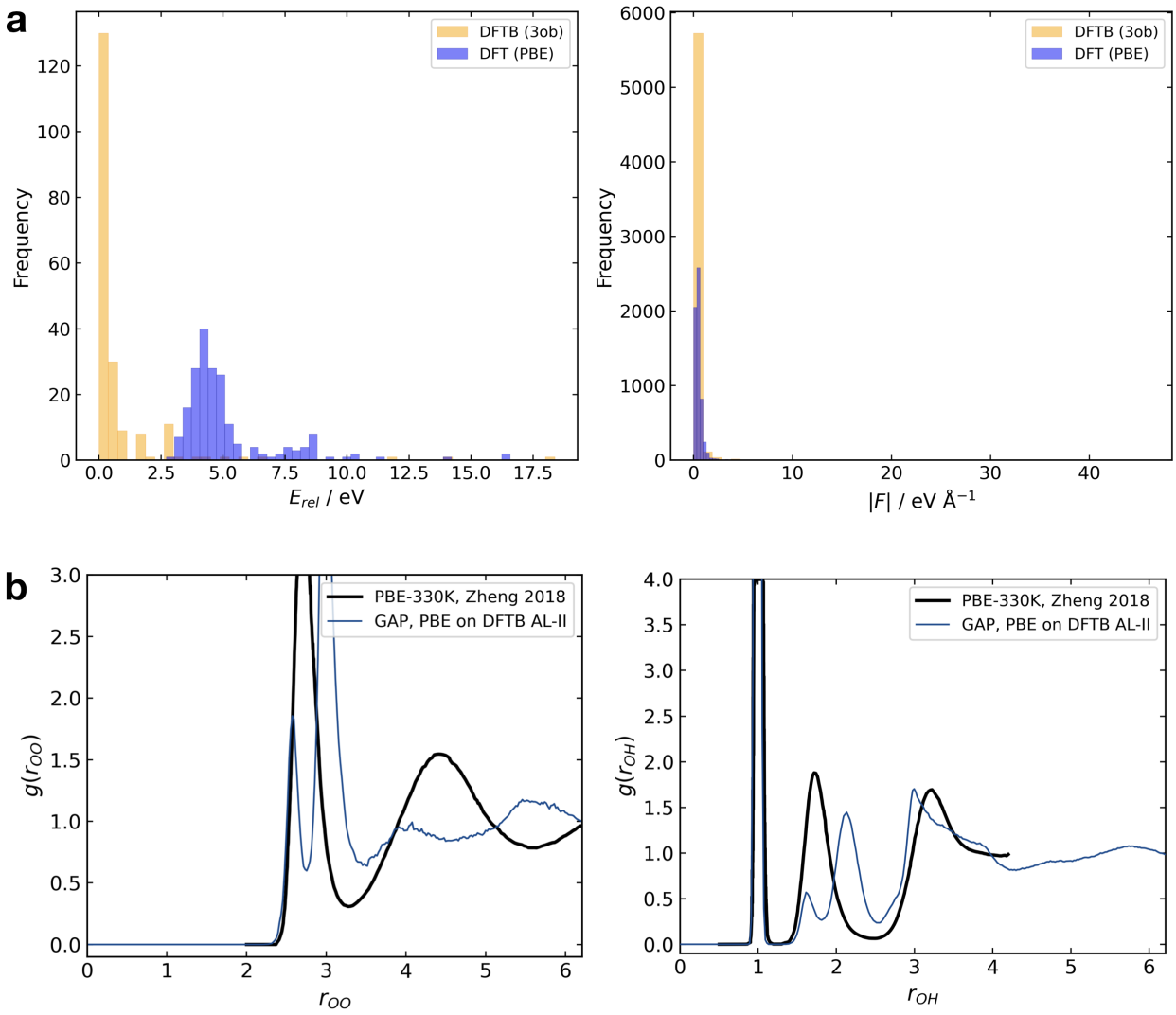


Figure 6.3.12. Attempted DFT (PBE/400 eV) uplift of DFTB GAP with single point energy and force evaluations on a set of configurations used to train a stable inter GAP (AL-I+I, Figure 6.2.1). Intramolecular energy evaluated using a DFT GAP trained using active learning at 1600 K with a 2 eV maximum energy threshold. (a) Energy and force histograms, where the DFT configurations are referenced to the lowest energy located in an active learning cycle at the DFT level. (b) Pair radial distribution functions calculated from a 10 ps, 300 K MD simulations with the inter (+intra) GAP trained on the DFTB configurations, compared to a PBE reference from ref. [487].

6.3.1 Other Solvents

To demonstrate the transferability of the intra+inter active learning training method to other solvents presented here are learning curves in τ_{acc} for a selection of small, commonly used, organic solvents. The accuracy of the potential is quantified by the final error metric following the full active learning, and as a quick validation the radial distribution functions are shown for all pairs in acetonitrile. Given the much higher dimensionality of the intramolecular PES in all of the solvents a dense grid over all coordinates is not possible (e.g. $N_{\text{atoms}} = 6$ in MeCN $= 8^{(3\times6)-6} \sim 10^{11}$ points) and active learning needs to be employed for the intra surface. Employing active learning on the intramolecular modes of water requires both a high temperature, as to sample the curvature of the PES at regions often sampled at 300 K, and an energy threshold to prevent high energy configurations entering the fit (Figure 6.3.14). Active learning at 1600 K then readily dissociates Cl radicals with a DFTB ground truth method.

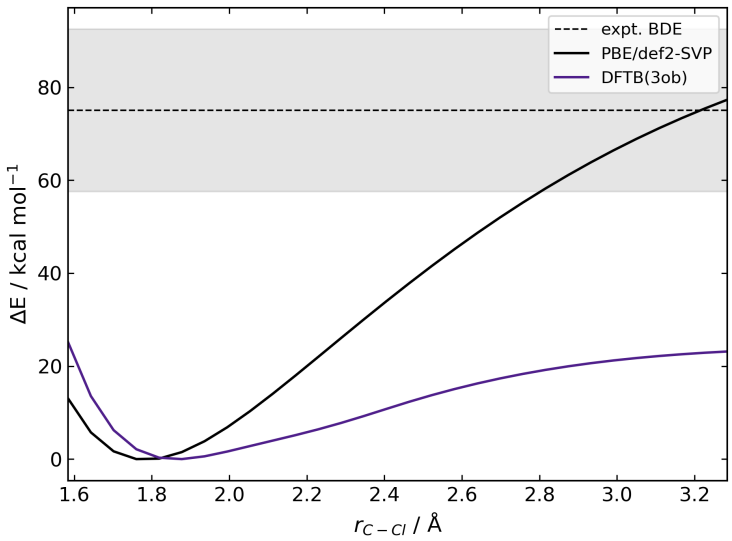


Figure 6.3.13. Comparison between DFT (PBE/def2-SVP) and DFTB energies for the C—Cl dissociation curve in CH_2Cl_2. Geometries from unrelaxed PES scan from the DFT geometry. Experimental data indicated with dash grey line from ref. [509].

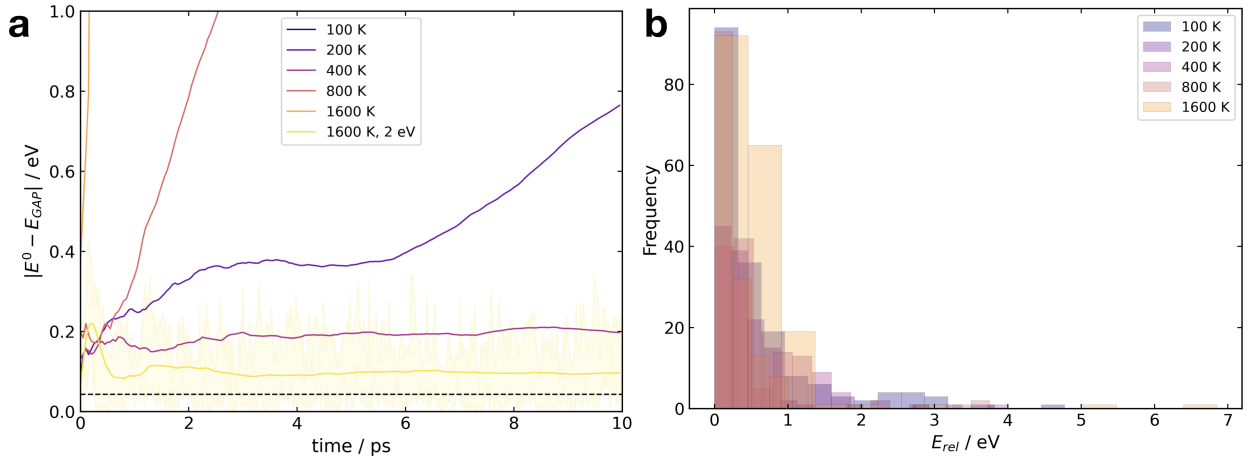


Figure 6.3.14. Absolute errors between GAP and the ground truth DFTB on GAP-MD trajectories propagated at 300 K of a water box. Intramolecular GAP(2b+3b) trained using active learning using intermediate GAP-MD at the quoted temperature. 1600 K, 2 eV also specifies an energy cut-off on GAP training data. Intermolecular GAP trained as Figure 6.2.1 (AL-I+I). Error ranges are generally not shown for clarity. Initial random configurations to start the active training loop are identical for each training run.

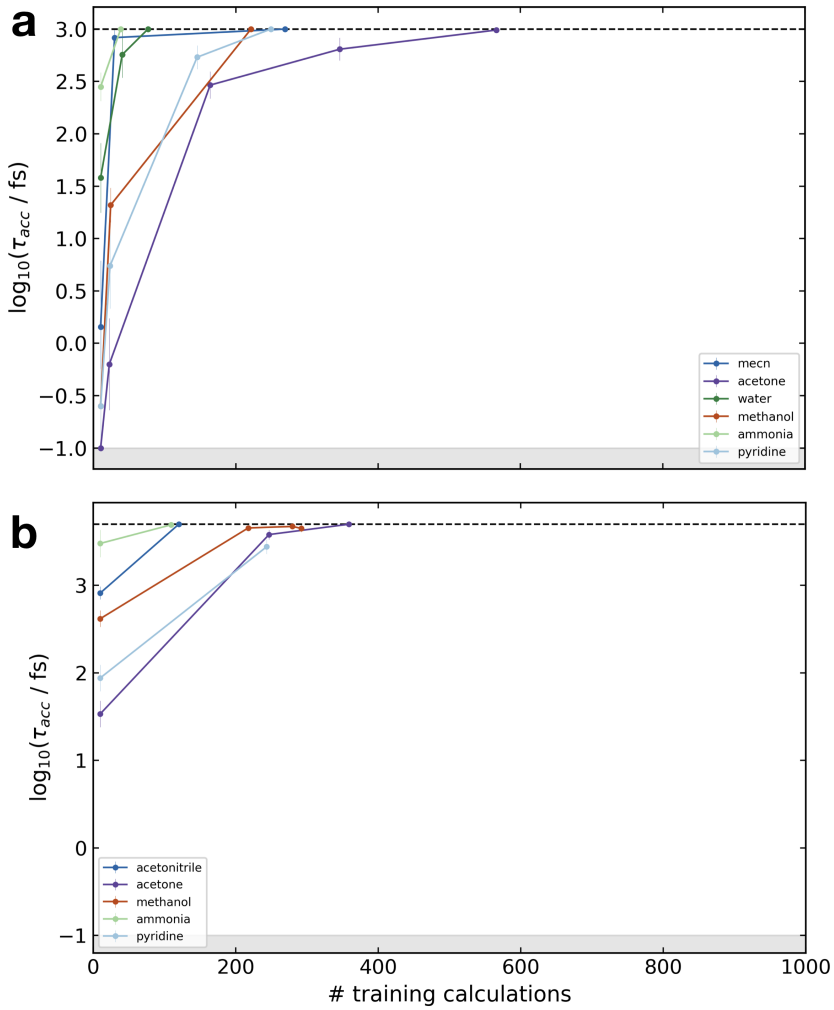


Figure 6.3.15. Learning curves for (a) intra- and (b) intermolecular components in different molecular solvents. Parameters for each system shown in full Supporting Information. 1600 K for all intramolecular active learning and 300 K for the intermolecular equivalent from 10 initial random normal displacements of all atoms ($\sigma = 0.05$ Å, $\mu = 0$ Å). Maximum τ_{acc} shown as dashed lines and $\min(\tau_{\text{acc}}) = 0.1$ fs for plotting. Error bars plotted as the standard error in the mean from 5 independent repeats. Active learning halted if no configurations have error above the threshold.

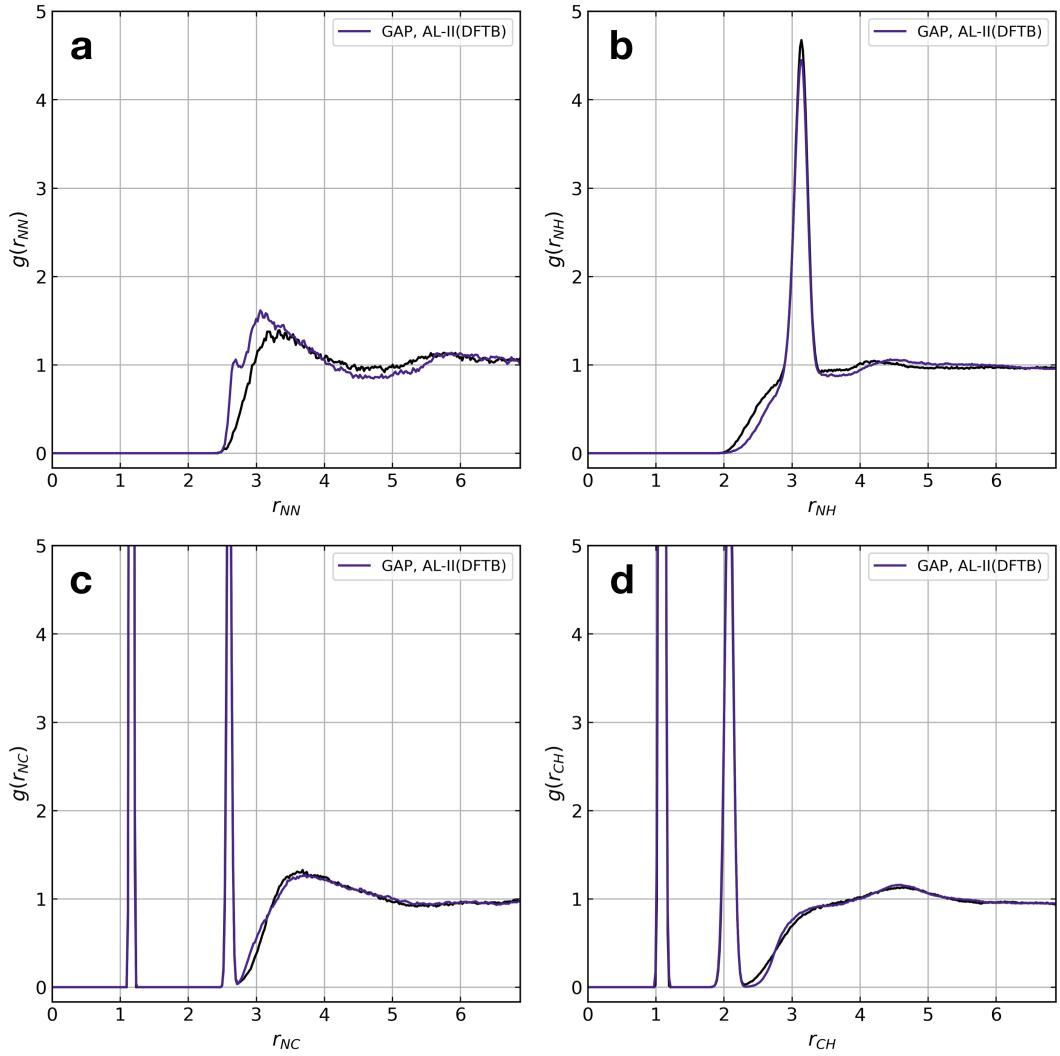


Figure 6.3.16. Radial distributions functions for acetonitrile generated using GAPs (purple) trained with active learning on inter and intramolecular degrees of freedom (as Figure 6.3.15) ground truth DFTB(3ob) level (black). Dynamics run for 30 ps at 300 K in a 13.8 Å length box with a time-step of 0.5 fs.

6.3.2 Reactions

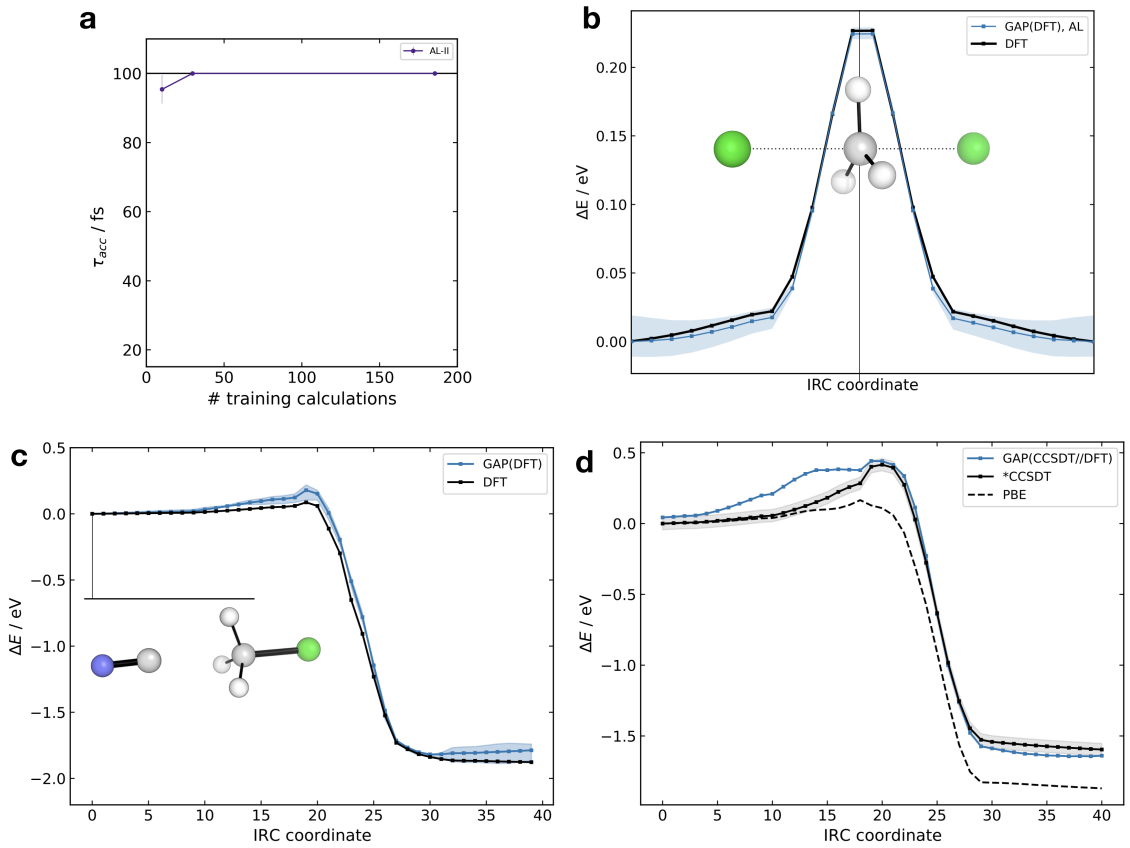


Figure 6.3.17. (a) Learning curve for TS originated dynamics in the gas phase for Cl^- + CH_3Cl → Cl^- + CH_3Cl. Number of ground truth evaluations does not include those used to find the initial transition state. SOAP descriptors with r_c = 3.5 Å on C and Cl. TS optimisation and energy/force evaluations performed with ORCA at the PBE/ma-def2-SVP level of theory. τ_{acc} calculated using a 2 fs time interval, 1 kcal mol^{-1} error threshold, 10 kcal mol^{-1} maximum total error to a maximum of τ_{acc} = 100 fs, as only short time dynamics are required from the TS. (b) IRC for Cl^- + CH_3Cl → Cl^- + CH_3Cl calculated in ORCA at the ground truth (PBE/ma-def2-SVP) and predicted with 5 trained GAPs. Average shown as the blue line and the range of predictions in blue. IRC configurations were not present in the training data. (c) As (b) for CN^- + CH_3Cl → Cl^- + CH_3CN but trained using uphill active learning i.e. without knowledge of the TS. 0.5 eV of energy was added to the breaking C–Cl bond and dynamics propagated for up to 500 fs at a temperature of 200 K. (d) Predicted IRC using the active-learnt GAP with configurations added from close to the minimum energy pathway with NEB relaxation (final geometry selected manually) using the GAP, energies and forces calculated, then re-predicted. Coupled cluster single point energies and numerical frequencies were then calculated on the 200 configurations and the IRC compared. *CCSD(T) ≡ DLPNO-CCSD(T)/ma-def2-TZVPP) energy values on the MP2/ma-def2-TZVPP.

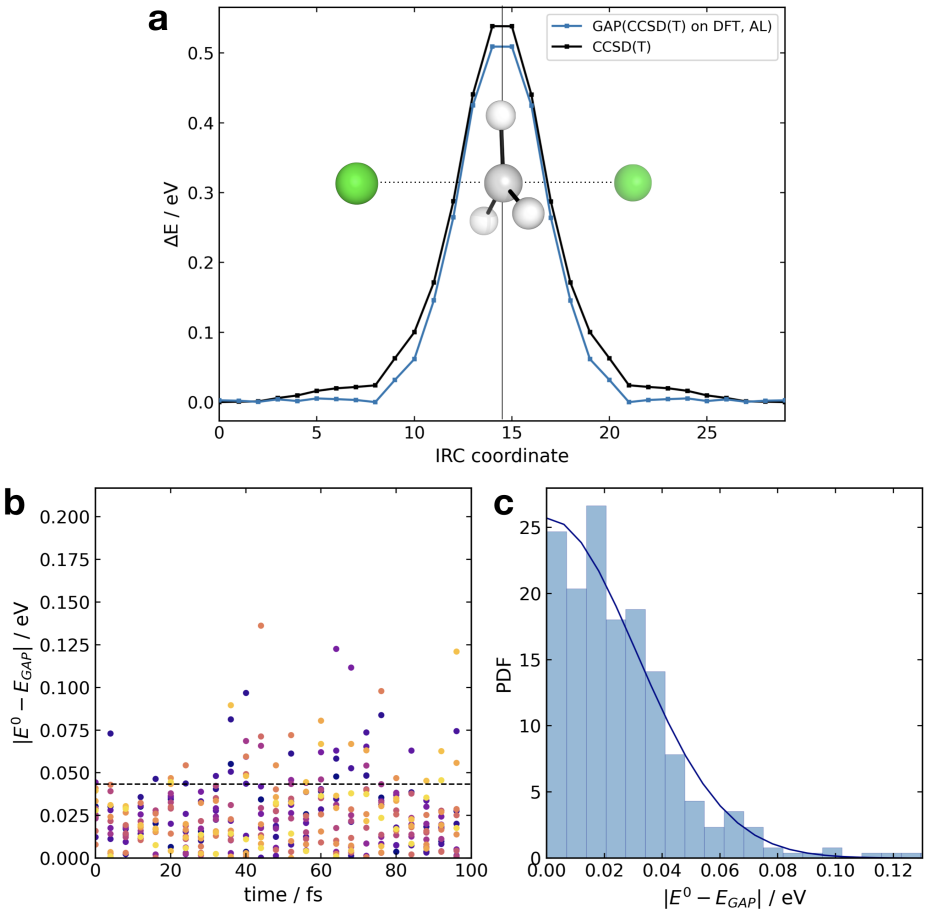


Figure 6.3.18. (a) Accuracy of a GAP trained on configurations generated from DFT (PBE/ma-def2-SVP) active learning (Figure 6.3.17) for Cl^- + CH_3Cl → Cl^- + CH_3Cl from the TS. The energy profile is compared to the values obtained with an IRC calculation at the MP2/ma-def2-TZVPP level of theory from a MP2 optimised TS. CCSD(T)/ma-def2-TZVPP energy calculations on DFT configurations used numerical gradients over 55 configurations. (b, c) Error in dynamics calculated from 300 K MD simulations propagated with the trained GAP from the MP2 transition state. SOAP only descriptors on C and Cl with $r_c^{\mathrm{SOAP}} = 3.5$ Å.

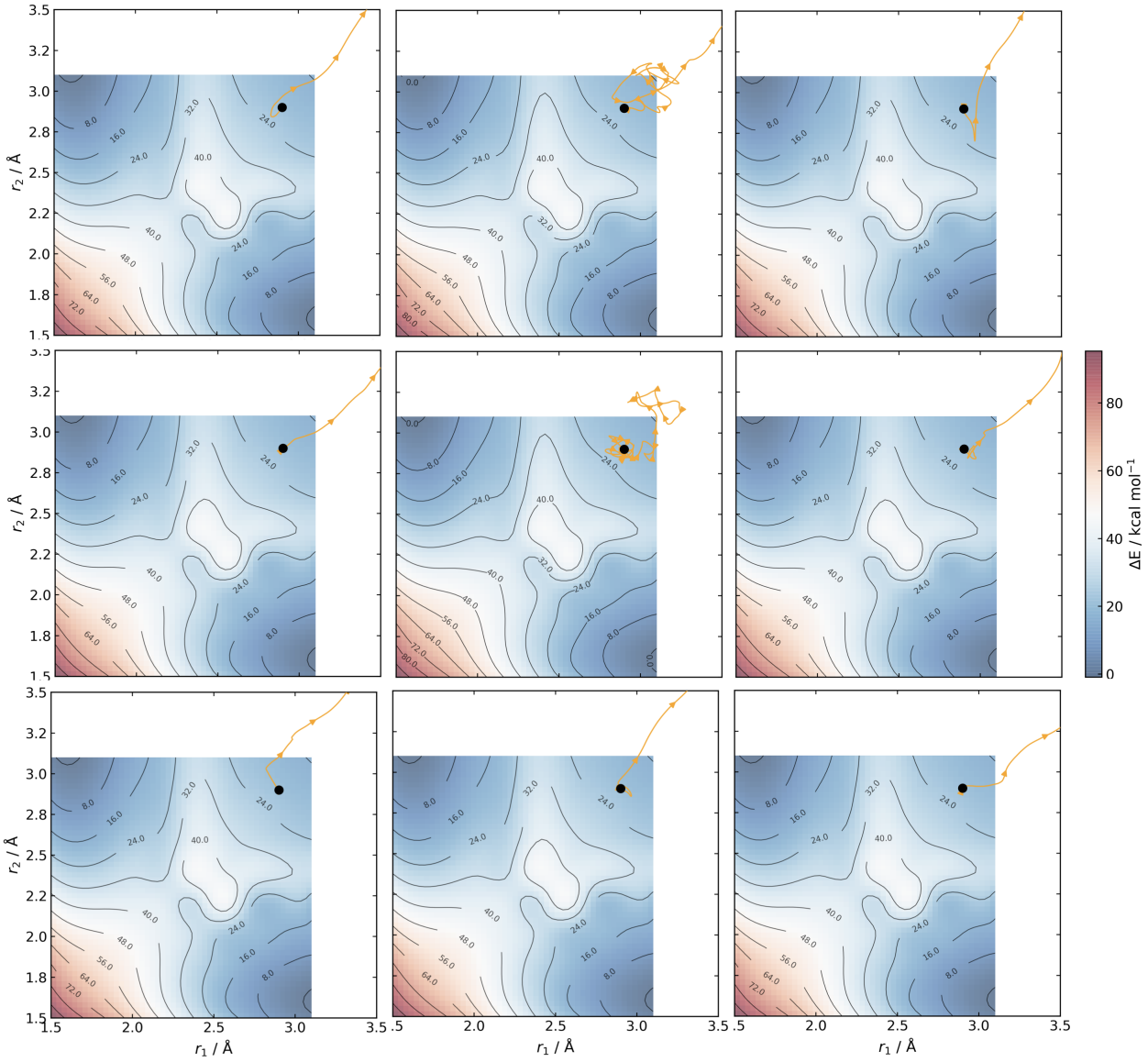


Figure 6.3.19. 2D PES along the forming bond distances (r_1, r_2) in the [4+2] dimerisation of cyclopentadiene as a function of time. GAP–propagated reactive dynamics (300 K, yellow lines) overlaid on the relaxed 2D PES. TS_1 (7N in ref. [496]) is shown as the black point. All trajectories lead to reactants (cyclopentadiene × 2) after ∼100 fs. GAP trained on ground truth B3LYP/def2-SVP. Active learning performed at 500 K initiated at $r_1, r_2 = 2.9$ Å.

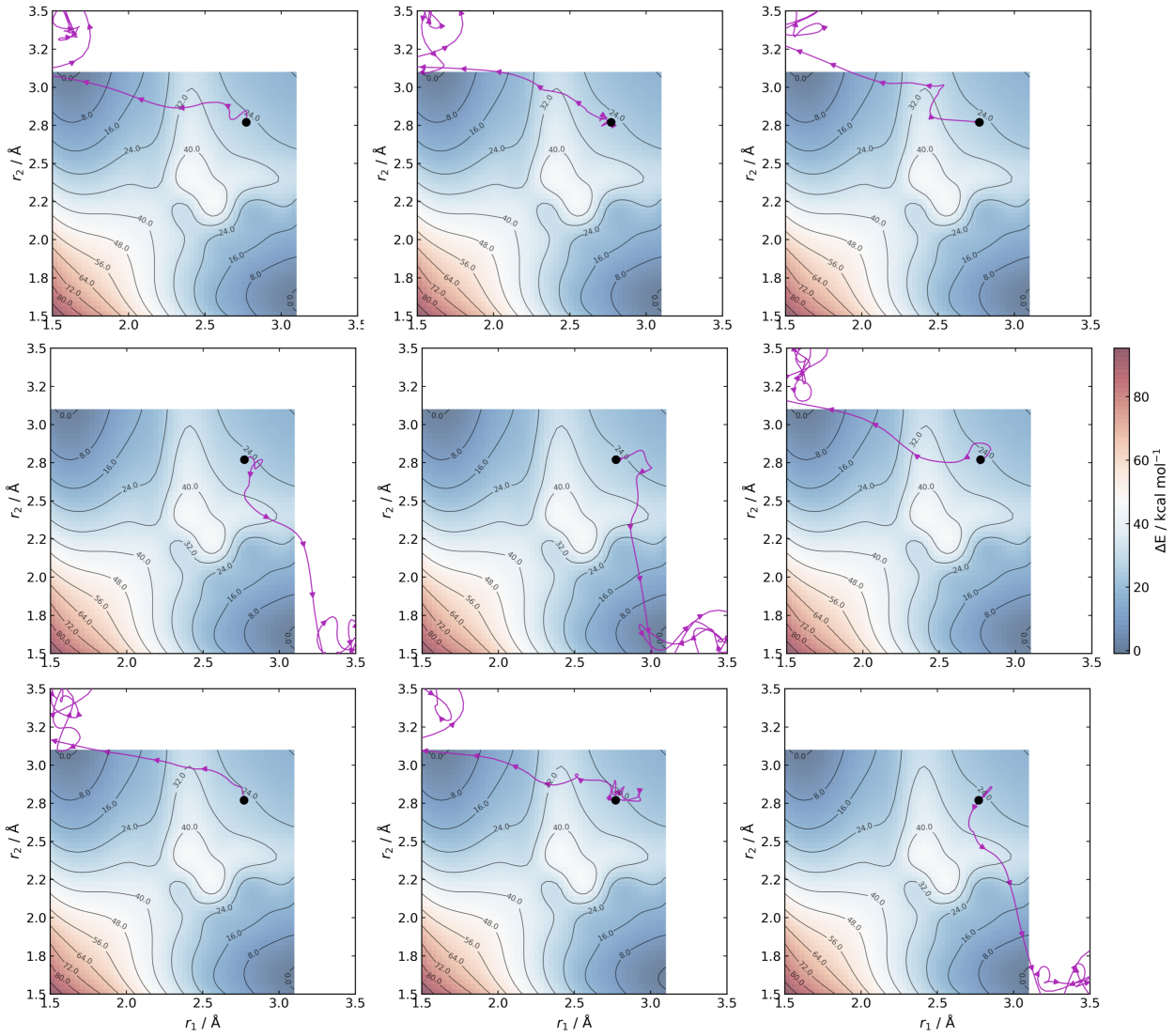


Figure 6.3.20. As Figure 6.3.19 but GAP active learning then dynamics initiated from TS_1' (r_1, r_2 = 2.77 Å).

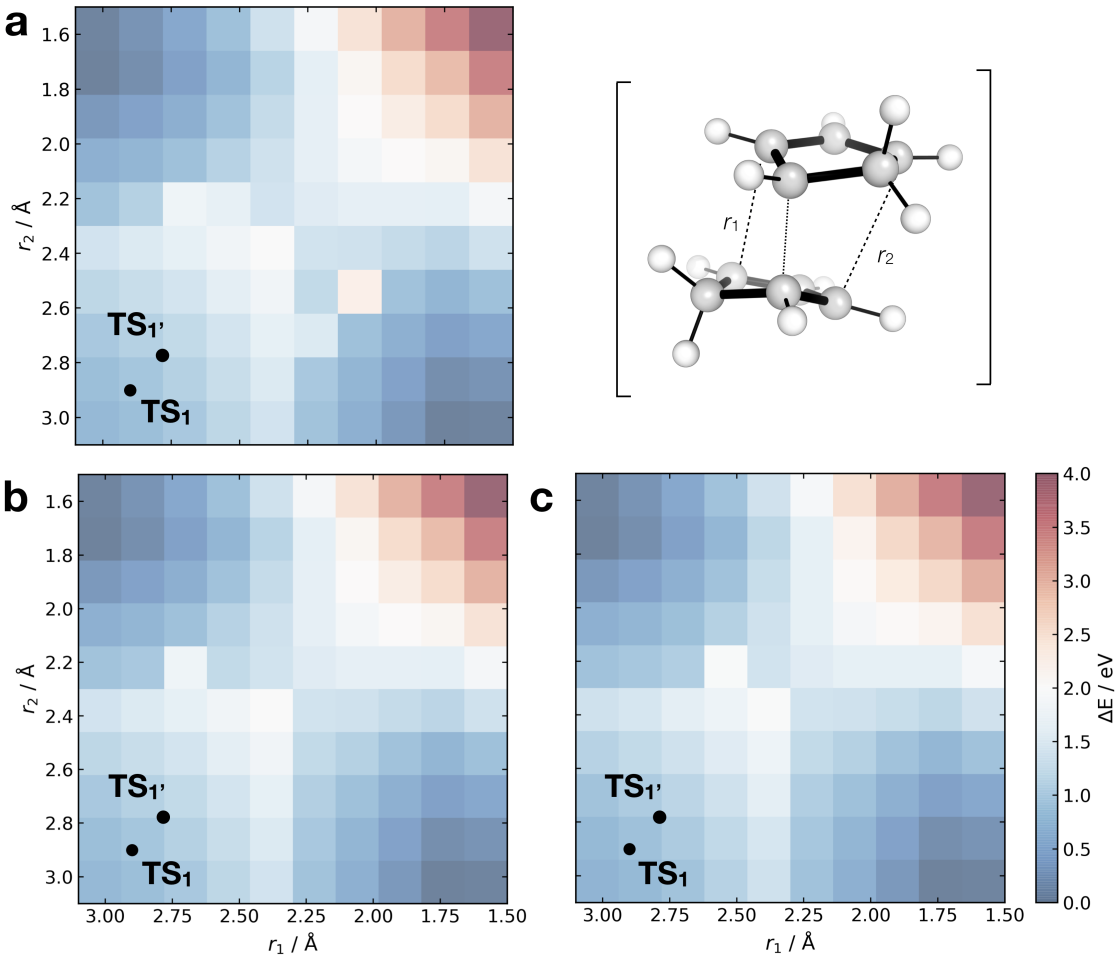


Figure 6.3.21. Comparison of 2D relaxed potential energy surfaces at (a) B3LYP/6-31G(d) as reported in ref. [496] and (b) a slightly improved level, B3LYP/def2-SVP.

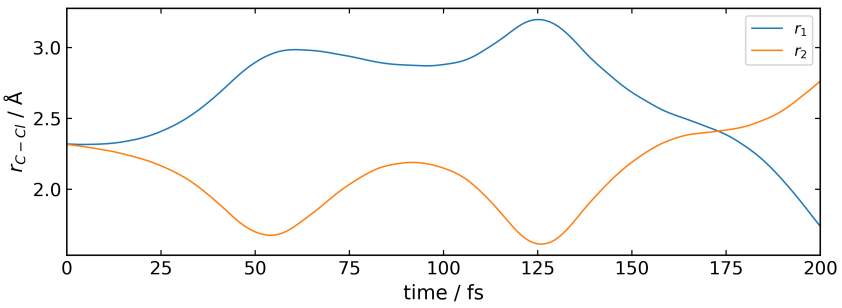


Figure 6.3.22. C-Cl distances as a function of time for one of the ten GAP-MD propagated from the TS of Cl+CH_3Cl in explicit water. A barrier recrossing event is observed at ∼170 fs.

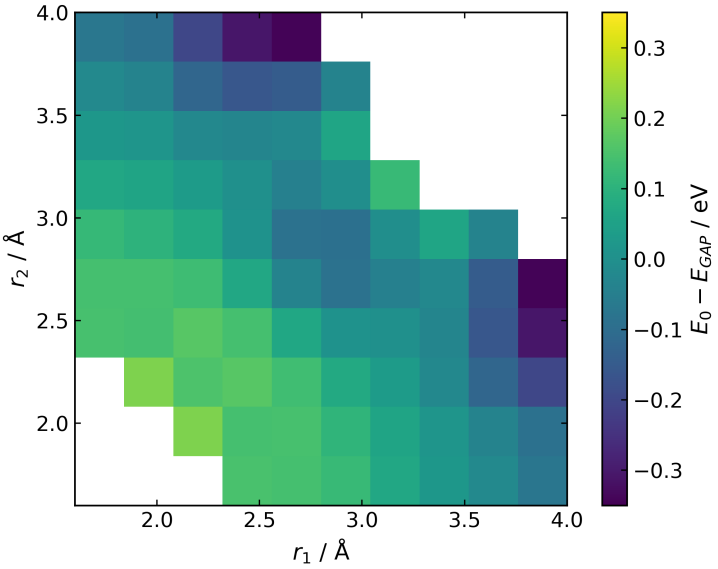


Figure 6.3.23. Error between GAP predicted and ground truth (CPCM(Water)-PBE/def2-SVP) referenced to the closest point on the surface to the TS ($r_1, r_2 = 2.4$ Å) i.e. the starting point for the GAP active learning at 1600 K. Regions above 2 eV on the ground truth surface masked, as the training explicitly does not include any configurations >2 eV from the minimum.

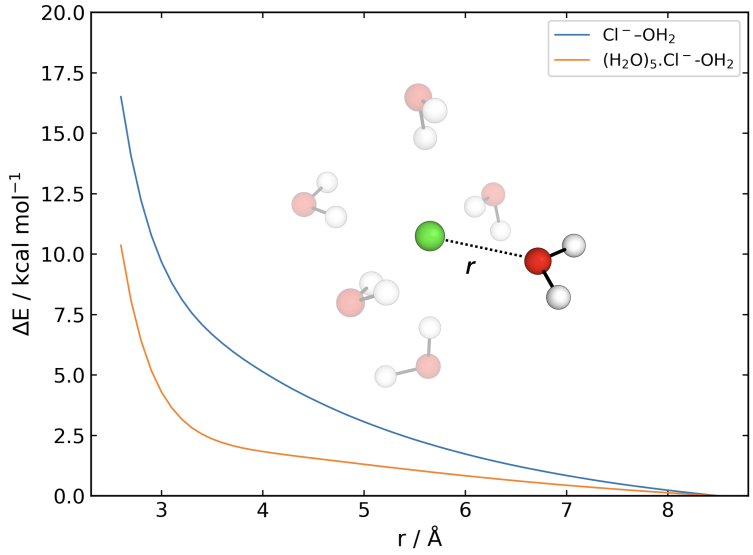


Figure 6.3.24. Gas-phase potential energy surfaces for $Cl^- \cdots H_2O$, with and without a first solvation shell at PBE/ma-def2-TZVP. Coordination number of chloride is 6 from ref. [510]. Initial partially solvated chloride generated by hand and optimised with ORCA constraining the Cl–O distance and the O–Cl–O angles to generate a close to octahedral geometry. The solvated ion shows a faster decaying potential; based on this model and $r_c^{\mathrm{SOAP}} = 4.5$ Å was chosen as a compromise between efficiency and accuracy.

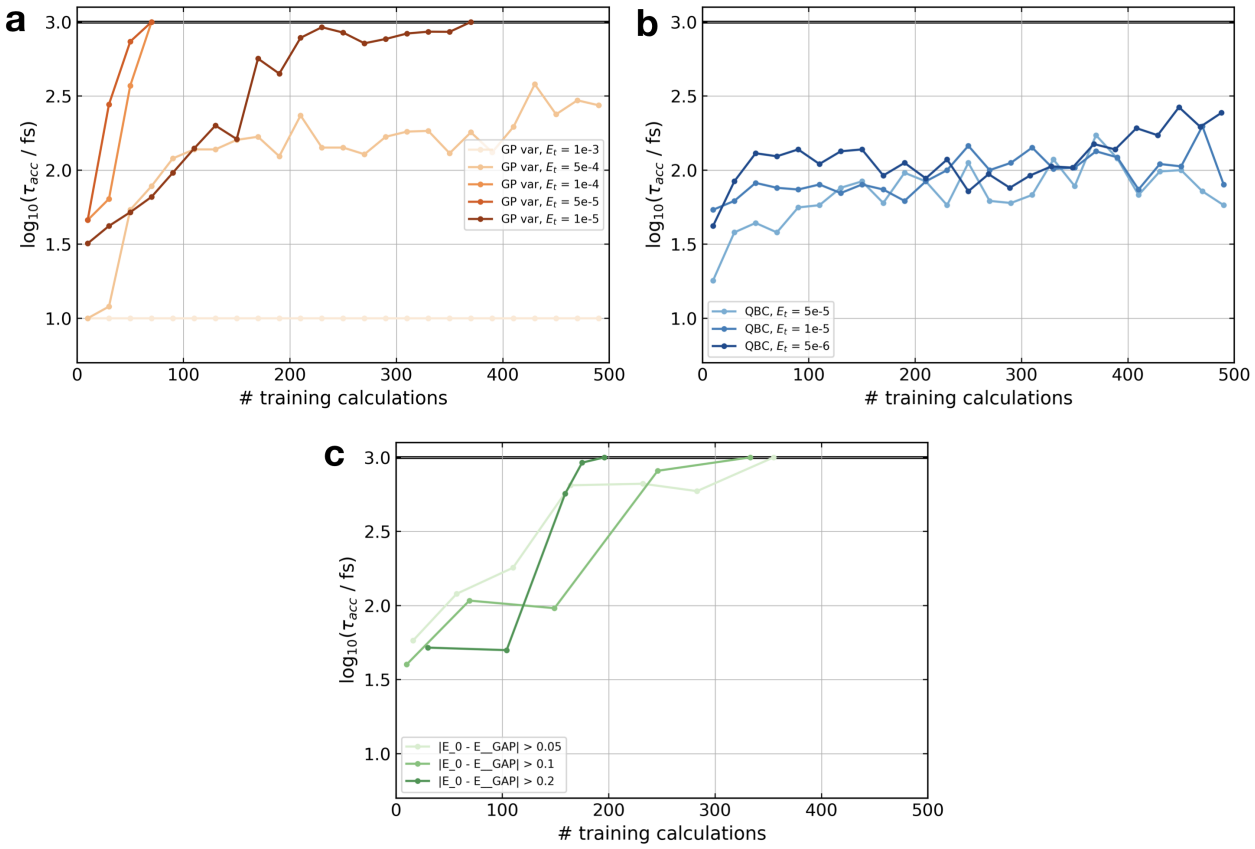


Figure 6.3.25. Comparison of selection strategies used in the 'active' learning loop for the intermolecular component of a water GAP trained in a 7 Å cubic box with 10 waters at the DFTB(3ob) level of theory. GAP-MD performed at 300 K with a 0.5 fs timestep for $n^3 + 2$ fs iterations sequentially until 1 ps of dynamics was performed. τ_{acc} used a 0.1 eV lower energy threshold and 1 eV total averaged over 5 random initial configurations (identical for each learning curve). (a) Adds a configuration when the maximum atomic energy variance predicted by the Gaussian Process exceeds a threshold E_T (in eV, where the max is taken over all the atoms in a frame of GAP-MD). (b) Query-by-committee (QBC), where a configuration is added if the standard deviation between GAPs trained on the same data (with different random noise) exceeds a threshold E_T (in eV). Values are chosen to span where the first frame from the iterative GAP-MD is chosen to the maximum 1 ps allowed. (c) Adds a configuration where the true difference of the total energy exceeds a threshold (in eV).

7 Conclusions

7.1 Modelling Catalytic Processes

How do we make better catalysts? Undoubtedly, computation has a central role to play in the discovery and development of existing catalysts; chemical space is simply too large for the experimental alternative. Although the ultimate goal of identifying the optimal catalyst for a given transformation remains elusive, employing calculations has already allowed chemists to understand, and even screen possible catalysts. This thesis focused on the latter and proposed methods that will enable faster predictions.

Modelling metallocage catalysis in Chapter 3 provided the motivation develop techniques to calculate catalytic proficiency within a small subspace relatively quickly. Using potential rather than free energies and transition state analogues, the catalytic proficiency of $[Pd_2L_4]^{4+}$ metallocages could be calculated in a few hundred CPUh. Computation uncovered an unintuitive flexibility that, at least partially, allows for catalysis.

To reduce the time-consuming setup time of metallocage calculations *cgbind* was developed as an open-source Python module and web-app. Accurate metallocage geometries ($\sim$ 1 Å RMSD to crystal structures) are obtained in seconds for various architectures. Rapidly generated features including the electrostatic potential surface and substrate addition provides

experimental colleagues to rapidly prototype cages prior to synthesis. Opportunities for improvement do, however, remain. For example, predicting binary binding affinities of a test set of $\sim$ 100 cage-substrate combinations yielded accuracy only just better than a random assignment (60%). This suggested that, like proteins, rigid docking of a substrate may not be sufficient to predict binding.[32]

To automate and accelerate the human-intensive search for conformational global minima and transition states that comprise catalytic cycles *autodE* was developed (Chapter 4) as an open-source and extensible approach to the problem. From SMILES strings 3D geometries are built using either a semi-deterministic, fully random build algorithm or an RDKit[294] method. A TS guess is obtained by aligning reactant and traversing a low-energy path after which the conformational space of TSs and minima are searched. The general applicability of *autodE* was illustrated in a range of examples, including cobalt and rhodium hydroformylation cycles. The *molfunc* open-source tool was developed alongside for molecular functionalisation, where 1D/3D molecular fragments are added to a core structure, providing an efficient approach to explore and optimise novel catalysts' scaffolds. While completely automated these platforms rely on standard methods to calculate free energies and currently do not account for any dynamic effects that may alter the predicted selectivity.

Towards more accurate free energy estimates, a novel method for calculating solution-phase translational entropy was outlined. The simple method is more faithful to the true environment and, by relying on a classical treatment of the partition function, is efficient to calculate. However, the improvement is only apparent for bimolecular reactions and is broadly a systematic shift making 'double differences' unaffected. The latter is generally the applicable domain in catalyst optimisation, where only the difference from a reference value is evaluated.

To study the dynamics of realistic (large/solvated) systems efficiently and with a view to obtain accurate free energy differences a method to train Gaussian Approximation Potentials was developed in Chapter 6. The active-learning based approach enables an efficient machine learnt potential to be obtained in a day and is able to accurately propagate dynamics at milliseconds per timestep rather than minutes or hours. These potentials and general ones like it will certainly make computational modelling and catalyst prediction more efficient and more accurate.

7.2 Future Directions

The future of computational catalyst design will rely on enhanced experimental collaboration and verification of computational predictions. An increasingly inverse approach to the problem will be required, alongside the development of more, accurate computational methods for energy and force evaluations. Employing a two-pronged strategy to both discover more diverse catalysts and by enumeration of human-guided libraries may be an effective approach. Machine learned models trained in a supervised way will continue to dominate the field and drive rapid predictions with perhaps a focus on reinforcement and 'self-play' strategies in the future. Increased use of automated methods both computational and synthetic will help to 'close-the-loop' and achieve highly effective catalysts for a specific transformation.

8 Appendices

8.1 ML Representations

8.1.1 SOAP

A SOAP vector[161] $\boldsymbol{p}$ is given by concatenating elements,

$$p_{nn'l}^{Z_1,Z_2} = \pi \sqrt{\frac{8}{2l+1}} \sum_m \left(c_{nlm}^{Z_1}\right)^* c_{n'lm}^{Z_2} \tag{8.1}$$

for every pair of atoms with atomic number Z and the coefficients (c_{nlm}) are of radial and angular basis functions centred a particular atom given by,

$$c_{nlm}^{Z} = \int_{\mathcal{R}} g_n(r) Y_{lm}(\phi, \theta) \rho^Z(\boldsymbol{r}) \, \mathrm{d}V \tag{8.2}$$

where the atomic density is given as a sum over 3D Gaussian centred on each atom i,

$$\rho^Z(\boldsymbol{r}) = \sum_{i \in \{Z\}} \exp\left(-\frac{|\boldsymbol{r} - \boldsymbol{R}_i|^2}{2\sigma_{\mathrm{at}}^2}\right) \tag{8.3}$$

A component of the neighbour density is then ($a := 1/2\sigma_{\text{at}}^2$),

$$\mathrm{e}^{-a|\boldsymbol{r}-\boldsymbol{R}_i|^2} = 4\pi \mathrm{e}^{-a(r^2+R_i^2)} i_l(2arR_i) \sum_{l'm'} Y_{l'm'}(\phi,\theta) Y^*_{l'm'}(\boldsymbol{R}_i) \tag{8.4}$$

inserting and using the orthogonality of the spherical harmonics integrated over theta and phi,

$$\begin{aligned} c^Z_{nlm} &= \int_{\mathcal{R}} g_n(r) Y_{lm}(\phi,\theta) \sum_{i\in\{Z\}} \exp\left(-a|\boldsymbol{r}-\boldsymbol{R}_i|^2\right) \mathrm{d}V \\ &= \sum_{i\in\{Z\}} \int_{\mathcal{R}} g_n(r) Y_{lm}(\phi,\theta) \exp\left(-a|\boldsymbol{r}-\boldsymbol{R}_i|^2\right) \mathrm{d}V \\ &= 4\pi \sum_{i\in\{Z\}} \sum_{l'm'} \int_{\mathcal{R}} g_n(r) Y_{lm}(\phi,\theta) \mathrm{e}^{-a(r^2+R_i^2)} i_l(2arR_i) Y_{l'm'}(\phi,\theta) Y^*_{l'm'}(\boldsymbol{R}_i) \mathrm{d}V \\ &= 4\pi \sum_{i\in\{Z\}} Y^*_{lm}(\boldsymbol{R}_i) \int_{\tilde{\mathcal{R}}} g_n(r) \mathrm{e}^{-a(r^2+R_i^2)} i_l(2arR_i) r^2 \, \mathrm{d}r \end{aligned} \tag{8.5}$$

where i_l is a modified spherical Bessel function of the first kind. Inserting a orthonormal radial basis as $g_n(r)$,

$$\begin{aligned} c^Z_{nlm} &= 4\pi \sum_{i\in\{Z\}} Y^*_{lm}(\boldsymbol{R}_i) \int_{\tilde{\mathcal{R}}} \sum_{n'}^{n_{\max}} \beta_{nn'l} r^l e^{-\alpha_{n'l} r^2} \mathrm{e}^{-a(r^2+R_i^2)} i_l(2arR_i) r^2 \, \mathrm{d}r \\ &= 4\pi \sum_{i\in\{Z\}} Y^*_{lm}(\boldsymbol{R}_i) \mathrm{e}^{-aR_i^2} \sum_{n'}^{n_{\max}} \beta_{nn'l} \int_{\tilde{\mathcal{R}}} r^{l+2} \mathrm{e}^{-(\alpha_{n'l}+a)r^2} i_l(2arR_i) \, \mathrm{d}r. \end{aligned} \tag{8.6}$$

Evaluating the integral where $b := \alpha_{n'l} + a$ and $c_i := 2aR_i$,

$$I_{iln'} = \int_0^\infty r^{l+2} \mathrm{e}^{-br^2} i_l(c_i r) \, \mathrm{d}r \tag{8.7}$$

using $i_\lambda(z) = i^{-\lambda} j_\lambda(iz)$ where from Gradshteyn and Ryzhik,[511]

$$\int_0^\infty x^{\lambda+2}\mathrm{e}^{-k_1x^2}j_\lambda(k_2x)\mathrm{d}x = \sqrt{\frac{\pi}{2}}\frac{k_2^\lambda}{(2k_1)^{\lambda+3/2}}\exp\left(-\frac{k_2^2}{4k_1}\right)$$

so the required integral is with $\lambda = l, k_1 = b, k_2 = ic_i$,

$$\begin{aligned} I_{iln'} &= \sqrt{\frac{\pi}{2}}\frac{c_i^l}{(2b)^{l+3/2}}\exp\left(\frac{c_i^2}{4b}\right) \\ &= \sqrt{\frac{\pi}{2}}\frac{(2aR_i)^l}{(2(\alpha_{n'l}+a))^{l+3/2}}\exp\left(\frac{(aR_i)^2}{\alpha_{n'l}+a}\right) \end{aligned} \tag{8.8}$$

so then,

$$c_{nlm}^Z = \sqrt{8\pi^3}\sum_{i\in\{Z\}}Y_{lm}^*(\boldsymbol{R}_i)\,\mathrm{e}^{-aR_i^2}\sum_{n'}^{n_{\max}}\beta_{nn'l}\frac{(2aR_i)^l}{(2(\alpha_{n'l}+a))^{l+3/2}}\exp\left(\frac{(aR_i)^2}{\alpha_{n'l}+a}\right). \tag{8.9}$$

Using real spherical harmonic functions,

$$Y_{l,m}(\theta,\phi) = \begin{cases} \sqrt{2}(-1)^m\,\mathrm{Im}\{Y_l^{|m|}(\theta,\phi)\} & \text{if } m<0 \\ 1 & \text{if } m=0 \\ \sqrt{2}(-1)^m\,\mathrm{Re}\{Y_l^m(\theta,\phi)\} & \text{if } m>0 \end{cases} \tag{8.10}$$

where,

$$Y_l^m(\theta,\phi) = \sqrt{\frac{(2l+1)(l-m)!}{4\pi(l+m)!}}P_l^m(\cos\theta)e^{im\phi}. \tag{8.11}$$

8.2 Solution Phase Entropy

8.2.1 Partition Functions

For completeness, the well-known integral approximation to the particle in a box PF is reproduced below. Provided the gaps between energy levels are small the sum can be replaced by a integral to give,

$$q_{\text{trans}} \approx \iiint_0^\infty dn_x\, dn_y\, dn_z\, \exp\Big\{ -\beta \cdot \frac{\hbar^2\pi^2}{2ml^2}(n_x^2 + n_y^2 + n_z^2)\Big\}. \tag{8.12}$$

Converting into spherical polar coordinates with $n_x^2 + n_y^2 + n_z^2 = n_r^2$ and integrating over n_θ and n_ϕ in the positive octant,

$$\begin{aligned} q_{\text{trans}} &\approx \frac{1}{8} \cdot 4\pi \int_0^\infty dn_r\, n_r^2 \exp\Big\{ -\beta \cdot \frac{\hbar^2\pi^2}{2ml^2} n_r^2\Big\} \\ &\approx \frac{\pi}{2} \cdot \frac{\sqrt{\pi}}{4}\Big(\frac{\hbar^2\pi^2\beta}{2ml^2}\Big)^{-3/2} \\ &\approx \pi^{3/2}\Big(\frac{ml^2}{2\hbar^2\pi^2\beta}\Big)^{3/2} \\ &\approx \Big(\frac{2\pi m k_B T}{h^2}\Big)^{3/2} V \end{aligned} \tag{8.13}$$

where the standard integral $\int_0^\infty dx\; x^2 e^{-ax^2} = \sqrt{\pi}/4a^{3/2}$ and $V = l^3$ have been used. Inter-

estingly, (8.13) is identical to that obtained classically (8.14).

$$\begin{aligned} q_{\text{trans, c}} &= \frac{1}{h^3}\int\int e^{-\beta H(\boldsymbol{p},\boldsymbol{x})}\, d\boldsymbol{p}^3 d\boldsymbol{x}^3 \\ &= \frac{1}{h^3}\int_{-\infty}^{\infty} e^{-\beta \boldsymbol{p}^2/2m}\, d\boldsymbol{p}^3 \int_0^l e^{-\beta\cdot 0}\, d\boldsymbol{x}^3 \\ &= \frac{1}{h^3}\Big\{\int_{-\infty}^{\infty} e^{-\beta {p_k}^2/2m}\, dp_k\Big\}^3 \Big\{\int_0^l dx_k\Big\}^3 \\ &= \frac{l^3}{h^3}\Big\{\sqrt{\frac{\pi}{1/(2k_BT)}}\Big\}^3 \\ &= \Big(\frac{2\pi m k_B T}{h^2}\Big)^{3/2} V. \end{aligned} \tag{8.14}$$

8.2.2 Classical Truncated Harmonic Oscillator

The classical entropy of the truncated harmonic oscillator with potential energy,

$$V(x) = \begin{cases} \frac{1}{2}m\omega^2 x^2 & \text{if} \quad x > -r_e \\ +\infty & \text{if} \quad x \le -r_e \end{cases} \tag{8.15}$$

is derived as follows. First the classical partition function is,[a]

$$\begin{aligned} Z_c &= \frac{1}{h}\int_{-\infty}^{\infty} dp \int_{-\infty}^{\infty} dx\; e^{-\beta E(x,p)} \\ &= \frac{1}{h}\int_{-\infty}^{\infty} dp\; e^{-\beta p^2/2m} \int_{-\infty}^{\infty} dx\; e^{-\beta V(x)} \\ &= \frac{1}{h}\int_{-r_e}^{\infty} dp\; e^{-\beta p^2/2m} \int_{-r_e}^{\infty} dx\; e^{-\beta m\omega^2 x^2/2} \\ &= \frac{1}{h}\cdot\frac{\sqrt{\pi}\,\operatorname{erfc}(-r_e\sqrt{\beta/2m})}{2\sqrt{\beta/2m}}\cdot\frac{\sqrt{\pi}\,\operatorname{erfc}(-r_e\sqrt{\beta m\omega^2/2})}{2\sqrt{\beta m\omega^2/2}} \\ &= \frac{\pi k_B T}{2h\omega}\operatorname{erfc}(-r_e\sqrt{\beta/2m})\operatorname{erfc}(-r_e\sqrt{\beta m\omega^2/2}) \end{aligned} \tag{8.16}$$

[a] Using $\int_a^{+\infty} e^{-bx^2} dx = \sqrt{\pi/b}\,\operatorname{erfc}(a\sqrt{b})/2$

where erfc(x) is the complementary error function. The derivative with respect to temperature for $a := r_e^2/2k_Bm$ and $b := r_e^2 m\omega^2/2k_B$ is,

$$\begin{aligned}\frac{dZ_c}{dT} &= \frac{d}{dT}\left\{\frac{\pi k_B T}{2h\omega}\,\mathrm{erfc}(-\sqrt{a/T})\,\mathrm{erfc}(-\sqrt{b/T})\right\} \\ &= \frac{k_B}{4\hbar\omega}\left(\mathrm{erfc}(-\sqrt{a/T})\,\mathrm{erfc}(-\sqrt{b/T}\,) - \frac{be^{-b/T}\,\mathrm{erfc}(-\sqrt{a/T})}{\sqrt{b\pi T}} - \frac{ae^{-a/T}\,\mathrm{erfc}(-\sqrt{b/T})}{\sqrt{a\pi T}}\right)\end{aligned} \tag{8.17}$$

8.2.3 Classical Morse Oscillator

The classical entropy of a Morse oscillator as a model of a chemical bond requires an upper bound to the total energy to prevent the configurational integral becoming unbounded. For a total energy,

$$E(p,x) = \frac{p^2}{2m} + D_e(1 - e^{-a(r-r_e)})^2 \tag{8.18}$$

where r is the bond length and r_e the equilibrium bond length. The maximum energy is the dissociation energy of the bond (D_e) such that the the momentum is bounded,

$$E(x,p) < D_e \qquad \Longrightarrow \qquad p_\pm = \pm\sqrt{2mD_e(1-(1-e^{-ax})^2)} \tag{8.19}$$

where x is the displacement from equilibrium, with $x > -\frac{1}{a}\ln(2)$ to ensure the momentum is real.

The classical partition function is then,[b]

$$
\begin{aligned}
Z_c &= \frac{1}{h}\int_{-\infty}^{\infty} dp \int_{-\infty}^{\infty} dx\; e^{-\beta E(x,p)} \\
&= \frac{1}{h}\int_{p_-}^{p_+} \int_{-\frac{1}{a}\ln(2)}^{\infty} e^{-\beta\{p^2/2m + D_e(1-e^{-ax})^2\}}\; dx\; dp \\
&= \frac{1}{h}\int_{p_-}^{p_+} \int_{-\frac{1}{a}\ln(2)}^{\infty} e^{-\beta p^2/2m} e^{-\beta D_e(1-e^{-ax})^2}\; dx\; dp \\
&= \frac{1}{h}\int_{-\frac{1}{a}\ln(2)}^{\infty} \int_{p_-}^{p_+} e^{-\beta p^2/2m} e^{-\beta D_e(1-e^{-ax})^2}\; dp\; dx \\
&= \frac{1}{h}\int_{-\infty}^{\infty} e^{-\beta D_e(1-e^{-ax})^2} \cdot \frac{\sqrt{\pi}\,\mathrm{erf}(p\sqrt{\beta/2m})}{\sqrt{\beta/2m}}\bigg|_{p_-}^{p_+} dx \\
&= \frac{2}{h}\sqrt{\frac{2m\pi}{\beta}} \int_{-\frac{1}{a}\ln(2)}^{\infty} e^{-\beta D_e(1-e^{-ax})^2}\; \mathrm{erf}(p_+\sqrt{\beta/2m})\; dx \\
&= \frac{2}{h}\sqrt{\frac{2m\pi}{\beta}} \int_{-\frac{1}{a}\ln(2)}^{\infty} e^{-\beta D_e(1-e^{-ax})^2}\; \mathrm{erf}\Big(\sqrt{\beta D_e(1-(1-e^{-ax})^2)}\Big)\; dx.
\end{aligned}
\tag{8.20}
$$

Although (8.20) has no analytical solution, the integrand is smooth allowing for efficient numerical evaluation (Figure 8.2.1).

8.2.4 Exponential Well

To derive the entropy and internal energy due to translation of a molecule in an 3D exponential well the temperature derivative of the partition function (Z_c^{exp}) is required. For,

$$
Z_c^{\mathrm{exp}} = \left(\frac{2m\pi k_B T}{h^2}\right)^{3/2} \int_0^{\infty} 4\pi r^2 e^{-k(e^{ar}-1)/k_B T}\; \mathrm{d}r := \Lambda(T) I(T) \tag{8.21}
$$

[b] Using $\int_{-a}^{+a} e^{-bx^2} dx = \sqrt{\pi/b}\,\mathrm{erf}(a\sqrt{b})$ and that erf is an even function.

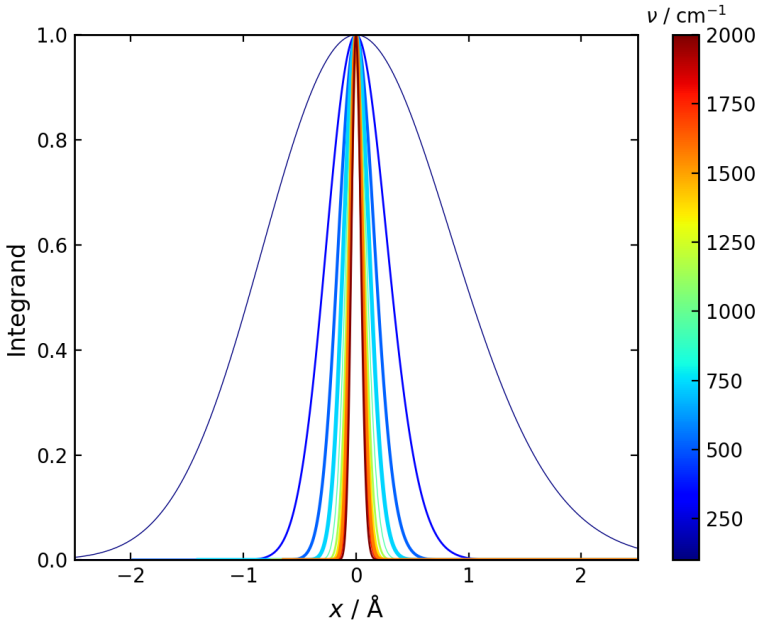


Figure 8.2.1. Integrand of Z_c vs x. $D_e = 400$ kJ mol^{-1}, $T = 298.15$ K, $m = 1$ amu.

the derivative is,

$$\begin{aligned}
\left(\frac{\partial Z_c^{\text{exp}}}{\partial T}\right)_{N,V} &= \left(\frac{\partial \Lambda}{\partial T}\right)_{N,V} I(T) + \Lambda(T)\left(\frac{\partial I}{\partial T}\right)_{N,V} \\
&= \Lambda(T)\frac{3}{2}T^{-1}I(T) + \Lambda(T)\int_0^\infty 4\pi r^2 \frac{k(e^{ar}-1)}{k_B T^2} e^{-k\beta(e^{ar}-1)}\,\mathrm{d}r \\
&= \frac{3}{2T}Z_c^{\text{exp}} + \frac{\Lambda(T)k\beta}{T}\int_0^\infty 4\pi r^2 e^{ar} e^{-k\beta(e^{ar}-1)}\,\mathrm{d}r \\
&\qquad - \frac{\Lambda(T)k\beta}{T}\int_0^\infty 4\pi r^2 e^{-k\beta(e^{ar}-1)}\,\mathrm{d}r \\
&= \frac{1}{T}\left(\frac{3}{2}Z_c^{\text{exp}} + \Lambda(T)k\beta\int_0^\infty 4\pi r^2 e^{ar} e^{-k\beta(e^{ar}-1)}\,\mathrm{d}r - k\beta Z_c^{\text{exp}}\right) \\
&= \frac{Z_c^{\text{exp}}}{T}\left(\frac{3}{2} - k\beta + \frac{\Lambda(T)k\beta}{Z_c^{\text{exp}}}\int_0^\infty 4\pi r^2 e^{ar} e^{-k\beta(e^{ar}-1)}\,\mathrm{d}r\right)
\end{aligned} \tag{8.22}$$

8.2.5 Test Sets

Test set **A** contains two reactions from the Morita Baylis–Hillman reaction for which experimental reaction ΔS values are available (Table 8.2.1, Figure 8.2.2).[363]

Reaction	T_{avg} / K	$T\Delta S$ / kcal mol^{-1}	ΔG / kcal mol^{-1}
R-MBH1	317.6	-7.9	-3.2
R-MBH2	328.4	-10.2	-3.9

Table 8.2.1. Thermodynamic data for **A**. No error is given. T_{avg} values are averages over the temperature range used to calculate ΔS.

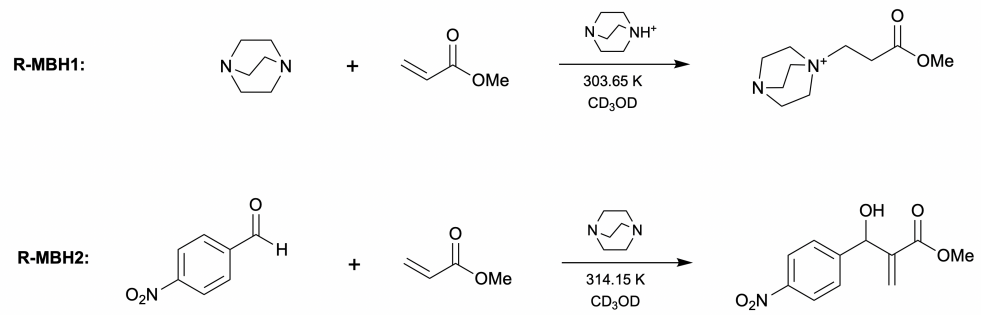


Figure 8.2.2. Reactions in test set **A**.

Test set **B** comprise entropies of activation for 10 S_N2 reactions collated in ref. [512] (Figure 8.2.3). These were chosen based on modest structural complexity, while spanning a wide range of $\Delta S^{\ddagger}$ values (Table 8.2.2).

Reaction	T / K	$\Delta H^{\ddagger}$ / kcal mol^{-1}	$T\Delta S^{\ddagger}$ / kcal mol^{-1}	$\Delta G^{\ddagger}$ / kcal mol^{-1}
R1	369.1	22.3	-3.9	26.0
R2	369.1	18.6	-6.1	24.4
R3	322.4	21.5	-1.9	23.4
R4	303.1	15.9	-3.7	19.6
R5	303.1	16.5	-2.1	18.6
R6	303.1	16.4	-4.9	21.3
R7	303.1	18.5	-2.0	20.5
R8	252.2	11.9	-3.1	15.0
R9	303.1	18.6	-2.4	20.9
R10	303.1	20.2	-1.3	21.5

Table 8.2.2. Activation parameters for test set **B**. Reactions are shown in Figure 8.2.3. Where errors are quoted: $\Delta S_{\text{err}} < \pm 2.5$ J K^{-1} mol^{-1}, $\Delta H_{\text{err}} < \pm 0.2$ kcal mol^{-1}.

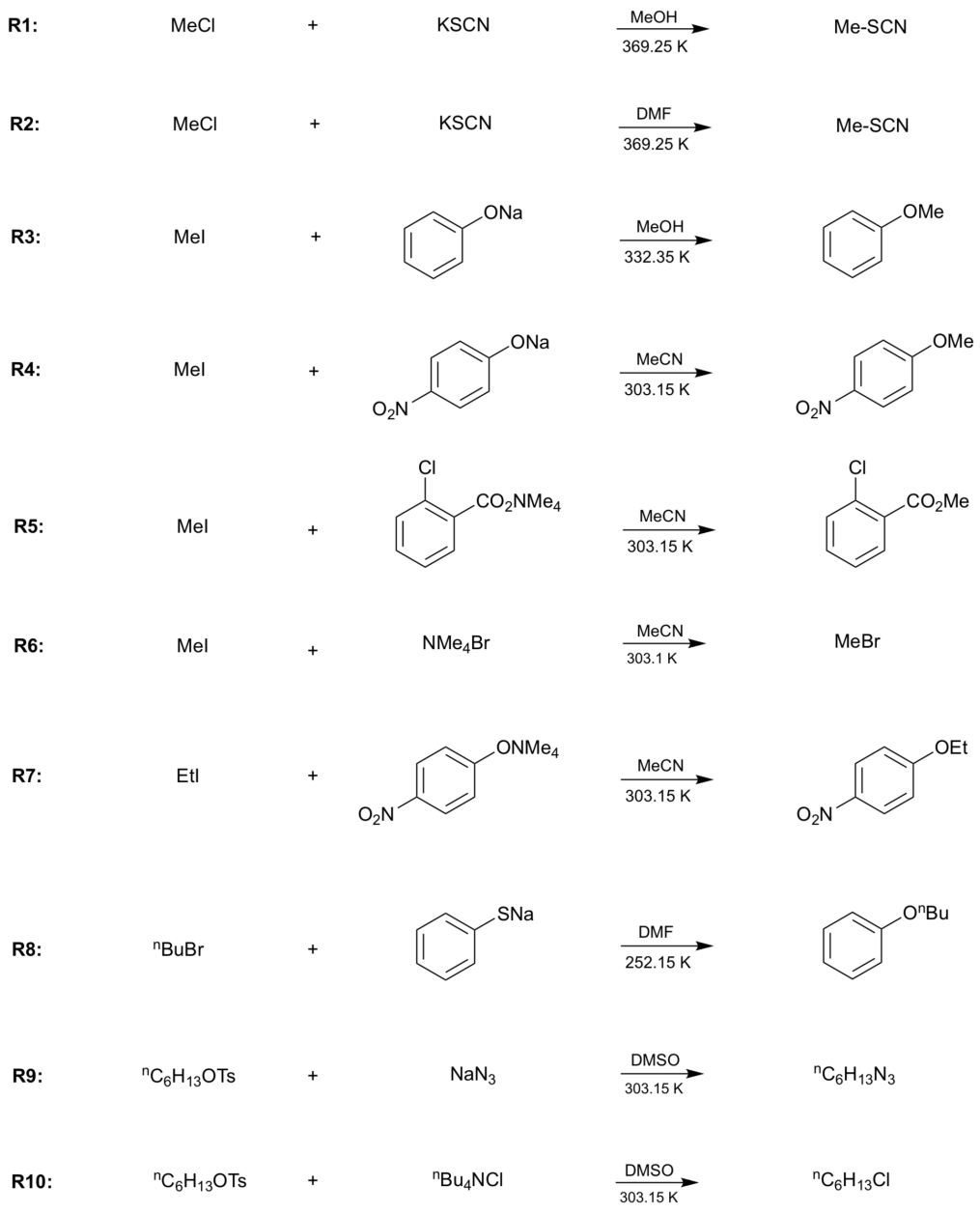


Figure 8.2.3. Reactions in test set **B**.

Bibliography

(1) United Nations Transforming Our World: The 2030 Agenda for Sustainable Development., 2015.

(2) Masson-Delmotte, V. et al. Global Warming of 1.5°C. An IPCC Special Report on the impacts of global warming of 1.5°C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change, sustainable development, and efforts to eradicate poverty., 2018.

(3) Roduner, E. *Chemical Society Reviews* **2014**, *43*, 8226–8239.

(4) Garcia-Martinez, J. *Angewandte Chemie International Edition* **2021**, *60*, 4956–4960.

(5) Miandad, R.; Rehan, M.; Barakat, M. A.; Aburiazaiza, A. S.; Khan, H.; Ismail, I. M. I.; Dhavamani, J.; Gardy, J.; Hassanpour, A.; Nizami, A.-S. *Frontiers in Energy Research* **2019**, *7*, 27.

(6) Poree, C.; Schoenebeck, F. *Accounts of Chemical Research* **2017**, *50*, 605–608.

(7) Nicolaou, K. C. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* **2014**, *470*, 20130690.

(8) McNally, A.; Prier, C. K.; MacMillan, D. W. C. *Science* **2011**, *334*, 1114–1117.

(9) Troshin, K.; Hartwig, J. F. *Science* **2017**, *357*, 175–181.

(10) Ahneman, D. T.; Estrada, J. G.; Lin, S.; Dreher, S. D.; Doyle, A. G. *Science* **2018**, *360*, 186–190.

(11) Meisner, J.; Zhu, X.; Martínez, T. J. *ACS Central Science* **2019**, *5*, 1493–1495.

(12) Dewyer, A. L.; Argüelles, A. J.; Zimmerman, P. M. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2017**, *8*, e1354.

(13) Habershon, S. *Journal of Chemical Theory and Computation* **2016**, *12*, 1786–1798.

(14) Ismail, I.; Stuttaford-Fowler, H. B.V. A.; Ashok, C. O.; Robertson, C.; Habershon, S. *The Journal of Physical Chemistry A* **2019**, *123*, 3407–3417.

(15) Robertson, C.; Habershon, S. *Catalysis Science & Technology* **2019**, *9*, 6357–6369.

(16) Kim, Y.; Kim, J. W.; Kim, Z.; Kim, W. Y. *Chemical Science* **2018**, *9*, 825–835.

(17) Ung Lee, J.; Kim, Y.; Kim, W. Y.; Oh, H. B. *Physical Chemistry Chemical Physics* **2020**, *22*, 5057–5069.

(18) McDermott, M. J.; Dwaraknath, S. S.; Persson, K. A. *Nature Communications* **2021**, *12*, DOI: 10.1038/s41467-021-23339-x.

(19) Rubino, G.; Tuffin, B. In *Rare Event Simulation using Monte Carlo Methods*; John Wiley and Sons, Inc.: Hoboken, New Jersey, USA, 2009; Chapter 1, pp 1–13.

(20) Yang, Y. I.; Shao, Q.; Zhang, J.; Yang, L.; Gao, Y. Q. *The Journal of Chemical Physics* **2019**, *151*, 070902.

(21) Laio, A.; Parrinello, M. *Proceedings of the National Academy of Sciences* **2002**, *99*, 12562–12566.

(22) Maeda, S.; Ohno, K.; Morokuma, K. *Physical Chemistry Chemical Physics* **2013**, *15*, 3683–3701.

(23) Uematsu, R.; Yamamoto, E.; Maeda, S.; Ito, H.; Taketsugu, T. *Journal of the American Chemical Society* **2015**, *137*, 4090–4099.

(24) Yoshimura, T.; Maeda, S.; Taketsugu, T.; Sawamura, M.; Morokuma, K.; Mori, S. *Chemical Science* **2017**, *8*, 4475–4488.

(25) Lavigne, C.; dos Passos Gomes, G.; Pollice, R.; Aspuru-Guzik, A. *ChemRxiv* **2020**, DOI: 10.26434/chemrxiv.13008500.v2.

(26) Wang, L.-P.; Titov, A.; McGibbon, R.; Liu, F.; Pande, V. S.; Martínez, T. J. *Nature Chemistry* **2014**, *6*, 1044–1048.

(27) Grimme, S. *Journal of Chemical Theory and Computation* **2019**, *15*, 2847–2862.

(28) Häse, F.; Aldeghi, M.; Hickman, R. J.; Roch, L. M.; Aspuru-Guzik, A. *arXiv* **2021**, DOI: arXiv:2003.12127.

(29) Shields, B. J.; Stevens, J.; Li, J.; Parasram, M.; Damani, F.; Alvarado, J. I. M.; Janey, J. M.; Adams, R. P.; Doyle, A. G. *Nature* **2021**, *590*, 89–96.

(30) Koinuma, H.; Takeuchi, I. *Nature Materials* **2004**, *3*, 429–438.

(31) Mennen, S. M. et al. *Organic Process Research & Development* **2019**, *23*, 1213–1242.

(32) Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. *Nature Reviews Drug Discovery* **2004**, *3*, 935–949.

(33) Guan, Y.; Ingman, V. M.; Rooks, B. J.; Wheeler, S. E. *Journal of Chemical Theory and Computation* **2018**, *14*, 5249–5261.

(34) Rosales, A. R.; Wahlers, J.; Limé, E.; Meadows, R. E.; Leslie, K. W.; Savin, R.; Bell, F.; Hansen, E.; Helquist, P.; Munday, R. H.; Wiest, O.; Norrby, P.-O. *Nature Catalysis* **2018**, *2*, 41–45.

(35) Muratov, E. N. et al. *Chemical Society Reviews* **2020**, *49*, 3525–3564.

(36) Toyao, T.; Maeno, Z.; Takakusagi, S.; Kamachi, T.; Takigawa, I.; ichi Shimizu, K. *ACS Catalysis* **2019**, *10*, 2260–2297.

(37) Kitchin, J. R. *Nature Catalysis* **2018**, *1*, 230–232.

(38) Friederich, P.; dos Passos Gomes, G.; Bin, R. D.; Aspuru-Guzik, A.; Balcells, D. *Chemical Science* **2020**, *11*, 4584–4601.

(39) Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. *ACS Central Science* **2018**, *4*, 268–276.

(40) Nigam, A.; Pollice, R.; Krenn, M.; dos Passos Gomes, G.; Aspuru-Guzik, A. *Chemical Science* **2021**, *12*, 7079–7090.

(41) Goldberg, D., *Genetic algorithms in search, optimization, and machine learning*; Addison-Wesley Publishing Company: Boston, Massachusetts, USA, 1989.

(42) Henle, J. J.; Zahrt, A. F.; Rose, B. T.; Darrow, W. T.; Wang, Y.; Denmark, S. E. *Journal of the American Chemical Society* **2020**, *142*, 11578–11592.

(43) Zahrt, A. F.; Rose, B. T.; Darrow, W. T.; Henle, J. J.; Denmark, S. E. *Reaction Chemistry & Engineering* **2021**, *6*, 694–708.

(44) Kirkpatrick, P.; Ellis, C. *Nature* **2004**, *432*, 823–823.

(45) Freeze, J. G.; Kelly, H. R.; Batista, V. S. *Chemical Reviews* **2019**, *119*, 6595–6612.

(46) Foscato, M.; Jensen, V. R. *ACS Catalysis* **2020**, *10*, 2354–2377.

(47) Weymuth, T.; Reiher, M. *International Journal of Quantum Chemistry* **2014**, *114*, 838–850.

(48) Chang, A. M.; Rudshteyn, B.; Warnke, I.; Batista, V. S. *Inorganic Chemistry* **2018**, *57*, 15474–15480.

(49) Norskov, J. K.; Abild-Pedersen, F.; Studt, F.; Bligaard, T. *Proceedings of the National Academy of Sciences* **2011**, *108*, 937–943.

(50) Tantillo, D. J. *Accounts of Chemical Research* **2016**, *49*, 1079–1079.

(51) Kozuch, S.; Shaik, S. *Accounts of Chemical Research* **2011**, *44*, 101–110.

(52) Anslyn, E. V.; Dougherty, D. A., *Modern Physical Organic Chemistry*; University Science Books: Mill Valley, California, USA, 2006.

(53) Mitsumori, S.; Zhang, H.; Cheong, P. H.-Y.; Houk, K. N.; Tanaka, F.; Barbas, C. F. *Journal of the American Chemical Society* **2006**, *128*, 1040–1041.

(54) Houk, K. N.; Cheong, P. H.-Y. *Nature* **2008**, *455*, 309–313.

(55) Cheong, P. H.-Y.; Zhang, H.; Thayumanavan, R.; Tanaka, F.; Houk, K. N.; Barbas, C. F. *Organic Letters* **2006**, *8*, 811–814.

(56) Houk, K.; Paddon-Row, M.; Rondan, N.; Wu, Y.; Brown, F.; Spellmeyer, D.; Metz, J.; Li, Y; Loncharich, R. *Science* **1986**, *231*, 1108–1117.

(57) Ardkhean, R.; Fletcher, S. P.; Paton, R. S. In *New Directions in the Modeling of Organometallic Reactions*; Springer International Publishing: 2020, pp 153–189.

(58) Planas-Iglesias, J.; Marques, S. M.; Pinto, G. P.; Musil, M.; Stourac, J.; Damborsky, J.; Bednar, D. *Biotechnology Advances* **2021**, *47*, 107696.

(59) Breslow, R.; Dong, S. D. *Chemical Reviews* **1998**, *98*, 1997–2012.

(60) Sanders, J. K. M. *Chemistry - A European Journal* **1998**, *4*, 1378–1383.

(61) Santamaria, J.; Martin, T.; Hilmersson, G.; Craig, S. L.; Rebek, J. *Proceedings of the National Academy of Sciences* **1999**, *96*, 8344–8347.

(62) Daver, H.; Harvey, J. N.; Rebek, J.; Himo, F. *Journal of the American Chemical Society* **2017**, *139*, 15494–15503.

(63) Daver, H.; Algarra, A. G.; Rebek, J.; Harvey, J. N.; Himo, F. *Journal of the American Chemical Society* **2018**, *140*, 12527–12537.

(64) Brea, O.; Daver, H.; Rebek, J.; Himo, F. *The Journal of Organic Chemistry* **2019**, *84*, 7354–7361.

(65) Welborn, V. V.; Li, W.-L.; Head-Gordon, T. *Nature Communications* **2020**, *11*, 415.

(66) Feng, Z.; Tantillo, D. J. *Journal of the American Chemical Society* **2021**, *143*, 1088–1097.

(67) Echenique, P.; Alonso, J. L. *Molecular Physics* **2007**, *105*, 3057–3098.

(68) Vuckovic, S.; Song, S.; Kozlowski, J.; Sim, E.; Burke, K. *Journal of Chemical Theory and Computation* **2019**, *15*, 6636–6646.

(69) Jensen, F., *Introduction to Computational Chemistry*; John Wiley and Sons, Inc.: Hoboken, New Jersey, USA, 2006.

(70) Thomas, L. H. *Mathematical Proceedings of the Cambridge Philosophical Society* **1927**, *23*, 542–548.

(71) Fermi, E. *Zeitschrift fur Physik* **1928**, *48*, 73–79.

(72) Haunschild, R.; Barth, A.; French, B. *Journal of Cheminformatics* **2019**, *11*, DOI: 10.1186/s13321-019-0395-y.

(73) Noorden, R. V.; Maher, B.; Nuzzo, R. *Nature* **2014**, *514*, 550–553.

(74) Hohenberg, P.; Kohn, W. *Physical Review* **1964**, *136*, B864–B871.

(75) Kanungo, B.; Zimmerman, P. M.; Gavini, V. *Nature Communications* **2019**, *10*, DOI: 10.1038/s41467-019-12467-0.

(76) Mardirossian, N.; Head-Gordon, M. *Molecular Physics* **2017**, *115*, 2315–2372.

(77) Goerigk, L.; Hansen, A.; Bauer, C.; Ehrlich, S.; Najibi, A.; Grimme, S. *Physical Chemistry Chemical Physics* **2017**, *19*, 32184–32215.

(78) Paz, J. L.; Hernandez, A. J., *New developments in quantum chemistry*; Transworld Research Network: Kerala, India, 2009.

(79) Kohn, W.; Sham, L. J. *Physical Review* **1965**, *140*, A1133–A1138.

(80) Lignères, V. L.; Carter, E. A. In *Handbook of Materials Modeling*; Springer Netherlands: 2005, pp 137–148.

(81) Perdew, J. P. In *AIP Conference Proceedings*, AIP: 2001.

(82) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Physical Review Letters* **1996**, *77*, 3865–3868.

(83) Becke, A. D. *Physical Review A* **1988**, *38*, 3098–3100.

(84) Lee, C.; Yang, W.; Parr, R. G. *Physical Review B* **1988**, *37*, 785–789.

(85) Tao, J.; Perdew, J. P.; Staroverov, V. N.; Scuseria, G. E. *Physical Review Letters* **2003**, *91*, 146401.

(86) Adamo, C.; Barone, V. *The Journal of Chemical Physics* **1999**, *110*, 6158–6170.

(87) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. *The Journal of Physical Chemistry* **1994**, *98*, 11623–11627.

(88) Becke, A. D. *The Journal of Chemical Physics* **1993**, *98*, 5648–5652.

(89) Zhao, Y.; Truhlar, D. G. *Theoretical Chemistry Accounts* **2007**, *120*, 215–241.

(90) Becke, A. D. *The Journal of Chemical Physics* **2014**, *140*, 18A301.

(91) Cohen, A. J.; Mori-Sánchez, P.; Yang, W. *Chemical Reviews* **2011**, *112*, 289–320.

(92) Puzder, A.; Dion, M.; Langreth, D. C. *The Journal of Chemical Physics* **2006**, *124*, 164105.

(93) Grimme, S. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2011**, *1*, 211–228.

(94) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. *The Journal of Chemical Physics* **2010**, *132*, 154104.

(95) Becke, A. D.; Johnson, E. R. *The Journal of Chemical Physics* **2005**, *123*, 154101.

(96) Grimme, S.; Ehrlich, S.; Goerigk, L. *Journal of Computational Chemistry* **2011**, *32*, 1456–1465.

(97) Goerigk, L.; Grimme, S. *Physical Chemistry Chemical Physics* **2011**, *13*, 6670.

(98) Wheeler, S. E.; Houk, K. N.; v. R. Schleyer, P.; Allen, W. D. *Journal of the American Chemical Society* **2009**, *131*, 2547–2560.

(99) Cremer, D. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2011**, *1*, 509–530.

(100) McKechnie, S.; Booth, G. H.; Cohen, A. J.; Cole, J. M. *The Journal of Chemical Physics* **2015**, *142*, 194114.

(101) Smith, J. S.; Nebgen, B. T.; Zubatyuk, R.; Lubbers, N.; Devereux, C.; Barros, K.; Tretiak, S.; Isayev, O.; Roitberg, A. E. *Nature Communications* **2019**, *10*, DOI: 10.1038/s41467-019-10827-4.

(102) Lee, J.; Head-Gordon, M. *Journal of Chemical Theory and Computation* **2018**, *14*, 5203–5219.

(103) Forsberg, B.; He, Z.; He, Y.; Cremer, D. *International Journal of Quantum Chemistry* **2000**, *76*, 306–330.

(104) Szalay, P. G.; Müller, T.; Gidofalvi, G.; Lischka, H.; Shepard, R. *Chemical Reviews* **2011**, *112*, 108–181.

(105) Řezáč, J.; Šimová, L.; Hobza, P. *Journal of Chemical Theory and Computation* **2012**, *9*, 364–369.

(106) Deustua, J. E.; Shen, J.; Piecuch, P. *The Journal of Chemical Physics* **2021**, *154*, 124103.

(107) Szabo, A.; Ostlund, N. S., *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory*, First; Dover Publications, Inc.: Mineola, 1996.

(108) Pritchard, B. P.; Altarawy, D.; Didier, B.; Gibson, T. D.; Windus, T. L. *Journal of Chemical Information and Modeling* **2019**, *59*, 4814–4820.

(109) Gulde, R.; Pollak, P.; Weigend, F. *Journal of Chemical Theory and Computation* **2012**, *8*, 4062–4068.

(110) Martin, J. M.; El-Yazal, J.; François, J.-P. *Molecular Physics* **1995**, *86*, 1437–1450.

(111) Sure, R.; Brandenburg, J. G.; Grimme, S. *ChemistryOpen* **2015**, *5*, 94–109.

(112) Zhang, J.; Valeev, E. F. *Journal of Chemical Theory and Computation* **2012**, *8*, 3175–3186.

(113) Kirschner, K. N.; Reith, D.; Heiden, W. *Soft Materials* **2020**, *18*, 200–214.

(114) Kestner, N. R.; Combariza, J. E. In *Reviews in Computational Chemistry*; John Wiley and Sons, Inc.: Hoboken, New Jersey, USA, 1999; Chapter 2, pp 99–132.

(115) Dyczmons, V. *Theoretica Chimica Acta* **1973**, *28*, 307–310.

(116) Häser, M.; Ahlrichs, R. *Journal of Computational Chemistry* **1989**, *10*, 104–111.

(117) Werner, H.-J.; Schütz, M. *The Journal of Chemical Physics* **2011**, *135*, 144116.

(118) Riplinger, C.; Neese, F. *The Journal of Chemical Physics* **2013**, *138*, 034106.

(119) Guo, Y.; Riplinger, C.; Becker, U.; Liakos, D. G.; Minenkov, Y.; Cavallo, L.; Neese, F. *The Journal of Chemical Physics* **2018**, *148*, 011101.

(120) Hasnip, P. J.; Refson, K.; Probert, M. I. J.; Yates, J. R.; Clark, S. J.; Pickard, C. J. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **2014**, *372*, 20130270.

(121) Sholl, D. S.; Steckel, J. A. In *Density Functional Theory*; John Wiley and Sons, Inc.: Hoboken, New Jersey, USA, 2009; Chapter 3, pp 49–81.

(122) Kratzer, P.; Neugebauer, J. *Frontiers in Chemistry* **2019**, *7*, 106.

(123) Christensen, A. S.; Kubař, T.; Cui, Q.; Elstner, M. *Chemical Reviews* **2016**, *116*, 5301–5337.

(124) Gaus, M.; Cui, Q.; Elstner, M. *Journal of Chemical Theory and Computation* **2011**, *7*, 931–948.

(125) Grimme, S.; Bannwarth, C.; Shushkov, P. *Journal of Chemical Theory and Computation* **2017**, *13*, 1989–2009.

(126) Addicoat, M. A.; Stefanovic, R.; Webber, G. B.; Atkin, R.; Page, A. J. *Journal of Chemical Theory and Computation* **2014**, *10*, 4633–4643.

(127) Sievers, C.; Noda, Y.; Qi, L.; Albuquerque, E. M.; Rioux, R. M.; Scott, S. L. *ACS Catalysis* **2016**, *6*, 8286–8307.

(128) Rush, L. E.; Pringle, P. G.; Harvey, J. N. *Angewandte Chemie* **2014**, *126*, 8816–8820.

(129) Harvey, J., *Computational chemistry*; Oxford University Press: Oxford, UK, 2018.

(130) Takano, Y.; Houk, K. N. *Journal of Chemical Theory and Computation* **2004**, *1*, 70–77.

(131) Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. *The Journal of Physical Chemistry B* **2009**, *113*, 6378–6396.

(132) Tomasi, J.; Mennucci, B.; Cammi, R. *Chemical Reviews* **2005**, *105*, 2999–3094.

(133) Szabo, A.; Ostlund, N. S., *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory*, First; Dover Publications Inc.: Mineola, New York, USA, 1996.

(134) Cusachs, L.; Politzer, P. *Chemical Physics Letters* **1968**, *1*, 529–531.

(135) Hirshfeld, F. L. *Theoretica Chimica Acta* **1977**, *44*, 129–138.

(136) Saha, S.; Roy, R. K.; Ayers, P. W. *International Journal of Quantum Chemistry* **2008**, *109*, 1790–1806.

(137) Fukui, K.; Yonezawa, T.; Shingu, H. *The Journal of Chemical Physics* **1952**, *20*, 722–725.

(138) Woodward, R. B.; Hoffmann, R. *Angewandte Chemie International Edition in English* **1969**, *8*, 781–853.

(139) Fukui, K. *Science* **1982**, *218*, 747–754.

(140) Albright, T. A.; Burdett, J. K.; Whangbo, M.-H., *Orbital interactions in chemistry, 2nd ed.* John Wiley and Sons, Inc.: Hoboken, New Jersey, USA, 2013.

(141) Cavallo, G.; Metrangolo, P.; Milani, R.; Pilati, T.; Priimagi, A.; Resnati, G.; Terraneo, G. *Chemical Reviews* **2016**, *116*, 2478–2601.

(142) Johnson, E. R.; Keinan, S.; Mori-Sánchez, P.; Contreras-García, J.; Cohen, A. J.; Yang, W. *Journal of the American Chemical Society* **2010**, *132*, 6498–6506.

(143) Laplaza, R.; Peccati, F.; Boto, R. A.; Quan, C.; Carbone, A.; Piquemal, J.-P.; Maday, Y.; Contreras-García, J. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2020**, *11*, e1497.

(144) Politzer, P.; Murray, J. S.; Peralta-Inga, Z. *International Journal of Quantum Chemistry* **2001**, *85*, 676–684.

(145) Hawkins, P. C. D. *Journal of Chemical Information and Modeling* **2017**, *57*, 1747–1756.

(146) Zhang, D. H.; Guo, H. *Annual Review of Physical Chemistry* **2016**, *67*, 135–158.

(147) Sweet, J. C.; Nowling, R. J.; Cickovski, T.; Sweet, C. R.; Pande, V. S.; Izaguirre, J. A. *Journal of Chemical Theory and Computation* **2013**, *9*, 3267–3281.

(148) Bussi, G.; Donadio, D.; Parrinello, M. *The Journal of Chemical Physics* **2007**, *126*, 014101.

(149) Iftimie, R.; Minary, P.; Tuckerman, M. E. *Proceedings of the National Academy of Sciences* **2005**, *102*, 6654–6659.

(150) Guvench, O.; MacKerell, A. D. In *Methods in Molecular Biology*; Humana Press: 2008, pp 63–88.

(151) Schlick, T. et al. *Annual Review of Biophysics* **2021**, *50*, DOI: `10.1146/annurev-biophys-091720-102019`.

(152) Boda, D.; Henderson, D. *Molecular Physics* **2008**, *106*, 2367–2370.

(153) Harrison, J. A.; Schall, J. D.; Maskey, S.; Mikulski, P. T.; Knippenberg, M. T.; Morrow, B. H. *Applied Physics Reviews* **2018**, *5*, 031104.

(154) Lewis-Atwell, T.; Townsend, P. A.; Grayson, M. N. *Tetrahedron* **2021**, *79*, 131865.

(155) Mitchell, T., *Machine Learning*; McGraw-Hill: New York City, New York, USA, 1997.

(156) Carleo, G.; Cirac, I.; Cranmer, K.; Daudet, L.; Schuld, M.; Tishby, N.; Vogt-Maranto, L.; Zdeborová, L. *Reviews of Modern Physics* **2019**, *91*, 045002.

(157) Cova, T. F.G. G.; Pais, A. A.C. C. *Frontiers in Chemistry* **2019**, *7*, 809.

(158) Behler, J.; Parrinello, M. *Physical Review Letters* **2007**, *98*, DOI: 10.1103/physrevlett.98.146401.

(159) Behler, J. *The Journal of Chemical Physics* **2011**, *134*, 074106.

(160) Smith, J. S.; Isayev, O.; Roitberg, A. E. *Chemical Science* **2017**, *8*, 3192–3203.

(161) Bartók, A. P.; Kondor, R.; Csányi, G. *Physical Review B* **2013**, *87*, 184115.

(162) Pozdnyakov, S. N.; Willatt, M. J.; Bartók, A. P.; Ortner, C.; Csányi, G.; Ceriotti, M. *Physical Review Letters* **2020**, *125*, 166001.

(163) Onat, B.; Ortner, C.; Kermode, J. R. *The Journal of Chemical Physics* **2020**, *153*, 144106.

(164) Thompson, A.; Swiler, L.; Trott, C.; Foiles, S.; Tucker, G. *Journal of Computational Physics* **2015**, *285*, 316–330.

(165) Huo, H.; Rupp, M. *arXiv* **2018**, DOI: arXiv:1704.06439.

(166) Braams, B. J.; Bowman, J. M. *International Reviews in Physical Chemistry* **2009**, *28*, 577–606.

(167) Van der Oord, C.; Dusson, G.; Csányi, G.; Ortner, C. *Machine Learning: Science and Technology* **2020**, *1*, 015004.

(168) Géron, A., *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow : concepts, tools, and techniques to build intelligent systems*; O'Reilly Media, Inc: Sebastopol, CA, 2019.

(169) Mueller, T.; Hernandez, A.; Wang, C. *The Journal of Chemical Physics* **2020**, *152*, 050902.

(170) Nielsen, M. A., *Neural networks and deep learning*; Determination press San Francisco, CA: 2015; Vol. 25.

(171) Bartók, A. P.; Csányi, G. *International Journal of Quantum Chemistry* **2015**, *115*, 1051–1057.

(172) Williams, C. K.; Rasmussen, C. E. Gaussian processes for regression: www.GaussianProcess.org/gp 1996.

(173) Völler, J.-S. *Nature Catalysis* **2020**, *3*, 774–774.

(174) Sepehrpour, H.; Fu, W.; Sun, Y.; Stang, P. J. *Journal of the American Chemical Society* **2019**, *141*, 14005–14020.

(175) Martí-Centelles, V.; Lawrence, A. L.; Lusby, P. J. *Journal of the American Chemical Society* **2018**, *140*, 2862–2868.

(176) Turcani, L.; Berardo, E.; Jelfs, K. E. *Journal of Computational Chemistry* **2018**, *39*, 1931–1942.

(177) Theil, E. C. *Current Opinion in Chemical Biology* **2011**, *15*, 304–311.

(178) Azuma, Y.; Zschoche, R.; Tinzl, M.; Hilvert, D. *Angewandte Chemie International Edition* **2016**, *55*, 1531–1534.

(179) Schmitt, D. L.; An, S. *Biochemistry* **2017**, *56*, 3184–3196.

(180) Rurup, W. F.; Snijder, J.; Koay, M. S. T.; Heck, A. J. R.; Cornelissen, J. J.L. M. *Journal of the American Chemical Society* **2014**, *136*, 3828–3832.

(181) Wiester, M. J.; Ulmann, P. A.; Mirkin, C. A. *Angewandte Chemie International Edition* **2011**, *50*, 114–137.

(182) Cook, T. R.; Zheng, Y.-R.; Stang, P. J. *Chemical Reviews* **2013**, *113*, 734–777.

(183) Pahima, E.; Zhang, Q.; Tiefenbacher, K.; Major, D. T. *Journal of the American Chemical Society* **2019**, *141*, 6234–6246.

(184) Zhang, Q.; Tiefenbacher, K. *Journal of the American Chemical Society* **2013**, *135*, 16213–16219.

(185) Ajami, D.; Rebek, J. *Accounts of Chemical Research* **2013**, *46*, 990–999.

(186) Kaanumalle, L. S.; Gibb, C. L. D.; Gibb, B. C.; Ramamurthy, V. *Journal of the American Chemical Society* **2005**, *127*, 3674–3675.

(187) Fang, Y.; Powell, J. A.; Li, E.; Wang, Q.; Perry, Z.; Kirchon, A.; Yang, X.; Xiao, Z.; Zhu, C.; Zhang, L.; Huang, F.; Zhou, H.-C. *Chemical Society Reviews* **2019**, *48*, 4707–4730.

(188) Caulder, D. L.; Powers, R. E.; Parac, T. N.; Raymond, K. N. *Angewandte Chemie International Edition* **1998**, *37*, 1840–1843.

(189) Fiedler, D.; Bergman, R. G.; Raymond, K. N. *Angewandte Chemie* **2004**, *116*, 6916–6919.

(190) Brown, C. J.; Bergman, R. G.; Raymond, K. N. *Journal of the American Chemical Society* **2009**, *131*, 17530–17531.

(191) Hart-Cooper, W. M.; Zhao, C.; Triano, R. M.; Yaghoubi, P.; Ozores, H. L.; Burford, K. N.; Toste, F. D.; Bergman, R. G.; Raymond, K. N. *Chemical Science* **2015**, *6*, 1383–1393.

(192) Kaphan, D. M.; Toste, F. D.; Bergman, R. G.; Raymond, K. N. *Journal of the American Chemical Society* **2015**, *137*, 9202–9205.

(193) Hastings, C. J.; Pluth, M. D.; Bergman, R. G.; Raymond, K. N. *Journal of the American Chemical Society* **2010**, *132*, 6938–6940.

(194) Pluth, M. D.; Bergman, R. G.; Raymond, K. N. *Science* **2007**, *316*, 85–88.

(195) Yoshizawa, M. *Science* **2006**, *312*, 251–254.

(196) Murase, T.; Nishijima, Y.; Fujita, M. *Journal of the American Chemical Society* **2011**, *134*, 162–164.

(197) Bolliger, J. L.; Belenguer, A. M.; Nitschke, J. R. *Angewandte Chemie International Edition* **2013**, *52*, 7958–7962.

(198) Cullen, W.; Misuraca, M. C.; Hunter, C. A.; Williams, N. H.; Ward, M. D. *Nature Chemistry* **2016**, *8*, 231–236.

(199) Cullen, W.; Metherell, A. J.; Wragg, A. B.; Taylor, C. G. P.; Williams, N. H.; Ward, M. D. *Journal of the American Chemical Society* **2018**, *140*, 2821–2828.

(200) Chakraborty, D.; Chattaraj, P. K. *The Journal of Physical Chemistry A* **2019**, *123*, 4513–4531.

(201) Goehry, C.; Besora, M.; Maseras, F. *ACS Catalysis* **2015**, *5*, 2445–2451.

(202) Welborn, V. V.; Head-Gordon, T. *The Journal of Physical Chemistry Letters* **2018**, *9*, 3814–3818.

(203) Norjmaa, G.; Maréchal, J.-D.; Ujaque, G. *Journal of the American Chemical Society* **2019**, *141*, 13114–13123.

(204) Lewis, J. E. M.; Gavey, E. L.; Cameron, S. A.; Crowley, J. D. *Chemical Science* **2012**, *3*, 778–784.

(205) Schmidt, A.; Casini, A.; Kühn, F. E. *Coordination Chemistry Reviews* **2014**, *275*, 19–36.

(206) Preston, D.; Barnsley, J. E.; Gordon, K. C.; Crowley, J. D. *Journal of the American Chemical Society* **2016**, *138*, 10578–10585.

(207) Vasdev, R. A. S.; Preston, D.; Crowley, J. D. *Chemistry - An Asian Journal* **2017**, *12*, 2513–2523.

(208) Preston, D.; Lewis, J. E. M.; Crowley, J. D. *Journal of the American Chemical Society* **2017**, *139*, 2379–2386.

(209) McMorran, D. A.; Steel, P. J. *Angewandte Chemie International Edition* **1998**, *37*, 3295–3297.

(210) Liao, P.; Langloss, B. W.; Johnson, A. M.; Knudsen, E. R.; Tham, F. S.; Julian, R. R.; Hooley, R. J. *Chemical Communications* **2010**, *46*, 4932.

(211) Clever, G.; Tashiro, S.; Shionoya, M. *Angewandte Chemie International Edition* **2009**, *48*, 7010–7012.

(212) Bloch, W. M.; Holstein, J. J.; Hiller, W.; Clever, G. H. *Angewandte Chemie International Edition* **2017**, *56*, 8285–8289.

(213) Zhu, R.; Bloch, W. M.; Holstein, J. J.; Mandal, S.; Schäfer, L. V.; Clever, G. H. *Chemistry - A European Journal* **2018**, *24*, 12976–12982.

(214) Li, R.-J.; Holstein, J. J.; Hiller, W. G.; Andréasson, J.; Clever, G. H. *Journal of the American Chemical Society* **2019**, *141*, 2097–2103.

(215) Chen, B.; Holstein, J. J.; Horiuchi, S.; Hiller, W. G.; Clever, G. H. *Journal of the American Chemical Society* **2019**, *141*, 8907–8913.

(216) Lewis, J. E.; Crowley, J. D. *Supramolecular Chemistry* **2013**, *26*, 173–181.

(217) McNeill, S. M.; Preston, D.; Lewis, J. E. M.; Robert, A.; Knerr-Rupp, K.; Graham, D. O.; Wright, J. R.; Giles, G. I.; Crowley, J. D. *Dalton Transactions* **2015**, *44*, 11129–11136.

(218) Jansze, S. M.; Wise, M. D.; Vologzhanina, A. V.; Scopelliti, R.; Severin, K. *Chemical Science* **2017**, *8*, 1901–1908.

(219) Chand, D. K.; Biradha, K.; Fujita, M. *Chemical Communications* **2001**, 1652–1653.

(220) Kishi, N.; Li, Z.; Yoza, K.; Akita, M.; Yoshizawa, M. *Journal of the American Chemical Society* **2011**, *133*, 11438–11441.

(221) Yamashina, M.; Sei, Y.; Akita, M.; Yoshizawa, M. *Nature Communications* **2014**, *5*, 4662.

(222) Schmidt, A.; Hollering, M.; Drees, M.; Casini, A.; Kühn, F. E. *Dalton Transactions* **2016**, *45*, 8556–8565.

(223) Duarte, F.; Bauer, P.; Barrozo, A.; Amrein, B. A.; Purg, M.; Åqvist, J.; Kamerlin, S. C. L. *The Journal of Physical Chemistry B* **2014**, *118*, 4351–4362.

(224) August, D. P.; Nichol, G. S.; Lusby, P. J. *Angewandte Chemie International Edition* **2016**, *55*, 15022–15026.

(225) Persson, I. *Pure and Applied Chemistry* **2010**, *82*, 1901–1917.

(226) Li, P.; Merz, K. M. *Chemical Reviews* **2017**, *117*, 1564–1686.

(227) Zhao, C.; Sun, Q.-F.; Hart-Cooper, W. M.; DiPasquale, A. G.; Toste, F. D.; Bergman, R. G.; Raymond, K. N. *Journal of the American Chemical Society* **2013**, *135*, 18802–18805.

(228) Sure, R.; Grimme, S. *Journal of Chemical Theory and Computation* **2015**, *11*, 3785–3801.

(229) Bootsma, A. N.; Wheeler, S. *ChemRxiv* **2019**, DOI: 10.26434/chemrxiv.8864204.v5.

(230) Grimme, S. *Chemistry - A European Journal* **2012**, *18*, 9955–9964.

(231) Caldararu, O.; Olsson, M. A.; Riplinger, C.; Neese, F.; Ryde, U. *Journal of Computer-Aided Molecular Design* **2016**, *31*, 87–106.

(232) Jensen, J. H. *Physical Chemistry Chemical Physics* **2015**, *17*, 12441–12451.

(233) Rizzi, A.; Murkli, S.; McNeill, J. N.; Yao, W.; Sullivan, M.; Gilson, M. K.; Chiu, M. W.; Isaacs, L.; Gibb, B. C.; Mobley, D. L.; Chodera, J. D. *Journal of Computer-Aided Molecular Design* **2018**, *32*, 937–963.

(234) Stewart, J. J. P. *Journal of Molecular Modeling* **2012**, *19*, 1–32.

(235) Uliana, M. P.; Servilha, B. M.; Alexopoulos, O.; de Oliveira, K. T.; Tormena, C. F.; Ferreira, M. A.; Brocksom, T. J. *Tetrahedron* **2014**, *70*, 6963–6973.

(236) Fringuelli, F., *The Diels Alder reaction : selected practical methods*; John Wiley and Sons, Inc.: Hoboken, New Jersey, USA, 2002.

(237) Corey, E. J. *Angewandte Chemie International Edition* **2002**, *41*, 1650–1667.

(238) Liu, Z.; Lee, J. H. Q.; Ganguly, R.; Vidović, D. *Chemistry - A European Journal* **2015**, *21*, 11344–11348.

(239) Martí-Centelles, V.; Duarte, F.; Lusby, P. J. *Israel Journal of Chemistry* **2018**, *59*, 257–266.

(240) Ess, D. H.; Houk, K. N. *Journal of the American Chemical Society* **2007**, *129*, 10646–10647.

(241) Schrodinger Schrödinger Release 2013-3., version 10.2, 2013.

(242) Frisch, M. J. et al. Gaussian˜09 Revision D.01., Gaussian Inc. Wallingford CT, 2009.

(243) Van der Spoel, D.; van Maaren, P. J.; Caleman, C. *Bioinformatics* **2012**, *28*, 752–753.

(244) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. *SoftwareX* **2015**, *1-2*, 19–25.

(245) Darden, T.; York, D.; Pedersen, L. *The Journal of Chemical Physics* **1993**, *98*, 10089–10092.

(246) Neese, F. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2017**, *8*, DOI: 10.1002/wcms.1327.

(247) Weigend, F.; Ahlrichs, R. *Physical Chemistry Chemical Physics* **2005**, *7*, 3297.

(248) Weigend, F. *Physical Chemistry Chemical Physics* **2006**, *8*, 1057–1065.

(249) Johnson, E. R.; Becke, A. D. *The Journal of Chemical Physics* **2006**, *124*, 174104.

(250) Zhao, Y.; Truhlar, D. G. *The Journal of Physical Chemistry A* **2005**, *109*, 5656–5667.

(251) Chai, J.-D.; Head-Gordon, M. *Physical Chemistry Chemical Physics* **2008**, *10*, 6615–6620.

(252) Riplinger, C.; Pinski, P.; Becker, U.; Valeev, E. F.; Neese, F. *The Journal of Chemical Physics* **2016**, *144*, 024109.

(253) Glendening, E. D.; Landis, C. R.; Weinhold, F. *Journal of Computational Chemistry* **2013**, *34*, 1429–1437.

(254) Lu, T.; Chen, F. *Journal of Computational Chemistry* **2011**, *33*, 580–592.

(255) Yoneya, M.; Yamaguchi, T.; Sato, S.; Fujita, M. *Journal of the American Chemical Society* **2012**, *134*, 14401–14407.

(256) Aaqvist, J.; Warshel, A. *Journal of the American Chemical Society* **1990**, *112*, 2860–2868.

(257) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *The Journal of Chemical Physics* **1983**, *79*, 926–935.

(258) Hirota, E.; Saito, S.; Endo, Y. *The Journal of Chemical Physics* **1979**, *71*, 1183–1187.

(259) Ribeiro, R. F.; Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. *The Journal of Physical Chemistry B* **2011**, *115*, 14556–14562.

(260) Tarumi, M.; Nakai, H. *Chemical Physics Letters* **2018**, *700*, 149–155.

(261) Nakai, H.; Ishikawa, A. *The Journal of Chemical Physics* **2014**, *141*, 174106.

(262) García-Simón, C.; Costas, M.; Ribas, X. *Chemical Society Reviews* **2016**, *45*, 40–62.

(263) Chen, L.; Chen, Q.; Wu, M.; Jiang, F.; Hong, M. *Accounts of Chemical Research* **2015**, *48*, 201–210.

(264) Pöthig, A.; Casini, A. *Theranostics* **2019**, *9*, 3150–3169.

(265) Burke, B. P.; Grantham, W.; Burke, M. J.; Nichol, G. S.; Roberts, D.; Renard, I.; Hargreaves, R.; Cawthorne, C.; Archibald, S. J.; Lusby, P. J. *Journal of the American Chemical Society* **2018**, *140*, 16877–16881.

(266) Therrien, B.; Süss-Fink, G.; Govindaswamy, P.; Renfrew, A. K.; Dyson, P. J. *Angewandte Chemie* **2008**, *120*, 3833–3836.

(267) Zava, O.; Mattsson, J.; Therrien, B.; Dyson, P. J. *Chemistry - A European Journal* **2010**, *16*, 1428–1431.

(268) Samanta, S. K.; Isaacs, L. *Coordination Chemistry Reviews* **2020**, *410*, 213181.

(269) Casini, A.; Woods, B.; Wenzel, M. *Inorganic Chemistry* **2017**, *56*, 14715–14729.

(270) Desmarets, C.; Ducarre, T.; Rager, M.; Gontard, G.; Amouri, H. *Materials* **2014**, *7*, 287–301.

(271) Chakrabarty, R.; Mukherjee, P. S.; Stang, P. J. *Chemical Reviews* **2011**, *111*, 6810–6918.

(272) Segura, J. L.; Royuela, S.; Mar Ramos, M. *Chemical Society Reviews* **2019**, *48*, 3903–3945.

(273) Erythropel, H. C. et al. *Green Chemistry* **2018**, *20*, 1929–1961.

(274) Neese, F.; Atanasov, M.; Bistoni, G.; Maganas, D.; Ye, S. *Journal of the American Chemical Society* **2019**, *141*, 2814–2824.

(275) Xu, J.; Hagler, A. *Molecules* **2002**, *7*, 566–600.

(276) Durand, D. J.; Fey, N. *Chemical Reviews* **2019**, *119*, 6561–6594.

(277) Greenaway, R. L.; Santolini, V.; Bennison, M. J.; Alston, B. M.; Pugh, C. J.; Little, M. A.; Miklitz, M.; Eden-Rump, E. G. B.; Clowes, R.; Shakil, A.; Cuthbertson, H. J.; Armstrong, H.; Briggs, M. E.; Jelfs, K. E.; Cooper, A. I. *Nature Communications* **2018**, *9*, 2849.

(278) Rosales, A. R.; Wahlers, J.; Limé, E.; Meadows, R. E.; Leslie, K. W.; Savin, R.; Bell, F.; Hansen, E.; Helquist, P.; Munday, R. H.; Wiest, O.; Norrby, P.-O. *Nature Catalysis* **2019**, *2*, 41–45.

(279) Hachmann, J.; Olivares-Amaya, R.; Atahan-Evrenk, S.; Amador-Bedolla, C.; Sánchez-Carrera, R. S.; Gold-Parker, A.; Vogt, L.; Brockway, A. M.; Aspuru-Guzik, A. *The Journal of Physical Chemistry Letters* **2011**, *2*, 2241–2251.

(280) Young, T. A.; Martí-Centelles, V.; Wang, J.; Lusby, P. J.; Duarte, F. *Journal of the American Chemical Society* **2019**, *142*, 1300–1310.

(281) Wilmer, C. E.; Leaf, M.; Lee, C. Y.; Farha, O. K.; Hauser, B. G.; Hupp, J. T.; Snurr, R. Q. *Nature Chemistry* **2012**, *4*, 83–89.

(282) Simon, C. M.; Mercado, R.; Schnell, S. K.; Smit, B.; Haranczyk, M. *Chemistry of Materials* **2015**, *27*, 4459–4475.

(283) Ong, S. P.; Richards, W. D.; Jain, A.; Hautier, G.; Kocher, M.; Cholia, S.; Gunter, D.; Chevrier, V. L.; Persson, K. A.; Ceder, G. *Computational Materials Science* **2013**, *68*, 314–319.

(284) Willems, T. F.; Rycroft, C. H.; Kazi, M.; Meza, J. C.; Haranczyk, M. *Microporous and Mesoporous Materials* **2012**, *149*, 134–141.

(285) Bucior, B. J.; Rosen, A. S.; Haranczyk, M.; Yao, Z.; Ziebel, M. E.; Farha, O. K.; Hupp, J. T.; Siepmann, J. I.; Aspuru-Guzik, A.; Snurr, R. Q. *Crystal Growth & Design* **2019**, *19*, 6682–6697.

(286) Rosen, A. S.; Notestein, J. M.; Snurr, R. Q. *Journal of Computational Chemistry* **2019**, *40*, 1305–1318.

(287) Miklitz, M.; Jiang, S.; Clowes, R.; Briggs, M. E.; Cooper, A. I.; Jelfs, K. E. *The Journal of Physical Chemistry C* **2017**, *121*, 15211–15222.

(288) Sturluson, A.; Huynh, M. T.; York, A. H. P.; Simon, C. M. *ACS Central Science* **2018**, *4*, 1663–1676.

(289) Kravchenko, O.; Varava, A.; Pokorny, F. T.; Devaurs, D.; Kavraki, L. E.; Kragic, D. *Journal of Chemical Information and Modeling* **2020**, *60*, 1302–1316.

(290) Miklitz, M.; Jelfs, K. E. *Journal of Chemical Information and Modeling* **2018**, *58*, 2387–2391.

(291) Harris, C. R. et al. *Nature* **2020**, *585*, 357–362.

(292) Virtanen, P. et al. *Nature Methods* **2020**, *17*, 261–272.

(293) Weininger, D. *Journal of Chemical Information and Modeling* **1988**, *28*, 31–36.

(294) Landrum, G., *RDKit: Open-source cheminformatics, v. 2019*; GitHub (https://github.com/rdkit/rdkit): 2019.

(295) Riniker, S.; Landrum, G. A. *Journal of Chemical Information and Modeling* **2015**, *55*, 2562–2574.

(296) Young, T. A.; Silcock, J. J.; Sterling, A. J.; Duarte, F. *Angewandte Chemie International Edition* **2021**, *60*, 4266–4274.

(297) Bannwarth, C.; Ehlert, S.; Grimme, S. *Journal of Chemical Theory and Computation* **2019**, *15*, 1652–1671.

(298) Stewart, J. J. P., *MOPAC2016*; Stewart Computational Chemistry: Colorado Springs, CO, USA, 2016.

(299) Valiev, M.; Bylaska, E.; Govind, N.; Kowalski, K.; Straatsma, T.; Van Dam, H.; Wang, D.; Nieplocha, J.; Apra, E.; Windus, T.; de Jong, W. *Computer Physics Communications* **2010**, *181*, 1477–1489.

(300) Kabsch, W. *Acta Crystallographica Section A* **1976**, *32*, 922–923.

(301) Young, D. C., *Computational Chemistry*; John Wiley and Sons, Inc.: Hoboken, New Jersey, USA, 2001, Appendix A. A.1.6. 330.

(302) Wales, D. J.; Doye, J. P. K. *The Journal of Physical Chemistry A* **1997**, *101*, 5111–5116.

(303) Zhou, H.-X.; Gilson, M. K. *Chemical Reviews* **2009**, *109*, 4092–4107.

(304) Jiang, C.; Jin, X.; Dong, Y.; Chen, M. *Journal of Chemical Information and Modeling* **2016**, *56*, 1132–1138.

(305) Rego, N.; Koes, D. *Bioinformatics* **2015**, *31*, 1322–1324.

(306) O'Boyle, N. M.; Vandermeersch, T.; Flynn, C. J.; Maguire, A. R.; Hutchison, G. R. *Journal of Cheminformatics* **2011**, *3*, 8.

(307) Ebejer, J.-P.; Morris, G. M.; Deane, C. M. *Journal of Chemical Information and Modeling* **2012**, *52*, 1146–1158.

(308) Spicer, R. L.; Stergiou, A. D.; Young, T. A.; Duarte, F.; Symes, M. D.; Lusby, P. J. *Journal of the American Chemical Society* **2020**, *142*, 2134–2139.

(309) Ronson, T. K.; Meng, W.; Nitschke, J. R. *Journal of the American Chemical Society* **2017**, *139*, 9698–9707.

(310) Smulders, M. M. J.; Zarra, S.; Nitschke, J. R. *Journal of the American Chemical Society* **2013**, *135*, 7039–7046.

(311) Ertl, P.; Schuffenhauer, A. *Journal of Cheminformatics* **2009**, *1*, 8.

(312) Kromann, J. C., *Calculate Root-mean-square deviation (RMSD) of Two Molecules Using Rotation. v.1.3.0*; GitHub (https://github.com/charnley/rmsd): 2019.

(313) Grimme, S.; Ehrlich, S.; Goerigk, L. *Journal of Computational Chemistry* **2011**, *32*, 1456–1465.

(314) Andrae, D.; Häußermann, U.; Dolg, M.; Stoll, H.; Preuß, H. *Theoretica Chimica Acta* **1990**, *77*, 123–141.

(315) Mantina, M.; Chamberlin, A. C.; Valero, R.; Cramer, C. J.; Truhlar, D. G. *The Journal of Physical Chemistry A* **2009**, *113*, 5806–5812.

(316) Boresch, S.; Tettinger, F.; Leitgeb, M.; Karplus, M. *The Journal of Physical Chemistry B* **2003**, *107*, 9535–9551.

(317) Mobley, D. L.; Gilson, M. K. *Annual Review of Biophysics* **2017**, *46*, 531–558.

(318) Armacost, K. A.; Riniker, S.; Cournia, Z. *Journal of Chemical Information and Modeling* **2020**, *60*, 1–5.

(319) Jensen, F., *Introduction to Computational Chemistry*; John Wiley and Sons, Inc.: Hoboken, New Jersey, USA, 1999.

(320) Woo, H.-J.; Roux, B. *Proceedings of the National Academy of Sciences* **2005**, *102*, 6825–6830.

(321) Lateef, A. B.; Hyne, J. B. *Canadian Journal of Chemistry* **1969**, *47*, 1369–1373.

(322) Low, P. S.; Somero, G. N. *Proceedings of the National Academy of Sciences* **1975**, *72*, 3014–3018.

(323) McQuarrie, D., *Statistical mechanics*; University Science Books: Sausalito, California, USA, 2000.

(324) Chan, L.; Morris, G. M.; Hutchison, G. R. *Journal of Chemical Theory and Computation* **2021**, *17*, 2099–2106.

(325) Pracht, P.; Grimme, S. *ChemRxiv* **2021**, DOI: 10.26434/chemrxiv.13626083.v1.

(326) Shan, X.; Burd, T. A. H.; Clary, D. C. *The Journal of Physical Chemistry A* **2019**, *123*, 4639–4657.

(327) Amzel, L. M. *Proteins: Structure, Function and Genetics* **1997**, *28*, 144–149.

(328) Liu, Z.; Patel, C.; Harvey, J. N.; Sunoj, R. B. *Physical Chemistry Chemical Physics* **2017**, *19*, 30647–30657.

(329) Gilson, M. K.; Irikura, K. K. *The Journal of Physical Chemistry B* **2010**, *114*, 16304–16317.

(330) Sumimoto, M.; Iwane, N.; Takahama, T.; Sakaki, S. *Journal of the American Chemical Society* **2004**, *126*, 10457–10471.

(331) Tanaka, R.; Yamashita, M.; Chung, L. W.; Morokuma, K.; Nozaki, K. *Organometallics* **2011**, *30*, 6742–6750.

(332) Deubel, D. V.; Lau, J. K.-C. *Chemical Communications* **2006**, 2451.

(333) Li, X.; Song, W.; Ke, X.; Xu, X.; Liu, P.; Houk, K. N.; liang Zhao, X.; Tang, W. *Chemistry - A European Journal* **2016**, *22*, 7079–7083.

(334) Tommaso, S. D.; Tognetti, V.; Sicilia, E.; Adamo, C.; Russo, N. *Inorganic Chemistry* **2010**, *49*, 9875–9883.

(335) Garza, A. J. *Journal of Chemical Theory and Computation* **2019**, *15*, 3204–3214.

(336) Herman, I. P., *Physics of the Human Body*; Springer International Publishing: 2016.

(337) Chakravorty, A.; Higham, J.; Henchman, R. H. *Journal of Chemical Information and Modeling* **2020**, *60*, 5540–5551.

(338) Ali, H. S.; Higham, J.; Henchman, R. H. *Entropy* **2019**, *21*, 750.

(339) General, I. J. *Journal of Chemical Theory and Computation* **2010**, *6*, 2520–2524.

(340) Zhao, Y. H.; Abraham, M. H.; Zissimos, A. M. *The Journal of Organic Chemistry* **2003**, *68*, 7368–7373.

(341) Khanna, V.; Ranganathan, S. *BMC Bioinformatics* **2009**, *10*, DOI: 10.1186/1471-2105-10-s15-s10.

(342) Schlitter, J. *Chemical Physics Letters* **1993**, *215*, 617–621.

(343) Karplus, M.; Kushick, J. N. *Macromolecules* **1981**, *14*, 325–332.

(344) Duan, L.; Liu, X.; Zhang, J. Z. *Journal of the American Chemical Society* **2016**, *138*, 5722–5728.

(345) Coates, J. In *Encyclopedia of Analytical Chemistry*; American Cancer Society: 2006.

(346) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G.E. M. *Journal of Computational Chemistry* **1997**, *18*, 1463–1472.

(347) Young, T.; Johnston-Wood, T.; Deringer, V. L.; Duarte, F. *Chemical Science* **2021**, DOI: 10.1039/d1sc01825f.

(348) Rumble, J., *CRC handbook of chemistry and physics : a ready-reference book of chemical and physical data*; CRC Press/Taylor & Francis Group: Boca Raton, Florida, USA, 2020.

(349) Buckingham, R. A. *Proceedings of the Royal Society A* **1938**, *168*, 264–283.

(350) Huang, Y.-S.; Thomann, H.-R. *Europhysics Letters* **2016**, *115*, 60001.

(351) Amore, P.; Fernández, F. M. *Physics Letters A* **2008**, *372*, 3149–3152.

(352) Ahnert, K.; Mulansky, M.; Simos, T. E.; Psihoyios, G.; Tsitouras, C.; Anastassi, Z. In *Odeint – Solving Ordinary Differential Equations in C++*, AIP: 2011.

(353) Sondhi, S. L.; Girvin, S. M.; Carini, J. P.; Shahar, D. *Reviews of Modern Physics* **1997**, *69*, 315–333.

(354) Grunwald, E.; Steel, C. *Journal of the American Chemical Society* **1995**, *117*, 5687–5692.

(355) Chodera, J. D.; Mobley, D. L. *Annual Review of Biophysics* **2013**, *42*, 121–142.

(356) Leung, D. H.; Bergman, R. G.; Raymond, K. N. *Journal of the American Chemical Society* **2008**, *130*, 2798–2805.

(357) Kruse, H.; Grimme, S. *The Journal of Chemical Physics* **2012**, *136*, 154101.

(358) Kutta, W. *Zeitschrift für angewandte Mathematik und Physik* **1901**, *46*, 435–453.

(359) Ahnert, K.; Mulansky, M. Boost ODE C++ library., `http://www.boost.org/`, 2021.

(360) Barone, V.; Cossi, M. *The Journal of Physical Chemistry A* **1998**, *102*, 1995–2001.

(361) Hourahine, B. et al. *The Journal of Chemical Physics* **2020**, *152*, 124101.

(362) Gaus, M.; Goez, A.; Elstner, M. *Journal of Chemical Theory and Computation* **2012**, *9*, 338–354.

(363) Plata, R. E.; Singleton, D. A. *Journal of the American Chemical Society* **2015**, *137*, 3811–3826.

(364) Wang, G.-W.; Sokolova, O. O.; Young, T. A.; Christodoulou, E. M. S.; Butts, C. P.; Bower, J. F. *Journal of the American Chemical Society* **2020**, *142*, 19006–19011.

(365) Branfoot, C.; Young, T. A.; Wass, D. F.; Pringle, P. G. *Dalton Transactions* **2021**, *50*, 7094–7104.

(366) Grimme, S.; Schreiner, P. R. *Angewandte Chemie International Edition* **2018**, *57*, 4170–4176.

(367) Cheng, G.-J.; Zhang, X.; Chung, L. W.; Xu, L.; Wu, Y.-D. *Journal of the American Chemical Society* **2015**, *137*, 1706–1725.

(368) Sameera, W. M. C.; Maseras, F. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2012**, *2*, 375–385.

(369) Simm, G. N.; Vaucher, A. C.; Reiher, M. *The Journal of Physical Chemistry A* **2019**, *123*, 385–399.

(370) Coley, C. W.; Eyke, N. S.; Jensen, K. F. *Angewandte Chemie International Edition* **2020**, *59*, 22858–22893.

(371) Dewyer, A. L.; Argüelles, A. J.; Zimmerman, P. M. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2018**, *8*, e1354.

(372) Zimmerman, P. M. *Journal of Computational Chemistry* **2015**, *36*, 601–611.

(373) Maeda, S.; Harabuchi, Y.; Takagi, M.; Taketsugu, T.; Morokuma, K. *The Chemical Record* **2016**, *16*, 2232–2248.

(374) Martínez-Núñez, E. *Journal of Computational Chemistry* **2015**, *36*, 222–234.

(375) Halgren, T. A.; Lipscomb, W. N. *Chemical Physics Letters* **1977**, *49*, 225–232.

(376) Peng, C.; Bernhard Schlegel, H. *Israel Journal of Chemistry* **1993**, *33*, 449–454.

(377) Dewar, M. J. S.; Healy, E. F.; Stewart, J. J. P. *Journal of the Chemical Society, Faraday Transactions 2* **1984**, *80*, 227.

(378) Elber, R.; Karplus, M. *Chemical Physics Letters* **1987**, *139*, 375–380.

(379) Dohm, S.; Bursch, M.; Hansen, A.; Grimme, S. *Journal of Chemical Theory and Computation* **2020**, *16*, 2002–2012.

(380) Suleimanov, Y. V.; Green, W. H. *Journal of Chemical Theory and Computation* **2015**, *11*, 4248–4259.

(381) Behn, A.; Zimmerman, P. M.; Bell, A. T.; Head-Gordon, M. *The Journal of Chemical Physics* **2011**, *135*, 224108.

(382) Gao, C. W.; Allen, J. W.; Green, W. H.; West, R. H. *Computer Physics Communications* **2016**, *203*, 212–225.

(383) Broadbelt, L. J.; Stark, S. M.; Klein, M. T. *Industrial & Engineering Chemistry Research* **1994**, *33*, 790–799.

(384) Rappoport, D.; Aspuru-Guzik, A. *Journal of Chemical Theory and Computation* **2019**, *15*, 4099–4112.

(385) Bergeler, M.; Simm, G. N.; Proppe, J.; Reiher, M. *Journal of Chemical Theory and Computation* **2015**, *11*, 5712–5722.

(386) Jacobson, L. D.; Bochevarov, A. D.; Watson, M. A.; Hughes, T. F.; Rinaldo, D.; Ehrlich, S.; Steinbrecher, T. B.; Vaitheeswaran, S.; Philipp, D. M.; Halls, M. D.; Friesner, R. A. *Journal of Chemical Theory and Computation* **2017**, *13*, 5780–5797.

(387) Chan, L.; Hutchison, G. R.; Morris, G. M. *Journal of Cheminformatics* **2019**, *11*, 32.

(388) Pracht, P.; Bohle, F.; Grimme, S. *Physical Chemistry Chemical Physics* **2020**, *22*, 7169–7192.

(389) Hagberg, A. A.; Schult, D. A.; Swart, P. J. In *Proceedings of the 7th Python in Science Conference*, ed. by Varoquaux, G.; Vaught, T.; Millman, J., Pasadena, CA USA, 2008, pp 11 –15.

(390) Crabtree, J. D.; Mehta, D. P. *Journal of Experimental Algorithmics* **2009**, *13*, 1.15.

(391) Dijkstra, E. W. *Numerische Mathematik* **1959**, *1*, 269–271.

(392) Goodman, J. M.; Silva, M. A *Tetrahedron Letters* **2003**, *44*, 8233–8236.

(393) Hamlin, T. A.; Swart, M.; Bickelhaupt, F. M. *ChemPhysChem* **2018**, *19*, 1315–1330.

(394) Ferro-Costas, D.; Martínez-Núñez, E.; Rodríguez-Otero, J.; Cabaleiro-Lago, E.; Estévez, C. M.; Fernández, B.; Fernández-Ramos, A.; Vázquez, S. A. *The Journal of Physical Chemistry A* **2018**, *122*, 4790–4800.

(395) Lee, C. W.; Taylor, B. L. H.; Petrova, G. P.; Patel, A.; Morokuma, K.; Houk, K. N.; Stoltz, B. M. *Journal of the American Chemical Society* **2019**, *141*, 6995–7004.

(396) Franke, R.; Selent, D.; Börner, A. *Chemical Reviews* **2012**, *112*, 5675–5732.

(397) Heck, R. F.; Breslow, D. S. *Journal of the American Chemical Society* **1961**, *83*, 4023–4027.

(398) Kégl, T. *RSC Advances* **2015**, *5*, 4304–4327.

(399) Decker, S. A.; Cundari, T. R. *Organometallics* **2001**, *20*, 2827–2841.

(400) Maeda, S.; Morokuma, K. *Journal of Chemical Theory and Computation* **2012**, *8*, 380–385.

(401) Jones, J. H. *Platinum Metals Review* **2000**, *44*.

(402) Armstrong, A.; Boto, R. A.; Dingwall, P.; Contreras-García, J.; Harvey, M. J.; Mason, N. J.; Rzepa, H. S. *Chemical Science* **2014**, *5*, 2057–2071.

(403) Mieusset, J.-L.; Brinker, U. H. *The Journal of Organic Chemistry* **2008**, *73*, 1553–1558.

(404) Domingo, L. R.; Sanz-Cervera, J. F.; Williams, R. M.; Picher, M. T.; Marco, J. A. *The Journal of Organic Chemistry* **1997**, *62*, 1662–1667.

(405) Schneebeli, S. T.; Hall, M. L.; Breslow, R.; Friesner, R. *Journal of the American Chemical Society* **2009**, *131*, 3965–3973.

(406) Neese, F. *Journal of Computational Chemistry* **2003**, *24*, 1740–1747.

(407) Neese, F.; Wennmohs, F.; Hansen, A.; Becker, U. *Chemical Physics* **2009**, *356*, 98–109.

(408) He, K.; Allen, W. D. *Journal of Chemical Theory and Computation* **2016**, *12*, 3571–3582.

(409) Havel, T. F. In *Encyclopedia of Computational Chemistry*; John Wiley and Sons, Inc.: Hoboken, New Jersey, USA, 2002.

(410) Ioannidis, E. I.; Gani, T. Z. H.; Kulik, H. J. *Journal of Computational Chemistry* **2016**, *37*, 2106–2117.

(411) Ounkham, W. L.; Frost, B. J. In *Encyclopedia of Inorganic and Bioinorganic Chemistry*; John Wiley and Sons, Inc.: Hoboken, New Jersey, USA, 2017, pp 1–26.

(412) Zhao, L.; Pan, S.; Holzmann, N.; Schwerdtfeger, P.; Frenking, G. *Chemical Reviews* **2019**, *119*, 8781–8845.

(413) In *IUPAC Compendium of Chemical Terminology*; IUPAC: Research Triagle Park, NC, 2020.

(414) Kwan, E. E.; Zeng, Y.; Besser, H. A.; Jacobsen, E. N. *Nature Chemistry* **2018**, *10*, 917–923.

(415) Jónsson, H.; Mills, G.; Jacobsen, K. W. In *Classical and Quantum Dynamics in Condensed Phase Simulations*, World Scientific: 1998.

(416) Henkelman, G.; Uberuaga, B. P.; Jónsson, H. *The Journal of Chemical Physics* **2000**, *113*, 9901–9904.

(417) Henkelman, G.; Jónsson, H. *The Journal of Chemical Physics* **2000**, *113*, 9978–9985.

(418) Martin-Gondre, L.; Crespos, C.; Larrégaray, P.; Rayez, J.; Conte, D.; van Ootegem, B. *Chemical Physics* **2010**, *367*, 136–147.

(419) Vaucher, A. C.; Reiher, M. *Journal of Chemical Theory and Computation* **2018**, *14*, 3091–3099.

(420) Larsen, A. H. et al. *Journal of Physics: Condensed Matter* **2017**, *29*, 273002.

(421) Pribram-Jones, A.; Gross, D. A.; Burke, K. *Annual Review of Physical Chemistry* **2015**, *66*, 283–304.

(422) Krongchon, K.; Busemeyer, B.; Wagner, L. K. *The Journal of Chemical Physics* **2017**, *146*, 124129.

(423) Zhang, J.; Zhang, H.; Wu, T.; Wang, Q.; van der Spoel, D. *Journal of Chemical Theory and Computation* **2017**, *13*, 1034–1043.

(424) Basdogan, Y.; Keith, J. A. *Chemical Science* **2018**, *9*, 5341–5346.

(425) Thiel, W., *Multiscale simulation methods in molecular sciences winter school; lecture notes*; Grotendorst, J., Attig, N., Blügel, S., Marx, D., Eds.; NIC: Jülich, Germany, 2009.

(426) Grajciar, L.; Heard, C. J.; Bondarenko, A. A.; Polynski, M. V.; Meeprasert, J.; Pidko, E. A.; Nachtigall, P. *Chemical Society Reviews* **2018**, *47*, 8307–8348.

(427) Stirling, A.; Nair, N. N.; Lledós, A.; Ujaque, G. *Chemical Society Reviews* **2014**, *43*, 4940–4952.

(428) Hassanali, A. A.; Cuny, J.; Verdolino, V.; Parrinello, M. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **2014**, *372*, 20120482.

(429) Bartók, A. P.; Payne, M. C.; Kondor, R.; Csányi, G. *Physical Review Letters* **2010**, *104*, 136403.

(430) Devereux, C.; Smith, J. S.; Davis, K. K.; Barros, K.; Zubatyuk, R.; Isayev, O.; Roitberg, A. E. *Journal of Chemical Theory and Computation* **2020**, *16*, 4192–4202.

(431) Frenkel, D.; Smit, B., *Understanding Molecular Simulation: From Algorithms to Applications*, 2nd ed.; Academic Press: Cambridge, Massachusetts, 2002.

(432) Lindorff-Larsen, K.; Maragakis, P.; Piana, S.; Eastwood, M. P.; Dror, R. O.; Shaw, D. E. *PLOS ONE* **2012**, *7*, ed. by Muller, D. J., e32131.

(433) Noé, F.; Tkatchenko, A.; Müller, K.-R.; Clementi, C. *Annual Review of Physical Chemistry* **2020**, *71*, 361–390.

(434) Unke, O. T.; Koner, D.; Patra, S.; Käser, S.; Meuwly, M. *Machine Learning: Science and Technology* **2020**, *1*, 013001.

(435) Khaliullin, R. Z.; Eshet, H.; Kühne, T. D.; Behler, J.; Parrinello, M. *Nature Materials* **2011**, *10*, 693–697.

(436) Sosso, G. C.; Miceli, G.; Caravati, S.; Giberti, F.; Behler, J.; Bernasconi, M. *The Journal of Physical Chemistry Letters* **2013**, *4*, 4241–4246.

(437) Niu, H.; Bonati, L.; Piaggi, P. M.; Parrinello, M. *Nature Communications* **2020**, *11*, 2654.

(438) Cheng, B.; Mazzola, G.; Pickard, C. J.; Ceriotti, M. *Nature* **2020**, *585*, 217–220.

(439) Deringer, V. L.; Bernstein, N.; Csányi, G.; Ben Mahmoud, C.; Ceriotti, M.; Wilson, M.; Drabold, D. A.; Elliott, S. R. *Nature* **2021**, *589*, 59–64.

(440) Ang, S. J.; Wang, W.; Schwalbe-Koda, D.; Axelrod, S.; Gómez-Bombarelli, R. *Chem* **2021**, *7*, 738–751.

(441) Cole, D. J.; Mones, L.; Csányi, G. *Faraday Discussions* **2020**, *224*, 247–264.

(442) Rufa, D. A.; Bruce Macdonald, H; Fass, J.; Wieder, M.; Grinaway, P. B.; Roitberg, A. E.; Isayev, O.; Chodera, J. D. *bioRxiv* **2020**, DOI: 10.1101/2020.07.29.227959.

(443) Gastegger, M.; Behler, J.; Marquetand, P. *Chemical Science* **2017**, *8*, 6924–6935.

(444) Li, J.; Reiser, P.; Boswell, B. R.; Eberhard, A.; Burns, N. Z.; Friederich, P.; Lopez, S. A. *Chemical Science* **2021**, *12*, 5302–5314.

(445) Behler, J. *Angewandte Chemie International Edition* **2017**, *56*, 12828–12840.

(446) Schütt, K. T.; Arbabzadah, F.; Chmiela, S.; Müller, K. R.; Tkatchenko, A. *Nature Communications* **2017**, *8*, 13890.

(447) Unke, O. T.; Meuwly, M. *Journal of Chemical Theory and Computation* **2019**, *15*, 3678–3693.

(448) Chmiela, S.; Tkatchenko, A.; Sauceda, H. E.; Poltavsky, I.; Schütt, K. T.; Müller, K.-R. *Science Advances* **2017**, *3*, e1603015.

(449) Shapeev, A. V. *Multiscale Modeling & Simulation* **2016**, *14*, 1153–1173.

(450) Zuo, Y.; Chen, C.; Li, X.; Deng, Z.; Chen, Y.; Behler, J.; Csányi, G.; Shapeev, A. V.; Thompson, A. P.; Wood, M. A.; Ong, S. P. *The Journal of Physical Chemistry A* **2020**, *124*, 731–745.

(451) Szlachta, W. J.; Bartók, A. P.; Csányi, G. *Physical Review B* **2014**, *90*, 104108.

(452) Deringer, V. L.; Csányi, G. *Physical Review B* **2017**, *95*, 094203.

(453) Bartók, A. P.; Kermode, J.; Bernstein, N.; Csányi, G. *Physical Review X* **2018**, *8*, 041048.

(454) Sivaraman, G.; Krishnamoorthy, A. N.; Baur, M.; Holm, C.; Stan, M.; Csányi, G.; Benmore, C.; Vázquez-Mayagoitia, Á. *npj Computational Materials* **2020**, *6*, 104.

(455) Mocanu, F. C.; Konstantinou, K.; Lee, T. H.; Bernstein, N.; Deringer, V. L.; Csányi, G.; Elliott, S. R. *The Journal of Physical Chemistry B* **2018**, *122*, 8998–9006.

(456) Dral, P. O.; Owens, A.; Dral, A.; Csányi, G. *The Journal of Chemical Physics* **2020**, *152*, 204110.

(457) Veit, M.; Jain, S. K.; Bonakala, S.; Rudra, I.; Hohl, D.; Csányi, G. *Journal of Chemical Theory and Computation* **2019**, *15*, 2574–2586.

(458) Deringer, V. L.; Caro, M. A.; Csányi, G. *Nature Communications* **2020**, *11*, 5461.

(459) Cheng, B.; Engel, E. A.; Behler, J.; Dellago, C.; Ceriotti, M. *Proceedings of the National Academy of Sciences* **2019**, *116*, 1110–1115.

(460) Schran, C.; Brezina, K.; Marsalek, O. *The Journal of Chemical Physics* **2020**, *153*, 104105.

(461) Podryabinkin, E. V.; Tikhonov, E. V.; Shapeev, A. V.; Oganov, A. R. *Physical Review B* **2019**, *99*, 064114.

(462) Artrith, N.; Behler, J. *Physical Review B* **2012**, *85*, 045439.

(463) Gubaev, K.; Podryabinkin, E. V.; Hart, G. L.; Shapeev, A. V. *Computational Materials Science* **2019**, *156*, 148–156.

(464) Yang, M.; Bonati, L.; Polino, D.; Parrinello, M. *Catalysis Today* **2021**, DOI: 10.1016/j.cattod.2021.03.018.

(465) Smith, J. S.; Nebgen, B.; Lubbers, N.; Isayev, O.; Roitberg, A. E. *The Journal of Chemical Physics* **2018**, *148*, 241733.

(466) Podryabinkin, E. V.; Shapeev, A. V. *Computational Materials Science* **2017**, *140*, 171–180.

(467) Zhang, L.; Lin, D.-Y.; Wang, H.; Car, R.; E, W. *Physical Review Materials* **2019**, *3*, 023804.

(468) Zhang, Y.; Wang, H.; Chen, W.; Zeng, J.; Zhang, L.; Wang, H.; E, W. *Computer Physics Communications* **2020**, *253*, 107206.

(469) Zhang, L.; Han, J.; Wang, H.; Car, R.; E, W. *Physical Review Letters* **2018**, *120*, 143001.

(470) Vandermause, J.; Torrisi, S. B.; Batzner, S.; Xie, Y.; Sun, L.; Kolpak, A. M.; Kozinsky, B. *npj Computational Materials* **2020**, *6*, 20.

(471) Jinnouchi, R.; Lahnsteiner, J.; Karsai, F.; Kresse, G.; Bokdam, M. *Physical Review Letters* **2019**, *122*, 225701.

(472) Jinnouchi, R.; Miwa, K.; Karsai, F.; Kresse, G.; Asahi, R. *The Journal of Physical Chemistry Letters* **2020**, *11*, 6946–6955.

(473) Zeng, J.; Cao, L.; Xu, M.; Zhu, T.; Zhang, J. Z. H. *Nature Communications* **2020**, *11*, 5713.

(474) Kearnes, S. *Trends in Chemistry* **2021**, *3*, 77–79.

(475) Cawley, G. C.; Talbot, N. L. *Journal of Machine Learning Research* **2010**, *11*, 2079–2107.

(476) Kramer, C.; Gedeck, P. *Journal of Chemical Information and Modeling* **2010**, *50*, 1961–1969.

(477) Li, Y.; Yang, J. *Journal of Chemical Information and Modeling* **2017**, *57*, 1007–1012.

(478) Chen, L.; Cruz, A.; Ramsey, S.; Dickson, C. J.; Duca, J. S.; Hornak, V.; Koes, D. R.; Kurtzman, T. *PLOS ONE* **2019**, *14*, ed. by Zhang, Y., e0220113.

(479) Kovács, D. P.; McCorkindale, W.; Lee, A. A. *Nature Communications* **2021**, *12*, 1695.

(480) Sheridan, R. P. *Journal of Chemical Information and Modeling* **2013**, *53*, 783–790.

(481) Pickard, C. J.; Needs, R. J. *Journal of Physics: Condensed Matter* **2011**, *23*, 053201.

(482) Bernstein, N.; Csányi, G.; Deringer, V. L. *npj Computational Materials* **2019**, *5*, 99.

(483) Mahoney, M. W.; Drineas, P. *Proceedings of the National Academy of Sciences* **2009**, *106*, 697–702.

(484) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. *Journal of Chemical Theory and Computation* **2015**, *11*, 2087–2096.

(485) Sukuba, I.; Chen, L.; Probst, M.; Kaiser, A. *Molecular Simulation* **2018**, *47*, 113–118.

(486) Wengert, S.; Csányi, G.; Reuter, K.; Margraf, J. T. *Chemical Science* **2021**, DOI: 10.1039/D0SC05765G.

(487) Zheng, L.; Chen, M.; Sun, Z.; Ko, H.-Y.; Santra, B.; Dhuvad, P.; Wu, X. *The Journal of Chemical Physics* **2018**, *148*, 164505.

(488) Li, P.; Merz, K. M. *Chemical Reviews* **2017**, *117*, 1564–1686.

(489) Ohtaki, H.; Radnai, T. *Chemical Reviews* **1993**, *93*, 1157–1204.

(490) Pratihar, S.; Ma, X.; Homayoon, Z.; Barnes, G. L.; Hase, W. L. *Journal of the American Chemical Society* **2017**, *139*, 3570–3590.

(491) Ess, D. H.; Wheeler, S. E.; Iafe, R. G.; Xu, L.; Çelebi-Ölçüm, N.; Houk, K. N. *Angewandte Chemie International Edition* **2008**, *47*, 7592–7601.

(492) Xie, J.; Hase, W. L. *Science* **2016**, *352*, 32–33.

(493) Szabó, I.; Czakó, G. *The Journal of Physical Chemistry A* **2017**, *121*, 9005–9019.

(494) Tirado-Rives, J.; Jorgensen, W. L. *The Journal of Physical Chemistry A* **2019**, *123*, 5713–5717.

(495) Lee, S.; Goodman, J. M. *Journal of the American Chemical Society* **2020**, *142*, 9210–9219.

(496) Caramella, P.; Quadrelli, P.; Toma, L. *Journal of the American Chemical Society* **2002**, *124*, 1130–1131.

(497) Hjorth Larsen, A. et al. *Journal of Physics: Condensed Matter* **2017**, *29*, 273002.

(498) Young, T.; Johnston-Wood, T. gap-train., 2020.

(499) Himanen, L.; Jäger, M. O.; Morooka, E. V.; Federici Canova, F.; Ranawat, Y. S.; Gao, D. Z.; Rinke, P.; Foster, A. S. *Computer Physics Communications* **2020**, *247*, 106949.

(500) Gaus, M.; Goez, A.; Elstner, M. *Journal of Chemical Theory and Computation* **2013**, *9*, 338–354.

(501) Mortensen, J. J.; Hansen, L. B.; Jacobsen, K. W. *Physical Review B* **2005**, *71*, 035109.

(502) Enkovaara, J et al. *Journal of Physics: Condensed Matter* **2010**, *22*, 253202.

(503) Zhang, Y.; Yang, W. *Physical Review Letters* **1998**, *80*, 890–890.

(504) Kühne, T. D. et al. *The Journal of Chemical Physics* **2020**, *152*, 194103.

(505) Neese, F. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2012**, *2*, 73–78.

(506) Berendsen, H.; van der Spoel, D.; van Drunen, R. *Computer Physics Communications* **1995**, *91*, 43–56.

(507) Morawietz, T.; Singraber, A.; Dellago, C.; Behler, J. *Proceedings of the National Academy of Sciences* **2016**, *113*, 8368–8373.

(508) Singraber, A.; Morawietz, T.; Behler, J.; Dellago, C. *Journal of Chemical Theory and Computation* **2019**, *15*, 3075–3092.

(509) deB Darwent, B., *Bond Dissociation Energies in Simple Molecules: NSRDS-NBS 31*; U.S. National Bureau of Standards.: Washington DC, 1970, p 20.

(510) Busch, S.; Pardo, L. C.; O'Dell, W. B.; Bruce, C. D.; Lorenz, C. D.; McLain, S. E. *Physical Chemistry Chemical Physics* **2013**, *15*, 21023.

(511) Gradshteyn, I. S.; Ryzhik, I. M., *Table of integrals, series, and products*, Seventh, Translated from the Russian, Translation edited and with a preface by Alan Jeffrey and Daniel Zwillinger, With one CD-ROM (Windows, Macintosh and UNIX); Elsevier/Academic Press, Amsterdam: 2015.

(512) Vlasov, V. M. *Russian Chemical Reviews* **2006**, *75*, 765–796.