

# **Computational modelling of ligand shape and interactions for medicines design**

A thesis submitted to the University of Manchester for the degree of  
Doctor of Philosophy  
in the Faculty of Medical and Human Sciences

**2016**

**Panichakorn Jaiyong**

**Manchester Pharmacy School**

# Contents

List of Figures.....	6
List of Tables.....	10
List of Abbreviations.....	12
Abstract.....	16
Declaration.....	17
Copyright Statement.....	17
Dedication.....	19
Acknowledgements.....	20
The Author.....	21
CHAPTER 1 Introduction .....	22
CHAPTER 2 Computational methods.....	30
2.1    Quantum mechanics (QM) .....	31
2.1.1    The Schrödinger equation .....	31
2.1.2    Basis sets .....	35
2.1.3 <i>Ab initio</i> quantum mechanical methods .....	38
2.1.4    Semiempirical quantum mechanical (SQM) methods .....	42
2.1.5    Density functional theory (DFT).....	45
2.1.6    Density functional tight binding (DFTB).....	48
2.2    Molecular Mechanics (MM).....	54
2.3    Quantum mechanics/Molecular mechanics (QM/MM) .....	55

2.4 Molecular docking.....	59
2.4.1 Conformational search methods.....	59
2.4.2 Scoring functions .....	61
2.5 Molecular classes of the hits.....	64
2.6 Virtual screening .....	66
CHAPTER 3 Identification of putative ligands for anti-inflammatory treatment.....	70
3.1 Inflammation .....	70
3.2 Boron-based inhibitors .....	74
3.3 Aims .....	84
3.4 Methods .....	85
3.4.1 Generation of BC23 conformations .....	85
3.4.2 Preparation of compound databases .....	86
3.4.3 ROCS shape-based screening .....	88
3.5 Results and discussion.....	90
3.5.1 DFT analysis of 2-APB analogue compounds .....	90
3.5.2 ROCS screening .....	99
3.6 Conclusions and outlook .....	108
CHAPTER 4 Protein-ligand docking guided by semiempirical quantum mechanical potentials.....	110
4.1 Introduction .....	110
4.2 Aims .....	118
4.3 Implementation of QM/MM MC module in Gauussion 09 .....	119

4.4	Computational details.....	123
4.4.1	Systems .....	123
4.4.2	GOLD docking setup .....	124
4.4.3	QM/MM MC docking parameters .....	125
4.5	Results and discussion.....	128
4.5.1	Implementation of in-house QM/MM MC module in the Gaussian 09 program.....	128
4.5.2	Monte Carlo parameter optimisation .....	132
4.5.3	QM/MM MC docking .....	143
4.6	Conclusions and outlook .....	167
CHAPTER 5 Assessment of quantum chemical methods for modelling $\beta$ -cyclohexatin conformation and its interaction with a single-layer graphene sheet: modelling the vicinal diol interaction in carbohydrates .....		174
5.1	Introduction .....	174
5.2	Aims .....	181
5.3	Computational methods.....	182
5.3.1	$\beta$ CD geometries and conformational energetics <i>in vacuo</i> .....	183
5.3.2	Intermolecular interactions at $\beta$ CD/graphene interface .....	183
5.4	Results and Discussion .....	185
5.4.1	Gas-phase $\beta$ CD conformation .....	185
5.4.2	$\beta$ CD/Graphene complex .....	199
5.5	Conclusions and outlook .....	206

CHAPTER 6 Conclusions .....	211
Appendix A.....	215
Appendix B.....	220
Appendix C.....	235
Appendix D.....	237
References.....	247

**Word Count: 57966**

# List of Figures

Figure 2.1 Illustration of ROCS algorithm. <sup>98</sup> .....	69
Figure 3.1 Production of active IL-1 $\beta$ cytokines in monocytic cells (adapted from Brough <i>et al.</i> <sup>101</sup> ). .....	72
Figure 3.2 Structure of dipeptidyl boronic acid, Xaa-boroPro.....	75
Figure 3.3 Cyclisation equilibrium of Xaa-boroPro. <sup>110</sup> .....	76
Figure 3.4 Acid-base equilibrium of benzoxaborole. <sup>119</sup> .....	80
Figure 3.5 Diphenylborinic acid esters. .....	82
Figure 3.6 Chemical structure of (i) 2-APB, (ii) BC7, and (iii) BC23. ....	84
Figure 3.7 The query compound of BC23 in (i) 2D chemical structure (ii) 3D geometries.....	86
Figure 3.8 Frontier orbital interactions. <sup>143</sup> .....	91
Figure 3.9 HOMO and LUMO energies (in kcal/mol) of five boron-based heterocyclic compounds. ....	92
Figure 3.10 Summary of DFT analysis of five 2-APB analogue compounds that can inhibit release of IL-1 $\beta$ .....	98
Figure 3.11 Hypothetical forms of biologically active BC23.....	99
Figure 3.12 2D structures and 3D overlays of top screened compound against BC23 ranked by different scoring functions: TanimotoCombo, ShapeTanimoto and ColorTanimoto.....	105
Figure 3.13 Selected compounds for testing by biological assay. ....	107
Figure 3.14 AN0128 forming (i) 5-membered ring and (ii) 6-membered ring.....	109

Figure 4.1 The preferred binding mode of His-Zn <sup>2+</sup> via in-plane sp <sup>2</sup> lone pair electrons of the N <sub><i>e</i></sub> . <sup>162</sup> .....	113
Figure 4.2 The binding site of hCA II in one-letter codes: E=glutamate, H=histidine, N=asparagine, Q=glutamine, and T=threonine.....	114
Figure 4.3 Structures of the synthesised sulphonamide derivatives. <sup>168</sup> .....	115
Figure 4.4 Hydrolytic deamination of cytidine via a tetrahedral transition state. <sup>169</sup> ....	116
Figure 4.5 Structures of cytidine deaminase (CDA) and its transition-state analogue inhibitors: zebularine (ZEB), 3,4-hydrate zebularine (ZEB-H <sub>2</sub> O), and 3,4-dihydrozebularine (DHZ). ....	117
Figure 4.6 Active species of zinc-coordinated hydroxide ion [Zn-(OH) <sup>-</sup> ] in the lowest-energy s1 configuration closest to a crystallographic structure. <sup>172</sup> .....	118
Figure 4.7 Structures of inhibitors for cytidine deaminase (i) 3,4-dihydrozebularine (DHZ) from PDB ID 1CTT and (ii) 3-deazacytidine (DAC) from PDB ID 1ALN; for human carbonic anhydrase II (iii) <i>N</i> -(2,3-difluoro-benzyl)-4-sulfamoyl-benzamide (FSB) from PDB ID 1G52 and (iv) dorzolamide (DZA) from PDB ID 1CIL.....	124
Figure 4.8 Structure of a chemical probe of sulfamoyl methanol used in optimising MC parameters. ....	132
Figure 4.9 Acceptance rate of sulfamoyl methanol binding to CAII using (i) 5.0 bohrs of Δ <i>x<sub>t</sub></i> and (ii) 10.0 bohrs of Δ <i>x<sub>t</sub></i> on 2k MC trajectories for values of <i>vdw<sub>cutoff</sub></i> at 2.5, 3.0, 3.5 or 4.0 bohrs and <i>d<sub>max</sub></i> at 2.0, 5.0 or 10.0 bohrs. ....	133
Figure 4.10 Acceptance rate of sulfamoyl methanol binding to CAII (i) 5.0 bohrs of Δ <i>x<sub>t</sub></i> and (ii) 10.0 bohrs of Δ <i>x<sub>t</sub></i> on MC trajectories of 1k, 2k and 4k using <i>d<sub>max</sub></i> at 2.0, 5.0, and 10.0 bohrs; <i>vdw<sub>cutoff</sub></i> was set at 3.0 bohrs.....	136
Figure 4.11 RMSD in Å of accepted poses for CDA receptor and a pair of consecutive poses that causes a sudden change in RMSD values. ....	139

Figure 4.12 RMSD in Å of accepted poses for hCA II receptor and poses that (i) cause a sudden change in RMSD values (ii) present similar structural orientation. ....	140
Figure 4.13 Volume sphere of docked poses explored by QM/MM MC scheme in the binding site of cytidine deaminase with (i) 3,4-dihydrozebularine (DHZ) from PDB ID 1CTT and (ii) 3-deazacytidine (DAC) from PDB ID 1ALN; for human carbonic anhydrase II with (iii) N-(2,3-difluoro-benzyl)-4-sulfamoyl-benzamide (FSB) from PDB ID 1G52 and (iv) dorzolamide (DZA) from PDB ID 1CIL. ....	142
Figure 4.14 Superposition of (i) top-scored GOLD docked pose (cyan stick), (ii) lowest energy QM/MM MC docked pose (pink stick), and (iii) lowest RMSD QM/MM MC docked pose (purple stick) onto the native conformation (grey stick) of DHZ (PDB ID 1CTT) at the binding site of CDA.....	152
Figure 4.15 Superposition of (i) optimised native pose (green stick) (ii) optimised lowest energy docked pose (pink stick), and (iii) optimised lowest RMSD docked pose (purple stick) onto the native pose (grey stick) of DHZ (PDB ID 1CTT) at the binding site of CDA. ....	153
Figure 4.16 Superposition of (i) top-scored GOLD docked pose (cyan stick), (ii) lowest energy QM/MM MC docked pose (pink stick), and (iii) lowest RMSD QM/MM MC docked pose (purple stick) onto the native conformation (grey stick) of DAC (PDB ID 1ALN) at the binding site of CDA. ....	156
Figure 4.17 Superposition of (i) optimised native pose (green stick) (ii) optimised lowest energy docked pose (pink stick), and (iii) optimised lowest RMSD docked pose (purple stick) onto the native pose (grey stick) of DAC (PDB ID 1ALN) at the binding site of CDA. ....	157
Figure 4.18 Superposition of (i) top-scored GOLD docked pose (cyan stick), (ii) lowest energy QM/MM MC docked pose (pink stick), and (iii) lowest RMSD QM/MM MC	

docked pose (purple stick) onto the native conformation (grey stick) of FSB (PDB ID 1G52) at the binding site of hCA II.....	161
Figure 4.19 Superposition of (i) optimised native pose (green stick) (ii) optimised lowest energy docked pose (pink stick), and (iii) optimised lowest RMSD docked pose (purple stick) onto the native pose (grey stick) of FSB (PDB ID 1G52) at the binding site hCA II.....	162
Figure 4.20 Superposition of (i) top-scored GOLD docked pose (cyan stick), (ii) lowest energy QM/MM MC docked pose (pink stick), and (iii) lowest RMSD QM/MM MC docked pose (purple stick) onto the native conformation (grey stick) of DZA (PDB ID 1CIL) at the binding site of hCA II.....	165
Figure 4.21 Superposition of (i) optimised native pose (green stick) (ii) optimised lowest energy docked pose (pink stick), and (iii) optimised lowest RMSD docked pose (purple stick) onto the native pose (grey stick) of DZA (PDB ID 1CIL) at the binding site of hCA II.....	166
Figure 5.1 $\beta$ CD geometries optimised at M06-2X/def2-TZVPP (i) cccw (ii) cccc (iii) cwccw (iv) cwcc (v) O-cwcc (vi) O-cccc .....	177
Figure 5.2 Structures and relative energies of 1,2 ethanediols (i) cw/tGg– (ii) cc/gGg– (iii) cw/tGg– optimised at MP2/def2-TZVPP (in grey) and PM6-DH2 (in cyan) and (iv) cc/gGg– optimised at MP2/def2-TZVPP (in grey) and PM6-DH2 (in cyan).....	192
Figure 5.3 The cwcc $\beta$ CD/C <sub>1006</sub> H <sub>88</sub> complex optimised at PM6-DH2 level of theory (C2C3 rim at graphene surface) .....	202
Figure 5.4 Interaction energies in kcal/mol of graphene $\beta$ CD/C <sub>1006</sub> H <sub>88</sub> computed at PM6-DH2, PM7 and DFTB3 levels of theory [Note: *open geometries collapsed to cccc] .....	203

## List of Tables

Table 3.1 Boronic acid inhibitors described in the literature.....	77
Table 3.2 Benzoxaborole inhibitors described in the literature.....	81
Table 3.3 Shape, symmetries and energies (in Hartrees) of the HOMO and LUMO of five boron-based heterocyclic compounds.....	93
Table 3.4 Mulliken partial atomic charges on boron atom (in $e$ ) and bond lengths of heterocyclic boron compounds derived from DFT calculations.....	96
Table 3.5 Top 25 hits of BC23 ranked by the <i>TanimotoCombo</i> scoring function.....	102
Table 3.6 Top 25 hits of BC23 ranked by the <i>ShapeTanimoto</i> scoring function.....	103
Table 3.7 Top 25 hits of BC23 ranked by the <i>ColorTanimoto</i> scoring function.....	104
Table 4.1 Comparison of standard two-layer ONIOM energies and nonstandard MC-based QM/MM energies in a FSB-hCA II complex (PDB ID 1G52). ....	131
Table 4.2 The number of van der Waals clashes between QM and MM atoms when $\Delta x_t$ was set at 5.0 and 10.0 bohrs for sampling sulfamoyl methanol on 2k MC trajectories in the hCA II site; $vdw_{cutoff}$ was set at 2.5, 3.0, 3.5, and 4.0 bohrs; $d_{max}$ was set at 2.0, 5.0, and 10.0 bohrs. ....	135
Table 4.3 Ligand poses generated by GA search algorithm and then ranked by empirical ChemPLP score compared to ONIOM energy-based score.....	145
Table 4.4 The complex energy (in kcal/mol) and RMSD (in Å) of the best poses scoring by the QM/MM function; Ligand poses sampling from two consecutive 4k MC runs and a single 10k MC run.....	148
Table 5.1 Relative gas-phase conformational energetics in kcal/mol of $\beta$ CD.....	187
Table 5.2 Relative gas-phase conformational energetics in kcal/mol of glucose. ....	190

Table 5.3 Relative gas-phase conformational energetics in kcal/mol of 1,2-ethanediols relative to cw/tGg <sup>-</sup> . The relative energy between cc/gGg <sup>-</sup> and cw/tGg <sup>-</sup> conformers, Ecc-cw, is in bold.....	193
Table 5.4 Geometrical data of cw/tGg <sup>-</sup> and cc/gGg <sup>-</sup> conformers of 1,2-ethanediols. <sup>a</sup> Distances in Å and angles in degrees.....	197
Table 5.5 Mulliken atomic partial charges (in e) of selected 1,2-ethanediols. ....	198
Table 5.6 Interaction energies (IEs) of C <sub>96</sub> H <sub>24</sub> with C2C3 rim of βCD complexes (in kcal/mol). BSSE-corrected IEs are given in parentheses.....	200
Table 5.7 Interaction energies (IEs) of βCD/C <sub>1006</sub> H <sub>88</sub> complexes in kcal/mol. Collapsed open structures labelled in italics. ....	204
Table 5.8 Interaction energy (IE) and overall binding energies (BE) of βCD/C <sub>1006</sub> H <sub>88</sub> complexes at C2C3 rim in gas-phase (g) and solution (aq) and binding energy corrected for cc bias (BE') at PM6-DH2 level, in kcal/mol. IE computed in regard to respective conformer; BE and BE' computed in regard to cccw conformer. Collapsed open structure labelled in italics. ....	206

## List of Abbreviations

2-APB	2-Aminoethyl diphenylborinate
2D	two dimensional
3D	three dimensional
ADME	absorption, distribution, metabolism and excretion
AM1	Austin Model 1
BE	binding energy
BMMUE	balanced mean mean unsigned error
CARD	caspase activation and recruitment
CC	coupled cluster
CCSD(T)/CBS	single- and double-electron excitations iteratively and triple-electron excitation perturbatively coupled cluster with complete basis sets
CD	cyclodextrin
CDA	cytidine deaminases
CI	configuration interaction
clogP	calculated octanol-water partition coefficient
CNDO	complete neglect of differential overlap
CNT	carbon nanotube
DAC	3-deazacytidine
DAMPs	danger-associated molecular patterns
DDP-IV	dipeptidyl peptidase IV
DFT	density functional theory

DFTB	density functional tight binding
DFT-D	dispersion-corrected DFT
DHZ	3,4-dihydrozebularine
DZ	double-zeta
DZA	dorzolamide
FEP	free energy perturbation
FSB	<i>N</i> -(2,3-difluoro-benzyl)-4-sulfamoyl-benzamide
GA	genetic algorithm
GOLD	genetic optimisation for ligand docking
hCA II	human carbonic anhydrase II
HF	Hartree-Fock
HOMO	highest occupied molecular orbital
HTS	high-throughput screening
IC <sub>50</sub>	a half maximal inhibitory concentration
IE	interaction energy
IL-1 $\beta$	interleukin-1 $\beta$
INDO	intermediate neglect of differential overlap
IOps	internal command options
LPS	lipopolysaccharide
LUMO	lowest unoccupied molecular orbital
MC	Monte Carlo
MD	molecular dynamics
MIC	minimum inhibitory concentration
MM	molecular mechanics
MMFF	Merck molecular force field

MNDO	modified neglect of diatomic overlap
MP2	second-order Møller-Plesset perturbation
MUE	mean unsigned error
MW	molecular weight
NACHT	central nucleotide binding domain
NCI	National Cancer Institute
NDDO	neglect of diatomic differential overlap
NLRs	nucleotide-binding, oligomerisation domain (nod)-like receptors
NMR	nuclear magnetic resonance spectroscopy
PAMPs	pathogen-associated molecular patterns
PDB	Protein Data Bank
PM3	Parametric Method number 3
PM6	Parametric Method number 6
PM7	Parametric Method number 7
PMIs	principal moments of inertia
PYD	N-terminal pyrin domain
QM	quantum mechanics
QSAR	quantitative structure-activity relationships
RMSD	root-mean-square deviation
RMSE	root-mean-square error
ROCS	rapid overlay of chemical structures
SCC	self-consistent charge
SCC-DFTB	self-consistent charge density functional tight-binding
SCF	self-consistent field
SLIM	shaped-based ligand matching with binding pocket

SQM	semiempirical quantum mechanics
STO	Slater-type orbitals
SZ	single-zeta
TLRs	Toll-like receptors
TZ	triple-zeta
USR	Ultrafast Shape Recognition
ZEB	zebularine

## Abstract

Computational methods have been extensively developed at various levels of approximation in recent years to model biomolecular interactions and for rational drug design. This research work aims to explore the feasibility of using quantum mechanical (QM) methods within the two broad categories of *in silico* ligand-based and structure-based drug design. First, density functional theory at the M06L level of theory was employed to examine structure-activity relationships of boron-based heterocyclic compounds, anti-inflammatory inhibitors targetting the interleukin-1 $\beta$  (IL-1 $\beta$ ) cytokine. Our findings from computed energies and shapes of the molecular orbitals provide understanding of electronic effects associated with the inhibitory activity. We also found that the boron atom, specifically its electrostatic polarity, appears to be essential for the anti-IL-1 $\beta$  activity as evidenced by the biological assay of non-boron analogues selected from the ligand-based virtual screening results. Secondly, we aimed to dock ligands at the active sites of zinc-containing metalloproteins with reasonable computational cost and with accuracy. Therefore, an in-house docking scheme based on a Monte Carlo sampling algorithm using the semiempirical PM6/AMBER force field scoring function was compiled for the first time within the Gaussian 09 program. It was applied to four test cases, docking to cytidine deaminase and human carbonic anhydrase II proteins. The docking results show the method's promise in resolving false-positive docking poses and improving the predicted binding modes over a conventional docking scheme. Finally, semiempirical QM methods which include dispersion and hydrogen-bond corrections were assessed for modelling conformations of  $\beta$ -cyclodextrin ( $\beta$ CD) and their adsorption on graphene. The closed *in vacuo*  $\beta$ CD cccw conformer was found to be in the lowest energy, in good agreement with previous *ab initio* QM studies. DFTB3, PM6-DH2 and PM7 methods were applied to model the intermolecular interactions of large  $\beta$ CD/graphene complexes, over a thousand atoms in size. We found that the binding preference of  $\beta$ CD on graphene is in a closed conformation via its C2C3 rim, agreeing with reported experimental and computational findings.

## **Declaration**

The author declares that no portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

## **Copyright Statement**

- i. The author of this thesis (including any appendices and/or schedules to this thesis) owns certain copyright or related rights in it (the “Copyright”) and she has given The University of Manchester certain rights to use such Copyright, including for administrative purposes.
- ii. Copies of this thesis, either in full or in extracts and whether in hard or electronic copy, may be made only in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has from time to time. This page must form part of any such copies made.
- iii. The ownership of certain Copyright, patents, designs, trademarks and other intellectual property (the “Intellectual Property”) and any reproductions of copyright

works in the thesis, for example graphs and tables (“Reproductions”), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.

iv. Further information on the conditions under which disclosure, publication and commercialisation of this thesis, the Copyright and any Intellectual Property University IP Policy (see <http://documents.manchester.ac.uk/display.aspx?DocID=24420>), in any relevant Thesis restriction declarations deposited in the University Library, The University Library’s regulations (see <http://www.library.manchester.ac.uk/about/regulations/>) and in The University’s policy on Presentation of Theses

---

Dedicated To My Parents

## Acknowledgements

First, I would like to express my sincere gratitude to my supervisor Dr. Richard A. Bryce for his continued support and patience throughout my PhD study. His guidance and immense knowledge has helped me in all the time of conducting research and writing of this thesis.

Besides my supervisor, I would like to thank my advisor Dr. Sally Freeman for valuable advice and encouragement. My sincere thanks also go to Dr. Mark A. Vincent, Dr. Daniel Mucs, Dr. Kepa K. Burusco-Goni, and Dr. Neil J. Bruce for their technical support and useful comments. I would like to acknowledge Research Computing Services at the University of Manchester for providing the assistance of IT services and the computational resources, particularly Pen Richardson for her excellent support.

I am grateful for the financial support that has been provided by the Ministry of Science, Royal Thai Government as well as Department of Chemistry, Faculty of Science and Technology, Thammasat University, Thailand for the leaving-to-study opportunity.

Also, I would like to thank all the people I have met during my stay in the UK, especially the past and present members of Dr. Bryce's Computational Biophysics and Drug Design group. Thank you for friendship I have received in Manchester Pharmacy School.

Special thanks go to Assoc. Prof. Ekasith Somsook for enlightening me by the first glance of computational research, and Assoc. Prof. Yuthana Tantirungrotechai for his support and encouragement to pursue my PhD study.

I thank my family, Dad and Mom who always give me unconditional love and care, no matter how I am. Thanks to my sisters for taking care of my parents during four years I have been away from home. Last, I would like to thank all friends and the people who have taken part in my PhD journey.

## The Author

Panichakorn Jaiyong holds a degree of Bachelor of Science in Chemistry with honours in the second class, obtained from Chiang Mai University, Thailand in 2001. In 2004, she graduated with Master of Science in Physical Chemistry from Mahidol University, Thailand.

## CHAPTER 1 Introduction

A newly developed medicine ideally should not only produce the desired therapeutic response with minimal side-effects, but also be a clearly better treatment than existing medicines. A drug design and discovery scheme involves many stages from target identification, target validation, hit discovery process, hit-to-lead phase, lead optimisation, preclinical development to clinical trials.

Drug discovery typically begins with identifying the druggable targets, for example, enzymes, genes, and nucleic acids. The identification of a target represents the association of the individual targets with a particular disease state. This also allows us to explore whether the mechanism-based side effects of a drug could occur by the target modulation. There is a less chance of undesirable side effects if a drug is highly selective activity to its target; therefore, the validation of target specificity and selectivity is of importance. Validation techniques could range from *in vitro* assays (i.e., on isolated cells, tissues or enzymes) to *in vivo* tests (i.e., on animals). Finding a balance between good activity of a drug at the desired target and minimal activity at the other targets becomes a challenge. This indeed affects the direction of a drug development scheme.

The subsequent stage after the target validation is the hit identification. During this stage, compound screening assays are carried out for identifying the *hits*, putative compounds that have the desired pharmacological activity when interact with the target. From the results of screening assays, the dose-response curves are obtained to generate a half maximal inhibitory concentration ( $IC_{50}$ ). This  $IC_{50}$  is commonly used to compare

a potency of candidate compounds. Typically, a potency of the hits can be in a range of 100 nM to 5  $\mu$ M.<sup>1</sup> More detail about molecular classes of the hits will be discussed in Chapter 2. The most promising series of the hits will then be refined to be more potent and selective compounds, so-called the *leads* in the stage of a hit-to-lead optimisation. The intensive structure-activity relationship (SAR) around the core structure is studied systematically to improve both pharmacodynamic (i.e., potency, affinity, efficacy and selectivity) and pharmacokinetic properties (i.e., absorption, distribution, metabolism and excretion (ADME)) of the lead compounds. The next phase is lead optimisation, aims to maintain those optimised properties while improving on structural deficiencies of the leads. Lead compounds at this stage will be declared as the preclinical candidates.

Preclinical and clinical phases are in a drug development scheme, conducted before a full-scale production. The promising drug candidates are required to be tested in the preclinical stage to ensure that they are safe and effective for administration in human clinical trials. This includes toxicity, pharmacology, and stability tests as well as drug metabolism and formulation studies. Based on the obtained knowledge, dosage, schedule and suitable administration of the drugs can be recommended for the first time use in human. After filing the investigational new drug application, the clinical trials are conducted in three or four phases. Phase I studies normally use less than a hundred of healthy volunteers to determine the drug's safety, dosage and possible side effects. These studies represent the first exposure to the drug in human, and provide knowledge to design studies in the next phase. In phase II, the drug is administered to hundreds of patients to assess drug efficacy for a specific disease. The studies in this phase identify the most effective dose showing the optimal benefit-risk profile of drug administration. Next, phase III studies are carried out on a large scale involving thousands of patients to

investigate the statistical data of administering the drug and the less common side effects. They are also used to confirm the findings of drug effectiveness and safety profile from previous studies. The studies in phase IV are post-approval trials that monitor the long-term use of the drug in specific patients. Rare side effects that could occur are also observed in this phase. Additional information regarding the safety profile, efficacy or manufacturing processes for the drug is usually required to be submitted.

Developing a new medicine from basic research into the market is a long and costly process. Each industrial project might initially screen hundreds of thousands up to millions of compounds for the hits and leads discovery. However, the estimated success rate for the preclinical testing is about 5 to 20 out of 5,000 to 10,000 compounds. Only 2 to 5 of the preclinical candidates is successfully transferred into clinical phase I, II and III.<sup>2</sup> A compound is finally approved and selected to market in an average of 10-15 years.<sup>3</sup> Although the overall cost of drug development is still controversial, the total capitalised cost of an approved drug has been estimated by using financial models in a range of US\$868 million to US\$1,778 million.<sup>2, 4</sup> To reduce costs and shorten these time-consuming processes, computational techniques are becoming established as powerful tools in drug discovery and development process.

Computational modelling has contributed in drug design and discovery through advances in virtual screening, predicting of pharmacological properties, and estimating the receptor-ligand interactions. Virtual screening can be used to filter compounds for suitable assays. More detail of virtual screening will be discussed in Chapter 2 and 3. To identify the promising lead compounds, computer-aided molecular design utilises the

structural knowledge of either the target or known ligands. A set of active ligands can be used to derive a correlation between significant the structural features, namely pharmacophores, and the observed biological activity. The changes in a bioactivity are related to molecular variations in a set of ligand compounds. The structure-activity relationship of a chemical series is useful, especially in the lead optimisation phase, for tuning the ADME and toxicity profile while maintaining the binding affinity. Both pharmacophore modelling and quantitative structure-activity relationship (QSAR) are the common computational techniques in ligand-based drug design. The concept of ligand-based molecular design using shape similarity forms the basis of the research work in Chapter 3.

In contrast, structure-based drug design provides an insight into the receptor-ligand interactions. It relies on the knowledge of the three dimensional (3D) structures derived from the experimental X-ray crystallography or nuclear magnetic resonance spectroscopy (NMR) as well as from homology modelling. The binding site with chemical functionalities, that interacts with a ligand is identified in the first step. From the binding site scaffold, a putative ligand can be built up by assembling individual atoms or molecular fragments in a stepwise manner. This ligand construction is called *de novo* design, able to identify novel ligand structures which may not be found in compound libraries. In addition to the *de novo* ligand design, molecular docking is one of the most well-known methods in structure-based drug design. Given the ligand geometries, it predicts possible binding modes of a ligand and estimates their binding affinity. Sampling of ligand conformations and scoring functions are the key success of docking algorithm (more detail will be discussed in Chapter 2). Docking of chemical libraries becomes an essential approach for structure-based lead generation. In Chapter

4, the in-house docking scheme is developed and aimed to improve the false-predicted docking results.

With the increasing computer power since the 1980s, computational modelling has had a significant impact on rational drug design and discovery. Marketed drug compounds which were discovered with assistance by computational approaches demonstrate the impact of computer-aided drug design on industrial drug discovery. Nelfinavir, amprenavir, lopinavir, zanamivir, oseltamivir, captopril and dorzolamide are examples of marketed drug compounds, discovered with the aid of structure-based drug design.<sup>5</sup> The first three compounds (nelfinavir, amprenavir, and lopinavir) are the HIV protease inhibitors, approved and marketed during the late 1990s. They alter a death sentence of HIV infected patients to a chronic condition. Zanamivir and oseltamivir are the neuraminidase inhibitors, marketed for the treatment of influenza. Captopril is targeted at the angiotensin converting enzyme and used to treat hypertension whereas dorzolamide is the inhibitor of carbonic anhydrase II used for the treatment of glaucoma. Other examples from the ligand-based drug design are norfloxacin, losartan and zolmitriptan.<sup>6</sup> Norfloxacin is the fluoroquinolone antibiotic, launched in 1984 while losartan was approved later to treat hypertension in 1995. Zolmitriptan is used for the acute treatment of migraine, leading its second generation of triptan that has sales of US\$352 million in 2005.

For applications in rational drug design, computational methods have been extensively developed at various levels of approximation but generally belong to the broad classes of molecular mechanics (MM) or quantum mechanics (QM). MM involves the application of classical mechanics to model an atom as its basic particle. Bonded

molecular interactions are treated as “springs” with equilibrium distances derived from experimental or calculated bond lengths. MM methods can rapidly provide three-dimensional (3D) molecular geometries in conformational analysis. The potential energy of systems is calculated using force fields. With currently available force fields, MM methods enables us to simulate dynamics of proteins up to millions of atoms; however, they are sometimes sensitive to the training sets used for developing the force field parameters and are typically unable to describe changes in electronic structure as a function of geometry. QM methods allow a realistic representation of the electronic polarisability of atoms and molecules which is otherwise inaccessible by the fixed point charge models used in classical force fields.

QM effects, e.g. bond breaking, bond formation, polarisation and charge transfer are important in describing the nature of interactions between molecules. They cannot be ignored in seeking to accurately model for molecular design. To account for these effects, QM methods have been employed at different levels of theory. *Ab initio* QM methods are based on the Schrödinger equation which describes how electrons in the molecule behave. The electron distribution is derived from mathematical wavefunctions which arise from solution of the Schrödinger equation. These methods are accurate but computationally expensive; therefore, they are only feasible for small-to-medium size systems. Some approximations can be made in solving integrals in the Schrödinger equation. This becomes the concept of semiempirical quantum mechanics (SQM) methods which compromise between computational accuracy and cost. Unlike the *ab initio* QM and SQM methods, density functional theory (DFT) methods derive the electron distribution directly from electron density functions. DFT methods are usually

faster than *ab initio*, but slower than SQM methods. More detail of the computational methods used in this work will be discussed in Chapter 2.

This research work presents the use of computational QM methods in two main categories of *in silico* ligand-based and structure-based drug design. In Chapter 3, the M06L density functional theory method was employed to examine structure-activity relationships of boron-based heterocyclic inhibitors against interleukin-1 $\beta$  cytokine involved in inflammation. One of the lead boron compounds was selected as the query for ligand-based virtual screening. The screens were carried out based on shape and chemical similarity between the lead molecule and a library of compounds. This aimed at exploring potential carbon-based analogues in ligand design of potential anti-inflammatory treatments.

In Chapter 4, we focus on the modelling of noncovalent interactions that occur in ligand-protein complexes. Despite recent advances in computational power and capacity, modelling of noncovalent interactions, e.g. dispersion, hydrogen bonding and hydrophobic effects, in biomolecular systems remains challenging due to the weak nature of interactions of less than 5 kcal/mol in strength and the complexity of the systems. This chapter demonstrates the use of a hybrid QM/MM energy function as a scoring function in structure-based molecular docking of zinc-containing metalloenzymes. To model ligand-protein systems, sampling of relevant conformational states is equally as important as accurately estimating the binding affinities. Therefore, an in-house docking scheme based on Monte Carlo sampling algorithm and semiempirical PM6/AMBER force field scoring function was compiled for the first time in the Gaussian 09 program. This was aimed to achieve an expedient docking tool for

studying ligand-metalloprotein interactions at a reasonable cost with sufficient accuracy. Its feasibility to be used for resolving the false-positive docking poses predicted by a commercially-available docking program was also investigated.

In Chapter 5, we focus on the assessment of SQM methods including dispersion and hydrogen-bond corrections for modelling conformations of  $\beta$ -cyclodextrin ( $\beta$ CD), a cyclic oligosaccharide.  $\beta$ CD displays a hydrophobic cavity and hydrophilic rims such that it is used as a pharmaceutical excipient; it is able to form water-soluble inclusion complexes with poorly soluble lipophilic drugs. Recently,  $\beta$ CD has been used to enhance the aqueous solubility of graphene via noncovalent functionalisation. Graphene has been extensively studied due to its incredible properties such as high strength, stiffness and surface area, high electrical and thermal conductivity with just one-atom thickness. The enhanced solubility in conjunction with other distinctive properties of graphene itself shows that the noncovalent  $\beta$ CD/graphene composite is a very promising material. It can be used as a biosensor via host-guest interactions at the graphene surface. In this chapter, some selected QM methods were applied to explore the structures of  $\beta$ CD/graphene complex and their intermolecular interactions.

## CHAPTER 2 Computational methods

Over recent decades, significant advances in computer hardware and computational chemistry software packages have been made and applied to compute molecular structures, properties and reactivity in a variety of chemical applications. No universal and transferable computational method exists for all system types. The choice of applied methodologies is influenced by the problem at hand, generally a trade-off between accuracy and computational cost. In terms of accuracy, the performance of developed methods is required to be evaluated prior to use. This can be achieved by comparing the computed data of the test sets with the experimental/biological data, for example, geometrical bond lengths, atomisation energies, heats of formation, vibrational frequencies, pKa, IC<sub>50</sub>, etc. Not only using the experimentally derived data, benchmark studies are often carried out against the computed data at the high-level *ab initio* methods.<sup>7</sup> Some computational methods such as semiempirical quantum mechanics (SQM) have used the training sets for parameterisation, the process to derive numerical values for the key parameters and/or functional forms for the approximated term. These methods can be high in accuracy for the test sets that are similar to the applied training sets. To obtain satisfied accuracy in computed results within a reasonable cost, the system size, computer power and theoretical sophistication of the selected methods are necessarily of concern.

This chapter will provide the overview of the computational methods used in this research covered in Chapters 3, 4 and 5.

## 2.1 Quantum mechanics (QM)

One grand challenge in computational chemistry is how to solve the Schrödinger equation for a system of interest. Quantum mechanics (QM) methods range from *ab initio* methods, semiempirical quantum mechanics (SQM) and density functional theory (DFT) approaches and can be distinguished by how they handle the electron-electron repulsion term in the Schrödinger equation. One such approach is the Hartree-Fock method that treats electron-electron interactions in an average and approximate way. It typically requires the numerical evaluation of a large number of integrals. *Ab initio* QM methods attempt to evaluate these integrals numerically, leading to more precise treatment of electron-electron interactions while SQM approaches set these integrals to zero or to values determined from experiments. Density functional theory focuses on electron probability density rather than on electronic wavefunctions.

### 2.1.1 The Schrödinger equation

The electronic properties of a system can be determined by solving the Schrödinger equation,

$$\hat{H}\Psi(\mathbf{r};\mathbf{R})=E\Psi(\mathbf{r};\mathbf{R}) \quad (2.1)$$

where  $E$  is the total energy and  $\Psi(\mathbf{r};\mathbf{R})$  is the many-electron wavefunction as a function of the coordinates of all the electrons and the nuclei, with  $\mathbf{r}$  and  $\mathbf{R}$  denoting electronic and nuclear degrees of freedom, respectively. In atomic units, the

Hamiltonian operator  $\hat{H}$ , described the kinetic and potential energy of the system, is given by,

$$\hat{H} = -\frac{1}{2} \sum_i^{N_e} \nabla_i^2 - \frac{1}{2M_A} \sum_A^{N_N} \nabla_A^2 - \sum_i^{N_e} \sum_A^{N_N} \frac{Z_A}{r_{Ai}} + \sum_{i \neq j}^{N_e} \frac{1}{r_{ij}} + \sum_{A \neq B}^{N_N} \frac{Z_A Z_B}{R_{AB}} \quad (2.2)$$

where  $i$  refers to electron  $i$ ,  $M_A$  is the mass ratio of nucleus  $A$  to that of an electron, and  $Z_A$  is the atomic charge of nucleus  $A$ . The first and the second terms in Equation 2.2 are respectively the kinetic energy of the electrons and the nuclei. The third term is the Coulomb attraction between the nuclei and the electrons. The fourth and the last terms are respectively the Coulomb repulsion among the electrons and among the nuclei.

For a numerical solution of the Schrödinger equation, the simplification known as the Born-Oppenheimer approximation is invoked to treat the electronic and the nuclear motions separately. Considering the magnitude of the masses of nuclei and electrons, the nuclei are much heavier than the electrons. The electrons see the heavy, slow-moving nuclei as almost stationary point charges. Upon every move of the nuclei, the electrons would immediately adapt themselves in an eigenstate. This assumption allows a composition of the wavefunction for a system as a product of nuclear and electronic wavefunctions. By assuming a fixed configuration of the nuclei under non-relativistic and time-independent conditions and omitting the nuclear kinetic energy, the electronic Schrödinger equation is formulated as expressed below

$$\hat{H}_{el} \Psi(\mathbf{r}_i) = E_{el} \Psi(\mathbf{r}_i) \quad (2.3)$$

The Hamiltonian  $\hat{H}_{el}$  for the electronic Schrödinger equation in Equation 2.3 can be expressed as a one-electron part  $\hat{H}^c$ , called the core Hamiltonian, as a sum of one-electron operators  $\hat{h}_i^c$  for each electron  $i$ , and a two-electron part  $\hat{G}^{ij}$ , as given in Equation 2.4.

$$\hat{H}_{el} = \hat{H}^c + \hat{G}^{ij} = \sum_i \hat{h}_i^c + \sum_{i \neq j} \frac{1}{r_{ij}} = -\frac{1}{2} \sum_i^{N_e} \nabla_i^2 - \sum_i^{N_e} \sum_A^{N_N} \frac{Z_A}{R_{Ai}} + \sum_{i \neq j}^{N_e} \frac{1}{r_{ij}} \quad (2.4)$$

To describe the  $N$ -electron system, we need a function guaranteeing antisymmetric behaviour consistent with the Pauli exclusion principle, which states that two electrons cannot have all their quantum numbers equal. The total electronic wave function must be antisymmetric with respect to interchange of any two electrons' coordinates. The requirement for antisymmetry of the electron wavefunction can be achieved by building it from the Slater determinant  $\Psi_{SD}$ , a determinant of spin orbitals.

$$\Psi_{SD} = \frac{1}{\sqrt{N!}} \begin{vmatrix} \psi_1(1) & \psi_2(1) & \cdots & \psi_N(1) \\ \psi_1(2) & \psi_2(2) & \cdots & \psi_N(2) \\ \vdots & \vdots & \ddots & \vdots \\ \psi_1(N) & \psi_2(N) & \cdots & \psi_N(N) \end{vmatrix} \quad (2.5)$$

The spin orbital is the product of a spatial function describing the radial and angular distribution and either the alpha or beta spin function, i.e.  $\psi_i \equiv |\psi_i\rangle \cdot |\alpha\rangle$ ;  $\bar{\psi}_i \equiv |\psi_i\rangle \cdot |\beta\rangle$ . The columns in a Slater determinant given in Equation 2.5 are single-electron wavefunctions which are orbitals while the electronic coordinates are along the rows. This Slater determinant is used for the eigenfunction in the electronic Schrödinger

equation. The solution of the electronic Schrödinger equation is obtained from guessing the initial electron wavefunction as a function of  $3n$  spatial and  $n$  spin coordinates. The trial wavefunction  $\Psi$ , as a linear combination of finite atomic orbital wavefunctions  $\psi_i$ , is written as

$$\Psi = \sum_{i=1}^N c_i \psi_i \quad (2.6)$$

where  $\psi_i$  is a member of the basis set of  $N$  wavefunctions, with associated coefficient  $c_i$ . This construction is known as the linear combination of atomic orbitals (LCAO) approach.

The electronic energy  $E_{el}$  of the trial wavefunction  $\Psi$  is now evaluated in Equation 2.7 as the expectation value of the Hamiltonian operator divided by the norm of the trial wavefunction.

$$E_{el} = \frac{\langle \Psi | \hat{H}_{el} | \Psi \rangle}{\langle \Psi | \Psi \rangle} \quad (2.7)$$

For the energy expression above, we can employ the variational principle, which states that the energy calculated from an approximate wavefunction always has an energy above or equal to the true ground state energy  $E_0$  (Equation 2.8);

$$E_{el} = \frac{\langle \Psi | \hat{H}_{el} | \Psi \rangle}{\langle \Psi | \Psi \rangle} \geq E_0 \quad (2.8)$$

Hence, the quality of the trial wavefunctions can be judged by their associated energies: the lower, the better. The best approximate wavefunction can be obtained by iteratively varying the coefficients within the given basis set until the lowest electronic energy is achieved. For this reason, this is referred to as the self-consistent-field (SCF) approach. This process is repeated for many different fixed nuclear configurations to give the electronic energy as a function of the positions of the nuclei. The nuclear configuration that corresponds to the minimum electronic energy is the equilibrium geometry of the molecule.

### 2.1.2 Basis sets

The basis set is the set of mathematical functions used to represent molecular orbitals. Each molecular orbital is constructed by a linear combination of atomic orbitals, typically a finite set of basis functions centred at each atomic nucleus. Typically the greater the number of basis functions used, the better described the electron distribution. This could be up to maximum flexibility when the basis set is expanded towards an infinite complete set of functions, the so-called *basis set limit*. The computational effort for integral evaluation formally scales as  $N^4$  depending on the number of functions applied. Therefore, the number as well as the type of functions appropriate for a system of interest has to be concerned for a balance between accuracy and computational feasibility. For computational purposes, basis functions are normally in a contracted form represented by a linear combination of a number of primitive Gaussian functions.

QM calculations are performed using a finite set of basis functions. A good basis set should (i) provide an adequate description of electronic behaviour required for obtaining

accurate results; (ii) systematically cover all of coordinate space: for larger basis sets, the higher accuracy would be obtained only if both the radial and angular parts can actually represent the coordinate space; and (iii) be in a compact form because the cost of calculation increases with the number of functions in the basis set. The main categories of basis functions in QM calculations are briefly discussed below.

### (i) Minimal basis sets

This is the simplest level of basis set containing a minimum number of basis functions necessary for describing electrons in a neutral atom. A single basis function per atomic orbital is used for each atom in a molecule, e.g. elements in the second period of the periodic table would have a basis set of five functions: two *s*- and three *p*-functions. The most common one is STO-*nG* where *n* is an integer, referred to the number of Gaussian primitive functions comprising a single Slater-type orbitals (STO).

### (ii) Multiple-zeta basis sets

Slater-type orbitals can be considered as the *single-zeta (SZ) basis* where zeta ( $\zeta$ ) stems from the exponent of its exponential component. They are usually inadequate to describe the deviation of electron distribution from spherical harmonics, resulting in the development of the *double-zeta (DZ) basis* with two independent functions per atomic orbital. They better describe changes during a molecular bond formation because basis functions with different exponents  $\zeta$  can be mixed in different ratios for different bonding orientations. Similarly, the *triple-zeta (TZ) basis* contains three times as many functions as the minimal basis, and so on. Examples of these basis sets are the

*correlation consistent basis sets* with the notation of cc-pV<sub>n</sub>Z ( $n = D$  (double), T (triple), Q (quadrupole), 5 and 6) developed by T. H. Dunning.<sup>8</sup>

### (iii) Split-valence basis sets

In valence bond theory, chemical bonding takes place by the valence electrons rather than the core ones. The core atomic orbitals in the split-valence basis sets are represented by a single contracted basis function while each valence atomic orbital is split into many arbitrary functions. The most widely used belongs to J. A Pople *et al.*<sup>9</sup> in a notation of  $k\text{-}lmnG$  where  $k$ ,  $l$  and  $m$  are integers. The integer  $k$  represents the number of Gaussian primitives used in a basis set for each core atomic orbital whereas the integers  $l$ ,  $m$  and  $n$  represent that for valence orbitals. If only  $l$  and  $m$  are stated, it is a valence-double-zeta basis set, e.g. 4-31G, 6-31G. If all  $l$ ,  $m$  and  $n$  are stated, it is a valence-triple-zeta, e.g. 6-311G.

Auxiliary functions are commonly introduced into basis sets regardless of their categories to provide accurate representations of electron distribution. The most common addition is polarisation functions denoted by an asterisk (\*) in Pople-style sets, e.g. 6-31G\*, or by “P” in multiple-zeta and other split-valence sets<sup>10</sup>, e.g. DZP and SVP. These basis sets include functions with one higher angular momentum than that required by valence atomic orbitals to describe the asymmetric distortion of the atomic orbitals in a molecular environment.<sup>11</sup> For instance, *p*-functions are added for the hydrogen whereas *d*-functions are added to a basis set with valence *p*-orbitals, and so on. In addition to polarisation functions, diffuse functions can be used to improve a treatment of electron density disperse from nuclei such as anions. These diffuse

functions have small exponents which can be optimised in many ways as they extend far from the nuclei.<sup>12</sup> The use of diffuse functions is denoted by a plus sign (+) in Pople-style sets, e.g. 6-31+G, or by “aug” in Dunning-style sets, e.g. aug-cc-pVDZ.

### **2.1.3 *Ab initio* quantum mechanical methods**

Computation of the electronic wavefunction derived from theoretical principles without the inclusion of experimental data is termed the *ab initio* QM approach. Generally, *ab initio* QM methods are feasible for small molecules. They need the Born-Oppenheimer approximation to greatly simplify solving the Schrödinger equation. The simplest *ab initio* QM method is the Hartree-Fock (HF) approach which solves a single determinant and takes the average Coulombic electron-electron repulsion into account, but not the explicit electron pair repulsion interactions. Electron correlation can be added by the post Hartree-Fock methods such as configuration interaction, coupled cluster and perturbation theory.

#### **2.1.3.1 The Hartree-Fock formalism**

Hartree-Fock (HF) theory is fundamental to electronic structure theory. It was developed to approximately solve the electronic Schrödinger equation based on the antisymmetrised product of one-electron wavefunctions, i.e. the Slater determinant, consisting of one spin orbital per electron. This is equivalent to the assumption that each electron moves independently from the others except that it feels the Coulomb repulsion due to the average positions of all other electrons. This Coulomb repulsion is also

accompanied by the exchange interaction as a consequence of the antisymmetric behaviour of electrons. The HF methods determine the set of spin orbitals which minimise the energy and give the best single determinant by introducing a set of one-electron operators for electron  $i$ , the so-called the Fock operator  $\hat{F}$ ,<sup>13</sup>

$$\hat{F}_i = \hat{h}_i^c + \sum_j^{occ} (2\hat{J}_j - \hat{K}_j) \quad (2.9)$$

where  $\hat{h}_i^c$  is the core Hamiltonian in similar fashion to the expression of the electronic Hamiltonian operator given in Equation 2.4,  $\hat{J}$  and  $\hat{K}$  are respectively the Coulomb and exchange operators, acting on all  $j$  occupied spin orbitals. The total energy (Equation 2.10) is not just a sum of molecular orbital energies  $\varepsilon_i$ ; it includes electron-electron repulsion accounted in an average fashion of a mean-field approximation to a potential energy of nuclear-nuclear repulsion  $V_{nn}$ , and becomes<sup>14</sup>

$$E_{HF} = \sum_{i=1}^{N_{el}} \varepsilon_i - \frac{1}{2} \sum_{ij}^{N_{el}} (J_{ij} - K_{ij}) + V_{nn} \quad (2.10)$$

The practical computation of this is carried out by transforming the differential equation of electronic Schrodinger equation into matrix algebra, called the Roothaan-Hall equations, in Equation 2.11.<sup>13</sup>

$$\mathbf{FC} = \mathbf{SC}\varepsilon \quad (2.11)$$

The  $\mathbf{S}$  matrix contains the elements from the overlap integrals between basis functions. The coefficients that appear in the LCAO used to simulate the molecular orbital are stored in the  $\mathbf{C}$  matrix. The  $\boldsymbol{\varepsilon}$  matrix is a diagonal matrix of orbital energies. The  $\mathbf{F}$  matrix contains the Fock matrix elements. Each element in the Fock matrix is derived from the Fock operator containing two parts: (i) the one-electron integrals, and (ii) a product of a density matrix and two-electron integrals. The latter term interprets the interaction of one electron with a field of the other ( $n-1$ ) electrons. The density matrix, derived from the probability density of electrons, describes the degree to which individual basis functions contribute to the many-electron wavefunction and how energetically important the Coulomb and exchange integrals should be.

In principle, we can find the orbital energies in the  $\boldsymbol{\varepsilon}$  matrix by solving the secular equations,

$$|\mathbf{F} - \boldsymbol{\varepsilon}\mathbf{S}| = 0 \quad (2.12)$$

then, use these energies to find the molecular orbital coefficients that make up the  $\mathbf{c}$  matrix by solving the Roothaan-Hall equations. The HF method follows the SCF procedure in which we firstly guess the orbital coefficients, and then construct an initial density matrix  $\mathbf{P}$  and a Fock matrix  $\mathbf{F}$ . The Fock matrix is further diagonalised to obtain eigenfunctions containing the new set of coefficients. This starting solution is continuously improved by several cycles with the improved coefficients. The process will be converged if the difference between  $\mathbf{P}_n$  and  $\mathbf{P}_{n-1}$  has become insignificant. As the density matrix modulates the field of the ( $n-1$ ) electrons experienced by one electron under consideration, this procedure iteratively optimises this field until consistency is achieved.

The main limitation of the HF formalism is the one-electron nature of the Fock operator. It evaluates the energy of an electron in each molecular orbital moving in the mean field of all other electrons. Each electron's motion can be described by a single-particle function (orbital) which does not rely explicitly on the instantaneous motions of the other electrons. Hence, the instantaneous repulsion between electrons is not included. Other than the electron exchange interaction, there is no attempt to take into account the electron correlation, which is the tendency of electrons to stay apart to minimise their repulsion. The neglect of electron correlation can have a strong impact on determining accurate wavefunctions for chemical properties. To improve the energy of a system, the use of an extended basis set is advisable. However, this improvement with the number of basis functions becomes gradually smaller and converging asymptotically to a value of which further addition of functions has no more effect on the energy. This HF energy when converged to a limit of an infinite basis set is called the *HF limit*. The energy difference between the HF limit and the true energy is the electron correlation energy. By considering this energy difference, *ab initio* correlated QM methods can overcome the HF limit.

### 2.1.3.2 Correlated methods

Post-HF methods take into account the interaction between electrons and normally use the HF wavefunction as a starting point for improvements. In the HF formalism previously described, a certain amount of electron correlation is already considered i.e. the electron exchange integrals describing the correlation between electrons with parallel spins. These exchange terms create the *Fermi correlation*, the correlation that prevents two parallel-spin electrons to be too close at the same position. On the other

hand, the *Coulomb correlation* is the opposite spin correlation for both intra- and inter-orbital contribution.

The common feature of *ab initio* correlated methods is the extension from one-determinant to multi-determinant for assigning electron wavefunctions. These additional spin determinants, usually called excited states, are constructed using virtual unoccupied eigenfunctions resulting from a Hartree-Fock calculation. This does not mean to perform independent calculations on real excited states, but actually is a mathematical tool to increase the flexibility for describing electron distribution and allow electrons to avoid each other in a better way, rather than a participation of excited electron in the ground state. Some correlated methods are configuration interaction (CI)<sup>15</sup>, coupled cluster (CC)<sup>16</sup> and Moller-Plesset MP2<sup>17</sup> approaches. We refer the reader to excellent reviews of these methods.<sup>18</sup> These methods are all more computationally expensive as well as more accurate than HF.

#### **2.1.4 Semiempirical quantum mechanical (SQM) methods**

Performing even an HF calculation without approximations for large molecules is too expensive. Its cost formally scales as  $N^4$  due to the number of electron integrals necessary to construct the Fock matrix. To reduce the computational cost to  $N^2$ , SQM methods have been developed based on molecular orbital theory and make use of integral approximations and parameterisations. To speed up the calculations, SQM methods generally consider only valence electrons explicitly by using a minimal basis set. The core electrons are approximated by reducing nuclear charge and by introducing

functions to model combined effects of nuclear and core electrons. As a result, these SQM methods are efficient to model a large molecule in a realistic manner.

The central assumption of SQM methods is the zero differential overlap approximation. This approximation neglects all products of basis functions that depend on the same electron coordinates but locate on different atoms. As a consequence, the overlap matrix  $S$  is reduced to a unit matrix; three-centre one-electron integrals are set to zero; and all three- and four-centre two-electron integrals are neglected. To compensate for errors caused by these approximations, the remaining integrals are approximated as empirical parameters calibrated against experimental or theoretical reference data.

Existing SQM models differ by further approximations made when evaluating one- and two-electron integrals and by the parameterisation philosophy. One of the early SQM approaches was the  $\pi$ -electron method, proposed by E. Hückel, which generates the molecular orbitals from the connectivity matrix of a molecule. Later, R. Hoffmann proposed the extended Hückel method to include all valence electrons in qualitative studies of inorganic and organometallic compounds. In this method, the basis functions are only used for calculating the overlap integrals; the others for the Fock matrix elements are based on empirical parameters.<sup>19</sup> These Hückel-type methods consider only one-electron integrals, therefore they are non-iterative. Two-electron interactions are later included explicitly in semiempirical self-consistent field methods. J. A. Pople, awarded the Nobel Prize in Chemistry in 1998, introduced a hierarchy of integral approximations for SQM methods that use (i) the core approximation to include valence electrons explicitly and (ii) the zero differential overlap approximation to reduce the number of multicentre integrals. These methods are the complete neglect of differential

overlap (CNDO), the intermediate neglect of differential overlap (INDO), and the neglect of diatomic differential overlap (NDDO). The extension from the CNDO model to the INDO model is accomplished by adding flexibility to describe the one-centre two-electron integrals. Likewise, the NDDO model improves over the INDO model by relaxing the constraints on the two-centre two-electron integrals in a similar fashion to that on the one-centre integrals in the INDO model. The NDDO model is a fundamental of most modern SQM models. The early modified NDDO models are the modified neglect of diatomic overlap (MNDO)<sup>20</sup>, the Austin Model 1 (AM1)<sup>21</sup> and the Parametric Method number 3 (PM3)<sup>22</sup>. Research along this line has been very active over the years. Several improved methods such as PM6 and PM7 have been proposed recently.<sup>23</sup>

Pure SQM methods can give accurate results for charge transfer and polarisation effects but may be problematic for describing the electron correlation effects arisen from van der Waals dispersion and hydrogen bond interactions. Some empirical MM-type corrections can be added either for dispersion<sup>24</sup> termed by “-D” or for hydrogen bonding<sup>25</sup> termed by “-H”, or for both termed by “-DH” e.g. PM3-D<sup>26</sup> and PM6-DH2<sup>27</sup>. Parameterisation for both corrections has been implemented in PM7, the most recent NDDO-based method.<sup>7d</sup>

Dispersion corrections in SQM methods follow Grimme’s pioneering work at the DFT level.<sup>24</sup> The dispersion corrected energy  $E_{disp}$  is given by

$$E_{disp} = -s_6 \sum_{i=1}^{N-1} \sum_{j=i+1}^N \frac{C_6^{ij}}{R_{ij}^6} f_{dmp}(R_{ij}) \quad (2.13)$$

where  $N$  is the number of atoms,  $C_6^{ij}$  denotes the dispersion coefficient for atom pair  $ij$ ,  $s_6$  is a global scaling factor, and  $R_{ij}$  is an interatomic distance. To avoid near-singularities for small distance, a damping function  $f_{dmp}$  is used as a function of the sum of atomic van der Waals radii  $R_{ij}$ .

### 2.1.5 Density functional theory (DFT)

The central focus of DFT is not the electron wavefunction but the electron density  $\rho$ . The wavefunction for a system of  $N$  electrons has  $4N$  variables (three spatial and one spin coordinate for each electron). While the complexity of a wavefunction exponentially increases with the number of electrons, that of the electron density with the same number of variables is independent of system size. It is a function of the positions of electrons, and determined by the square of an electronic wavefunction  $\psi_i$  (Equation 2.14).

$$\rho(\mathbf{r}) = \sum_i^N |\psi_i(\mathbf{r})|^2 \quad (2.14)$$

This gives the total number of electrons  $N$  as  $N = \int \rho(\mathbf{r}) d\mathbf{r}$ . The number of electrons together with the external potential (see below) defines the Hamiltonian for an  $N$ -electron system.

Modern DFT began with the Hohenberg-Kohn formulations<sup>28</sup> applied to a system consisting of electrons moving under the influence of an external potential. Its first theorem states that the external potential is a unique functional of electron density  $\rho(\mathbf{r})$ ,

and so is the derived total energy. Therefore, all components of the many-particle ground state energy can be expressed as a functional of the electron density. Later, W. Kohn and L. J. Sham<sup>29</sup> proposed the idea of calculating the kinetic energy of a non-interacting reference system with the same electron density as the interacting one occurring in reality. They considered the Hamiltonian analogous to the HF formalism as a sum of one-electron operators for non-interacting electrons using the Slater determinant of the individual one-electron eigenfunctions  $\psi_i$ . The kinetic energy for the non-interacting system  $T_s$  can be expressed as<sup>30</sup>

$$T_s = -\frac{1}{2} \sum_i^N \langle \psi_i | \nabla^2 | \psi_i \rangle \quad (2.15)$$

It is of course not equal to the true kinetic energy of the interacting system. The Kohn-Sham theory accounted for this energy difference by introducing a separate functional  $E_{XC}$ , the so-called exchange-correlation energy<sup>31</sup> given as

$$E_{xc}[\rho(\mathbf{r})] \equiv (T[\rho(\mathbf{r})] - T_s[\rho(\mathbf{r})]) + (E_{ee}[\rho(\mathbf{r})] - J[\rho(\mathbf{r})]) \quad (2.16)$$

where  $T[\rho(\mathbf{r})]$  and  $T_s[\rho(\mathbf{r})]$  are respectively the kinetic energy of the real and non-interacting systems;  $E_{ee}[\rho(\mathbf{r})]$  represents the electronic interactions and  $J[\rho(\mathbf{r})]$  is the classical Coulomb interaction. The first term on the right hand side in Equation 2.16 can be considered as the kinetic correlation energy between the fictitious non-interacting system and the real one whereas the last term contains all non-classical effects, i.e. self-interaction correction, electron exchange and correlation.

The total ground-state energy expression<sup>32</sup> based on the Kohn-Sham formulation as a functional of the electron density becomes

$$E[\rho(\mathbf{r})] = T_s[\rho(\mathbf{r})] + J[\rho(\mathbf{r})] + E_{xc}[\rho(\mathbf{r})] + E_{Ne}[\rho(\mathbf{r})] \quad (2.17)$$

where the last term defines electron-nuclei interactions. To minimise the energy in Equation 2.17 above, we have to find the orbitals  $\psi$  that satisfy the equation 2.18 below,

$$h_i^{KS}\psi_i = \varepsilon_i\psi_i \quad (2.18)$$

where  $h_i^{KS}$  is the Kohn-Sham one-electron operator,<sup>33</sup> defined as

$$h_i^{KS} = -\frac{1}{2}\nabla^2 + \frac{1}{2}\int \frac{\rho(\mathbf{r}')}{|\mathbf{r}_i - \mathbf{r}'|} d\mathbf{r}' - \sum_A \frac{Z_A}{|\mathbf{r}_i - \mathbf{R}_A|} + V_{xc}(\mathbf{r}) \quad (2.19)$$

The exchange-correlation potential  $V_{xc}(\mathbf{r})$  can be calculated once we know the exchange-correlation energy by forming the functional derivative below.

$$V_{xc}(\mathbf{r}) = \frac{\delta E_{xc}[\rho(\mathbf{r})]}{\delta \rho(\mathbf{r})} \quad (2.20)$$

This requires to be solved iteratively and self-consistently. First, we guess the electron density  $\rho(\mathbf{r})$  from a superposition of atomic electron probability densities. Then, the exchange-correlation potential  $V_{xc}$  is calculated by assuming the approximate exchange-correlation energy from the initial guess of the electron density and evaluated

by the functional derivative. Next, the Kohn-Sham equations in Equation 2.18 are solved for an initial set of Kohn-Sham orbitals,  $\psi_i$  that satisfy the equation 2.18. This set of derived orbitals is further used to obtain a better approximation of the electron probability density. The process is carried out iteratively until the energy density is converged within the specified tolerance.

The DFT method is a non-exact formalism, in contrast to the HF method. The HF methods exactly define operators and solve relevant equations exactly. In contrast, DFT methods give non-exact solutions by manipulating the non-exact operator. Kohn-Sham density functional theory is now the most widely used method for calculating electronic structure in condensed matter physics and quantum chemistry. Its success mainly results from significant robustness, providing reasonably accurate predictions within the scaling behaviour of  $N^3$  for computational time, instead of  $N^4$  in the HF method, where  $N$  is the number of basis functions.<sup>33</sup> However, the most crucial problem of DFT still remains, in searching for suitable approximations of  $V_{XC}$  and  $E_{XC}$ .

### **2.1.6 Density functional tight binding (DFTB)**

Tight binding theory was first developed in a field of computational material science to calculate atomic and electronic structures, energies and forces for a large condensed phase system. Empirical tight-binding methods use a minimal atomic orbital basis to approximate the treatment of Kohn-Sham orbitals with a few empirical parameters, representing the counterpart of semiempirical methods. The density functional tight binding (DFTB) approach incorporates tight binding principles and includes a treatment of electron correlation as in a DFT framework, in contrast to the NDDO-type methods,

which start from solving the Hartree-Fock density matrix and introduce correlation effects solely through parametrisation.

The total energy of DFTB methods is the energy expression of Kohn-Sham density functional theory (Equation 2.17 in section 2.1.5) as a functional of charge density. Instead of finding the electronic density  $\rho(\vec{r})$  that minimises the total energy, a reference density  $\rho_0$  is introduced and perturbed by some density fluctuation  $\delta\rho$ ,

$$\rho(\vec{r}) = \rho_0(\vec{r}) + \delta\rho(\vec{r}) \quad (2.21)$$

where the reference density  $\rho_0$  is given by a superposition of atom-like densities  $\rho_A$  centred on the nuclei  $A$ ,

$$\rho_0(\vec{r}) = \sum_A^N \rho_A^A(\vec{r}_A) \quad (2.22)$$

Different DFTB models are derived by expanding the DFT total energy with respect to the electronic density fluctuation  $\delta\rho$  around the reference density  $\rho_0$  in a Taylor series and truncated to the third order as shown in Equation 2.23,<sup>34</sup>

$$E[\rho] = E^0[\rho_0] + E^1[\rho_0, \delta\rho] + E^2[\rho_0, (\delta\rho)^2] + E^3[\rho_0, (\delta\rho)^3] \quad (2.23)$$

The DFTB1 model, originally called DFTB, is introduced by the expansion to the first order, and neglects the second and higher order terms. The total energy for this model (Equation 2.24) is determined by non-self-consistently solving the atomic orbital

Hamiltonian in which the generalised eigenvalues have to be diagonalised only once.<sup>35</sup>

Hence, it is computationally fast.

$$E^i[\rho_0] = \sum_i^{occ} \langle \Psi_i | \hat{H}_0 | \Psi_i \rangle + E_{rep} \quad (2.24)$$

The first term in Equation 2.24 is the energy contributed from the effective Kohn-Sham Hamiltonian  $\hat{H}_0$  applied to Kohn-Sham orbital  $\Psi_i$ . The second term is a short-range repulsive energy  $E_{rep}$  which is a sum of short-range core-core repulsion and DFT double counting of spin spatial orbitals in two-electron integrals. For short interatomic distances where the nuclear repulsion dominates, the  $E_{rep}$  term in equation 2.24 can be derived from repulsive pairwise potentials.<sup>36</sup> For long interatomic distances, the repulsion energy does not decay to zero, but to a constant value depending on its atomic contributions as in Equation 2.25.<sup>37</sup>

$$\lim_{R_{AB} \rightarrow \infty} E_{rep}[\rho_0] = \sum_A^N E_{rep}[\rho_0^A] \quad (2.25)$$

The  $\sum_A^N E_{rep}[\rho_0^A]$  term is assumed so that  $E_{rep}$  depends only on the two-centre potential.

In practice, this term is obtained by fitting two-body potentials to the difference of total energies between a full DFT calculation and the electronic part of a DFTB calculation with respect to the bond length of interest in a suitable training set.<sup>38</sup> This DFTB1 model is recommended for systems in which charge transfer between atoms is small such as for hydrocarbons and fullerene clusters, such that higher-order terms can be neglected.<sup>39</sup>

The self-consistent charge density functional tight-binding (SCC-DFTB) method, recently renamed DFTB2, has been developed for better describing electronic systems that require a charge balance between atoms such as heteronuclear molecules and polar semiconductors.<sup>37</sup> This model considers the second-order term of the total energy expanded around the reference electronic density. The second-order correction due to charge fluctuations is approximated by decomposing the electronic density fluctuation  $\delta\rho$  in a series of radial and angular functions. The charge fluctuation decays quickly when the distance increases from the corresponding centre. By truncating the multipole terms of the charge expansion after the monopole term, the expression of the total charge in a system is given by

$$\int \delta\rho(\vec{r}) = \sum_A^N \Delta q_A \quad (2.26)$$

where  $\Delta q_A = q_A - q_A^0$  denotes the atomic net charge, i.e. the negative Mulliken charge of atom  $A$ . The SCC-DFTB total energy<sup>40</sup> becomes

$$\begin{aligned} E^{SCC-DFTB} &= E^2[\rho_0] \\ &= \sum_i^{occ} \langle \Psi_i | \hat{H}_0 | \Psi_i \rangle + E_{rep} + \frac{1}{2} \sum_{AB} \gamma_{AB} \Delta q_A \Delta q_B \end{aligned} \quad (2.27)$$

The extension from the DFTB1 model for the total energy is the last term in Equation 2.27 representing the long-range Coulomb interactions between two spherical charge distributions, i.e.  $\Delta q_A$  and  $\Delta q_B$  located at atom  $A$  and  $B$ . The analytic function  $\gamma_{AB}$  is

calculated for every atom pair by using an interpolated formula in Equation 2.28 depending on the interatomic distances  $R_{AB}$ , and the atomic Hubbard-like parameters  $\gamma_{AA}$  and  $\gamma_{BB}$  (also see Equation 2.29).<sup>41</sup>

$$\gamma_{AB} = \frac{1}{R_{AB}} - S[R_{AB}, \gamma_{AA}, \gamma_{BB}] \quad (2.28)$$

The first term in Equation 2.28 presents the  $R^{-1}$  behaviour corrected at the limit of large interatomic distances  $R_{AB}$ . The second term is a short-range function  $S$  to ensure a convergence at  $R_{AB} = 0$ . For  $R_A = R_B$ , the function  $\gamma_{AB}$  turns out to be  $\gamma_{AA}$  which gives the self-interaction of atom A. This function  $\gamma_{AA}$  is related to the chemical hardness  $\eta_A$ <sup>42</sup> or the Hubbard parameter  $U_A$  shown below.

$$\gamma_{AA} = 2\eta_A = U_A = \frac{\partial^2 E_A}{\partial^2 q_A} \quad (2.29)$$

The  $\gamma$ -function can be expressed as the second derivative of the total energy with respect to the atomic charge. It can be seen that the atomic charge is inversely proportional to its chemical hardness, intuitively, the more polarisable donor atoms such as anions, a smaller chemical hardness it is.<sup>43</sup> The  $\gamma$ -function is typically calculated from the difference between the ionisation potential and the electron affinity. Due to practical problems related to non-existence of some anions or missing experimental validation for the electron affinities,<sup>37</sup> the Hubbard parameters and its related parameters have to be derived to improve a description of interatomic electrostatics.<sup>44</sup> The option to modify

this function becomes a default and a key concern in the next generation of developed model, DFTB3.<sup>7b</sup>

DFTB3 is the recently developed model considering the third-order term in a series expansion of the total energy, similar to the development of the DFTB2 model. The third-order term describes the change in chemical hardness  $\eta_A$  of an atom with respect to its charge state. This is shown in the last term in Equation 2.30 which introduces the new parameter  $\Gamma_{AB}$ , the derivative of the  $\gamma$ -function with respect to atomic charge.<sup>45</sup> The DFTB3 model can overcome some problematic approximations for the  $\gamma$ -function in DFTB2 model and is more robust in describing the dependence of system properties on atomic charge.<sup>34</sup>

$$\begin{aligned} E^{DFTB3} &= E^3[\rho_0] \\ &= \sum_i^{occ} \langle \Psi_i | \hat{H}_0 | \Psi_i \rangle + E_{rep} + \frac{1}{2} \sum_{AB} \gamma_{AB} \Delta q_A \Delta q_B + \frac{1}{3} \sum_{AB} \Delta q_A^2 \Delta q_B \Gamma_{AB} \end{aligned} \quad (2.30)$$

In brief, the development of DFTB models concerns the improvement of the effective Coulomb repulsion in self-consistent field and tight-binding formalism. These models have very good computational performance for organic and biomolecules in terms of accuracy and cost. They could be as accurate as DFT methods but two to three orders faster in some cases.<sup>43</sup> However, there are some limitations: (i) the neglect of a treatment for three-centre and four-centre electron integrals for more accurate results in solving the exchange-correlation energy, and (ii) inheriting the numerical problems of DFT as their predecessor, e.g. grid-based approaches which rely on weighting functions for partitioning 3D exchange-correlation integrals into a sum of one-centre atomic contributions, leading to numerical inaccuracies and noise.

## 2.2 Molecular Mechanics (MM)

Molecular mechanics methods consider an atom with a given nuclear configuration as a discrete unit which can be joined together with other atoms via bonds. As electrons are not treated explicitly but together with the nuclei as effective atoms, standard MM methods are inadequate for studying bond forming/breaking. Bonding information has to be input explicitly rather than obtained from solving the electronic Schrödinger equation.

The total energy expressed in MM methods is a function of the nuclear coordinates. The potential energy function contains both the intra- and intermolecular interactions that obey classical physics as described in a standard form below.<sup>46</sup>

$$E_{MM} = \sum_{bonds} k_R (R - R_{eq})^2 + \sum_{angles} k_\theta (\theta - \theta_{eq})^2 + \sum_{dihedrals} \frac{V_n}{2} [1 + \cos(n\phi - \gamma)] \\ + \sum_{i < j}^{atoms} \left( \frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^6} \right) + \sum_{i < j}^{atoms} \frac{Q_i Q_j}{\epsilon R_{ij}} \quad (2.31)$$

The energy contribution in Equation 2.31 is a sum of the harmonic terms and non-bonded interactions, i.e. van der Waals and electrostatic interactions. The parameters  $k_R$ ,  $R_{eq}$ ,  $k_\theta$ ,  $\theta_{eq}$ ,  $V_n$ ,  $\gamma$ ,  $A_{ij}$ ,  $B_{ij}$ , and the charge  $Q$  are parametrised to reproduce experimental properties or higher level computational data. A specific set of these derived parameters is termed a *force field*. More sophisticated force fields have been developed by including the higher terms and/or cross terms for anharmonic vibration as well as a better treatment of valence angles and dihedrals. Another development for more

accurate force fields is a consideration of explicit electronic polarisation, lone pairs of electrons, solvation and long-range interactions which are important for the study of biological macromolecules.<sup>47</sup>

MM methods have both advantages and limitations. Their main advantage is computational speed, i.e. faster than Hartree-Fock or density functional theory by several orders of magnitude. Therefore, treatment of large systems up to millions of atoms such as proteins and DNA becomes feasible. MM methods are also used to investigate dynamics of atoms. Although there are a large number of empirical parameters for parameterisation, these methods can be very accurate for the system they have been parameterised for. However, unique parameterisation may be required for each chemical system. Among a variety of force fields available for use, no universal force field is applicable for best describing chemical properties. Drawbacks of MM methods include a lack of parameters for a number of elements and missing quantum effects. For these reasons, the active centres corresponding to chemical reactivity of macromolecules must often be modelled at the QM level of theory. This leads to the development of hybrid QM/MM methods. One such method is the combined QM/MM described in the next section.

### 2.3 Quantum mechanics/Molecular mechanics (QM/MM)

As discussed above, with currently available force fields, MM methods can rapidly compute the molecular energy and predict molecular geometry in conformational analysis; however, they are unable to describe the details of bond forming/breaking or

electron transfer processes which occur during chemical reactions. QM methods are therefore the method of choice for such cases, but they are only applicable to small-size systems due to their computational cost. For large systems, the entire system sometimes cannot be treated by the electronic structure methods. The combination of QM and MM methods, denoted as QM/MM, has been designed for modelling such cases at a reasonable cost with satisfied accuracy. The hybrid QM/MM approach, first published by A. Warshel and M. Levitt,<sup>48</sup> has been developed based on the idea of partitioning large chemical systems into (i) an electronically important region, which contains a small number of atoms, treated by a relevant QM level of theory, and (ii) the remainder described by a suitable force field. The total Hamiltonian<sup>49</sup> for the entire system in the QM/MM framework is

$$\hat{H} = \hat{H}^{QM} + \hat{H}^{MM} + \hat{H}^{QM/MM} \quad (2.32)$$

where  $\hat{H}^{QM}$  is the Hamiltonian for the QM region,  $\hat{H}^{MM}$  is the empirical force field for the MM region, and  $\hat{H}^{QM/MM}$  is the Hamiltonian that describes the interactions between those two regions. The simplest form of Hamiltonian  $\hat{H}^{QM/MM}$  at the interface contains both electrostatic terms, in which MM atoms interact with both QM electrons and QM nuclei, and van der Waals terms for all of the MM and QM atoms.<sup>50</sup>

The treatment for both QM and MM regions is carried out analogous to their corresponding isolated systems except for at the partition MM-QM boundary. The coupling region between the QM charge density and the point charge model used in the MM region can be handled at different levels of sophistication. The interactions in this region consist of the bonded interactions at the boundary, and the non-bonded

electrostatic and van der Waals interactions. The first concern is the treatment of covalent bonds crossing the boundary. Ideally, partitioning a system should not cut any covalent bonds to ensure the completeness of the QM subsystem. This implies that no dangling bonds are permitted. To deal with this problem, the most commonly used approach uses the *link atom*, typically a hydrogen atom but not always. This additional atom is placed along covalent bonds cut at the boundary to saturate and truncate the QM region. It is considered as a regular QM atom included in the QM part of calculations. Different approaches such as capping potentials<sup>51</sup> and frozen orbitals<sup>52</sup> have also been proposed to treat dangling bonds.

The second concern is the treatment of electrostatic interactions between the QM and MM regions. Two main schemes have been introduced according to how the MM charges are embedded in the QM calculation.<sup>53</sup> The simplest is *mechanical embedding* that treats interactions between those two regions at the MM level. The concept of charge models, e.g. rigid atomic point charge or bond dipoles, used in MM region is also applied to the QM atoms. This scheme is straightforward at reasonable cost; however, it has some drawbacks. For instance, the accurate MM parameter sets may be suitable for elements in the MM region but could be problematic for elements in the QM region, e.g. metals at the binding site. Also, the use of a single set of MM parameters could lead to a serious error due to the changing QM charge distribution during the progress of a chemical reaction. Such perturbation is completely ignored by this simplest scheme. To avoid these shortcomings, *electrostatic embedding* accounts for the interactions between the polarisable QM density and MM charges by incorporating the fixed MM point charges as one-electron terms in the QM Hamiltonian. The electrostatic interactions at the boundary are treated at the QM level.

This provides a more accurate description than the mechanical embedding scheme. The electronic structure of the QM region can adapt to changes in charge distribution polarised by a molecular environment. No MM charge model is required to model the QM region. This scheme is the most popular approach currently in use for a treatment of electrostatic interactions in the coupling region. Apart from these two main schemes, a *polarised embedding* scheme can be used by introducing a polarisable charge model in the MM region to respond to the QM electric field.<sup>53c, 54</sup>

The last concern is QM-MM van der Waals coupling.<sup>55</sup> The van der Waals terms, typically described by the Lennard-Jones potential, can estimate dispersion attractions that fall off as  $r^{-6}$ . It also prevents a molecular collapse due to strong repulsions at short distances. Every QM atom contributes to van der Waals interactions with all MM atoms but the closest ones to the boundary do the most. This requires the optimisation of Lennard-Jones parameters for the QM atoms.<sup>55-56</sup>

In brief, the combined QM/MM methods integrate the best distinctive advantage of efficient MM force fields with the accuracy of QM approaches. Various sophisticated levels of QM levels of theory can be applied to a chemically reactive region which is influenced by its environment treated at the MM level. The frontier between the QM and MM regions requires special treatment of covalent bonds and non-bonded interactions. QM/MM methods have been widely used to study the enzymatic reactivity. For a study of protein-ligand binding, the computationally relevant techniques will be presented in the next section.

## 2.4 Molecular docking

Docking plays a key role in rational drug design. It predicts the preferred bound orientation of small-molecule drug candidates to their protein targets. The binding affinities and activities of protein-ligand bound complexes can be estimated subsequently for lead optimisation.

A docking scheme requires 3D protein and ligand structures as input. The protein structure and known ligand are typically determined using a biophysical technique such as X-ray crystallography or NMR spectroscopy. The structure of potential ligands can be selected from various classes of a molecular or fragment database depending on required properties. Their 3D geometries are modelled prior to docking. The success of docking approaches relies on both the conformational search algorithm for ligand-receptor geometries and the scoring function for predicting the binding affinity as discussed below.

### 2.4.1 Conformational search methods

Conformational search algorithms can be divided into two broad categories: deterministic and stochastic methods. *Deterministic (or systematic search) algorithms* enumerate exhaustively all degrees of freedom, for example, incremental construction and grid search approaches, so that they are generally possible at certain discrete intervals. The deterministic methods using an accurate force field can be very effective for small molecules with limited conformational flexibility, but may not be suitable for

molecules with a higher number of rotatable bonds or ring closures.<sup>57</sup> For flexible molecules with up to 10 rotatable bonds, *stochastic (or random search) methods* are very useful.<sup>58</sup> These methods consider a whole molecular geometry in a sampling process by using random variables to explore molecular chemical space.

Stochastic methods that have been widely used are distance geometry,<sup>57</sup> genetic algorithms,<sup>59</sup> Monte Carlo (MC), molecular dynamics (MD), and simulated annealing. In *Monte Carlo (MC)* sampling, random changes in structure are attempted by rotations in small increments about the selected dihedral angles or other geometrical changes. The newly generated conformations associated to each trial move are validated according to the Metropolis criterion.<sup>60</sup> They are in a continuous trajectory, namely a Markov chain. Because of this continuity, a large number of conformations can be sampled for the entire conformational space. However, there is no guarantee of finding conformers in the global energy minimum because some regions of the conformational space may be explored more thoroughly than others.

*Simulated annealing* is an approach in which the temperature of a system is adjustable. During a sampling process, the simulation temperature slowly decreases over a period of time so that only high energy conformations are explored in the initial stages (either using MC or MD simulation). Those conformers with sufficient energy are able to cross the energy barriers to other local minima. The ability to sample crossing over the energy minima results in greater performance in finding the lowest energy conformer. Due to the fact that less favourable conformers become more energetically inaccessible near the end stage of sampling process, the lower energy conformers can be more populated.

## 2.4.2 Scoring functions

Along with a conformational search algorithm, successful structure-based drug design lies in accurate estimation of the binding affinity of docked poses in a receptor pocket. The estimated binding energy is derived by various mathematical methods, collectively called *scoring functions*. These scoring functions are used as a selection criterion for ligand candidates. The best candidate should have the strongest binding affinity. Fortunately, the absolute binding constant is not necessary because relative values are sufficient to determine the best ligand pose among all the others. To calculate the relative binding constant, free energy perturbation (FEP) is a reliable scoring function but too computationally demanding for routinely scoring docked poses. FEP combined with MD or MC simulations can calculate the relative free energy of binding to the same drug target within 1-2 kcal/mol of the experimental value, for favourable cases.<sup>61</sup>

The ideal scoring function would be able to accurately calculate binding free energies; however, existing scoring functions are not sufficiently capable. They typically assume that binding affinity can be described as a sum of independent interaction terms. This can be reliable for a simplification of the full receptor-ligand interaction by representing general terms such as hydrogen bonds and hydrophobic effects, and omitting more complex aspects such as solvation, polarisation and entropy. These issues are addressed in the advanced scoring functions developed for use in virtual screening and fragment-based drug design.<sup>62</sup>

Most scoring functions can be classified into three main types: force field-based, empirical and knowledge-based scoring functions. *Force field-based scoring functions*

depend on classical molecular mechanics energy functions which represent the relationship between a molecular structure and its energy (Equation 2.31 in section 2.2). They approximate the binding free energy of protein-ligand complexes by summing up the electrostatic and van der Waals interaction energies. These interaction energy terms are generally applicable; however, they have two weaknesses.

Firstly, the van der Waals interaction is described by the Lennard-Jones potential from which strong repulsion of significant steric clashes can arise at short interatomic distances. This term is sensitive to small differences in packing. However, to compensate for potential induced fit of protein and ligand, this term is often scaled down in docking. Secondly, polar interaction can be overemphasized by the applied force fields, but it might be compensated by incorporation of desolvation energies. It is worth noting that solvation is usually taken into account by using either a distance-dependent dielectric function or the continuum electrostatics solvent models According to the presented results of an assessment of nine scoring functions for studying protein-ligand interactions in docking, steric repulsion is more important than electrostatic interaction for identifying near-native poses of decoys.<sup>63</sup> It can be inferred that the treatment of solvation effects might have a minor effect on the effectiveness of force fields for elucidation of near-native configurations.

The next class is *empirical scoring functions*. They have been designed based on the idea that binding free energies can be approximated by a sum of individual uncorrelated interactions, e.g. hydrogen bonding, binding entropy, ionic and lipophilic interactions. The coefficients of each term are obtained from regression analysis of experimental binding constants of a training set of protein-ligand complexes.<sup>64</sup> Due to the dependence

on experimental data sets in regression analysis, different weighted terms can be subsequently obtained. Combining different weighted coefficients obtained from different fitted scoring functions into the united value becomes a challenge. However, the benefit of these empirical scoring functions is a use of experimental binding energies in parameterisation.

Lastly, *knowledge-based scoring functions* are derived from statistical analysis of the known 3D structures of protein-ligand complexes without any reference to the experimentally determined binding affinities. The binding free energies of these complexes are estimated by a sum of protein-ligand atom pair interactions. Those interactions are counted by the frequency of occurring individual contacts between the studied protein and ligand atom types on their energetic bindings. A preferred contact that occurs frequently tends to be not random; hence, this indicates an attractive interaction. A consideration of frequency distributions for a dataset of interacting molecules can then be mapped to atom-pair potential sets. In other words, the atom-pair potentials extract structural information from protein-ligand complexes and employ the Boltzmann law to transform the atom pair preferences into distance-dependent pairwise potentials.<sup>65</sup> On the one hand, knowledge-based scoring functions attempt to implicitly capture binding effects that are difficult to model explicitly. These scoring functions are, on the other hand, limited to the known sets of protein-ligand complex experimental structures.

In addition to these three main categories of scoring function, we have employed the energy-based QM/MM scoring function to predict the binding affinities between ligands and their enzyme receptors at the active sites. The concepts of QM/MM scheme is

mentioned earlier in Chapter 2. More detail of this QM/MM scoring function will be discussed in Chapter 4.

## 2.5 Molecular classes of the hits

A process for the development of a compound into a marketed drug involves several stages: target identification, target validation, hit and lead identification, lead optimisation, preclinical stage and clinical trials.<sup>3</sup> Lead identification is a key stage to find synthetically accessible chemicals that exhibit potency and specificity against a biological target associated with the disease of interest. This exploration can be achieved either by experimental screening, traditional high-throughput screening (HTS) or computational screening (or virtual screening) to discover the putative compounds, the *hits*. The hits can be divided into three categories: drug-like, lead-like or fragment-like compounds, each of which differently facilitate the next stage of optimisation.

The main purpose of assembling compounds as chemical libraries for screening is usually to identify leads, which may not be entirely drug-like initially. According to the “Rule of Five” of drug-likeness, first introduced by C. A. Lipinski,<sup>66</sup> the orally active drugs should not violate the following criteria: (i) hydrogen-bond donors  $\leq 5$ ; (ii) hydrogen-bond acceptors  $\leq 10$ ; (iii) the molecular weight (MW)  $\leq 500$  daltons; and (iv) the calculated logarithm of the octanol-water partition coefficient ( $c\log P$ )  $\leq 5$ . Otherwise, they tend to have poor absorption or permeation. The compliance of this rule for non-oral drugs has also been addressed.<sup>67</sup>

Some drug-like hits that break the Lipinski's rule of five can cause failure in a subsequent stage of drug development. Therefore, the lead-like compounds are a good choice for optimisation because they are smaller in size than drug-like molecules and typically exhibit an optimal binding affinity towards protein targets. Compared to drug-like compounds, the lead-like hits should have less molecular complexity (low MW, fewer rings and rotatable bonds), less hydrophobicity (low clog P values), lower polarisability (less calculated molecular refractivity), and less drug-likeness (low drug-like scores).<sup>68</sup> Those differences underline the importance of an increase in molecular weight and of incorporating more lipophilic moieties in the development of leads to drugs.

Although a number of high-quality lead compounds can be successfully optimised in a drug discovery process, their lipophilic nature can be problematic. To overcome such problems, fragmentation of leads into small pieces or even into discrete functional groups is a good choice. Fragments are superior to drug-sized hits for ultimately obtaining good clinical candidates because they are simpler with less functional groups, but still have physicochemical properties suitable for drug development.<sup>69</sup> They generally obey "the Rule of Three", i.e. MW  $\leq$  300, hydrogen-bond donors or acceptors  $\leq$  3, and moderately lipophilic ( $\text{clogP} \leq 3$ ). In addition, the number of rotatable bonds  $\leq$  3 and the polar surface area  $\leq 60 \text{ \AA}^2$ , can be considered as an optional criteria to select the fragment.<sup>70</sup> Moreover, high *binding efficiency* of the fragment towards its specific protein target, i.e. its averaged free energy per heavy atom,<sup>71</sup> is often used to indicate the quality of fragment hits.<sup>72</sup> Eventually, when the fragment hits are discovered, they are chemically modified by applying suitable hit-to-lead techniques<sup>73</sup> to become leads.

The probability of identifying fragment hits for a known target relies significantly on the lower molecular complexity of a fragment according to Hann's model.<sup>74</sup> This model suggests that the probability of good binding to a receptor decreases rapidly as the complexity of the ligand increases, because obtaining a match is harder than a mismatch. This hypothesis implies that fragments have higher hit rates in the screening due to their small size. However, the limited size and less complex nature of fragments can cause very low affinity binding, in the μM to mM range. High aqueous solubility is essential to identify such weak interaction as well as screening at higher concentration with detection possible by biophysical techniques<sup>75</sup> such as X-ray crystallography<sup>76</sup>, NMR<sup>77</sup> and mass-spectrometry<sup>78</sup>.

For further optimisation, the discovered hits should ideally possess desired properties in terms of potency, selectivity and pharmacokinetic properties. They should have (i) relatively simple chemical features for rapid follow-up and further optimisation; (ii) a structure-activity correlation; (iii) no toxicophores; and (iv) appropriate pharmacokinetic properties related to drug absorption, distribution, metabolism and excretion.<sup>68b</sup>

## 2.6 Virtual screening

Virtual screening<sup>79</sup> is the *in silico* analogue of traditional high-throughput screening,<sup>80</sup> and is sometimes referred to as molecular database screening. This technique uses computational methods to identify the hits (described in earlier section 2.5). From the discovered hits using virtual screening methods, bioactive molecules can be designed

and optimised.<sup>81</sup> Using knowledge of a biological target's structure, the binding mode of the screened compounds at the active site can be predicted and evaluated by using docking algorithms and scoring functions (described earlier in section 2.4). This is called structure-based virtual screening.<sup>64</sup>

When the 3D structure of a receptor is not available, knowledge of active ligands that interact with the protein receptor can be useful for virtual screening. The identity of these ligands can be hypothesised from knowledge of substrates, known inhibitors, pharmacophores and structure-activity relationships. The concept that molecules with similar structures are likely to share similar properties forms the basis of ligand-based virtual screening.<sup>82</sup> For this, a reference ligand called the *query* is required to perform searching of the compound database. It is necessary to generate the complementary structural image of the binding site by using the bioactive conformation of the query. However, in the case that the bioactive conformation of a query ligand bound to a protein may not be available, a calculated low-energy conformation can be generated and used as the active conformation for the query.

The concept of shape-matching can be applied for ligand-based virtual screening. Candidate ligands in a compound database are superimposed onto the query molecule in order to identify ligands that best match the volume and disposition of query functional groups. Compounds that possess a comparable shape and chemistry to known reference ligands are assumed to show similar activities. Unfortunately, active compounds considerably different in size from the query are likely to be missed during this type of screen. This high false negative rate still remains the major problem of shape-based similarity screening, whereas receptor-based screening usually suffers from a high false

positive rate, i.e. inactive compounds are incorrectly assigned to be active.<sup>83</sup> It has also been reported that ligand-based shape matching has higher enrichments and is superior to receptor-based virtual screening for the rational design of HIV-1 therapeutics.<sup>84</sup>

Various shape-based methods, for example, SQW,<sup>85</sup> catShape,<sup>86</sup> PhaseShape,<sup>87</sup> Shapelets,<sup>88</sup> USR (Ultrafast Shape Recognition)<sup>89</sup> and SLIM (shaped-based ligand matching with binding pocket)<sup>90</sup> are available for ligand-based virtual screening. Another robust ligand-based tool for virtual screening is ROCS (Rapid Overlay of Chemical Structures).<sup>83, 91</sup> A chemical force field implemented in ROCS appears to be of great benefit to its performance.<sup>83, 91</sup> Recently, Hawkins *et al.*<sup>92</sup> performed screening using ROCS, comparing with published data sets on receptor-based virtual screening via a docking protocol. A direct comparison from their results showed that ROCS performs at least as well as and often better than docking programs. Another study also suggested that ROCS can be an effective protocol to detect Type II protein kinase inhibitors and distinguish from those of type I.<sup>93</sup> This study also showed that using ROCS with a consensus query is not strongly dependent on conformer generator and performs at least as well as pharmacophore filters. Although ROCS may be more computationally demanding and more dependent on conformational sampling compared to pharmacophore-based virtual screening, it tends to increase the accuracy for larger conformational ensembles.<sup>94</sup>

The approach in ROCS uses atom-centred Gaussian functions<sup>95</sup> to describe molecular shape. The centre of mass of the query and a database molecule is calculated prior to their superposition with respect to the aligned axes of their principal moments of inertia (Figure 2.1). The volume overlay is derived from the overlaps of their Gaussian

functions,<sup>96</sup> which are parameterised according to the hard-sphere volume of heavy atoms. It is subsequently optimised by applying a solid-body optimisation algorithm to find the maximum volume overlap. A database molecule which occupies the highest volume overlap of Gaussian functions is considered the best match. To obtain the best match, screened compounds are ranked by using two distinctive scoring functions for either shape similarity (a *ShapeTanimoto* score) or chemical pattern similarity (a *ColorTanimoto* score). A combination of both shape and functional group matching is represented by a *TanimotoCombo* score. Overall, 3D shape-based similarity implemented in ROCS is not only able to search the database compounds similar in shape to the query, but it can also identify potential screened compounds with similar functional groups to the query but in different scaffolds.<sup>97</sup>

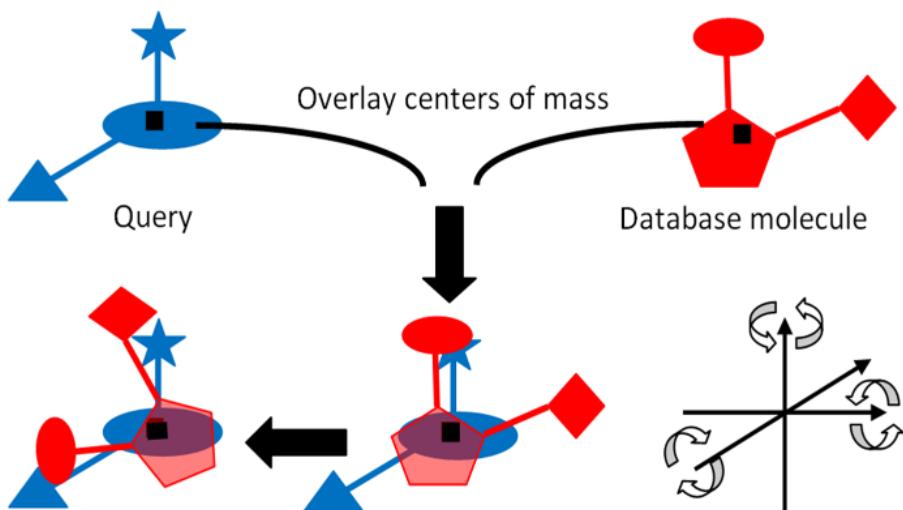


Figure 2.1 Illustration of ROCS algorithm.<sup>98</sup>

# CHAPTER 3 Identification of putative ligands for anti-inflammatory treatment

In this chapter, ligand-based virtual screening is used to predict potential anti-interleukin-1 $\beta$  (IL-1 $\beta$ ) cytokine agents for treatment of inflammatory diseases. The query ligands in this work are based on existing leads which are heterocyclic boron-containing compounds. We firstly introduce the role of the IL-1 $\beta$  cytokine in inflammation and the reported use of boronic acids and some of their ester derivatives as inhibitors of a number of enzymes. In the result section, we report a structure-activity analysis of the existing boron-based anti-IL-1 $\beta$  inhibitors and finally discuss the results of ligand-based virtual screening.

## 3.1 Inflammation

Inflammation is indeed an important target area for new therapies and is the specific area of interest for this chapter's ligand design work. Therefore, we consider in more detail the processes of inflammation here. Inflammation involves the innate and adaptive immune systems and is a complex vascular response to harmful stimuli. These stimuli include pathogens, damaged cells, and irritants for eliminating infection and repairing injured tissue. The initial and immediate defences that protect the host from an attack by other organisms in a generic, non-specific manner are the response from the *innate immune system* and governed by specific immune cells, i.e. macrophages.

In contrast, the *adaptive immune system* responds to pathogens in a specific manner by two major types of lymphocytes: B cells and T cells. The active B cells are involved in production of antibodies. Apart from being antigens themselves, B cells can alternatively be activated by the helper T cells. The helper T cells are not able to kill infected cells or clear invaders, but play an important role in immune functions by signalling other cells through the release of various cytokines including the pro-inflammatory interleukin-1 $\beta$  (IL-1 $\beta$ ). Understanding the role of IL-1 $\beta$  in inflammatory processes and its biological regulation is a key step towards identifying a novel anti-inflammatory therapeutic agent.

A cytokine is a small cell-signalling protein that functions as an immunomodulating agent. They can fall into two different classes: (1) *pro-inflammatory cytokines* that accelerate inflammation, either directly by themselves, or by regulating other cytokines. This is in contrast to (2) *anti-inflammatory cytokines* that act mainly by inhibiting production of pro-inflammatory cytokines or counteracting the biological effects of pro-inflammatory mediators. IL-1 $\beta$  is a pro-inflammatory cytokine that is generated at the site of injury. Its precursor is an inactive pro-IL-1 $\beta$  expressed in response to pathogen-associated molecular patterns (PAMPs) or damage-associated molecular patterns (DAMPs). IL-1 $\beta$  can also be enhanced by second messenger signalling from increased levels of extracellular ATP that bind to the P2X7 purinergic receptor.<sup>99</sup> This will decrease the amount of intracellular potassium, resulting in an oligomerisation of intracellular proteins to form what is termed an “inflammasome”.<sup>100</sup> The NLRP3 inflammasome, for example, participates in progression of some non-infectious diseases such as gout, type II diabetes, and Alzheimer’s disease when inappropriate activation

occurs. It has become an important therapeutic target with many classes of identified small-molecule inhibitors.<sup>66</sup>

The NLRP3 inflammasome consists of the NLRP3 scaffold, the ASC adaptor (or the PYCARD domain) and pro-caspase-1 (see Figure 3.1). The NLRP3 scaffold has three structural units of PAMP/DAMP-sensing C-terminal leucine-rich repeats (LRR), a central nucleotide binding (NACHT) domain, and an N-terminal pyrin (PYD) domain. The PYD domain of NLRP3 interacts with that of the ASC adaptor protein as shown in Figure 3.1. The caspase activation and recruitment (CARD) domain of the ASC adaptor protein also binds to the CARD domain of pro-caspase-1.

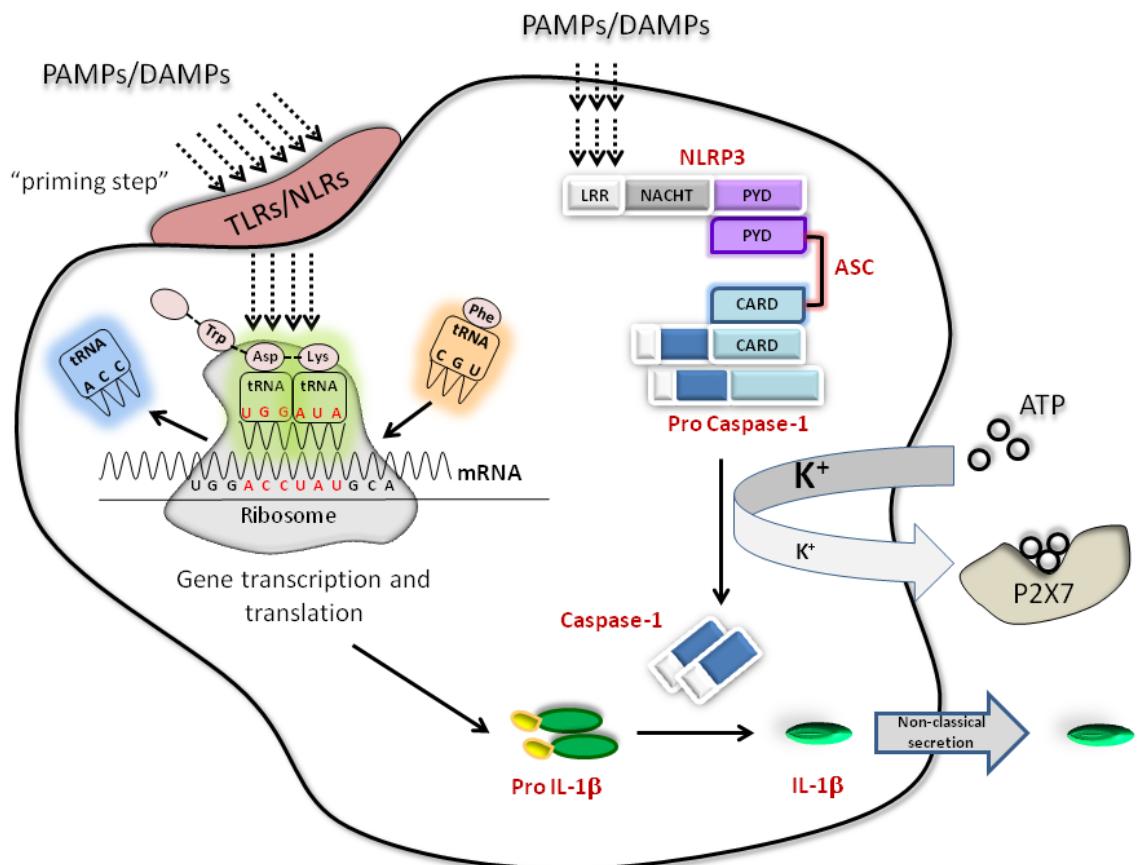


Figure 3.1 Production of active IL-1 $\beta$  cytokines in monocytic cells (adapted from Brough *et al.*<sup>101</sup>).

Although NLRP3 can be activated by pathogens, PAMPs, DAMPs and environmental irritants, the mechanisms leading to NLRP3 inflammasome activation are a controversial issue. The possible models have been reviewed by Tschopp *et al.*<sup>102</sup> Overall, the NLRP3 inflammasome facilitates the conversion of pro-caspase-1, the inactive enzyme precursor, to the active caspase-1. The active caspase-1 subsequently cleaves pro-IL-1 $\beta$  to the active IL-1 $\beta$  cytokine. In acute brain injury, reviewed by Brough *et al.*, processing of IL-1 $\beta$  takes place in M1 microglia and initiates a pro-inflammatory response.<sup>101</sup>

After production and activation of active IL-1 $\beta$ , the final step in regulation of IL-1 $\beta$  is the secretion into an extracellular compartment. Secretion of proteins, in general, depends on either classical or non-classical pathways. For the classical secretory pathway, the majority of secretory proteins are first translocated across the endoplasmic reticulum membrane, and then transported further to the plasma membrane through the Golgi apparatus. However, the non-classical secretory pathway is taken by some proteins that lack a conventional hydrophobic signal sequence, so-called “leaderless proteins”. IL-1 $\beta$  belongs to this class of protein. Therefore, it is believed to be secreted via a non-classical pathway. The mechanism of an IL-1 $\beta$  release has not been established yet, but possible models have been postulated based on a non-classical secretory pathway.<sup>103</sup>

IL-1 $\beta$  plays a key role in acute and chronic inflammation as well as autoimmune disorders. Blocking IL-1 $\beta$  has been currently a standard therapy for “autoinflammatory” syndromes<sup>100</sup> and common diseases mediated by IL-1 $\beta$ . These diseases, for example, are rheumatoid arthritis, smoldering multiple myeloma, post-myocardial infarction heart

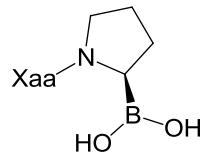
failure and osteoarthritis. In addition, blocking IL-1 $\beta$  can be used to treat metabolic disorders and to induce an inflammatory response to tissue injury in the absence of infection termed “sterile inflammation”.<sup>104</sup> Sterile inflammation responses occur in some acute and chronic diseases such as gout, type 2 diabetes, Alzheimer’s disease and ischemia. Inhibiting IL-1 $\beta$  has been reported to be protective in experimental models of stroke, traumatic brain injury, excitotoxicity and subarachnoid haemorrhage.<sup>101</sup>

### 3.2 Boron-based inhibitors

Boron is in the class of metalloids, exhibiting properties somewhat of a cross between metals and non-metals. It readily forms six-valence compounds, rather than the usual eight as predicted by the octet rule. With the exception of fewer than eight valence electrons, many boron compounds are electron-deficient; therefore, they are often reactive. Boron is also a trivalent metal, unlike a true metal; it has a strong electron affinity due to its vacant 2p-orbital. Boron compounds act as Lewis acids which easily react with nucleophiles and form tetrahedral boron adducts from their trigonal planar geometry.

In addition to such unique chemical properties of boron, dietary boron is the nutrient that can support immune system. It can act as a signal suppressor regulating activities of enzymes involved in an inflammatory process to reduce symptoms of inflammatory diseases.<sup>105</sup> As such, there has recently been interest in exploring compounds which contain this element as potential novel therapeutics. One such class of compounds are based on boronic acids. Several studies report the biological activity of boronic acids

towards theirs targets, leading to a treatment of related diseases as shown in Table 3.1. For example, bortezomib is a dipeptidyl boronic acid, commonly known as Valcade®. It inhibits the 26S proteasome and was approved in May 2003 in the United States for the treatment of multiple myeloma.<sup>106</sup> Another example of a dipeptidyl boronic acid is the Xaa-boroPro class where the ‘boroPro’ refers to the proline in which the C-terminal carboxylate is replaced by a boronic acid group, and the ‘Xaa’ is any unblocked amino acid (Figure 3.2). Its biological target is dipeptidyl peptidase IV (DPP-IV), a serine protease enzyme that cleaves dipeptides from peptides containing either L-proline or L-alanine. It is currently a validated target for the treatment of type 2 diabetes.



**Figure 3.2 Structure of dipeptidyl boronic acid, Xaa-boroPro.**

The Xaa-boroPro dipeptidyl boronic acid shown in Figure 3.2, as an inhibitor of DDP-IV (Table 3.1), can reversibly interconvert to a cyclic form analogous to a diketopiperazine as a function of pH. Both open and cyclic species were successfully isolated in a stable and analytically pure form, and were characterized by Snow *et al.*<sup>107</sup> The open chain is favoured at low pH, but undergoes a cyclisation at higher pH (Figure 3.3). The rate of cyclisation is slow relative to a conformational change ( $t_{1/2} \approx 30$  min at physiological pH).<sup>108</sup> At physiological pH, the equilibrium favours the cyclic form over the open chain by more than two orders of magnitude, but significantly reduces its inhibitory effect. However, the loss of the pharmacological activity with time is helpful to design a ‘soft drug’ which deactivates in a predictable and controlled manner after exerting its

therapeutic effects. The prodrug of Val-boroPro (Table 3.1) activated by and targeted to DPP-IV has been designed based on the soft drug effect.<sup>108</sup> Another example of using the soft drug effect is the prodrug of Phe-boroLeu dipeptidyl boronic acid (Table 3.1) for an inhibition of the proteasome enzyme.<sup>109</sup>

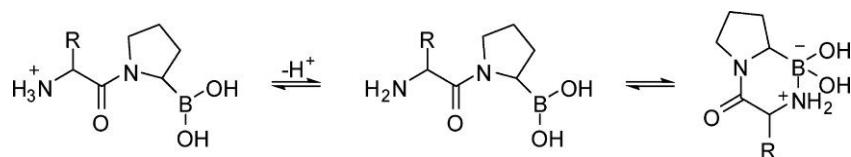
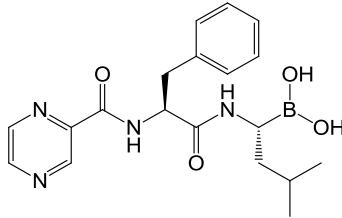
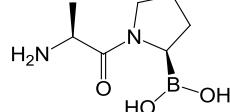
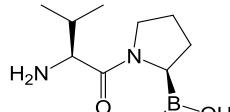
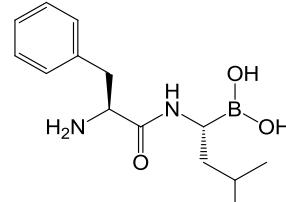


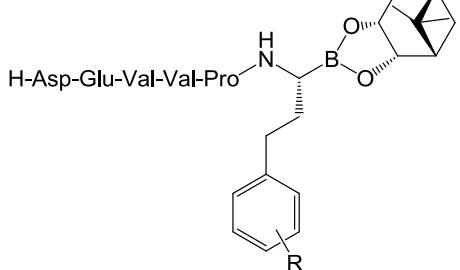
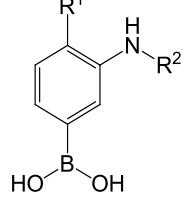
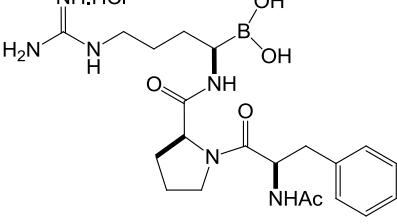
Figure 3.3 Cyclisation equilibrium of Xaa-boroPro.<sup>110</sup>

The unshared electrons of the amino group at the N-terminus of the Xaa-boroPro dipeptidyl boronic acid play a key role in an inhibiting efficacy. The cyclic form of Xaa-boroPro dipeptidyl boronic acid is inactive compared to the open chain because the electron pair of the N-terminal nitrogen is unavailable for the nucleophilic attack at the boron or for binding to the target. Although the open chain is more potent, it is not possible practically to remove the cyclic form. Therefore, there has been an attempt to synthesize conformationally restricted  $\gamma$ -lactam boronate derivatives analogous to Xaa-boroPro dipeptidyl boronic acid.<sup>110</sup> In addition to lactam boronates, 4-substituted boroproline dipeptides have also been synthesized and characterized, as well as tested for their inhibitory potency. These synthesized compounds are found to be potent DPP-IV inhibitors.

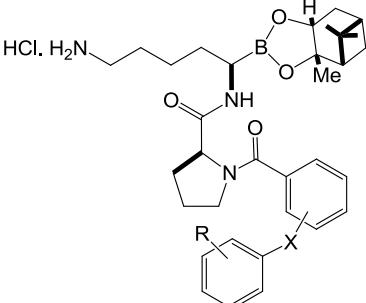
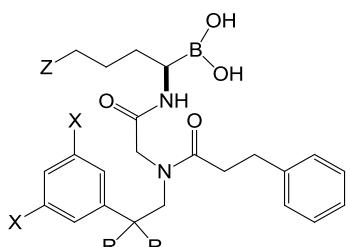
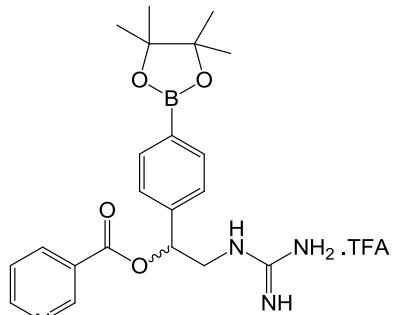
**Table 3.1 Boronic acid inhibitors described in the literature.**

<b>Receptor</b>	<b>Inhibitor</b>	<b>Treatment</b>
26S Proteasome <sup>111</sup>	Bortezomib (Velcade®, PS341)  $K_i$ for proteasome 20S = 0.62 nM	Cancer (refractory multiple myeloma)
Dipeptidyl peptidase IV (DDP-IV) <sup>112</sup>	Ala-boroPro  $K_i$ = 0.027 nM	Type 2 diabetes
	Val-boroPro  $K_i$ = 0.18 nM	Phase III clinical trials for cancer
Proteasome <sup>109</sup>	Phe-boroLeu  $K_i$ = 140 nM	Tumours

**Table 3.1 Boronic acid inhibitors described in the literature (*continued*).**

<b>Receptor</b>	<b>Inhibitor</b>	<b>Treatment</b>
NS3 serine protease <sup>113</sup>	 <p><math>K_i = 0.008 \text{ nM}</math> (<math>R = 4\text{-CF}_3</math>)</p>	Hepatitis C virus (HCV)
$\beta$ -lactamase <sup>114</sup>	 <p><math>K_i = 26 \text{ nM}</math></p> <p><math>R^1 = \text{COOH}; R^2 = \text{---S(=O)(=O)cyclohexyl---S(=O)(=O)cyclohexyl---}</math></p>	$\beta$ -Lactam antibiotics, Bacterial infection
Thrombin <sup>115</sup> Factor Xa (fXa) <sup>116</sup>	 <p><math>K_i = 0.04 \text{nM}</math> for thrombin, 9 nM for fXa</p>	Antithrombotic therapy, Antiblood coagulant, Haemostasis

**Table 3.1 Boronic acid inhibitors described in the literature (*continued*).**

<b>Receptor</b>	<b>Inhibitor</b>	<b>Treatment</b>
Thrombin <sup>115</sup> Factor Xa (fXa) <sup>116</sup>	 <p><math>K_i = 0.07 \text{ nM}</math> for thrombin  <math>(R = 2\text{-CF}_3, X = \text{meta-CH}_2)</math></p>  <p><math>(Z = \text{NHC}(=\text{NH})\text{NH}_2, X = \text{H},)</math>  <math>R = \text{CH}_2, K_i = 0.06 \text{ nM}</math> for thrombin,  <math>46 \text{ nM}</math> for fXa</p>	Antithrombotic therapy, Antiblood coagulant, Haemostasis
	 <p><math>\text{IC}_{50} \text{ fXIa} = 1.4 \mu\text{M}</math></p>	

The use of boronic acid-containing compounds is, however, limited by the distinctive properties of the aliphatic carbon-boron bond. Reactivity with unwanted targets and metabolic instability can cause elimination of the boron atom and therefore loss of biological activity.<sup>117</sup> This can be overcome by utilising the reactive *p*-orbital of the boron, i.e. by incorporating the boron into a structure containing an aromatic, fused aromatic or heteroaromatic ring. One such example is benzoxaborole, containing a benzene ring fused with an oxaborole heterocycle (Figure 3.4). It has higher affinities for diol motifs, especially sugars, than its boronic acid-based parent, phenylboronic acid.<sup>118</sup> Under aqueous condition, the trigonal planar geometry of the boron in benzoxaborole undergoes an acid-base equilibrium with its conjugate base that possesses a tetrahedral geometry (Figure 3.4).<sup>119</sup> This acid-base equilibrium has  $pK_a$  of 7.3 lower than that of phenylboronic acids ( $pK_a$  of 8.7). The measured  $pK_a$  values of benzoxaborole can be tuned by introducing substituent groups on the heterocyclic ring.<sup>120</sup> This modification, therefore, alters its sugar binding properties. The biological activities of some benzoxaborole compounds are given in Table 3.2

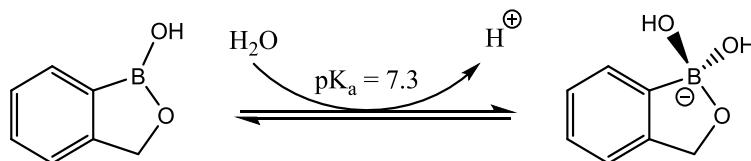
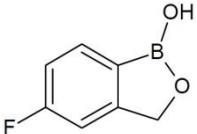
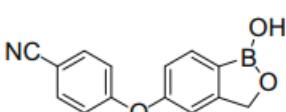
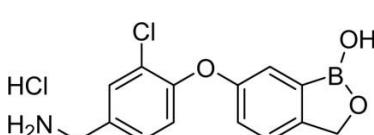
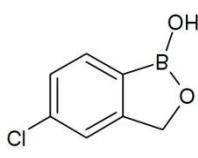
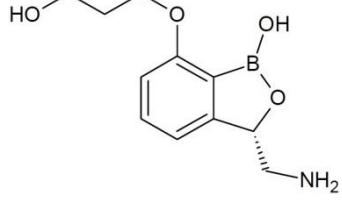
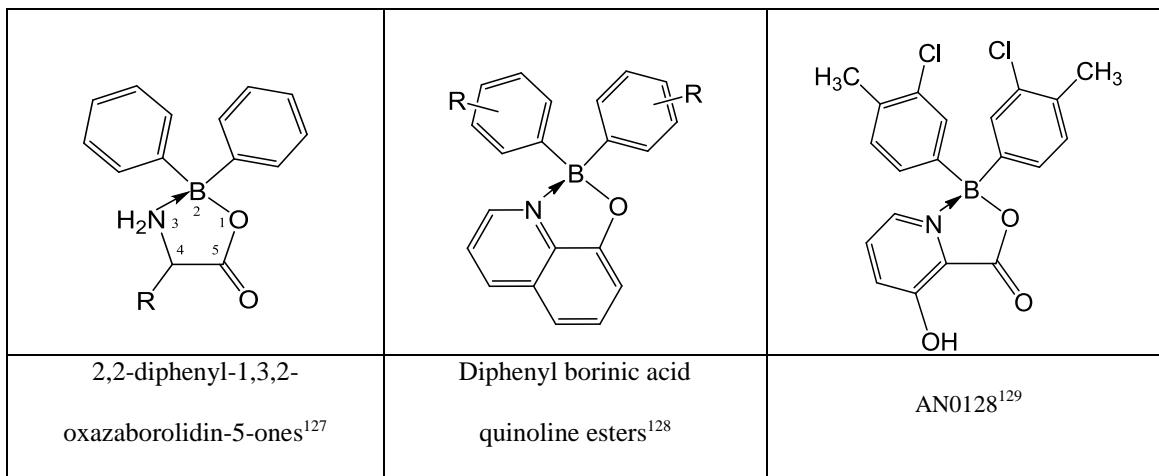


Figure 3.4 Acid-base equilibrium of benzoxaborole.<sup>119</sup>

**Table 3.2 Benzoxaborole inhibitors described in the literature.**

<b>Target/Receptor</b>	<b>Inhibitor</b>	<b>Treatment</b>
Aminoacyl-tRNA synthetase (ARS) <sup>121</sup>	AN2690 	Antifungal agent on Phase 3 clinical development for a treatment of onychomycosis <sup>122</sup>
Phosphodiesterase (PDE4), Targeting tumour necrosis factor- $\alpha$ (TNF- $\alpha$ ) cytokine, and interferon- $\gamma$ (IFN- $\gamma$ ) cytokine <sup>123</sup>	AN2728 	Antiinflammatory agent on Phase 2b clinical trial for a treatment of psoriasis and atopic dermatitis <sup>122</sup>
TNF- $\alpha$ , IL-1 $\beta$ and IL-6 cytokines <sup>124</sup>	AN3485 	Anti-inflammatory agent
Leucyl-tRNA synthetase (LeuRS) <sup>125</sup>	AN2718 	Antifungal agent that completed Phase 1 clinical trial for a treatment of onychomycosis and skin fungal infections <sup>122</sup>
Leucyl-tRNA synthetase (LeuRS) <sup>126</sup>	AN3365/GSK052 	Antibacterial agent (against Gram-negative bacteria)

In addition to boronic acid and benzoxaborole, esters of borinic acid have also exhibited interesting biological activities. The borinic acid ester, namely 2,2-diphenyl-1,3,2-oxazaborolidin-5-ones (Figure 3.5) contains the B-N bond and the  $\alpha$ -amino acid moiety on the heterocyclic ring. Velasco *et al.* showed that the borinic acid ester of glycine is the most highly potent cytotoxicity in murine L5178Y lymphoma cell line among ten amino acid derivatives that were synthesized.<sup>127</sup> They also noted that the modification at the fourth position of the ring decreases the compound potency. This five-membered heterocyclic ring is an active moiety responsible for inducing apoptosis, a programmed cell death that genetically controls morphological and biochemical cell phenomena. The induction of apoptosis is considered as one of the most interesting therapeutic strategies to specifically target cancer cells.



**Figure 3.5 Diphenylborinic acid esters.**

As another example, the boron atom of borinic acid quinoline esters appears to be essential for the antibacterial activity against various Gram positive and Gram negative bacteria, with the minimum inhibitory concentration (MIC) on the order of  $\mu\text{g}/\text{ml}$ . The compounds in this series, shown in Figure 3.5, were designed by Benkovic *et al.* to

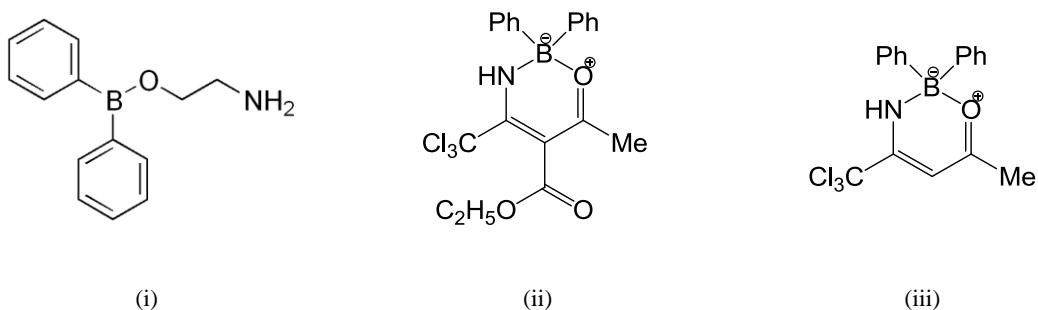
target a DNA methyltransferase enzyme from a Gram-negative  $\alpha$ -proteobacteria and a menaquinone methyltransferase enzyme from a Gram-positive bacteria. The in vitro testing of these compounds against their targets have shown that the chloro substituent group of the borinic acid moiety is important for uncompetitive inhibitory activity, as compared with the fluoro or the methoxy group.<sup>128</sup>

Borinic acid picolinate esters have been studied for the treatment of inflammatory diseases, specifically atopic dermatitis and acne. AN0128 (Figure 3.5) was identified as a novel boron-containing antibacterial agent with anti-inflammatory activity. This compound has been synthesised by Baker *et al.* as a stable compound and can be stored at room temperature.<sup>129</sup> According to Baker's study, AN0128 was the most potent compound of the series, having the best combination of anti-bacterial and anti-inflammatory activities. It contains the chloro substituent group on both phenyl rings, similar to Benkovic *et al.*'s results above. The diphenyl boronic acid moiety is a key motif for AN0128's activity against Gram positive bacteria. From a SAR study, the 3-hydroxyl group on the picolinic ring is essential for its activity.

Interestingly, the compound AN0128 also has activity against pro-inflammatory cytokines.<sup>130</sup> Cytokines are small cell-signalling protein molecules, referred to as an immunomodulating agent (more fully discussed below). AN0128's inhibition of release of pro-inflammatory interleukin-1 $\beta$  cytokines from human peripheral blood mononuclear cells was recorded as a percent of untreated control. Cyclohexamide was used as a positive control. Its activity has meant that it has progressed to phase II clinical trials for the treatment of dermatological disease.

### 3.3 Aims

Anti-inflammatory drug research has been a developing area.<sup>131</sup> The overall goal of this chapter is to use computational methods to identify potent novel compounds as anti-IL-1 $\beta$  agents for treatment of inflammatory diseases. A recently discovered lead for inhibiting release of IL-1 $\beta$  is 2-aminoethyl diphenylborinate (2-APB) in Figure 3.6(i) (Brough *et al.*, unpublished work). This was found to inhibit 75% of IL-1 $\beta$  release in bone marrow derived macrophages (unpublished data), so that the release of IL-1 $\beta$  in the presence of 40  $\mu$ M 2-APB is 25%. In the assay, lipopolysaccharide (LPS) treated (1  $\mu$ g/ml, 4h) bone marrow derived macrophages were incubated with 40  $\mu$ M of 2-APB for 15 min before subsequent stimulation with the NLRP3 inflammasome ATP activator (5 mM, 1h). The release of IL-1 $\beta$  in collected supernatants were quantified by ELISA Kit and then normalized to the ATP with no inhibitor (100%).



**Figure 3.6 Chemical structure of (i) 2-APB, (ii) BC7, and (iii) BC23.**

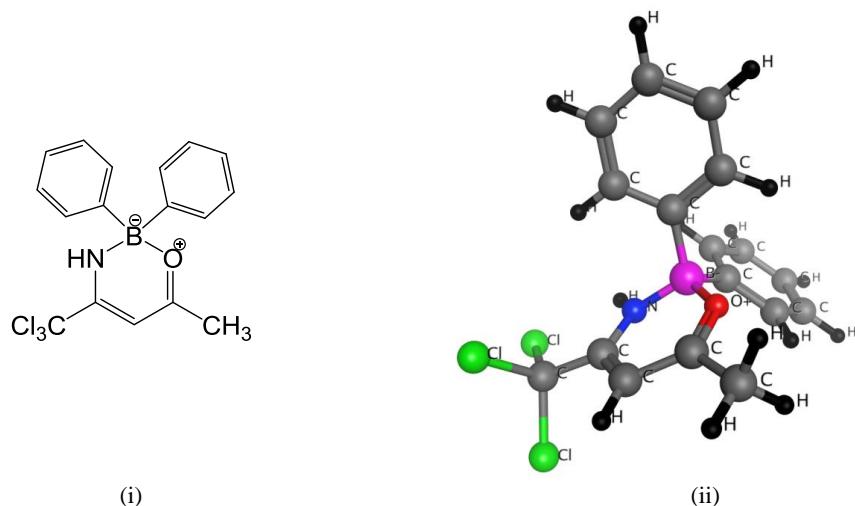
A series of derivatives of 2-APB were assayed (Brough and Freeman, unpublished work). Of these, compounds BC7 and BC23 were found to be more efficient than the lead, with only 5% and 10% of release of IL-1 $\beta$ , respectively (Figure 3.6(ii) and (iii)). The first specific aim of this work therefore was to obtain greater understanding of the

structure-activity relationship for these 2-APB derivatives. Consequently, a quantum mechanical study using density functional theory was performed to compare these boron-containing active compounds with other compounds in the series that had moderate or no inhibitory activity on release of IL-1 $\beta$ . Secondly, a ligand-based virtual screening was performed against the selected active BC23 compound. Although BC7 was the most potent in the series in bone-derived murine macrophages, the second most potent BC23 was found to be more potent than BC7 in human THP-1 monocytes. It was also readily synthesised compared to BC7. Therefore, ligand screening against BC23 aimed to identify potent carbon analogues from commercially available compounds, with similar shape and/or chemical functionality and suitable for lead optimisation.

## 3.4 Methods

### 3.4.1 Generation of BC23 conformations

Initial 3D molecular models of BC23 (Figure 3.7) were created using MOE 2011.10.<sup>132</sup> These were then optimised using the Powell conjugate gradients molecular mechanics method and the Tripos force field implemented in Sybyl 8.0.<sup>133</sup> The MM optimised geometries were quantum mechanically optimised using Gaussian 09<sup>134</sup> with the M06L density functional<sup>135</sup> and 6-31G\* basis set. The atomic partial point charges of those optimised conformations were also calculated by Mulliken population analysis. The molecular charge distribution was then visualized using GaussView5.0.<sup>136</sup> These QM geometries were subsequently used for structure-activity analysis and as search queries for virtual screening.



**Figure 3.7** The query compound of BC23 in (i) 2D chemical structure (ii) 3D geometries.

### 3.4.2 Preparation of compound databases

The ZINC database<sup>137</sup> provides various subsets of commercially available compounds for ligand discovery and virtual screening. These compounds are in the physiologically relevant protonated and tautomeric forms. Unsuitable compounds, such as peroxides, large insoluble molecules, large peptides and highly reactive reagents, were primarily removed by applying the filtering rules. It is noteworthy that compounds containing an atom other than hydrogen, carbon, nitrogen, oxygen, fluorine, sulfur, phosphorus, silicon, chlorine, bromine and iodine cannot be found in this database due to the limitations of using the Merck molecular force field 94 (MMFF94) by ZINC in preparation for its database.<sup>137</sup> Therefore, no boron-containing compounds are in this database.

Compounds in the ZINC database are generally classified by physicochemical properties and subcategorized by a delivery time from purchase. Regarding

physicochemical criteria, the *drug-like subset*,<sup>132</sup> which has been filtered using the Lipinski's rule of five, is the largest class. It contains approximately 26 million compounds. The *lead-like subset*<sup>138</sup> is a smaller subset containing about 11 million compounds whereas the smallest one is the *fragment-like subset*<sup>69</sup> containing 3 million fragments. However, both the *lead-like* and *fragment-like subsets* are more widely used for screening than the *drug-like subset*.

Each of three subsets above are subdivided by a delivery time from purchase into the *Now*, *Clean*, *Standard* and *JFK* subsets. The *Now* subsets contain only compounds that are in stock for rapid delivery within 2 weeks. In *Clean* subsets, stricter filtering rules have been applied so that compounds such as thiols, aldehydes, and Michael acceptors have been removed because they may be potentially problematic in some assays. *Standard subsets* contain most compounds for general screening purpose with 0-10 week(s) delivery time. These subsets have a large number of compounds because they include compounds in the *Now* and *Clean* subsets. Other compounds that are not in the *Now*, *Clean* or *Standard* subsets can be found in *JFK* subsets.

In this study, compounds in the *lead-like* subset (filtering criteria:  $250 \leq \text{MW} \leq 350$ , calculated  $\log P \leq 3.5$  and rotatable bonds  $\leq 7$ ) were screened against BC23, the second most active compound assayed. The subset of *leadsNow* was downloaded on 25 June 2012, containing 1,943,551 molecules in 15 directories (the 20 April 2012 update). Prior to shape-based screening, multiple conformations of each compound were generated by using Omega.<sup>139</sup>.

### 3.4.3 ROCS shape-based screening

ROCS (Rapid Overlay of Chemical Structures)<sup>140</sup> is a ligand-based tool for virtual screening according to similarity in molecular shape and chemical pattern between a query ligand and database compound (more detail in section 2.6). After input preparations, candidate compounds in the *leadsNow* subset were screened against BC23. The top 100 ranked hits of each ZINC sub-directory were extracted and subsequently merged into a single list to depict the top 25 rankings. Finally, two-dimensional (2D) structures and three-dimensional (3D) molecular overlays of the top compounds in the best list were examined by using Vida.<sup>141</sup>

A list of top hits was made when the screened compounds had been ranked by selected scoring functions. The basic scoring function in ROCS is the *ShapeTanimoto* score that quantifies the molecular volume overlap between a query and database compounds based on Gaussian functions. Screened compounds with the optimal shape overlap, however, are possibly aligned in an unfavourable way because of their distinct dimensions and mismatch in chemical functional groups.<sup>83</sup> Therefore, chemical functional groups could be used for considering the matches, instead of or alongside shape similarity. This type of scoring function, the *ColorTanimoto* score that quantifies chemical pattern similarity, is also implemented in ROCS. In this study, a combination of both shape and chemical similarities, namely the *TanimotoCombo* score, was applied, as well as *ShapeTanimoto* and *ColorTanimoto*. The hit lists ranked by these three scoring functions were obtained and compared.

Both *ShapeTanimoto* and *ColorTanimoto* scores are derived from Tanimoto's equation<sup>96</sup>. These scores are in a range of 0 to 1, where 0 stands for no similarity and 1 for identical overlap. Hence, this gives the range of 0 to 2 for the *TanimotoCombo* score because both scores are equally combined. The *ColorTanimoto* is a newer version of the *ScaledColor* score. The *ScaledColor* score is calculated by dividing a hit's actual colour score, namely the *ColorScore* by the colour score of the query to provide a relative value for a chemical similarity to the query. A combination of the *ScaledColor* and the *ShapeTanimoto* results in the *ComboScore*. In addition to the scores derived from Tanimoto's equation, ROCS also provides other scoring functions derived from the Tversky equation.<sup>67</sup> These scores for each rank of screened compounds will also be reported (see Table 3.5 - 3.7 in section 3.5.2 ROCS screening results). However, we will mainly focus on three scoring functions: the *TanimotoCombo*, *ShapeTanimoto* and *ColorTanimoto* scores.

To refine the shape superposition, chemical similarity is quantified by using the colour force fields with the key chemical functional groups. Six chemical features were matched during a screen: hydrophobe, hydrogen bond donor, hydrogen bond acceptor, anion, cation and ring. The hydrophobe feature is available for describing hydrophobic groups such as phenyl groups. The hydrogen bond donors and hydrogen bond acceptors were defined by using the extended version of the original hydrogen bond model introduced by Mills and Dean.<sup>142</sup> The interactions among these functional groups were described in two built-in colour force fields—the ImplicitMillsDean and the ExplicitMillsDean force fields. Both types have a similar overall interaction model. By using the ImplicitMillsDean, anions and cations are defined according to a simple pKa model assuming a pH of 7. The assignment of ionic charge is automated regardless of a

protonation state setting in the input structure of a query or database compounds whereas, in ExplicitMillsDean, the protonation or charge state of each molecule is explicitly interpreted as its actual state. However, a user-defined colour force field can be an option by using the checkcff utility and applying it in a specified format.

### 3.5 Results and discussion

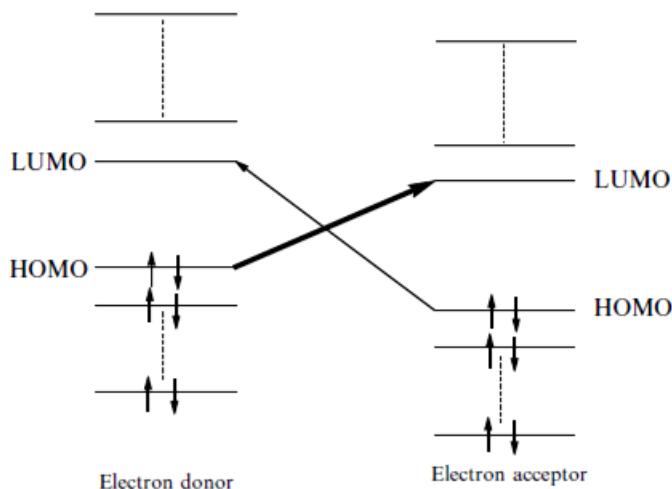
A series of five inhibitors (see the structures in Table 3.3) were examined in this study; each is a heterocyclic compound possessing tetravalent boron. These compounds are the analogues of the known lead, 2-aminoethyl diphenylborinate (2-APB). 2-APB can inhibit by 75% the release of interleukin-1 $\beta$  (IL-1 $\beta$ ) in bone marrow-derived macrophages (see section 3.3 for details about the biological assay). They have similar structures except for the substituent groups in a heterocyclic ring. However, their inhibitory effect on release of IL-1 $\beta$  shown in Table 3.3 varies from the most active BC7 (5% of release of IL-1 $\beta$ ) to the inactive BC21 (100% of release of IL-1 $\beta$ ). To explore potential relationships between the structures and the inhibitory activities of these compounds, calculations based on density function theory were performed. The following sections will discuss the results of DFT calculations and present the top ranked compounds obtained from virtual screens for the selected compound of BC23.

#### 3.5.1 DFT analysis of 2-APB analogue compounds

Here we explore a potential relationship between the electronic structure of the inhibitory compounds and their biological activity. This section will focus on using

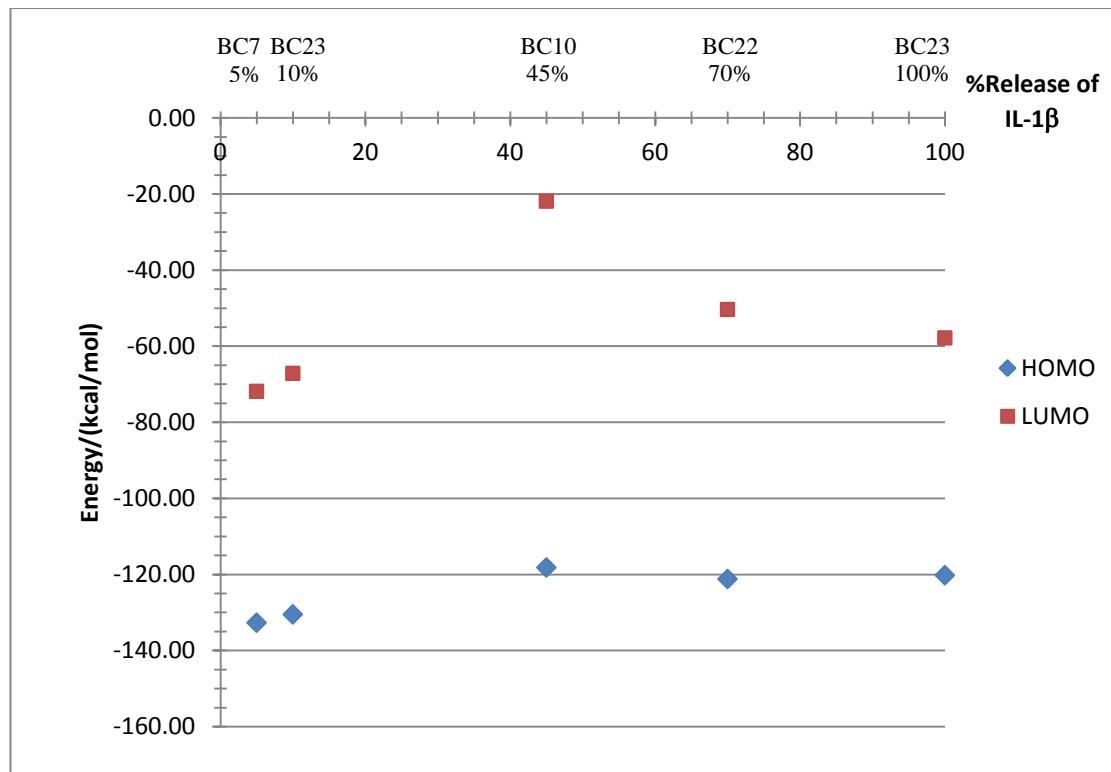
DFT for determining the energies and shape of the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO) of the five boron-based heterocyclic compounds (Table 3.3), as well as their charge distributions.

Two fundamental aspects for describing chemical reactivity are electron transfer effects and electrostatic interactions. Electron transfer effects relate to frontier orbitals and the ability of a chemical group to be a good donor or acceptor when a reaction takes place. Frontier molecular orbital theory can predict whether a reaction will take place by focusing on the shape and the symmetry of HOMO and LUMO. When the HOMO of one molecule (as an electron donor) and the LUMO of another molecule (as an electron acceptor) have the same shape and phase, the electrons can transfer from the HOMO to the LUMO and form a chemical bond. A strong chemical bond occurs when the energy gap between the HOMO of the electron donor and the LUMO of the electron acceptor is small.<sup>143</sup> Therefore, a good electron donor should have a high HOMO whereas a good electron acceptor should have a low LUMO (Figure 3.8). The smaller this energy gap is, the easier the chemical bond will be to form.



**Figure 3.8** Frontier orbital interactions.<sup>143</sup>

The structure and energetics of BC7, BC10, BC21, BC22 and BC23 were computed at the M06L/6-31G\* level of theory (Figure 3.9 and Table 3.3). From the calculated frontier orbital energies, the two lowest LUMO energies are  $-0.114509\text{ E}_h$  and  $-0.106986\text{ E}_h$  for BC7 and BC23, respectively, which are the two most active compounds (Table 3.3). Because BC7 has the lowest calculated LUMO energy ( $-71.9\text{ kcal/mol}$  in Figure 3.9 or  $-0.114509\text{ E}_h$  in Table 3.3), it is predicted to be the best electron acceptor in this series of compounds. However the LUMO energy does not track linearly with IL-1 $\beta$  release. In all five compounds, the LUMO is delocalised dominantly on the heterocyclic ring and over the substituent groups but not the phenyl rings.

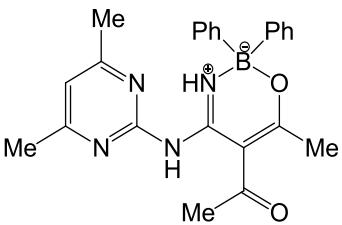
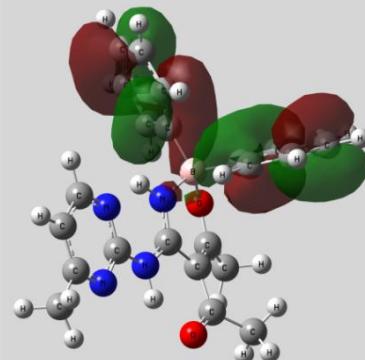
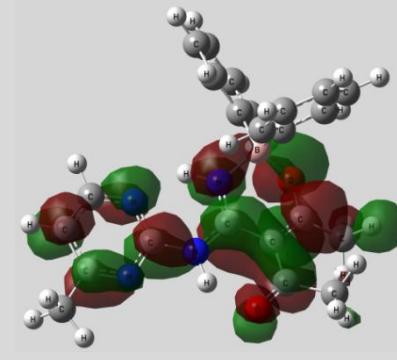
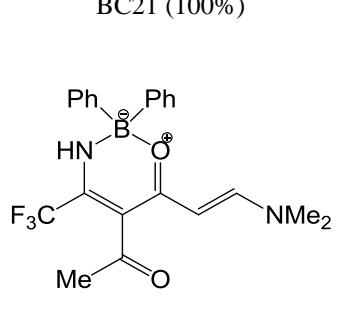
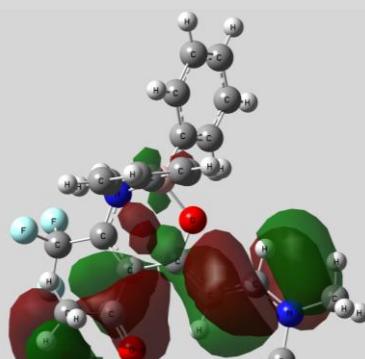
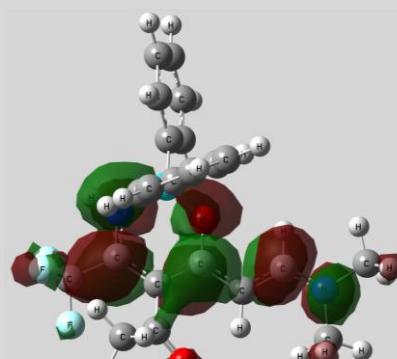


**Figure 3.9** HOMO and LUMO energies (in kcal/mol) of five boron-based heterocyclic compounds.

**Table 3.3 Shape, symmetries and energies (in Hartrees) of the HOMO and LUMO of five boron-based heterocyclic compounds.**

Compounds (% release of IL-1 $\beta$ )	Energy / (Hartree)	
	HOMO	LUMO
BC7 (5%) 	-0.211445 	-0.114509 
BC23 (10%) 	-0.207955 	-0.106986 
BC10 (45%) 	-0.188309 	-0.034867 

**Table 3.3 Shape, symmetries and energies of the HOMO and LUMO of five boron-based heterocyclic compounds (*continued*).**

Compounds (% release of IL-1 $\beta$ )	Energy / (Hartree)	
	HOMO	LUMO
BC22 (70%)  	-0.193120  	-0.080221  
BC21 (100%)  	-0.191559  	-0.092138  

Interestingly, the HOMO energies of the two most active compounds, BC7 and BC23, are respectively  $-0.211445\text{ E}_\text{h}$  ( $-132.7\text{ kcal/mol}$  in Figure 3.9) for BC7 and  $-0.207955\text{ E}_\text{h}$  ( $-130.5\text{ kcal/mol}$  in Figure 3.9) for BC23, which are the lowest energies of the five compounds. This suggests these compounds are the least nucleophilic. Electrons in the HOMO of the active boron compounds (BC7, BC23, BC10 and BC22) delocalise only

on the phenyl rings (Figure 3.9). This is in contrast to mainly delocalisation on the substituent group at *meta*-carbon of the heterocyclic ring of the inactive BC21. For further information on electronic structure, we next consider analysis of charge distribution.

Mulliken partial charges were computed for all atoms in the five compounds at the M06L/6-31G\* level of theory (Table 3.4). The calculated Mulliken charge on the boron atom in BC7, BC23, BC10, BC22 and BC21 is 0.595, 0.584, 0.639, 0.590 and 0.606  $e$ , respectively. It is positive in value for all compounds (see Appendix A for a representation of a molecular charge distribution). Therefore, these boron compounds are likely to be electrophilic at the boron centre. This also supports the idea that they are an electron acceptor, as suggested by frontier orbital analysis above. The positive charge on the boron was also predicted from *ab initio* calculations using the HF/6-31G\* basis set on substituted 2-aminoethylborinates by Höpfl *et al.*<sup>144</sup> This study also stated, based on the first observation by Hoffmann,<sup>145</sup> that the electrostatic charge on the boron in the five-membered 2-aminoethylborinate coordinate compounds studied is always positive. This is due to the low electronegativity of the boron atom, reported as 3.88 compared to values of, for example, 7.18 for H, 7.62 for N and 7.45 for the methyl group. Here, the M06 functional was used because it is a recently derived functional which has demonstrated accuracy in modelling noncovalent interactions and applications in organometallic and inorganometallic chemistry.<sup>135, 146</sup> Furthermore, it can reproduce the experimental trend in the coordinate B-N bond lengths and bond enthalpies of methyl-substituted aminoboranes.<sup>147</sup>

**Table 3.4 Mulliken partial atomic charges on boron atom (in  $e$ ) and bond lengths of heterocyclic boron compounds derived from DFT calculations.**

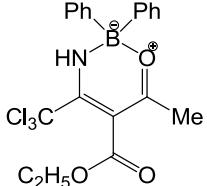
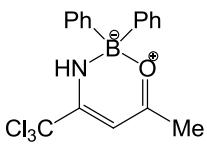
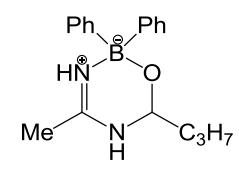
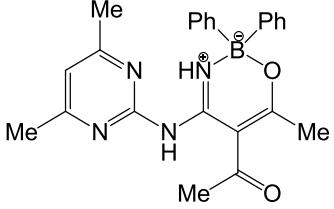
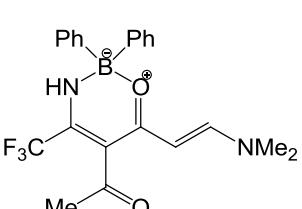
%release of IL-1 $\beta$	Compounds		Boron Mulliken charge	Bond length / ( $\text{\AA}$ )	
				B-N	B-O
5	BC7		+0.595	1.577	1.562
10	BC23		+0.584	1.593	1.551
45	BC10		+0.639	1.622	1.485
70	BC22		+0.590	1.559	1.543
100	BC21		+0.606	1.566	1.541

Table 3.4 shows the B-N and B-O bond lengths measured from the quantum mechanically optimised structures of all five BC compounds. The optimal B-N and B-O bond lengths of the energy minimised geometries are comparable for all compounds except BC10, the moderately active compound. For example, the measured B-N bond length of the two most active molecules, BC7 and BC23, is respectively 1.577 and

1.593 Å whereas it is 1.622 Å in BC10 (Table 3.4). A similar trend was found for the measured B-O bond length: it is respectively 1.562 and 1.551 Å in BC7 and BC23 whereas it is 1.485 Å in BC10. The difference in B-N and B-O bond lengths of BC10 is likely due to the  $sp^3$  hybridization of the carbon bonded to the oxygen in contrast to  $sp^2$  in the other four BC compounds. Although the B-N bond in BC10 is slightly longer and the B-O bond slightly shorter than those bonds in the other compounds, these two bonds are close to the experimental values obtained from the crystal structures of 2-aminoethyl diphenylboronate: by using X-ray crystallography direct method and full-matrix least-square refinement, its B-N distance was 1.653(3) Å for monoclinic and 1.657(3) Å for orthorhombic crystals; the B-O distance was 1.484(3) Å for monoclinic and 1.478(3) Å for orthorhombic symmetry.<sup>148</sup> The B-N bond is longer than the B-O bond in both the quantum mechanically optimised BC structures and the crystal structure of 2-aminoethyl diphenylboronate.

From a plot of Mulliken charge on the boron and the B-N and B-O bond lengths against their inhibitory efficacy as the percent release of IL-1 $\beta$  (Figure 3.10), however, it is clear there is not a simple trend with inhibitory activity. This was also true for analysis of frontier orbitals (Figure 3.8). Nevertheless, the most active compound had the lowest lying LUMOs, situated on the heterocyclic ring, and may point to a role in inhibition at the boron centre, either covalently or noncovalently.

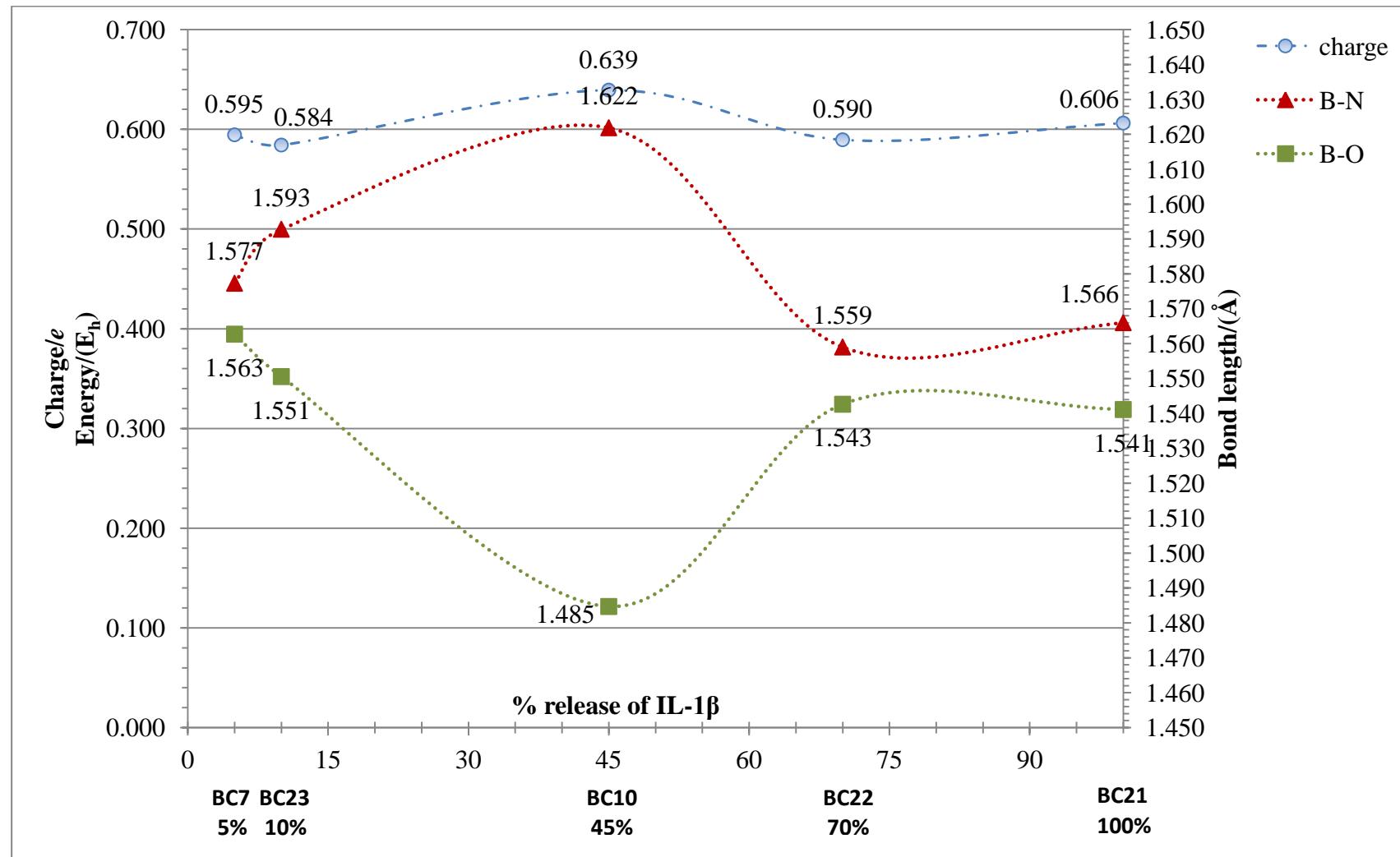
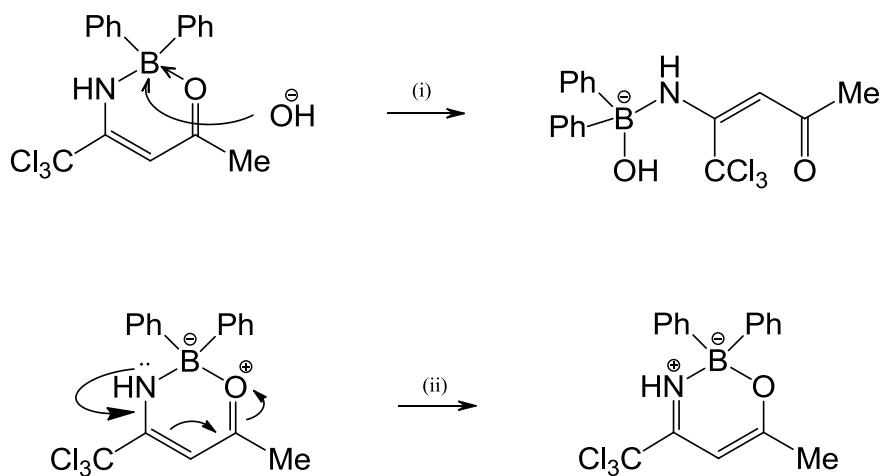


Figure 3.10 Summary of DFT analysis of five 2-APB analogue compounds that can inhibit release of IL-1 $\beta$ .

From the preliminary studies by Dr. Sally Freeman and Mr. Alex Baldwin, the active boron compounds show biological activities in aqueous condition. It is possible that the product in an open chain form, arising from nucleophilic attack by the solvent (Figure 3.11(i)), could contribute to the observed activity. However, the X-ray structures of the more potent BC23 derivatives (from the informal communication with Mr. Alex Baldwin; unpublished data) confirm the cyclic arrangement, indicating that the resonance structures of BC23 could be possible as shown in Figure 3.11(ii).



**Figure 3.11 Hypothetical forms of biologically active BC23 (from the informal communication with Mr. Alex Baldwin).**

### 3.5.2 ROCS screening

It was then of interest to further explore the uniqueness of the boron atom in inhibition of IL-1 $\beta$  release. Consequently, this section focuses on using a virtual screen to identify commercially available non-boron compounds, of analogous shape and/or chemical reactivity to the query compound BC23, which then could be assayed for their potency

as anti-IL-1 $\beta$ -cytokine agents. Virtual screening was carried out by using the ROCS program.

For compound scoring, there was a choice of scoring functions in ROCS. The *TanimotoCombo* score is the recommended scoring function in ROCS, as an equally weighted combination of the *ShapeTanimoto* and *ColorTanimoto* scores. The *ShapeTanimoto* score ranks the screened molecules based on the shape overlap while the *ColorTanimoto* score matches the “colour atoms” for chemical similarity. Both the *ShapeTanimoto* and the *ColorTanimoto* scores are in a range of 0 for no similarity to 1 for identical overlap, and so the *TanimotoCombo* score ranges from 0 to 2. However, other scoring functions are also provided in ROCS for use (see section 2.6 and 3.4.3 for more details about ROCS and its scoring functions).

Tables 3.5, 3.6 and 3.7 show the scores of the top 25 lead-like compounds screened against BC23 when they are ranked by the *TanimotoCombo*, *ShapeTanimoto* and *ColorTanimoto*, respectively (see Appendix B for 2D structures and 3D overlays on the query BC23). The best match of BC23 ranked by using the *TanimotoCombo* score is the compound ZINC02389343 (Table 3.5). It has the highest *TanimotoCombo* score of 1.758 while its *ShapeTanimoto* and *ColorTanimoto* scores are 0.798 and 0.960, respectively. Its chemical structure and 3D overlay on its query, the BC23, are presented in Figure 3.12. This screened compound has two phenyl rings similar to BC23; the *N,N*-dimethylamino group occupies the position of the oxygen in BC23 where the disposition of its oxygen of the tetrahydrofuran is at the position of the *meta*-carbon of BC23.

ZINC02389343 is compared to the best match ranked by using the *ShapeTanimoto* score, ZINC04590678 (Table 3.6). As can be seen from the overlay of ZINC04590678 on the query BC23 in Figure 3.12, this best shape-matching compound is more similar in shape to BC23 than the best match ranked by using *TanimotoCombo* score, ZINC02389343. It has the higher *ShapeTanimoto* score of 0.894; therefore, it can resemble BC23's shape better than the best match of the *TanimotoCombo*'s rank. However, it has a poorer *ColorTanimoto* score of 0.477, resulting in the lower *TanimotoCombo* score of 1.371.

During the ROCS ranking process, it is possible that more than one hit can have the same score. Therefore, the screened compounds ranked firstly by a given scoring function were secondarily ranked by using the *Overlap* score, the absolute value of a volume overlap between a screened molecule and a query. For example, the best match by using the *ColorTanimoto* score (Table 3.7) is ZINC71856537 which has the *ColorTanimoto* score of 0.993, the same as the second rank, ZINC71856536; but it has a higher overlap score of 660.630. However, these two top chemical-matched structures are diastereoisomers, different in the configuration of the chlorophenyl group (Figure 3.12).

**Table 3.5 Top 25 hits of BC23 ranked by the *TanimotoCombo* scoring function.**

No.	Compound Name	ID No.	Tanimoto Combo	Shape Tanimoto	Color Tanimoto	Overlap	Ref Tversky	RefColor Tversky	RefTver skyCombo	FitTversky	FitColor Tversky	FitTversky Combo	Scaled Color	Color Score	Rank	
<b>1</b>	ZINC02389343_0	1103	<b>1.758</b>	0.798	0.960	<b>848.233</b>	0.853	0.980	1.833	0.926	0.980	1.905	0.980	1.778	-2.939	1
<b>2</b>	ZINC06002617_3	3	<b>1.657</b>	0.674	0.983	<b>777.248</b>	0.781	0.991	1.772	0.831	0.991	1.823	0.991	1.665	-2.974	1
<b>3</b>	ZINC48289632_0	1203	<b>1.634</b>	0.678	0.956	<b>784.131</b>	0.787	0.978	1.765	0.830	0.978	1.807	0.978	1.655	-2.933	1
<b>4</b>	ZINC04970524_105	503	<b>1.634</b>	0.683	0.950	<b>758.584</b>	0.765	0.974	1.739	0.866	0.974	1.840	0.974	1.658	-2.923	1
<b>5</b>	ZINC48289630_16	1204	<b>1.632</b>	0.669	0.963	<b>778.599</b>	0.782	0.981	1.763	0.822	0.981	1.804	0.981	1.650	-2.944	2
<b>6</b>	ZINC04970522_100	504	<b>1.625</b>	0.673	0.953	<b>751.416</b>	0.757	0.976	1.733	0.857	0.976	1.833	0.976	1.648	-2.927	2
<b>7</b>	ZINC04970408_23	505	<b>1.624</b>	0.675	0.950	<b>754.525</b>	0.760	0.974	1.734	0.857	0.974	1.831	0.974	1.649	-2.922	3
<b>8</b>	ZINC04970406_35	506	<b>1.623</b>	0.673	0.950	<b>754.027</b>	0.760	0.974	1.734	0.855	0.974	1.829	0.974	1.647	-2.923	4
<b>9</b>	ZINC15780439_17	703	<b>1.614</b>	0.691	0.923	<b>744.458</b>	0.752	0.960	1.712	0.895	0.960	1.855	0.960	1.651	-2.880	1
<b>10</b>	ZINC27974081_69	303	<b>1.612</b>	0.623	0.989	<b>747.059</b>	0.750	0.994	1.744	0.786	0.994	1.780	0.994	1.617	-2.983	1
<b>11</b>	ZINC27974084_83	304	<b>1.609</b>	0.621	0.988	<b>746.127</b>	0.749	0.994	1.743	0.784	0.994	1.778	0.994	1.615	-2.981	2
<b>12</b>	ZINC71856536_13	704	<b>1.600</b>	0.608	0.993	<b>701.950</b>	0.708	0.996	1.704	0.811	0.996	1.807	0.996	1.604	-2.989	2
<b>13</b>	ZINC72152461_12	403	<b>1.592</b>	0.671	0.921	<b>766.133</b>	0.771	0.959	1.729	0.839	0.959	1.798	0.959	1.630	-2.876	1
<b>14</b>	ZINC00225590_4	203	<b>1.590</b>	0.666	0.924	<b>757.512</b>	0.762	0.960	1.723	0.841	0.960	1.801	0.960	1.627	-2.881	1
<b>15</b>	ZINC00156167_2	603	<b>1.587</b>	0.664	0.923	<b>707.565</b>	0.717	0.960	1.677	0.900	0.960	1.860	0.960	1.624	-2.880	1
<b>16</b>	ZINC71856537_12	705	<b>1.574</b>	0.584	0.990	<b>684.605</b>	0.690	0.995	1.685	0.791	0.995	1.786	0.995	1.579	-2.985	3
<b>17</b>	ZINC05132639_1	706	<b>1.573</b>	0.846	0.727	<b>887.992</b>	0.892	0.966	1.858	0.943	0.746	1.689	0.982	1.828	-2.946	4
<b>18</b>	ZINC05663305_0	707	<b>1.572</b>	0.873	0.699	<b>974.192</b>	0.971	0.944	1.915	0.897	0.729	1.626	0.960	1.833	-2.879	5
<b>19</b>	ZINC01591149_0	708	<b>1.571</b>	0.873	0.698	<b>947.314</b>	0.947	0.944	1.890	0.918	0.729	1.646	0.959	1.832	-2.878	6
<b>20</b>	ZINC05132640_2	103	<b>1.568</b>	0.845	0.723	<b>887.702</b>	0.892	0.963	1.855	0.941	0.744	1.685	0.979	1.824	-2.938	1
<b>21</b>	ZINC01675359_3	709	<b>1.564</b>	0.848	0.716	<b>873.180</b>	0.879	0.958	1.836	0.961	0.740	1.701	0.974	1.822	-2.921	7
<b>22</b>	ZINC37629199_3	903	<b>1.562</b>	0.843	0.719	<b>880.390</b>	0.885	0.960	1.844	0.947	0.741	1.688	0.976	1.819	-2.927	1
<b>23</b>	ZINC23398363_131	507	<b>1.561</b>	0.606	0.955	<b>693.153</b>	0.700	0.977	1.677	0.819	0.977	1.796	0.977	1.583	-2.931	5
<b>24</b>	ZINC03627349_0	104	<b>1.559</b>	0.842	0.717	<b>870.435</b>	0.876	0.958	1.834	0.957	0.740	1.696	0.974	1.816	-2.922	2
<b>25</b>	ZINC03627351_3	305	<b>1.559</b>	0.831	0.728	<b>864.211</b>	0.869	0.967	1.836	0.950	0.746	1.696	0.983	1.814	-2.948	3

**Table 3.6 Top 25 hits of BC23 ranked by the *ShapeTanimoto* scoring function.**

No.	Compound Name	ID No.	TanimotoCombo	Shape Tanimoto	Color Tanimoto	Overlap	Ref Tversky	RefColor Tversky	RefTversky skyCombo	FitTversky	FitColor Tversky	FitTversky Combo	Scaled Color	Combo Score	Color Score	Rank
<b>1</b>	ZINC04590678_0	403	1.371	<b>0.894</b>	0.477	<b>930.092</b>	0.933	0.923	1.856	0.956	0.497	1.453	0.969	1.863	-2.908	<b>1</b>
<b>2</b>	ZINC67818675_34	1103	1.205	<b>0.891</b>	0.315	<b>965.993</b>	0.965	0.871	1.835	0.921	0.33	1.251	0.958	1.848	-2.873	<b>1</b>
<b>3</b>	ZINC04006536_8	303	1.113	<b>0.891</b>	0.221	<b>873.764</b>	0.881	0.614	1.495	1.013	0.257	1.27	0.665	1.556	-1.995	<b>1</b>
<b>4</b>	ZINC01672685_2	603	1.168	<b>0.887</b>	0.282	<b>911.506</b>	0.915	0.628	1.544	0.966	0.338	1.304	0.66	1.546	-1.979	<b>1</b>
<b>5</b>	ZINC71851201_133	304	1.160	<b>0.886</b>	0.273	<b>918.778</b>	0.922	0.613	1.535	0.958	0.33	1.288	0.644	1.53	-1.931	<b>2</b>
<b>6</b>	ZINC57292122_64	1203	1.217	<b>0.886</b>	0.331	<b>878.840</b>	0.886	0.642	1.528	1	0.406	1.406	0.664	1.549	-1.991	<b>1</b>
<b>7</b>	ZINC03592562_7	404	1.294	<b>0.884</b>	0.410	<b>940.314</b>	0.941	0.909	1.85	0.935	0.428	1.363	0.97	1.853	-2.909	<b>2</b>
<b>8</b>	ZINC18143161_14	1303	1.063	<b>0.882</b>	0.182	<b>881.205</b>	0.888	0.597	1.484	0.992	0.207	1.199	0.666	1.548	-1.999	<b>1</b>
<b>9</b>	ZINC02620374_30	503	1.291	<b>0.881</b>	0.410	<b>963.302</b>	0.962	0.909	1.871	0.913	0.428	1.341	0.969	1.85	-2.908	<b>1</b>
<b>10</b>	ZINC21984511_2	1204	1.450	<b>0.880</b>	0.570	<b>926.314</b>	0.928	0.937	1.865	0.945	0.592	1.537	0.968	1.848	-2.903	<b>2</b>
<b>11</b>	ZINC01700895_12	1205	1.158	<b>0.880</b>	0.278	<b>882.134</b>	0.888	0.622	1.51	0.99	0.335	1.324	0.653	1.533	-1.959	<b>3</b>
<b>12</b>	ZINC02629053_31	3	1.238	<b>0.879</b>	0.359	<b>983.185</b>	0.979	0.895	1.874	0.895	0.375	1.271	0.969	1.848	-2.908	<b>1</b>
<b>13</b>	ZINC03572839_12	405	1.289	<b>0.879</b>	0.410	<b>942.855</b>	0.943	0.909	1.852	0.928	0.428	1.356	0.97	1.849	-2.909	<b>3</b>
<b>14</b>	ZINC01611635_1	1206	1.359	<b>0.879</b>	0.480	<b>937.683</b>	0.939	0.926	1.865	0.933	0.499	1.432	0.972	1.852	-2.917	<b>4</b>
<b>15</b>	ZINC68602906_21	504	1.237	<b>0.878</b>	0.359	<b>973.034</b>	0.97	0.895	1.865	0.902	0.375	1.277	0.969	1.847	-2.908	<b>2</b>
<b>16</b>	ZINC40160367_61	103	1.162	<b>0.878</b>	0.285	<b>948.601</b>	0.948	0.633	1.582	0.922	0.341	1.263	0.665	1.543	-1.995	<b>1</b>
<b>17</b>	ZINC67756807_8	4	1.137	<b>0.877</b>	0.260	<b>932.016</b>	0.933	0.849	1.783	0.936	0.272	1.208	0.962	1.839	-2.887	<b>2</b>
<b>18</b>	ZINC67678162_8	803	1.192	<b>0.876</b>	0.317	<b>956.427</b>	0.955	0.875	1.83	0.913	0.332	1.245	0.962	1.838	-2.887	<b>1</b>
<b>19</b>	ZINC01608670_4	604	1.441	<b>0.876</b>	0.565	<b>909.977</b>	0.913	0.932	1.845	0.956	0.589	1.545	0.963	1.839	-2.889	<b>2</b>
<b>20</b>	ZINC05185509_102	406	0.959	<b>0.876</b>	0.083	<b>905.035</b>	0.909	0.299	1.208	0.96	0.104	1.064	0.334	1.209	-1.001	<b>4</b>
<b>21</b>	ZINC69570522_90	506	1.194	<b>0.875</b>	0.318	<b>949.210</b>	0.949	0.623	1.572	0.919	0.394	1.313	0.644	1.519	-1.932	<b>4</b>
<b>22</b>	ZINC67787792_28	505	1.125	<b>0.875</b>	0.250	<b>933.988</b>	0.935	0.822	1.758	0.932	0.264	1.196	0.932	1.807	-2.796	<b>3</b>
<b>23</b>	ZINC04123460_12	407	1.140	<b>0.873</b>	0.267	<b>987.837</b>	0.983	0.714	1.697	0.886	0.299	1.186	0.774	1.647	-2.321	<b>5</b>
<b>24</b>	ZINC05663305_0	703	1.572	<b>0.873</b>	0.699	<b>974.192</b>	0.971	0.944	1.915	0.897	0.729	1.626	0.96	1.833	-2.879	<b>1</b>
<b>25</b>	ZINC67642529_103	203	1.156	<b>0.873</b>	0.283	<b>970.007</b>	0.967	0.856	1.823	0.899	0.297	1.196	0.956	1.829	-2.868	<b>1</b>

**Table 3.7 Top 25 hits of BC23 ranked by the *ColorTanimoto* scoring function.**

No.	Compound Name	ID No.	Tanimoto Combo	ShapeTani moto	Color Tanimoto	Overlap	Ref Tversky	RefColor Tversky	RefTver skyCombo	FitTversky	FitColor Tversky	FitTversky Combo	Scaled Color	Color Score	Rank	
<b>1</b>	ZINC71856537_14	704	1.545	0.552	<b>0.993</b>	<b>660.630</b>	0.666	0.997	1.663	0.763	0.997	1.760	0.997	1.549	-2.990	2
<b>2</b>	ZINC71856536_15	703	1.539	0.545	<b>0.993</b>	<b>655.443</b>	0.661	0.997	1.658	0.757	0.997	1.754	0.997	1.542	-2.990	1
<b>3</b>	ZINC27974081_69	303	1.612	0.623	<b>0.989</b>	<b>747.059</b>	0.750	0.994	1.744	0.786	0.994	1.780	0.994	1.617	-2.983	1
<b>4</b>	ZINC27974084_73	304	1.547	0.559	<b>0.988</b>	<b>697.920</b>	0.701	0.994	1.695	0.734	0.994	1.728	0.994	1.553	-2.982	2
<b>5</b>	ZINC06002617_3	3	1.657	0.674	<b>0.983</b>	<b>777.238</b>	0.781	0.991	1.772	0.831	0.991	1.822	0.991	1.665	-2.974	1
<b>6</b>	ZINC48289630_16	1203	1.632	0.669	<b>0.963</b>	<b>778.556</b>	0.782	0.981	1.763	0.822	0.981	1.804	0.981	1.650	-2.944	1
<b>7</b>	ZINC02389343_2	1103	1.646	0.685	<b>0.961</b>	<b>776.977</b>	0.781	0.980	1.761	0.848	0.980	1.828	0.980	1.665	-2.940	1
<b>8</b>	ZINC48289632_15	1204	1.546	0.585	<b>0.961</b>	<b>717.214</b>	0.720	0.980	1.700	0.758	0.980	1.738	0.980	1.565	-2.940	2
<b>9</b>	ZINC23398363_131	503	1.561	0.606	<b>0.955</b>	<b>693.153</b>	0.700	0.977	1.677	0.819	0.977	1.796	0.977	1.583	-2.931	1
<b>10</b>	ZINC23398366_63	504	1.534	0.581	<b>0.954</b>	<b>674.687</b>	0.681	0.976	1.657	0.797	0.976	1.774	0.976	1.557	-2.929	2
<b>11</b>	ZINC04970524_36	506	1.591	0.639	<b>0.953</b>	<b>727.428</b>	0.733	0.976	1.709	0.832	0.976	1.808	0.976	1.615	-2.927	4
<b>12</b>	ZINC04970408_14	508	1.588	0.635	<b>0.953</b>	<b>726.813</b>	0.732	0.976	1.708	0.827	0.976	1.803	0.976	1.611	-2.927	6
<b>13</b>	ZINC04970406_24	507	1.570	0.617	<b>0.953</b>	<b>715.022</b>	0.720	0.976	1.696	0.811	0.976	1.787	0.976	1.593	-2.927	5
<b>14</b>	ZINC04970522_104	505	1.561	0.608	<b>0.953</b>	<b>706.215</b>	0.712	0.976	1.688	0.806	0.976	1.782	0.976	1.584	-2.928	3
<b>15</b>	ZINC72152462_10	403	1.529	0.603	<b>0.926</b>	<b>717.527</b>	0.722	0.962	1.683	0.785	0.962	1.747	0.962	1.564	-2.885	1
<b>16</b>	ZINC00225590_4	203	1.590	0.666	<b>0.924</b>	<b>757.417</b>	0.762	0.961	1.723	0.841	0.961	1.801	0.961	1.627	-2.882	1
<b>17</b>	ZINC00156167_2	603	1.587	0.664	<b>0.923</b>	<b>707.565</b>	0.717	0.960	1.677	0.900	0.960	1.860	0.960	1.624	-2.880	1
<b>18</b>	ZINC15780439_23	705	1.546	0.623	<b>0.923</b>	<b>699.105</b>	0.706	0.960	1.666	0.840	0.960	1.800	0.960	1.583	-2.880	3
<b>19</b>	ZINC36753082_48	604	1.457	0.535	<b>0.922</b>	<b>634.716</b>	0.641	0.959	1.601	0.764	0.959	1.724	0.959	1.495	-2.878	2
<b>20</b>	ZINC72152461_12	404	1.592	0.671	<b>0.921</b>	<b>766.133</b>	0.771	0.959	1.729	0.839	0.959	1.798	0.959	1.630	-2.876	2
<b>21</b>	ZINC01664747_42	509	1.516	0.600	<b>0.917</b>	<b>757.303</b>	0.757	0.957	1.714	0.742	0.957	1.699	0.957	1.556	-2.870	7
<b>22</b>	ZINC15780508_31	706	1.460	0.545	<b>0.915</b>	<b>708.645</b>	0.709	0.956	1.665	0.702	0.956	1.658	0.956	1.501	-2.867	4
<b>23</b>	ZINC27954343_28	204	1.533	0.631	<b>0.903</b>	<b>711.567</b>	0.718	0.949	1.667	0.838	0.949	1.787	0.949	1.579	-2.847	2
<b>24</b>	ZINC27954347_27	205	1.520	0.618	<b>0.902</b>	<b>702.542</b>	0.709	0.949	1.658	0.828	0.949	1.777	0.949	1.567	-2.846	3
<b>25</b>	ZINC05422905_7	510	1.469	0.578	<b>0.891</b>	<b>717.463</b>	0.720	0.942	1.662	0.746	0.942	1.688	0.942	1.520	-2.827	8

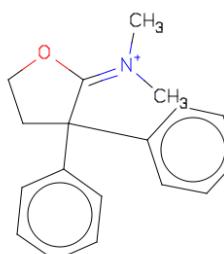
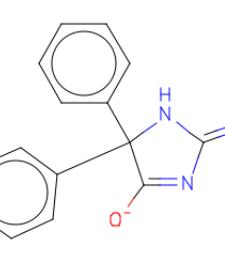
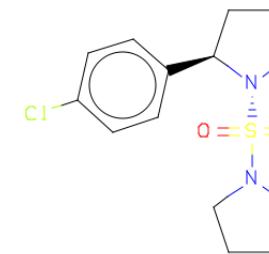
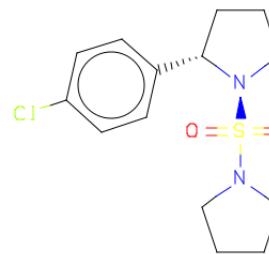
TanimotoCombo Rank 1	ShapeTanimoto Rank1	ColorTanimoto Rank 1	ColorTanimoto Rank 2
<p>ZINC02389343_0</p>  <p>1.758<sup>a</sup> 0.798<sup>b</sup> 0.960<sup>c</sup> 848.233<sup>d</sup></p>	<p>ZINC04590678_0</p>  <p>1.371<sup>a</sup> 0.894<sup>b</sup> 0.477<sup>c</sup> 930.092<sup>d</sup></p>	<p>ZINC71856537_14</p>  <p>1.545<sup>a</sup> 0.552<sup>b</sup> 0.993<sup>c</sup> 660.630<sup>d</sup></p>	<p>ZINC71856536_15</p>  <p>1.539<sup>a</sup> 0.545<sup>b</sup> 0.993<sup>c</sup> 655.443<sup>d</sup></p>

Figure 3.12 2D structures and 3D overlays of top screened compound against BC23 ranked by different scoring functions: TanimotoCombo, ShapeTanimoto and ColorTanimoto.

<sup>a</sup>TanimotoCombo, <sup>b</sup>ShapeTanimoto, <sup>c</sup>ColorTanimoto, and <sup>d</sup>Overlap scores

To follow up the virtual screening predictions, the biological assays of seven selected compounds from both the *TanimotoCombo* and *ShapeTanimoto* ranked lists were carried out (see structures in Figure 3.13). All of the compounds, except the best rank in shape,<sup>137</sup> were obtained from the National Cancer Institute (NCI)<sup>149</sup>. Based on availability, the two selected compounds in the *TanimotoCombo* ranking were rank number 19, ZINC01591149, containing 1,3-thiazinane-2,4-quinone; and rank number 29, ZINC01595186, containing the lactone of furan-2-one (Figure 3.13).

The best match in shape, ZINC04590678 in Figure 3.13, is also of interest for the assay. It has the highest *ShapeTanimoto* score (0.894) and moderate *ColorTanimoto* score (0.477) and contains a basic motif of the imidazole ring, with electrophilic carbonyl and cyano groups. Apart from the best rank in shape, four other shape-matching compounds were selected for the assay (Figure 3.13): (i) rank number 11, ZINC01700895 containing the ester of carbamic acid; (ii) rank number 14, ZINC01611635 containing imidazole and thiazine fused rings; (iii) rank number 19, ZINC01608670 containing a tetrahydropyran-2-one ring; and (iv) rank number 31, ZINC05698963 containing an aminourea motif.

When assayed, none of the seven procured compounds showed inhibiting activity of IL-1 $\beta$  release, supporting the hypothesis that the boron centre is crucial in the structure.

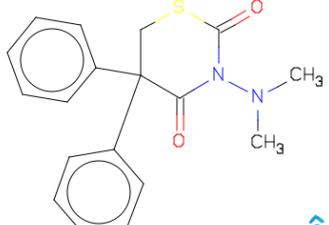
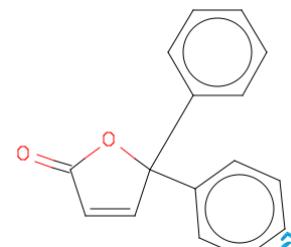
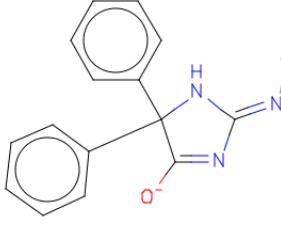
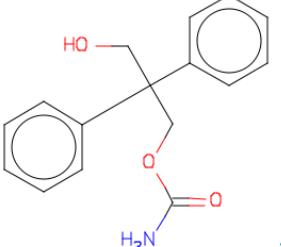
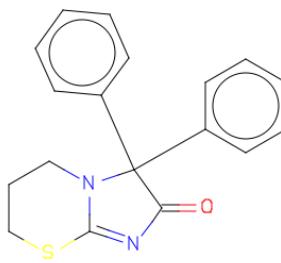
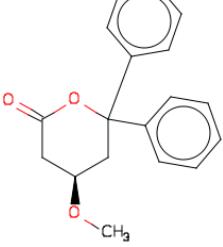
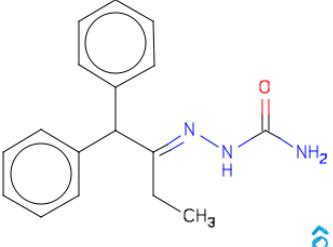
TanimotoCombo hits			
Rank 19		Rank 29	
ZINC01591149_0		ZINC01595186_4	
	<b>1.571<sup>a</sup></b> 0.873 <sup>b</sup> 0.698 <sup>c</sup>		<b>1.539<sup>a</sup></b> 0.847 <sup>b</sup> 0.692 <sup>c</sup>
ShapeTanimoto hits			
Rank 1		Rank 11	
ZINC04590678_0		ZINC01700895_12	
	<b>1.371<sup>a</sup></b> <b>0.894<sup>b</sup></b> 0.477 <sup>c</sup>		<b>1.158<sup>a</sup></b> <b>0.880<sup>b</sup></b> 0.278 <sup>c</sup>
Rank 14		Rank 19	
ZINC01611635_1		ZINC01608670_4	
	<b>1.359<sup>a</sup></b> <b>0.879<sup>b</sup></b> 0.480 <sup>c</sup>		<b>1.441<sup>a</sup></b> <b>0.876<sup>b</sup></b> 0.565 <sup>c</sup>
Rank 19		Rank 31	
ZINC05698963_9			
	<b>1.116<sup>a</sup></b> <b>0.873<sup>b</sup></b> 0.243 <sup>c</sup>		

Figure 3.13 Selected compounds for testing by biological assay.

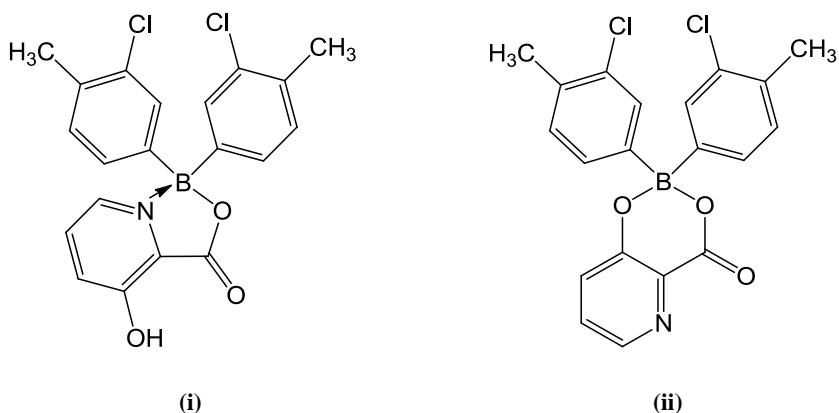
<sup>a</sup>TanimotoCombo, <sup>b</sup>ShapeTanimoto, <sup>c</sup>ColorTanimoto and <sup>d</sup>Overlap scores

### 3.6 Conclusions and outlook

According to unpublished work by Brough and Freeman at the University of Manchester, boron compounds have been identified as inhibiting IL-1 $\beta$  release. Five of these heterocyclic boron compounds, with different substituent groups and showing a range of inhibitory efficacy, were selected for structure-activity analysis using quantum mechanics. The M06L functional and 6-31G\* basis set was used to determine the shape of the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO). It was found that the LUMO shape among the active and inactive compounds are similar. However the HOMO of the active compounds are on the phenyl rings in contrast to on the substituent group and the heterocyclic ring of the inactive compound. From the DFT-based orbital energies, the most active BC7 compound has the lowest LUMO energy, suggesting it to be the best electron acceptor. This observation may link to the partially positive charge on the boron calculated using Mulliken population analysis at the same level of theory.

The boron-analogue of BC23 was used as the query for a virtual screen for non-boron-based lead-like compounds available from the ZINC database. The hits were ranked separately by three different scoring functions. The shape-matched compounds were ranked by the *ShapeTanimoto* score whereas the chemically matched compounds were ranked by the *ColorTanimoto* score. The best list for the hits is obtained by using the *TanimotoCombo* score, a combined scoring function considering both shape and chemical similarity. Subsequently, the biological assay of seven selected compounds (two from *TanimotoCombo*'s rank and five from *ShapeTanimoto*'s rank) were carried out. None showed activity for inhibiting the release of IL-1 $\beta$ .

Potentially, the electrostatic charge on the boron considerably contributes to the inhibitory activity of boron compounds (given none of the selected screened compounds has inhibiting activity on the release of IL-1  $\beta$ ). Partial charge of the boron ideally should be incorporated into the screens. The alternative explicit colour force field would be a good start for further work because it allows for any arbitrary atomic charge states to be included. More research is required to determine the atomic charge distribution of all seven biologically tested compounds at the same level of DFT theory. Due to the fact that the carbon atom is slightly higher in electronegativity than the boron, the calculated partial charge at the carbon of the tested compounds would be expected to be less positive. The charge distribution around the C-N bond could be expected to be polarised slight differently than that of the B-N bond in the queries. The calculated partial charge of the heterocyclic carbon of these tested compounds should be compared to those calculated values of the boron in the BC series as well as to those of the active compound, AN0128<sup>129</sup> published in the literature (see Figure 3.14). Also, it would be interesting to study the partial charge distribution of the carbon and nitrogen analogues of BC23 to gain better understanding of the importance of the boron in its anti-inflammatory activity.



**Figure 3.14** AN0128 forming (i) 5-membered ring and (ii) 6-membered ring.

# CHAPTER 4 Protein-ligand docking guided by semiempirical quantum mechanical potentials

## 4.1 Introduction

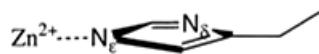
In structure-based drug design for screening potential lead compounds, docking tools have become essential for a given 3D protein structure derived from experimental data or homology modelling. A docking program is often used to determine the binding modes of inhibitors at the active sites of enzymes. Despite a number of docking programs being currently available for use,<sup>150</sup> there appears to be no universal one program suitable for application across the diversity of protein-ligand systems. The success rate across different docking programs is in a range of 35-60% considering a typical standard of root-mean-square deviation (RMSD) less than 2 Å relative to the crystallographic binding mode.<sup>151</sup> The key issue for successful docking relies on both conformational sampling algorithms for generating ligand poses and scoring functions for estimating the binding affinities at the active site.<sup>152</sup> The sampling algorithm in a docking program should be able to cover a conformational space of both ligand and the protein active sites whereas the predicted binding affinities of the ligand poses should be sufficiently accurate relative to experimental data, e.g. X-ray crystallographic poses. More details on conformational sampling techniques and classes of scoring functions are previously described in Chapter 2.

Docking algorithms are designed to simulate interactions between potential ligands and protein targets. Among several classes of protein targets, docking ligands to metal-containing sites in the design of metalloenzyme inhibitors presents challenges for modelling, for example due to charge transfer occurring between ligand, surrounding protein and metal atoms. Charge distribution in metalloenzymes appears to be inaccurately described by a force field-based model due to the limitation of a fixed atom-centred partial point charge model;<sup>153</sup> this is despite the claim of the universal force field<sup>154</sup> developed for drug-design applications. Ultimately, all atoms contributing to metal-binding in metalloproteins should be treated quantum mechanically for predicting a correct binding mode. A computational study taking into account polarisation effects also indicates that a quantum mechanical treatment is essential for accurate modelling of inhibitor-protein complexes with diverse charge-charge interactions, i.e. highly variable polarised charges of protein atoms upon binding different inhibitors.<sup>155</sup>

Due to difficulties in defining an accurate force field for a metal ion present in a protein, a combined quantum mechanical/molecular mechanical (QM/MM) method is an attractive alternative for modelling the binding affinities of metalloprotein-ligand complexes. The QM/MM method treats the ligand and part of the protein residues at a catalytic site quantum mechanically. The rest of the system, including the solvent environment, is modelled by an empirical force field (please see more details in QM/MM method in Chapter 2). The intrinsic advantage of a QM/MM hybrid potential is its potential to capture charge fluctuations and polarisation effects for modelling electrostatic interactions, as compared to an MM energy function. A QM/MM energy function can be used for scoring ligand poses in docking not only at a metal-containing

binding site, but also at a polar site or a hydrophobic site with  $\pi-\pi$  interactions in particular. The accuracy of QM/MM docking of ligands to these types of site (to an RMSD of less than 2 Å) was improved over a force field scoring function when the surrounding protein atoms are defined in the QM region.<sup>156</sup> A QM/MM scheme was also successfully employed in studying noncovalent binding interactions of fluorinated benzenesulfonamide inhibitors in the binding site of carbonic anhydrase II, in which dispersion plays a key role.<sup>7d, 157</sup> In this Chapter, we will consider the implementation of a QM/MM docking program and its application to predicting the binding mode of ligands in the active sites of two zinc-containing metalloenzymes. In the following section, we consider the function and structure of the two metalloenzymes studied.

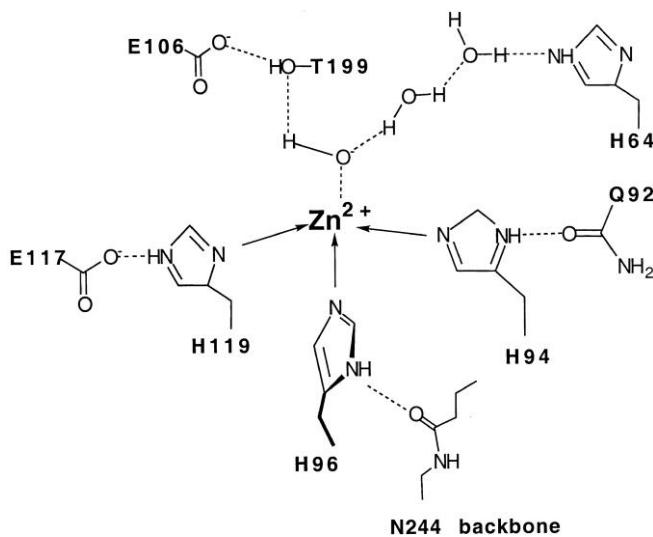
A large number of metalloenzymes contain zinc metal, the second most abundant trace element in most higher animals.<sup>158</sup> The zinc(II) ion ( $Zn^{2+}$ ) has a filled  $d$  orbital ( $d^{10}$ ) so that it functions as a Lewis acid to accept an electron pair, rather than engaging in redox reactions.<sup>159</sup> It is therefore a redox-stable metal cofactor with multiple possible coordination geometries. Although the binding geometries of zinc metalloenzyme have been reported in trigonal bipyramidal and octahedral geometries, a slightly distorted tetrahedral geometry has been observed most often.<sup>159</sup> For a tetrahedral geometry,  $Zn^{2+}$  coordinates with three or four amino side chains of the protein. Histidine (His) appears to be frequently observed in zinc-containing catalytic sites, coordinated through its  $N_c$  atom of the imidazole side chain via the in-plane  $sp^2$  lone pair electrons, rather than  $N_\delta$  (Figure 4.1).<sup>159</sup>



**Figure 4.1** The preferred binding mode of His-Zn<sup>2+</sup> via in-plane sp<sup>2</sup> lone pair electrons of the N<sub>c</sub>.<sup>159</sup>

One family of zinc metalloenzymes are the carbonic anhydrases. These exist in multiple forms known as isozymes. Its multiple forms differ in amino acid sequence but catalyse the same chemical reaction. The selectivity against different isoforms stems from different interaction patterns in which the inhibitors bind via van der Waals and polar interactions.<sup>160</sup> Up to 16 carbonic anhydrase isozymes have been found in human tissues<sup>161</sup> for catalysis of reversible hydration of carbon dioxide to bicarbonate.<sup>162</sup> This enzyme is implicated in various physiological processes such as maintaining an acid-base equilibrium in blood and tissues, and facilitating transport of carbon dioxide and protons across biological membranes.

The isozyme of relevance to this study is human carbonic anhydrase II (hCA II). Its active site is well conserved and forms a deep conical cleft, approximately 15 Å in depth.<sup>159</sup> The catalytic Zn<sup>2+</sup> site presents a distorted tetrahedral geometry of the hCAII–Zn<sup>2+</sup>–OH intermediate (Figure 4.2), coordinating with the three imidazole groups of His94, His96 and His112. These zinc-coordinating moieties also form a hydrogen bond network with other hydrophobic residues surrounding the binding site. Although the zinc-coordinating histidine residues do not have a direct function during catalysis, they are conserved for stabilising the electrostatic environment that occurs at the Zn<sup>2+</sup> site and at the negatively charged transition state.<sup>163</sup>

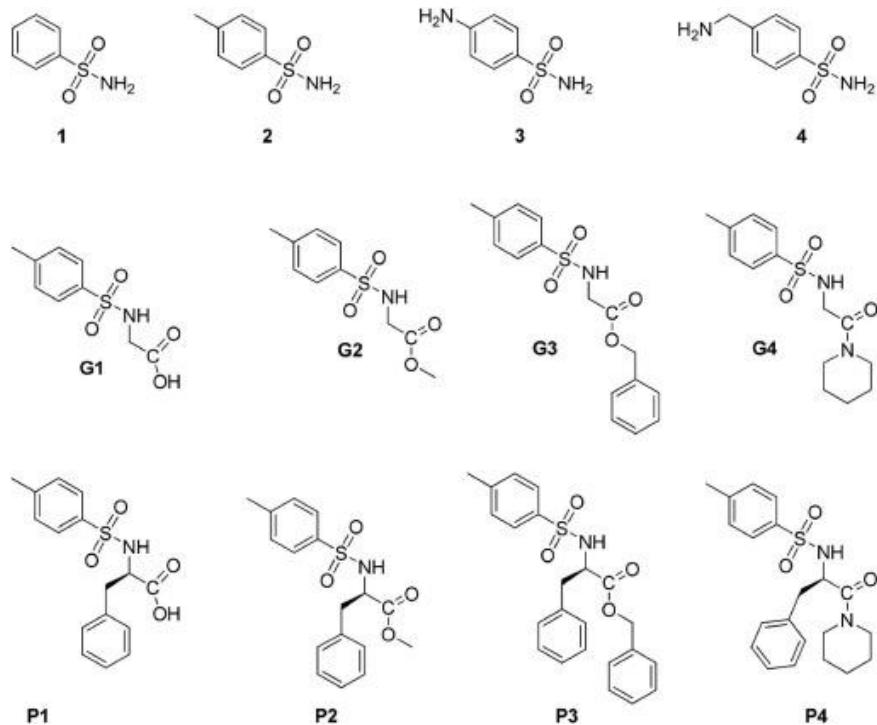


**Figure 4.2** The binding site of hCA II in one-letter codes: E=glutamate, H=histidine, N=asparagine, Q=glutamine, and T=threonine.

The hCA II inhibitors can be categorised into four main classes.<sup>164</sup> The first class is the zinc binders such as metal binding anions, sulfonamides and their derivatives, i.e. sulfamates, sulfamides, etc. The second class contains compounds such as phenol, anchoring to  $Zn^{2+}$  that coordinates with water or hydroxide ion. Slightly different from the second class, the third class of polyamine inhibitors such as spermine and spermidine rather bind to the water or hydroxide ion coordinated to the  $Zn^{2+}$ . The last class contains compounds such as coumarins and thiocoumarins that bind at the active site and block the entrance to the cavity of the active site; their inhibition mechanism does not involve the  $Zn^{2+}$ .

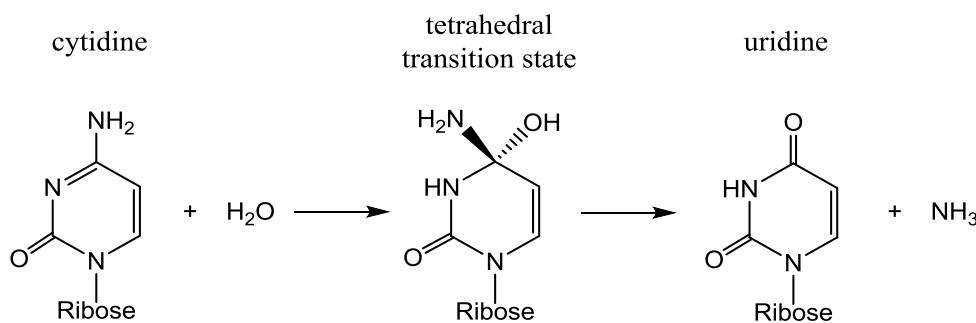
Among the classes of hCA II inhibitors mentioned above, arylsulfonamide ligands in the first class of the zinc binder are attractive for binding studies in drug design due to their particular structural features and physicochemical properties. Sulfonamides make good ligands when bound to hCA II because they can mimic aspects of the carbon

dioxide and bicarbonate ion substrates. In addition, this ligand class has a high binding affinity based on interactions of the sulfonamide nitrogen and oxygen atoms: the nitrogen atom of the sulfonamide group is ionised upon binding to displace zinc-bound water for a direct interaction with the  $Zn^{2+}$ . It also forms a hydrogen bond to the hydroxyl group of Thr199, while the backbone amine group of Thr199 interacts with the sulfonamide oxygen.<sup>160</sup> There was an attempt to synthesise derivatives of arylsulfonamide (compounds G1-G4 and P1-P4, Figure 4.3) by modifying chemical groups at the sulfonamide nitrogen.<sup>165</sup> However, the inhibitory activity of the synthesised compounds was not better than their parent compound **2** in Figure 4.3. The contacts between the sulfonamide aryl ring and hydrophobic residues at the catalytic zinc site additionally contribute to the binding affinity.



**Figure 4.3** Structures of the synthesised sulphonamide derivatives.<sup>165</sup>

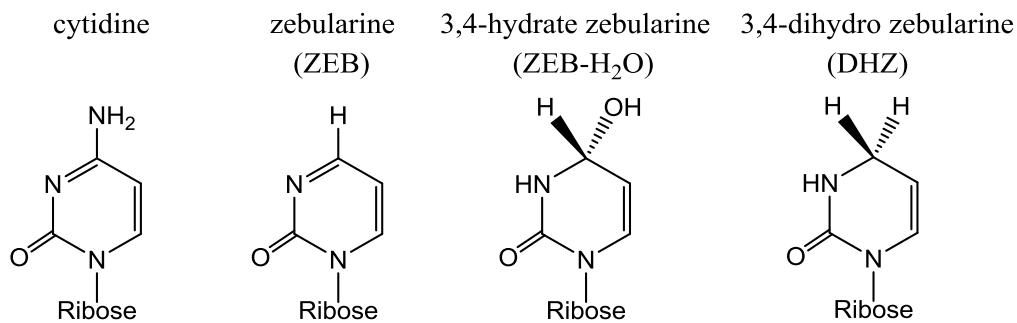
A second family of zinc-metalloenzymes that is considered in this chapter are cytidine deaminases (CDA), enzymes that accelerates hydrolytic deamination of cytidine, producing uridine via a tetrahedral transition state (Figure 4.4).



**Figure 4.4** Hydrolytic deamination of cytidine via a tetrahedral transition state.<sup>166</sup>

A CDA inhibitor that resembles the transition state of CDA is zebularine (ZEB) with  $K_i$  of  $2.5 \times 10^{-7}$  M, which undergoes an equilibrium with 3,4-hydrate zebularine (ZEB- $\text{H}_2\text{O}$ ) with  $K_{eq}$  of  $4.7 \times 10^{-6}$  (Figure 4.5).<sup>167</sup> The ZEB- $\text{H}_2\text{O}$  is generally generated in solution by a nucleophilic attack of water on ZEB. It appears to strongly inhibit CDA with higher affinity than that of its analogue 3,4-dihydrozebularine (DHZ, Figure 4.5) by a factor of  $10^7$ - $10^8$ .<sup>167</sup> However, DHZ itself is stable compared to ZEB- $\text{H}_2\text{O}$  due to a hydrogen substituent group at the C4 position on the six-membered ring. Xiang *et al.* demonstrated from their crystallographic data that there is a trapped water in the binding site of CDA·DHZ complex.<sup>168</sup> They suggested that the presence of the trapped water in the CDA·DHZ complex is a close analogy with the Michaelis complex for the deamination reaction. Also, their findings revealed that the position of the six-membered ring of DHZ is 1.0 Å away from the zinc compared to that of the ZEB- $\text{H}_2\text{O}$ .

inhibitor without any conformational change in the enzyme. This was considered to facilitate hydrogen bonds from DHZ to Glu104 and Ala103 residues.

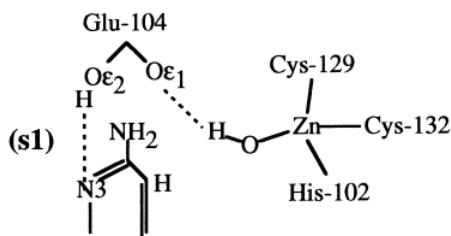


**Figure 4.5 Structures of cytidine deaminase (CDA) and its transition-state analogue inhibitors: zebularine (ZEB), 3,4-hydrate zebularine (ZEB-H<sub>2</sub>O), and 3,4-dihydrozebularine (DHZ).**

In the active site of CDA, the unsaturated Zn<sup>2+</sup> ion is coordinated with the thiolate sulfur atoms of Cys129 and Cys132 and with an imidazole nitrogen atom of His102. The unoccupied orbital of Zn<sup>2+</sup> is available for an incoming substrate to promote the deamination. This occurs via a hydrogen bonding network that forms among the enzyme-substrate complex, Glu104 and water as proposed based on quantum chemical studies by Matsubara *et al.*<sup>166</sup> They suggested that the migration of proton dissociated from water is a key process.

However, the mechanism of hydration catalyzed by zinc metalloenzymes is most likely via a direct attack of zinc-coordinated hydroxide ion, Zn-(OH)<sup>-</sup>, rather than zinc-coordinated water, Zn-H<sub>2</sub>O. Lee *et al.* performed linear-scaling quantum computations using PM3 Hamiltonian on the active site of CDA·H<sub>2</sub>O complex containing either Zn-H<sub>2</sub>O or possible configurations of Zn-(OH)<sup>-</sup>.<sup>169</sup> Their results showed that Zn-(OH)<sup>-</sup> in

the “s1” configuration (see Figure 4.6) is the lowest energy geometry that is closest to the crystallographic structure. Therefore, it was predicted to be an active intermediate formed prior to a nucleophilic attack of the ligand. Glu104 is then protonated at O<sub>ε2</sub>, and forms a hydrogen bond to the N3 of the cytidine ligand.



**Figure 4.6 Active species of zinc-coordinated hydroxide ion  $[Zn-(OH)^{-}]$  in the lowest-energy s1 configuration closest to a crystallographic structure.<sup>169</sup>**

## 4.2 Aims

Comparative evaluation of a range of docking programs has revealed that many popular scoring functions are able to discriminate the correct docked poses relative to the experimentally observed conformation, but the correlation with experimental binding affinity, of metalloproteins in particular, still remains problematic.<sup>170</sup> This chapter examines the ability of a Monte Carlo docking scheme using a QM/MM potential to discriminate the correct poses of inhibitors relative to their crystallographic conformations in the binding sites of zinc-containing metalloenzymes hCA II and CDA. An earlier version of QM/MM Monte Carlo docking scheme implemented by D. Mucs in the Gaussian 03 program used BH&H/6-31G\* method for evaluation of docking.<sup>171</sup> However this approach was limited by the computational cost of the QM method.

Consequently an aim of this work was to implement the MC docking approach for a SQM scoring function within the more recent Gaussian 09 program.

Therefore, the QM/MM MC module, using the Markov chain approach of Metropolis, was implemented in Gaussian 09 in conjunction with the semiempirical PM6 method and AMBER force field. The performance of this QM/MM MC module to identify nonnative ligand poses that resemble the native one was evaluated based upon the root-mean-square deviation (RMSD) between the predicted binding poses of ligand and the crystallographic bound ligand configuration.

### **4.3 Implementation of QM/MM MC module in Gaussian 09**

The QM/MM MC module implemented in Gaussian 09 comprises of two schemes: the standard ONIOM energy calculation in the program package release and the MC routine for searching ligand conformations. The MC scheme, first coded by Dr. Daniel Mucs, was implemented in Overlay 1 and integrated with the QM/MM module of Gaussian 03. It has three main components: the main executable shell, the unit that controls the dynamic memory allocation for calculations and all subroutines for performing the MC search, for example, a random number generator and a conformation generator. In this work, these three main components were modified and compiled for Gaussian 09.

A Gaussian 09 input deck generally consists of a series of lines in an ASCII format. Following this basic structure of a Gaussian input file with some modifications, our input file used for the in-house QM/MM MC module mainly contains 4 sections: Link 0

commands, a route card, molecular system specification, and QM/MM and MC parameters. The first section in the input file begins with the Link 0 commands specifying technical detail for job submission such as the amount of dynamic memory, the number of processors for shared memory parallel execution, location of checkpoint file, etc.

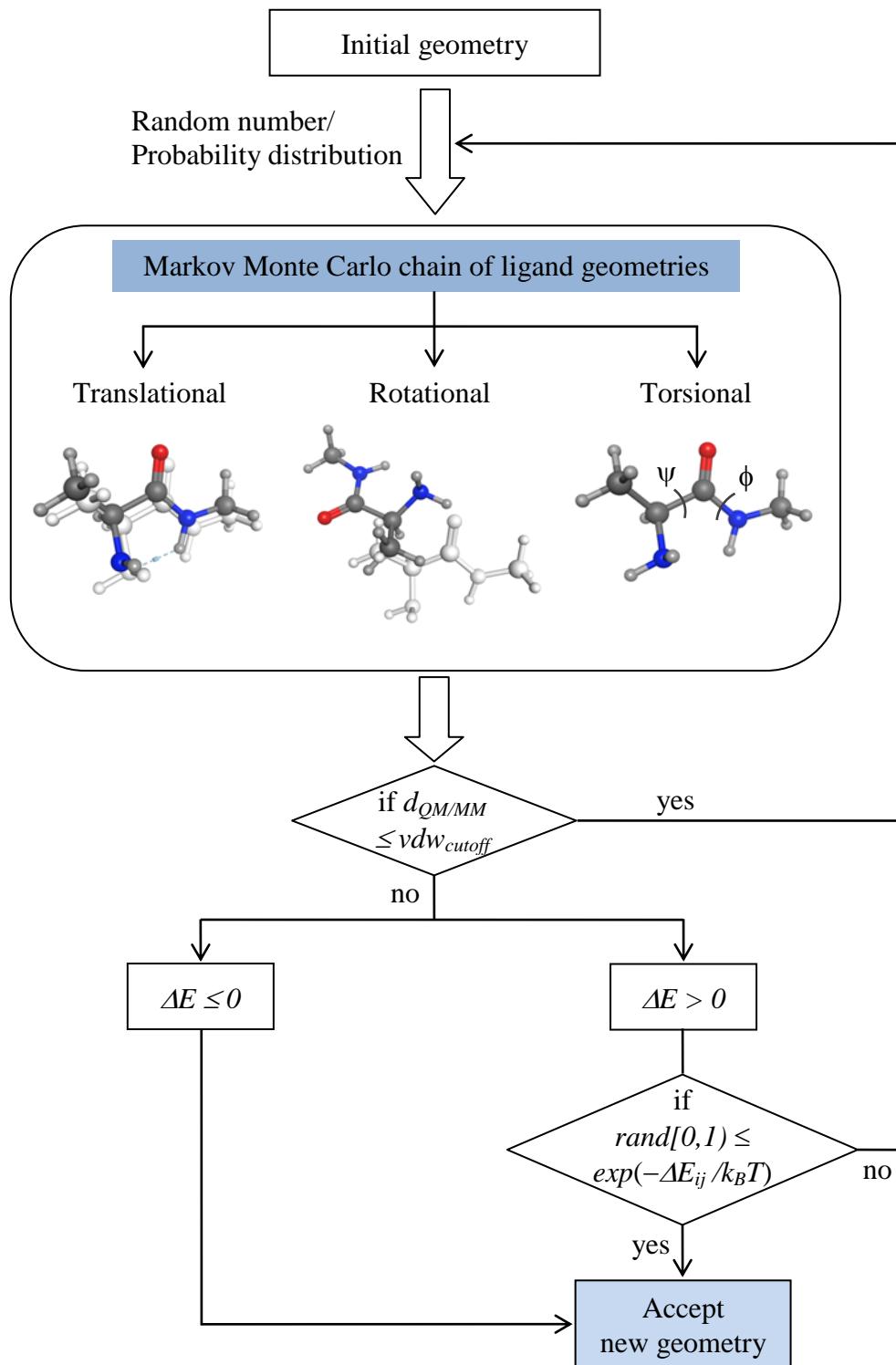
The second section of an input file is the nonstandard route card of Gaussian internal command options (IOps). This route section controls a sequence of program execution illustrated in Scheme 4.1. When the QM/MM MC docking scheme is invoked, a full standard ONIOM is firstly computed for any initial structure of the system. Then, the Metropolis MC sampling routine is called to generate a Markov chain of configurations. Instead of randomly choosing configurations then weighting them based on the Boltzmann energy distribution  $\exp(-\Delta E/kT)$ , this modified MC scheme chooses Boltzmann-weighted configurations and weights them evenly. Random perturbations (translation, rotation and torsion angle rotations) are selectively applied to a protein-bound ligand with equal probability using uniformly distributed random numbers. These random numbers, generated by a default Fortran function *gran()*, range between 0 and 1. Changing the weighting factor for choosing any perturbation in particular is also possible.

During MC sampling, an assumption made is that all atoms of the protein residues are kept rigid. Only the ligand, the zinc ion and protein atoms that noncovalently bind to the zinc are defined in the QM region. The rest of the protein is treated at the MM level of theory. A maximum boundary of a minimum distance between any QM and MM atoms is set as a filtering criterion to avoid steric clashes at the protein-ligand interface. A

ligand configuration that survives this criterion during the search is subject to an energy evaluation. The change in energy of the system  $\Delta E$ , is calculated: if the move of a ligand pose in a Markov chain brings the system to a new state that has the same or lower energy, i.e.  $\Delta E \leq 0$ , this move will be allowed with its new ligand position. However, if  $\Delta E > 0$ , the move and new ligand pose will be accepted only if the probability of choosing that state  $\exp(-\Delta E/kT)$  is equal or higher than the transition probability represented by a random number.

The third part of an input file defines atomic coordinates and partial charge. Each atom is labelled for a treatment in a high or low region to which QM or MM level of theory is applied. The dangling bonds at the partition boundary are also treated using the hydrogen atom as a link atom with the electrostatic embedding scheme (see more detail in section 2.3 of QM/MM method). The following part of the input file contains additional MM parameters for the QM atoms. The last part is a specification of some MC parameters that are the number of MC steps, the number of QM atoms, the translational step size and a torsional tree that explicitly defines rotatable bonds in pairs.

When the program is executed, the initial coordinates of a system is firstly written in a pdb format. After a calculation is complete, all moves according to the number of MC steps set are written in two separate files: one for only the accepted moves and one for rejected moves. The energy and  $\Delta E$  for each configuration can be found in these two files as well as a standard Gaussian log file.



Scheme 4.1 QM/MM MC docking scheme.

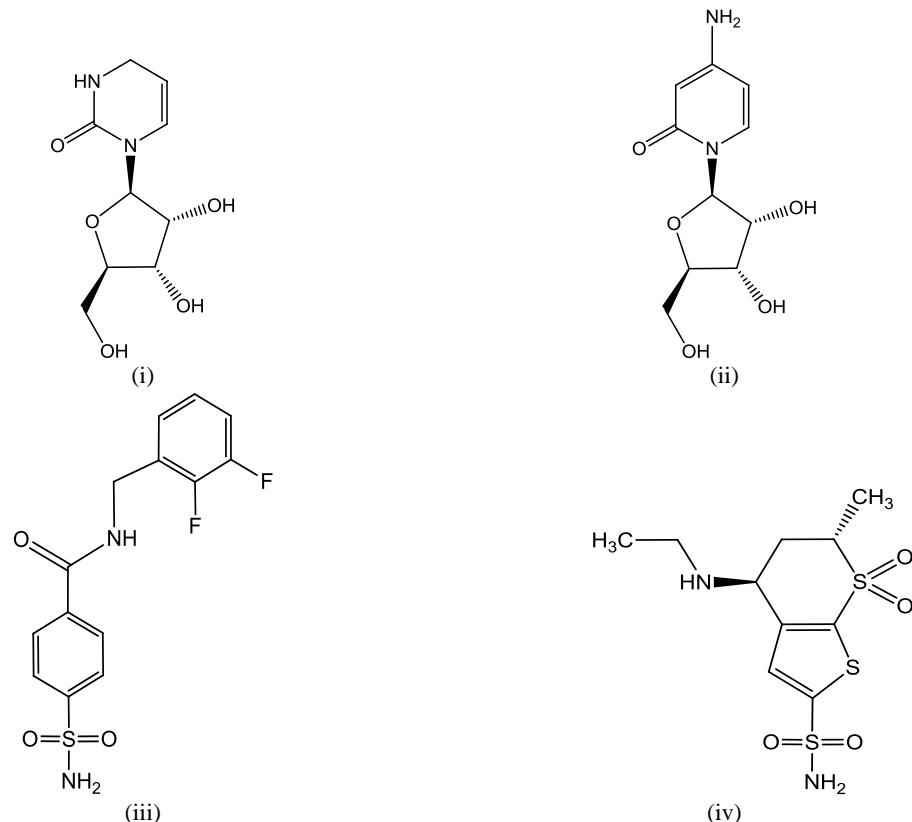
## 4.4 Computational details

### 4.4.1 Systems

For both conventional docking using GOLD suite and our in-house QM/MM MC docking scheme in Gaussian 09, we focus on two receptors: cytidine deaminase (CDA) and human carbonic anhydrase II (hCA II) with two inhibitors for each. All crystallographic structures were downloaded from the Protein Data Bank at [www.rcsb.org](http://www.rcsb.org). Both inhibitors for the CDA system are glycosylated pyrimidine-analogues of the cytidine substrate. One of them is the dihydro analogue of zebularine (DHZ), namely 3,4-dihydro-1*H*-pyrimidin-2-one nucleoside (Figure 4.7(i)), from PDB ID 1CTT with resolution of 2.1 Å.<sup>168</sup> It has a  $K_i$  of  $3 \times 10^{-5}$  M, intermediate between that of cytidine ( $\sim 10^{-4}$  M) and hydrated zebularine ( $\sim 10^{-12}$  M).<sup>73, 168</sup> The other ligand is 3-deazacytidine (DAC) from PDB ID 1ALN<sup>172</sup> with resolution of 2.3 Å (Figure 4.7(ii)). Its structure is the analogue of cytidine substrate, replacing the nitrogen on the cytosine base of cytidine with the carbon. A primary amine group at position 4 on the ring is a remarkable precursor for the leaving ammonia molecule. Considering the amine group of DAC, this ligand is a stronger nucleophile than the DHZ ligand in the first CDA system with PDB ID 1CTT in which an amide group is on the six-membered ring.

For hCA II, the fluorinated derivative of 4-(aminosulfonyl)-*N*-phenylmethylbenzamide inhibitor was retrieved from PDB ID 1G52<sup>18a</sup> with resolution of 1.8 Å. It is named *N*-(2,3-difluoro-benzyl)-4-sulfamoyl-benzamide (FSB) (Figure 4.7(iii)). It has an increased enzyme-inhibitor affinity by a factor of 10 over its parent inhibitor.<sup>74</sup> Another inhibitor for hCA II studied in this work is a clinically used sulfonamide, so-called

dorzolamide (DZA)<sup>173</sup> with  $K_i$  of 0.372 nM.<sup>40</sup> It is named (4*S*-trans)-4-(ethylamino)-5,6-dihydro-6-methyl-4*H*-thieno[2,3-*b*]thiopyran-2-sulfonamide-7,7-dioxide. The structure was retrieved from PDB ID 1CIL<sup>40</sup> with resolution of 1.6 Å.



**Figure 4.7** Structures of inhibitors for cytidine deaminase (i) 3,4-dihydrozebularine (DHZ) from PDB ID 1CTT and (ii) 3-deazacytidine (DAC) from PDB ID 1ALN; for human carbonic anhydrase II (iii) *N*-(2,3-difluoro-benzyl)-4-sulfamoyl-benzamide (FSB) from PDB ID 1G52 and (iv) dorzolamide (DZA) from PDB ID 1CIL.

#### **4.4.2 GOLD docking setup**

A hundred poses of ligand for each system mentioned above was generated based on a genetic algorithm (GA) available in the Genetic Optimisation for Ligand Docking.

(GOLD) suite. The key parameter in the GA that controls coverage of conformational space is the number of genetic operations performed, i.e. crossover, migration and mutation. By default, the search efficiency is set to 100% in which GOLD will attempt to apply around 30,000 GA operations for a ligand with five rotatable bonds. For a highly flexible ligand, it is recommended to set this at 200%. The 100,000 GA operations can be slow but deliver high predictive accuracy. In this study, the maximum number of operations applied over the course of a GA run was set to 125,000 to ensure fully covered sampling of our ligands.

GOLD suite offers several choices of empirical scoring function such as GoldScore, ChemScore and Astex Statistical Potential. ChemPLP pairwise linear potential is an empirical scoring function optimised for pose prediction, and is five times faster in speed than the original GoldScore scoring function of GOLD.<sup>174</sup> The ChemPLP scoring function takes into account the hydrogen bonding term obtained from the ChemScore function as well as multiple linear potentials to model van der Waals and repulsive terms. It has been reported that it gives the highest success rates for pose prediction against the test sets and becomes the default scoring function in recent versions of GOLD (version 5.1 and later). Analysis of docked poses was performed based on the calculated RMSD in coordinates of nonhydrogen atoms from crystal structures.

#### 4.4.3 QM/MM MC docking parameters

In order to evaluate suitable parameters for the QM/MM MC docking algorithm, chains of ligand poses were created by MC random moves: spatial translations, rotations about a centre of mass and torsional rotations. The translational movement was previously set

to be between 0 and 5 bohrs in the previous version of the QM/MM MC module compiled in Gaussian 03. For this work, values of the maximum translational step size ( $\Delta x_t$ ) of 2, 5 and 10 bohrs were considered for a trajectory of 2k MC steps using a polar fragment of sulfamoyl methanol as a probe in the binding site of hCA II. The probe, Zn<sup>2+</sup> ion and hydrophobic imidazole rings of His94, His96 and His119 protein residues were defined as the QM region. Both types of rotation were carried out using randomly selected Euler angles between 0-180°.

To avoid the energy penalty in QM/MM calculations, two following issues were considered for sufficient sampling. The first one is the position in which to locate the docked ligand pose in the binding site. The distance limit between the Zn<sup>2+</sup> and the centre of mass of a ligand pose ( $d_{max}$ ) was studied for values of 2.0, 5.0 and 10.0 bohrs. If any generated poses are placed too close from the Zn<sup>2+</sup> beyond the  $d_{max}$  value, they will be rejected without evaluating the system energy of the complex. Another issue is steric clashes between any QM and MM atoms. The van der Waals distance cutoff ( $vdw_{cutoff}$ ) was used to discard generated ligand poses in which any atom clashes with protein residues. The  $vdw_{cutoff}$  values examined were 2.5, 3.0, 3.5 and 4.0 bohrs.

From these studies, the optimised values, 5 bohrs for  $\Delta x_t$ , 5 bohrs for  $d_{max}$ , and 3.0 bohrs for  $vdw_{cutoff}$ , were used to investigate the MC acceptance rates of different lengths of the Markov chain. Previous work using the QM/MM MC docking algorithm at the BH&H/6-31G\* level of theory on hCA II and CDA receptors used 1k of MC steps.<sup>171</sup> It was suggested that longer sampling is desirable to ensure thoroughly sampling in a conformational space. This work, therefore, considered 1k, 2k, and 4k MC steps for sampling the binding poses of sulfamoyl methanol at the hCA II binding site.

Different MC approaches can be applied for sampling ligand poses. In parallel sampling, the same initial ligand pose is used for multiple sampling in parallel with different random seeds for sampling. In contrast, a consecutive approach explores the ligand poses in a single long run. According to earlier findings, the consecutive sampling method was recommended for use.<sup>171</sup> In this study, two runs of 4k MC steps were consecutively performed for each of the four CDA and hCA II complexes. Longer consecutive sampling of 10k MC steps was also carried out for comparison.

The binding modes of generated ligand poses were evaluated by a QM/MM energy potential. The two-layer ONIOM method<sup>53a</sup> in Gaussian 09<sup>175</sup> was applied for computing the QM/MM energies of bound complexes. The semiempirical PM6 level of theory was used for the QM description whereas parameters for the MM environment were modelled from the AMBER force field (constructed using the TAO<sup>176</sup> utility). Electrostatic interactions between QM and MM regions were treated at the QM level by the electronic embedding QM/MM coupling scheme. In this way, the partial charges of MM atoms are incorporated into the QM Hamiltonian so that QM wavefunctions are allowed to be polarized. This provides a better description of electrostatic QM/MM coupling.

Ligand nonelectrostatic parameters and partial charges of ligand atoms were determined by using the Antechamber module of Amber12<sup>177</sup> with the AM1-BCC charge scheme. For CDA model systems, the QM region includes the Zn<sup>2+</sup> ion, inhibitor and thiol groups of Cys129 and Cys132 as well as the OH<sup>-</sup> ion from deprotonation of a water molecule participating in catalysis<sup>166, 168</sup>. The QM definition for hCA II model systems

covers the deprotonated sulfonamide inhibitor revealed in the literature<sup>178</sup> as the active form, Zn<sup>2+</sup> ion, and imidazole rings of His94, His96, and His119. All other crystallographic water for all systems studied were removed prior to addition of hydrogen atoms and a treatment of protonation state of proteins using Leap.

## 4.5 Results and discussion

### 4.5.1 Implementation of in-house QM/MM MC module in the Gaussian 09 program

Prior to integration of the MC search routine into Gaussian 09, standard ONIOM calculations with electronic embedding scheme were verified using Gaussian 09. Throughout this chapter, the “ONIOM energy” is referred to the two-layer QM/MM system energy, computed by the standard ONIOM routine in Gaussian 09. In this section, only for hCA II in the receptor with PDB ID 1G52 was used with different definitions of the QM region, i.e. defined as only ligand atoms, ligand atoms with the Zn<sup>2+</sup> or ligand atoms with the Zn<sup>2+</sup> and the imidazole rings of zinc-bound histidine residues.

There was no discrepancy between the standard ONIOM energy and the QM/MM MC scheme implemented in Gaussian 09 for all QM definitions applied (Table 4.1). Since the QM/MM MC scheme had been implemented only in Gaussian 03, it was previously used only at HF and DFT levels for the QM description. Considering the computational cost, the semiempirical PM6 level of theory was applied for describing the QM region of the FSB·hCA II complex. QM/MM calculations at the PM6/Amber level of theory

were completed in about an hour on Intel SSE4.2 or Intel AVX 4 cores for each calculation with 4 GB of memory per core (Table 4.1). Calculations at the HF/6-31G\*/AMBER level of the same system could take up to about a week.

Computational cost and the performance of the docking scheme both depends on defining an appropriate QM region. On the one hand, the number of QM atoms should be reduced to save computational cost. On the other hand, it is desirable to include in the QM region all atoms that contribute to the key binding interactions, to hopefully increase docking accuracy. For metalloproteins, the QM region should ideally include the ligand, metal and coordinating protein atoms. It has been demonstrated that, when only the ligand is defined in the QM region, a docking and QM/MM rescoring scheme applied to metal-containing proteins failed in seven of 14 test cases in predicting docked poses with RMSD within 2.0 Å.<sup>156</sup> However, this failure was reduced to only one case when coordinating protein atoms were included along with the metal ion and ligand in the QM region. Previous study using the earlier version of the QM/MM MC module also found effects from different QM definitions on predicting the lowest-energy pose for hCA II:<sup>171</sup> specifically, it was noteworthy that a higher than 70% acceptance rate of MC sampling (Table 4.1) can be achieved when imidazole side groups of zinc-bound histidine residues of hCA II were included in the QM region, regardless of the level of theory applied in that study.

In brief, this section demonstrates that the QM/MM energies calculated from the nonstandard QM/MM MC scheme implemented in Gaussian 09 are valid compared to standard ONIOM energies. Second, the QM/MM MC scheme was adapted to use the semiempirical PM6/AMBER hybrid potential; for this, the MC conformational

sampling routine was implemented in Gaussian 09 for the first time. By using this scheme, the MC acceptance rate at higher than 70% in a test case could be achieved within an hour of run time. Finally, we note that evidence points to the importance of designating zinc-coordinating protein residues in the QM region together with the Zn<sup>2+</sup> ion and the ligand for the purposes of accuracy.

In the next section, optimisation of MC parameters will be discussed.

**Table 4.1 Comparison of standard two-layer ONIOM energies and nonstandard MC-based QM/MM energies in a FSB·hCA II complex (PDB ID 1G52).**

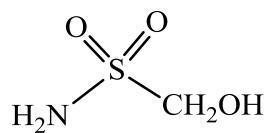
QM region		Charge coupling	Standard ONIOM routine	QM/MM MC	QM/MM MC	% accepted ratio	CPU time / (d:hr:min)
			$E_h$	initial pose	lowest-energy pose		
HF/ 6-31G*	lig	EE	-1471.889753	-1471.889753	-1471.946606	39	0:14:36
	lig + Zn	EE	-3249.444464	-3249.444464	-3249.485130	32	1:18:01
		scaling 500 <sup>b</sup>	-3249.444464				
		scaling 5000 <sup>c</sup>	-3249.406398	-3249.406398	-3249.430700	37	0:17:19
	lig + Zn + 3His <sup>a</sup>	EE	-3923.829414	-3923.829414	-3923.901640	75	6:14:10
		scaling 5000 <sup>c</sup>	-3923.824841	-3923.829414	-3923.875776	62	3:16:29
PM6	lig	EE	-5.071367	-5.071366	-5.114456	32	0:01:08
	lig + Zn	EE	-5.090449	-5.090449	-5.126698	36	0:01:05
	lig + Zn + 3His <sup>a</sup>	EE	-5.370173	-5.370173	-5.400998	72	0:01:39

<sup>a</sup>only the imidazole rings were defined in QM region<sup>b</sup>default scaling factor to turn off MM charges within two bonds of the QM region; the rest are unscaled as 0.2×5=1<sup>c</sup>scaling factor to turn off MM charges within three bonds of the QM region and to leave the rest unscaled

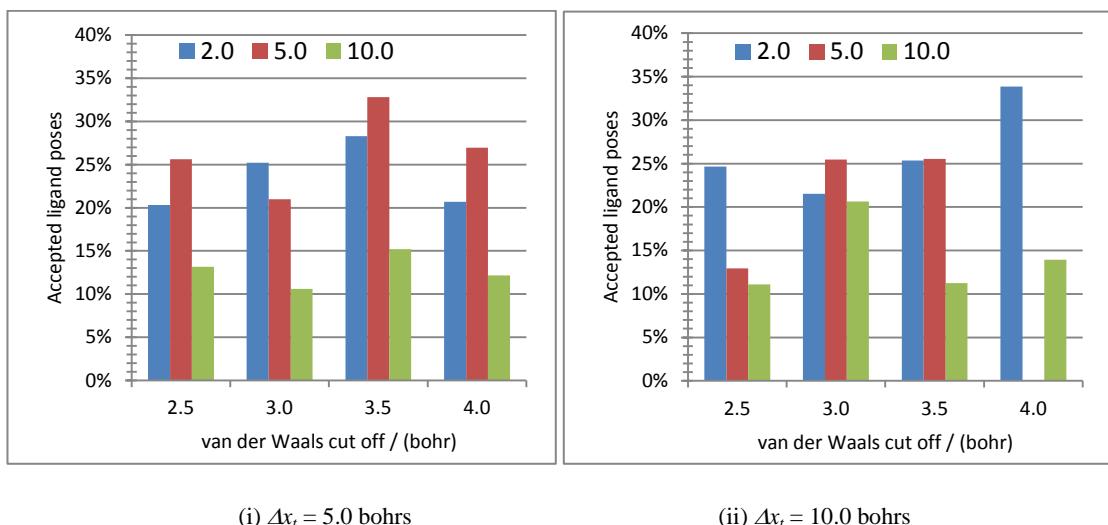
### 4.5.2 Monte Carlo parameter optimisation

This section focuses on identifying the optimal Monte Carlo parameters to use for sampling ligand poses in a subsequent stage. The parameters that require optimisation are the van der Waals cutoff ( $vdw_{cutoff}$ ), the translational step size ( $\Delta x_t$ ), the maximum boundary of a minimum distance between the  $Zn^{2+}$  ion and centre of mass of the ligand ( $d_{max}$ ), and the number of steps in the Markov chain.

On trajectories of 2k attempted MC moves,  $vdw_{cutoff}$  was set to values of 2.5, 3.0, 3.5 and 4.0 bohrs while  $d_{max}$  was fixed at 2.0, 5.0 and 10.0 bohrs. The assessment of these two parameters was carried out using  $\Delta x_t$  of 5.0 and 10.0 bohrs. The polar small molecule, sulfamoyl methanol (Figure 4.8), a representative ligand fragment, was used as a probe in the binding site of hCA II throughout MC parameter optimisation. The probe,  $Zn^{2+}$  ion, and imidazole side chains of His94, His96 and His119 protein residues were defined in the QM region. Figure 4.9 shows the percentage of accepted conformations of the ligand based on the Boltzmann energy distribution during MC sampling (see Table C1 for tabulated values).



**Figure 4.8 Structure of a chemical probe of sulfamoyl methanol used in optimising MC parameters.**



**Figure 4.9 Acceptance rate of sulfamoyl methanol binding to CAII using (i) 5.0 bohrs of  $\Delta x_t$  and (ii) 10.0 bohrs of  $\Delta x_t$ , on 2k MC trajectories for values of  $vdw_{cutoff}$  at 2.5, 3.0, 3.5 or 4.0 bohrs and  $d_{max}$  at 2.0, 5.0 or 10.0 bohrs.**

Typically, the efficiency of MC sampling method is only known after several long simulations. It is, therefore, difficult to determine the optimal acceptance rate. An approximate acceptance ratio of 50% or higher is presumably the best. However, depending on a free energy landscape of a system, acceptance rates could be targeted at around 30% if a system possesses a smooth free energy landscape.<sup>179</sup> For systems with a rugged free energy landscape, these could even shift towards smaller acceptance rates. This is due to the fact that larger step sizes may cross the energy barrier more easily.<sup>179</sup> Practically, acceptance ratios should be adjusted such that the attempted moves are accepted at 20-50% to obtain reasonable convergence behavior.<sup>180</sup> For optimisation of MC parameters on 2k trajectories, the highest accepted rate was achieved (33.9%) when using 10.0 bohrs for  $\Delta x_t$ , 4.0 bohrs for  $vdw_{cutoff}$ , and 2.0 bohrs for  $d_{max}$  (Figure 4.9(ii)). Comparable to the highest acceptance rate, the second top rate was 32.8% when using 5.0 bohrs for both  $\Delta x_t$  and  $d_{max}$ , and 3.5 bohrs for  $vdw_{cutoff}$  (Figure 4.9(i)).

It can be seen from Figure 4.9(ii) that 10.0 bohrs of  $\Delta x_t$  can be problematic in some cases. For instance, the top acceptance rate was 33.9% obtained with  $d_{max}$  of 2.0 bohrs and  $vdw_{cutoff}$  of 4.0 bohrs. It surprisingly dropped to 0.1% when the wider range of 5.0 bohrs for  $d_{max}$  was applied at the same setup. Only one pose out of 2k was accepted in this case. This could be from the conflict arisen from the constraint set up that seeks for the  $Zn^{2+}$ -ligand geometries for this case within 5.0 bohrs but the QM-MM distances higher than 4.0 bohrs were rejected. Therefore, 5.0 bohrs of  $\Delta x_t$  was used for subsequent sampling of inhibitor conformations.

In addition to  $\Delta x_t$ , steric clashes due to van der Waals repulsion is also of concern. When a newly generated ligand pose is placed in a binding pocket, the maximum boundary of a minimum distance allowed between any QM and MM atoms is defined as a  $vdw_{cutoff}$ . Any QM-MM interatomic distance higher than this value will result in discarding the pose without a QM/MM energy evaluation in the subsequent stage. Figure 4.9 shows the tendency to achieve a high acceptance rate when using the  $vdw_{cutoff}$  values of 3.5 or 4.0 bohrs. In general, the higher the  $vdw_{cutoff}$  value, the higher the number of van der Waals repulsions between QM and MM atoms (Table 4.2). For example, using 4.0 bohrs of  $vdw_{cutoff}$  (with 10.0 bohrs for  $\Delta x_t$  and 2.0 bohrs for  $d_{max}$ ), the highest ligand acceptance ratio can be achieved at 33.9% (Figure 4.9(ii)). However, this raised the number of van der Waals repulsions (Table 4.2) to three times higher than that found at a  $vdw_{cutoff}$  of 3.0 bohrs, and roughly five times higher than that at 2.5 bohrs, given the same setup for other parameters. On the other hand, it is also important to note that technical issues such as convergence failure could occur during sampling if nonbonded atoms are too close in contact according to the 12-6 Lennard-Jones

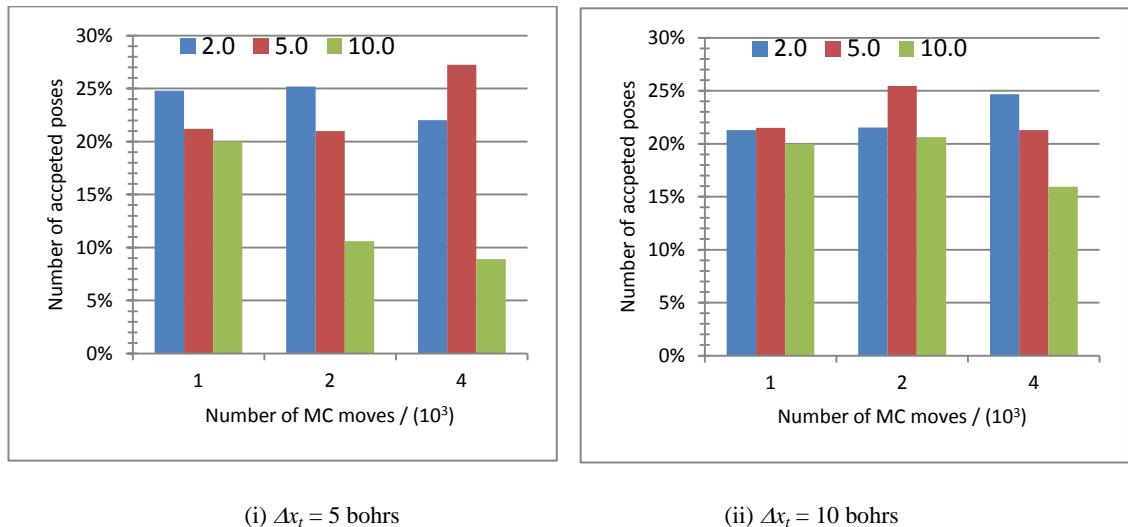
potential. Therefore, the use of an intermediate  $vdw_{cutoff}$  value, of 3.0 bohrs, was applied for subsequent docking.

**Table 4.2 The number of van der Waals clashes between QM and MM atoms when  $\Delta x_t$  was set at 5.0 and 10.0 bohrs for sampling sulfamoyl methanol on 2k MC trajectories in the hCA II site;  $vdw_{cutoff}$  was set at 2.5, 3.0, 3.5, and 4.0 bohrs;  $d_{max}$  was set at 2.0, 5.0, and 10.0 bohrs.**

$vdw_{cutoff}$ / (bohrs)	$d_{max}$ / (bohrs)	Translational step size ( $\Delta x_t$ ) / (bohrs)	
		5.0	10.0
2.5	2.0	140	249
	5.0	81	282
	10.0	98	227
3.0	2.0	511	384
	5.0	229	394
	10.0	492	372
3.5	2.0	791	817
	5.0	1560	1057
	10.0	864	1437
4.0	2.0	1937	1186
	5.0	712	6755
	10.0	668	509

Using 3.0 bohrs for  $vdw_{cutoff}$ , the acceptance rates of ligand poses on the Markov chains of 1k, 2k, and 4k were observed (Figure 4.10). For this,  $d_{max}$  was set at 2.0, 5.0 or 10.0 bohrs, while  $\Delta x_t$  was taken as 5.0 or 10.0 bohrs. The highest rate found was 27.3%, for the 4k Markov chain using 5.0 bohrs for  $d_{max}$  (Figure 4.10(i), see Table C2 for tabulated

values). This is a higher rate than that of 21.0% for the 2k Markov chain in Figure 4.9(i). The sampling efficiency for a 10k Markov chain was also computed as discussed further below.



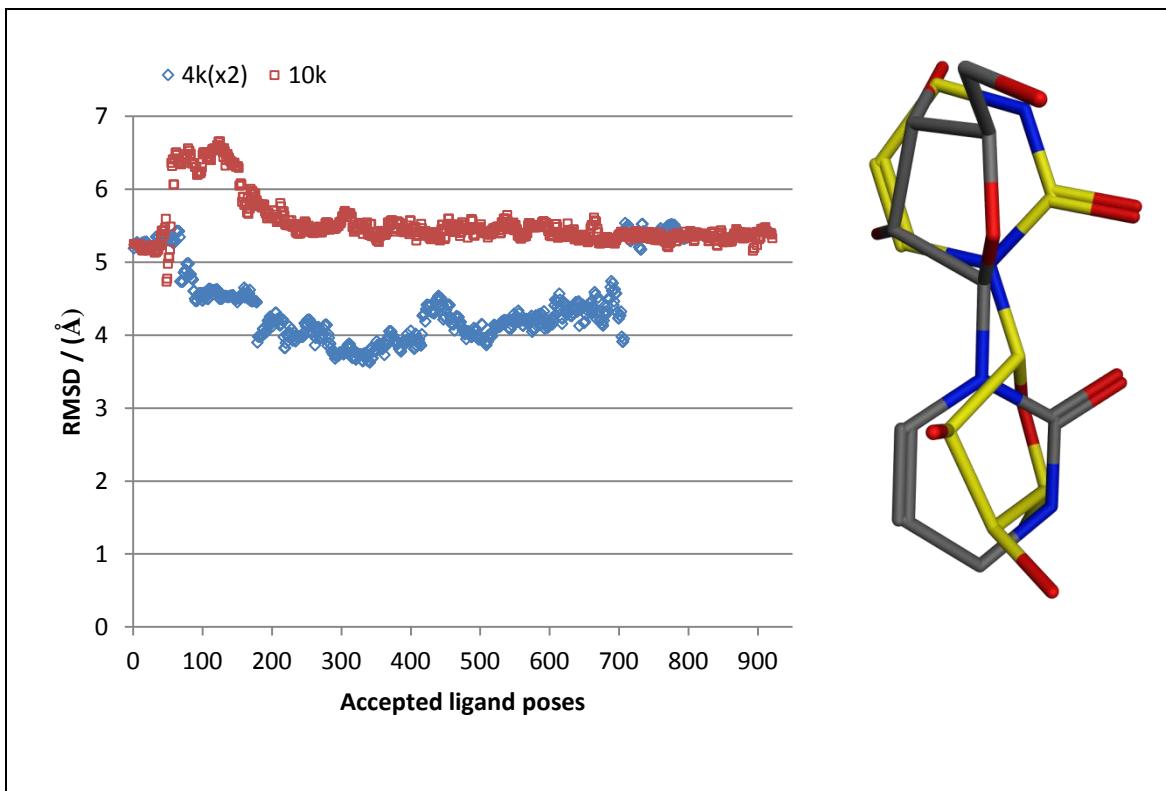
**Figure 4.10 Acceptance rate of sulfamoyl methanol binding to CAII (i) 5.0 bohrs of  $\Delta x_t$  and (ii) 10.0 bohrs of  $\Delta x_t$  on MC trajectories of 1k, 2k and 4k using  $d_{max}$  at 2.0, 5.0, and 10.0 bohrs;  $vdw_{cutoff}$  was set at 3.0 bohrs.**

Another question is how best to create the Markov chain during sampling. In his earlier work, D. Mucs performed a test on parallel versus consecutive sampling for the 1,2-difluorobenzene ligand in CAII receptor.<sup>171</sup> In his study, ten *parallel* runs of 100 MC steps using an ab initio QM/MM potential were carried out with the same starting structure but with different random seeds. This was compared to the result obtained from a single *consecutive* run of 1k MC steps. His results showed that the lowest energy pose of the ligand was found in the consecutive run if the zinc-coordinating side chains of the protein were included in the QM region. Otherwise, the lowest energy pose was found from the parallel runs. Depending on the QM definition, those two lowest energy ligand poses found in parallel and consecutive MC runs were geometrically different in

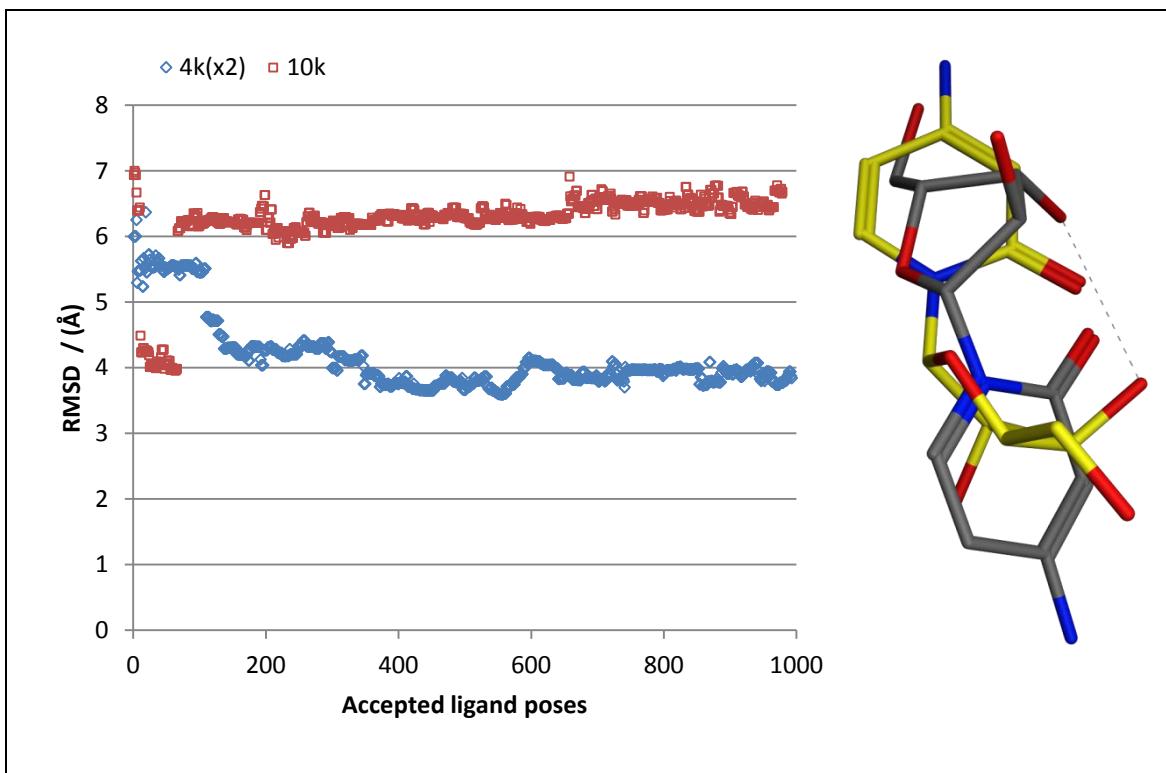
the orientation of the difluorine-substituted phenyl ring. However, they converged to the same energy minimum after optimisation. Both types of MC run were able to explore ligand poses in a similar phase space volume at the active site of CAII.

In D. Mucs' study, the maximum number of MC steps for each run was 1k.<sup>171</sup> The highest possible number of MC steps, given the computational expense of the QM potential, was recommended for use.<sup>171</sup> In this work, for each of the four ligands (for cytidine deaminase, DHZ and DAC; for human carbonic anhydrase II, FSB and DZA, Figure 4.7), two 4k sampling runs were compared with a single sampling run of 10k steps. In general, both types of MC sampling were able to explore the poses that lower RMSD values relative to their initial poses. The initial RMSD was respectively 5.2 and 5.7 Å for DHZ and DAC inhibitors of CDA. The two consecutive 4k runs were able to sample the poses with the RMSD around 4 Å for both inhibitors of CDA (Figure 4.11). In the case of both FSB and DZA inhibitors of hCA II, poses with RMSD less than 2 Å were found, starting from the initial pose with the RMSD of 4.6 Å and 5.7 Å for FSB and DZA, respectively (Figure 4.12). Interestingly, there was a jump in RMSD values on the single 10k trajectory for both ligands of CDA. Figure 4.11 illustrates two consecutive poses of the ligand that caused a jump in RMSD values due to incorrect orientation. The 6-membered ring was spatially interchanged with the 5-membered ribofuranose ring, resulting in higher RMSD values by approximately 2 Å (Figure 4.11). However, the ring-interchanged ligand geometries were more favourable in the QM/MM complex energy by 7.3 and 2.1 kcal/mol (data not shown) for DHZ and DAC, respectively.

A plot of calculated RMSD values over the 10k trajectories for FSB, the inhibitor of hCA II (PDB ID 1G52) shows some fluctuations (Figure 4.12(i)) due to rotation about single bonds present in the structure. This ligand consists of a benzenesulfonamide fragment connected to a difluoro-substituted phenyl ring via an amide bond. The difluorophenyl ring was able to flip around in a solvent-exposed region by rotation around the N-C bond. Three possible alignments of this difluorophenyl ring for example, as shown in Figure 4.12(i), exhibit the RMSD values of 1.7, 2.5 and 4.3 Å. Binding to hCA II, the QM/MM energy of their complexes differed over a range of 4 kcal/mol (data not shown). In contrast to the FSB ligand (PDB ID 1G52), the conformation of the DZA ligand (PDB ID 1CIL) is more conserved because of the rigidity of its fused aromatic heterocyclic ring. As a result, the RMSD values of ligand poses steadily decreased during the beginning of sampling until there was no significant change subsequently. A pair of poses are depicted at an RMSD of 2 Å for both (Figure 4.12(ii)). Their complexes with hCA II give a QM/MM energy difference of 1 kcal/mol (data not shown).

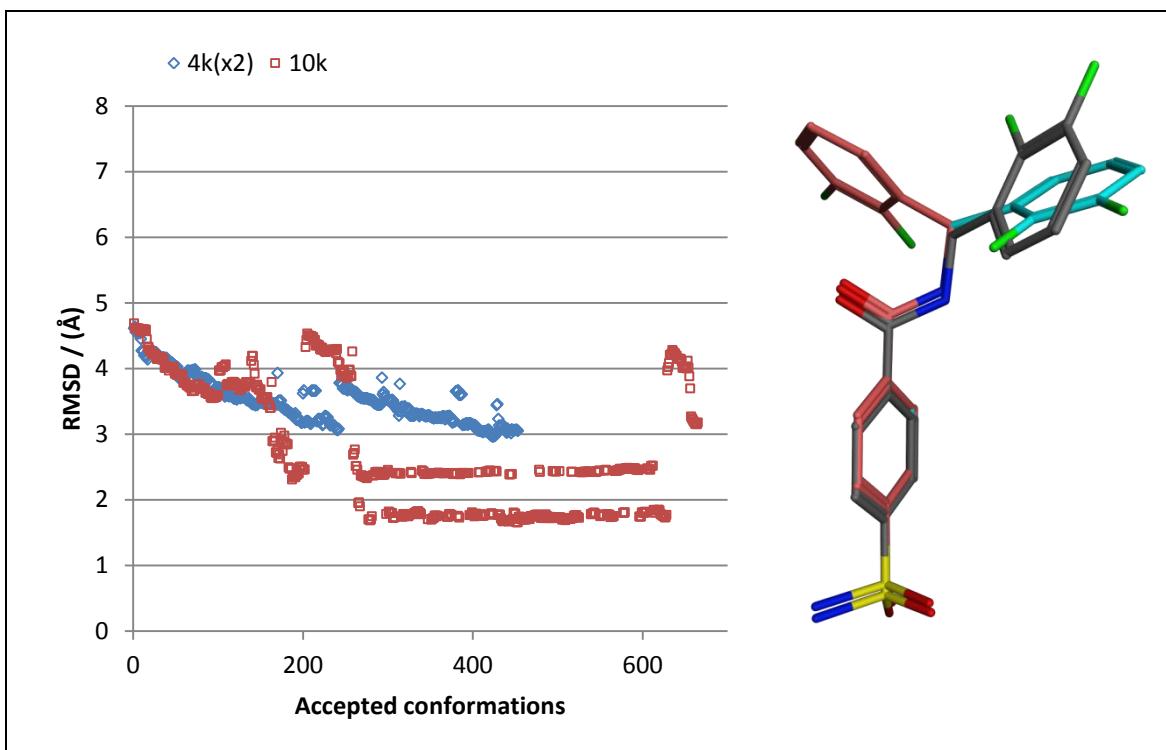


(i) MC poses of DHZ ligand (PDB ID 1CTT) sampling by two of 4k consecutive runs (blue) and a single 10k run (red). Poses with jumping RMSD from 4.7 to 6.4 Å are respectively shown in grey and yellow.

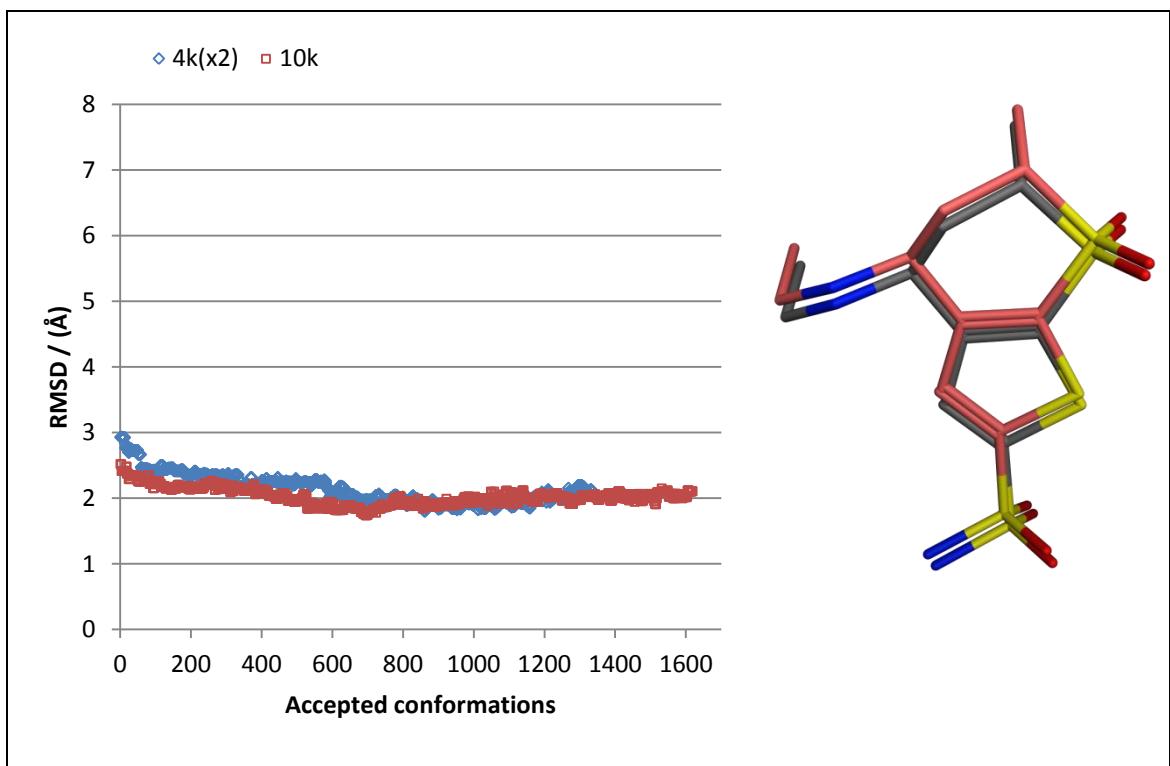


(ii) MC poses of DAC ligand (PDB ID 1ALN) sampling by two of 4k consecutive runs (blue) and a single 10k run (red). Poses with jumping RMSD from 4.0 to 6.1 Å are respectively shown in grey and yellow.

**Figure 4.11 RMSD in Å of accepted poses for CDA receptor and a pair of consecutive poses that causes a sudden change in RMSD values.**



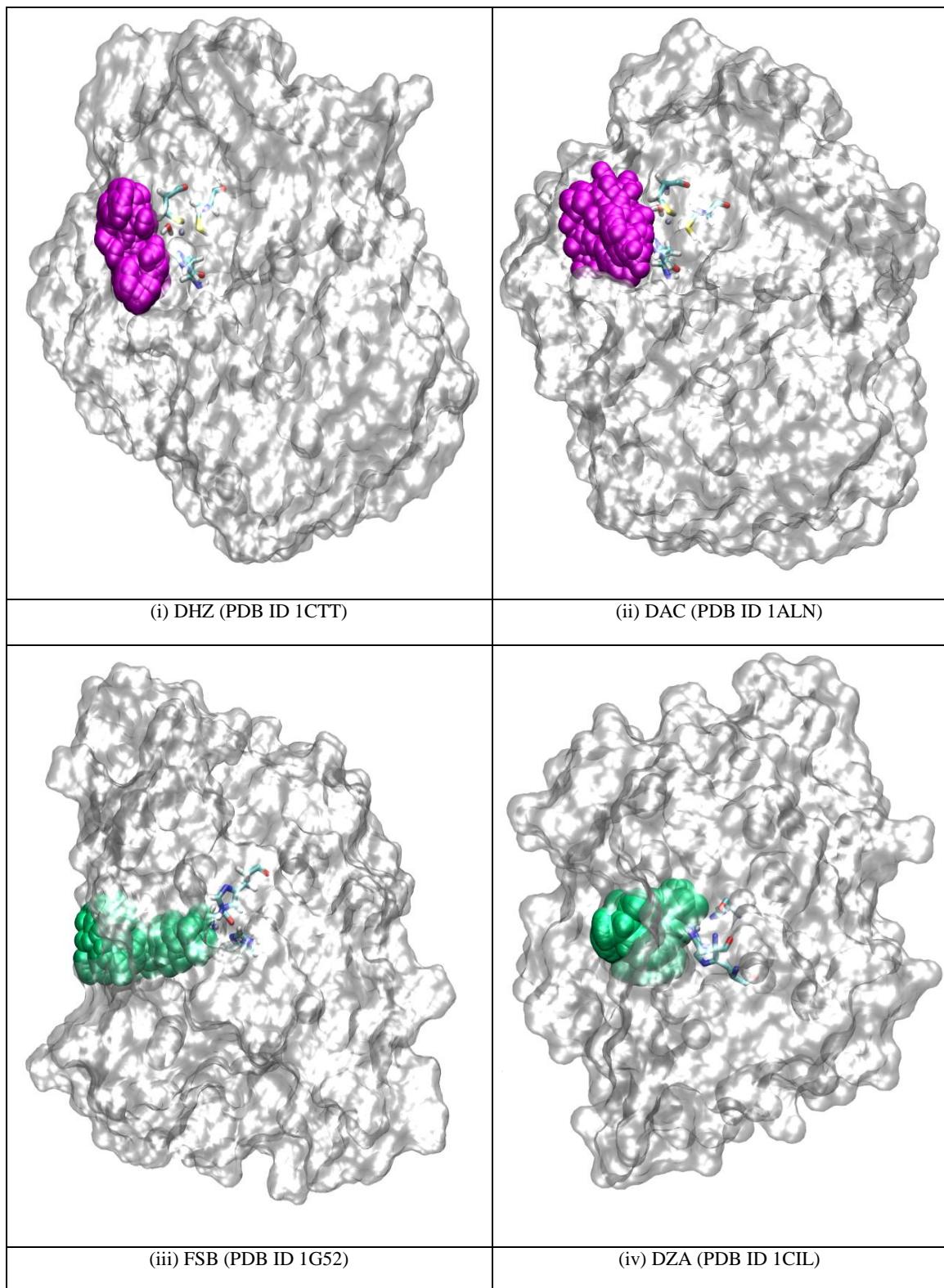
(i) MC poses of FSB ligand (PDB ID 1G52) sampling by two of 4k consecutive runs (blue) and a single 10k run (red). Poses with jumping RMSD from 1.7 to 2.5 and 4.3 Å are respectively shown in grey, cyan and peach.



(ii) MC poses of DZA ligand (PDB ID 1CIL) sampling by two of 4k consecutive runs (blue) and a single 10k run (red). Two consecutive poses with RMSD of 1.7 and 2.0 Å are respectively shown in grey and peach.

**Figure 4.12 RMSD in Å of accepted poses for hCA II receptor and poses that (i) cause a sudden change in RMSD values (ii) present similar structural orientation.**

To sum up for this subsection, a set of MC parameters of 5.0 bohrs for  $\Delta x_t$ , 3.0 bohrs for  $vdw_{cutoff}$  and 3.0 bohrs for  $d_{max}$  for docking were obtained by using a sulfomoyl methanol as a probe in the binding site of hCA II. These parameters were derived based mainly on assessing calculated RMSD values of docked poses with respect to the crystallographic conformations. Using this parameter set, MC search by either two 4k runs or a single 10k run were evaluated for their ability to explore conformations of two inhibitors binding to CDA and two to hCA II. During sampling of 10k steps, incorrect structural alignment of MC poses presented a shift in RMSD values for three cases out of four inhibitors studied. In addition, superposition of all accepted poses sampling from a single 10k run for each inhibitor at the corresponding binding site (Figure 4.13) shows reasonable coverage of phase space by the MC algorithm. In the next section, energy-based scoring of these generated conformations by QM/MM function will be discussed.



**Figure 4.13 Volume sphere of docked poses explored by QM/MM MC scheme in the binding site of cytidine deaminase with (i) 3,4-dihydrozebularine (DHZ) from PDB ID 1CTT and (ii) 3-deazacytidine (DAC) from PDB ID 1ALN; for human carbonic anhydrase II with (iii) N-(2,3-difluoro-benzyl)-4-sulfamoyl-benzamide (FSB) from PDB ID 1G52 and (iv) dorzolamide (DZA) from PDB ID 1CIL.**

### 4.5.3 QM/MM MC docking

For accurately predicting the binding modes of protein-ligand complexes, docking tools should be able to efficiently sample ligand conformations and reasonably score them. Prior to QM/MM MC docking, a hundred ligand poses in the binding site of both CDA and hCA II receptors were generated based on a genetic algorithm (GA) and then ranked by the ChemPLP scoring function available in the GOLD suite, a commercial docking tool. This score is dimensionless, describing how well a ligand pose is aligned in a binding site considering criteria such as ligand shape and ligand interactions with protein residues. The higher the score, the closer the match to the true pose the docked pose is likely to be in principle. In practice, false prediction can occur as observed in this work for CDA systems. The ChemPLP scoring function in the GOLD suite was unable to correctly rank ligand poses; the top-scored poses had RMSDs of 5.2 Å and 5.7 Å for DHZ and DAC, respectively, compared to the crystal structures (Table 4.3); the lowest-scored poses (from the top 100 saved poses from GOLD) had RMSD values of 4.3 Å and 1.3 Å, respectively.

Using the misdocked results obtained from GOLD, a standard ONIOM interaction energy was first computed for the top- and bottom-ranked poses of all four systems studied in this work (Table 4.3). This rescoring appears to be useful only in the case of the DAC-CDA system (PDB ID 1ALN) although the chemical structures of the inhibitors in both CDA complexes are similar. The ONIOM interaction energy was able to distinguish the good ligand pose of DAC with RMSD of 1.3 Å, with a value of –960.1 kcal/mol (Table 4.3) compared to –335.7 kcal/mol for the bad pose (RMSD of 5.7 Å).

The two hCA II inhibitors are more dissimilar than the two CDA ligands, such that they have only a sulfonamide fragment in common. One of them is FSB (PDB ID 1G52) containing flexible rotatable bonds. Another one is DZA (PDB ID 1CIL) possessing a rigid aromatic heterocyclic ring (Figure 4.7). For the FSB ligand, the top-scored and bottom-scored docked poses obtained by the GOLD suite differ by a score of 5.5; their RMSD values differ by only 0.2 Å (Table 4.3). There is difficulty to infer whether or not the ONIOM energy-based rescoring is helpful in this case. The energies of the complexes with those two ligand poses also differ from each other only by 7 kcal/mol. In the case of DZA, the best pose predicted by the GOLD Suite has a ChemPLP score of 57.19 and RMSD of 5.7 Å (Table 4.3). Only in this case from all four systems studied in this work can GOLD correctly distinguish the higher from lower RMSD pose. When the ONIOM energy was computed for the complexes with the top- and bottom-scored poses of DZA, such rescoring fails to distinguish the lowest-RMSD ligand pose.

We note that the QM/MM energy discrepancy in Table 4.3 between the top- and bottom-scored Gold poses is due to a strong electrostatic interaction. For example, the QM/MM energy of DAC-CDA with the top-scored ligand pose was unexpectedly higher than that with the bottom-scored pose by 624.2 kcal/mol. For the top-ranked DAC pose, one of the hydroxyl oxygen of the ribofuranose ring was pointing toward the Zn<sup>2+</sup> ion at a distance of 2.37 Å. Its Mulliken partial charge was positive with the value of 0.34 *e*, compared with the negative charge of -2.12 *e* at the same atom of the bottom-ranked pose. As a result, the positive charge-charge electrostatic interaction between the hydroxyl oxygen and the Zn<sup>2+</sup> ion is responsible for the very high QM/MM energy. This

could contribute to the failure of using QM/MM rescoring without the minisation scheme.

Another issue in rescoring using the ONIOM energy could relate to the existence of multiple minima of similar energy on the QM/MM energy surface, as discussed by Fong *et al.*<sup>181</sup> They have reported that for docking to HIV-1 protease, the polarization energy ( $\Delta E_{pol}$ ) component of QM/MM binding potentials is relatively insensitive to changes in RMSD of docked poses from native ones. Although in their studies  $\Delta E_{pol}$  comprised 12-18% of a total protein-ligand electrostatic interaction, it affects the accurate discrimination of native from non-native poses by distorting the smooth topology of binding energy profile required for efficient docking.

**Table 4.3 Ligand poses generated by GA search algorithm and then ranked by empirical ChemPLP score compared to ONIOM energy-based score.**

Receptor	Ligand (PDB ID)	GOLD Suite		Gaussian 09
		ChemPLP Score	RMSD / (Å)	ONIOM energy / (kcal/mol)
Cytidine deaminase (CDA)	DHZ (1CTT)	54.07	5.2	-997.2
		39.75	4.3	-871.4
	DAC (1ALN)	52.24	5.7	-335.7
		46.57	1.3	-960.1
Human carbonic anhydrase II (hCA II)	FSB (1G52)	65.48	4.6	-3583.4
		59.98	4.4	-3576.4
	DZA (1CIL)	57.19	5.7	-2520.0
		28.87	7.1	-3212.2

To improve the docking results obtained from GOLD, the MC conformational search method and QM/MM energy scoring function were integrated (see Section 4.3) and applied to these top-scored poses that were misdocked by GOLD. For comparison, two separate 4k MC runs and one 10k MC run were used for docking each ligand (Table 4.4).

For all four systems studied, for the 4k and 10k runs, the QM/MM MC algorithm was able to identify docked poses that were more favourable in QM/MM complex energy than the initial GOLD poses; and poses with smaller RMSD values than the GOLD poses (Table 4.4). However, as suggested by the single point ONIOM calculations above, the lowest energy and the lowest RMSD poses from the same MC runs were not necessarily identical. For example, the lowest energy pose of DHZ ligand docked to the CDA protein from the first trajectory of 4k steps was  $-1031.8$  kcal/mol in QM/MM interaction energy, with an RMSD of  $4.3$  Å (4k-run1 in Table 4.4). This is  $8.0$  kcal/mol lower in energy than the lowest RMSD pose (with a value of  $3.6$  Å) sampled from the same trajectory. Such difference in complex energy between the lowest energy and the lowest RMSD poses on each trajectory was in the range of  $2\text{--}25$  kcal/mol depending on the system studied.

In terms of the difference in sampling by the 4k and 10k runs, the longer sampling of 10k MC steps only improved sampling, i.e. lower RMSD values of ligand conformations and lower energy of the complex, of the two hCA II complexes but not for CDA ligand docking. For hCA II, a single run of 10k MC steps performed better compared to the two 4k runs in these respects. For example, the lowest RMSD pose of the FSB ligand from its 10k trajectory had an RMSD of  $1.7$  Å which was  $1.3$  Å lower in

value compared with the lowest RMSD pose of the same ligand obtained from the 4k trajectories (Table 4.4). For the same FSB system, the complex with the lowest energy pose from the 10k trajectory had a value of -3617.9 kcal/mol, which was 5 kcal/mol lower in energy than that from the 4k trajectories. In the case of DZA, the energy improvement of the 10k run over the 4k runs was about 8-9 kcal/mol, while the lowest RMSD found was not improved significantly. There is no clear evidence here that longer MC trajectories give better results for CDA systems. For example, the lowest energy conformation of docked DHZ from the 10k run gave a complex energy of -1033.9 kcal/mol (Table 4.4). This energy value is in between that of -1031.8 and -1035.0 kcal/mol for the lowest energy conformations found by the two 4k runs. For this 10k trajectory, the RMSD value of the lowest RMSD pose was 4.7 Å, somewhat higher than that of 3.6 and 3.9 Å found from the 4k runs.

**Table 4.4 The complex energy (in kcal/mol) and RMSD (in Å) of the best poses scoring by the QM/MM function; Ligand poses sampling from two consecutive 4k MC runs and a single 10k MC run.**

Receptor	Ligand (PDB ID)	Initial pose from GOLD docking		Number of MC steps	Lowest energy conformer			Lowest RMSD conformer		
		RMSD / (Å)	Complex energy / (kcal/mol)		MC step#	RMSD / (Å)	Complex energy / (kcal/mol)	MC step#	RMSD / (Å)	Complex energy / (kcal/mol)
Cytidine deaminase (CDA)	DHZ (1CTT)	5.2	-997.2	4k-run1	1767	4.3	-1031.8	3332	3.6	-1023.8
				4k-run2	3473	5.4	-1035.0	997	3.9	-1022.2
				10k	1667	5.5	-1033.9	401	4.7	-1017.6
	DAC (1ALN)	5.7	-335.7	4k-run1	3257	3.9	-1035.5	3790	3.6	-1033.6
				4k-run2	51	3.8	-1036.0	749	3.6	-1027.7
				10k	3524	6.2	-1032.2	738	4.0	-1020.4
Human carbonic anhydrase II (hCA II)	FSB (1G52)	4.6	-3583.4	4k-run1	1032	3.8	-3612.6	3957	3.1	-3594.4
				4k-run2	85	3.8	-3612.9	3516	3.0	-3593.8
				10k	2925	4.4	-3617.9	6473	1.7	-3592.6
	DZA (1CIL)	5.7	-2520.0	4k-run1	3474	2.0	-3402.9	3772	1.9	-3401.3
				4k-run2	397	2.0	-3403.9	805	1.8	-3401.8
				10k	765	2.2	-3411.8	3954	1.7	-3403.3

Turning now to look at the protein-ligand geometries of the best docked MC QM/MM poses listed in Table 4.4, this will be compared with active site interactions of the native pose of the inhibitor and the top-scored GOLD docked poses. As discussed above, the lowest energy ligand poses predicted by the QM/MM MC docking algorithm were different from the lowest RMSD poses. To observe whether they converge to the same minima, their optimised poses were superimposed onto the optimised native conformation. Protein-ligand interactions of those optimised poses are also discussed further

The key interactions of the DHZ inhibitor at the binding site of CDA (PDB ID 1CTT) come from both the six-membered and ribofuranose rings. Of the six-membered ring, the NH of the amide group interacts with the carboxyl side group of Glu104 (grey ligand in Figure 4.14 panels). The oxygen on the ribofuranose ring interacts with the amide proton of the Asn89 side chain. There is also the interaction between one of the ribose hydroxyl groups with the carboxyl side group of Glu91.

The top-scored GOLD pose deviates from this native conformation with an RMSD of 5.2 Å (cyan, Figure 4.14(i)). It vertically rotates to some degree so that two rings are interchanged. This geometry is stabilised by hydrogen bonds between the ligand hydroxyl groups and the carbonyl group of Thr127. One of these hydroxyl groups also forms a hydrogen bond with the zinc-coordinating water molecule. When this top-scored GOLD pose binds to CDA, the QM/MM energy of its protein-ligand complex was -997.2 kcal/mol.

This energy was improved by 37.8 kcal/mol when using the QM/MM MC docking module (Table 4.4). The alignment of this lowest energy QM/MM MC docked pose, found from the 4k runs, is a little closer, in terms of volume overlap, to the native conformation (pink, Figure 4.14(ii)) than its parent GOLD docked pose but with no improvement in RMSD, with a value of 5.4 Å. Both poses are considered misdocked; positions of both six-membered and ribofuranose rings were exchanged with respect to the crystal conformation (Figure 4.14(i) and (ii)). However, this orientation facilitates hydrogen bonds forming between hydroxyl groups of ribofuranose ring and the carbonyl group of Thr127 (or zinc-coordinating water), instead of Glu91 and Asn89 residues as in the case of the native conformation.

The QM/MM MC scheme is able to sample the lower RMSD pose for DHZ, with a value of 3.6 Å from the 4k runs (Figure 4.14(iii)). The position of both six-membered and ribofuranose rings in the binding site were improved over that of the lowest energy QM/MM MC docked pose and the top-scored GOLD docked pose. Considering interactions of the lowest RMSD pose binding to CDA, the amide proton of the six-membered ring forms a hydrogen bond to the carbonyl group of Thr127 rather than the carboxyl side group of Glu104 as seen in the native conformation. Overall, the ligand overlaps the correct rings but upside down relative to the native pose. However, the QM/MM energy of the complex with this lowest RMSD pose was 11.2 kcal/mol higher than that of the complex with the lowest QM/MM energy pose. The lowest energy and lowest RMSD docked poses for this system therefore locate in different energy minima.

Then, the native, GOLD-docked and QM/MM-docked ligand-protein complexes were each energy optimised using the QM/MM potential (Figure 4.15). Optimisation of the

lowest energy and the lowest RMSD docked complexes leads to lowering of the QM/MM energy by 40.7 and 53.8 kcal/mol respectively. However, these poses are now found to be 17.4 and 15.5 kcal/mol higher in energy than the optimised native pose (Figure 4.15(i)). Furthermore, optimisation does not improve the agreement of the docked poses with the native geometry for either the lowest energy or the lowest RMSD pose (Figure 4.15(ii) and (iii) compared with Figure 4.14(ii) and (iii)). The results highlight the distinctiveness of the minima found by docking from the native state. Interestingly however it does suggest that the QM/MM potential, *with optimisation*, is capable of placing the native geometry as the lowest energy minimum, for this ligand at least, although it was not found by these docking runs.

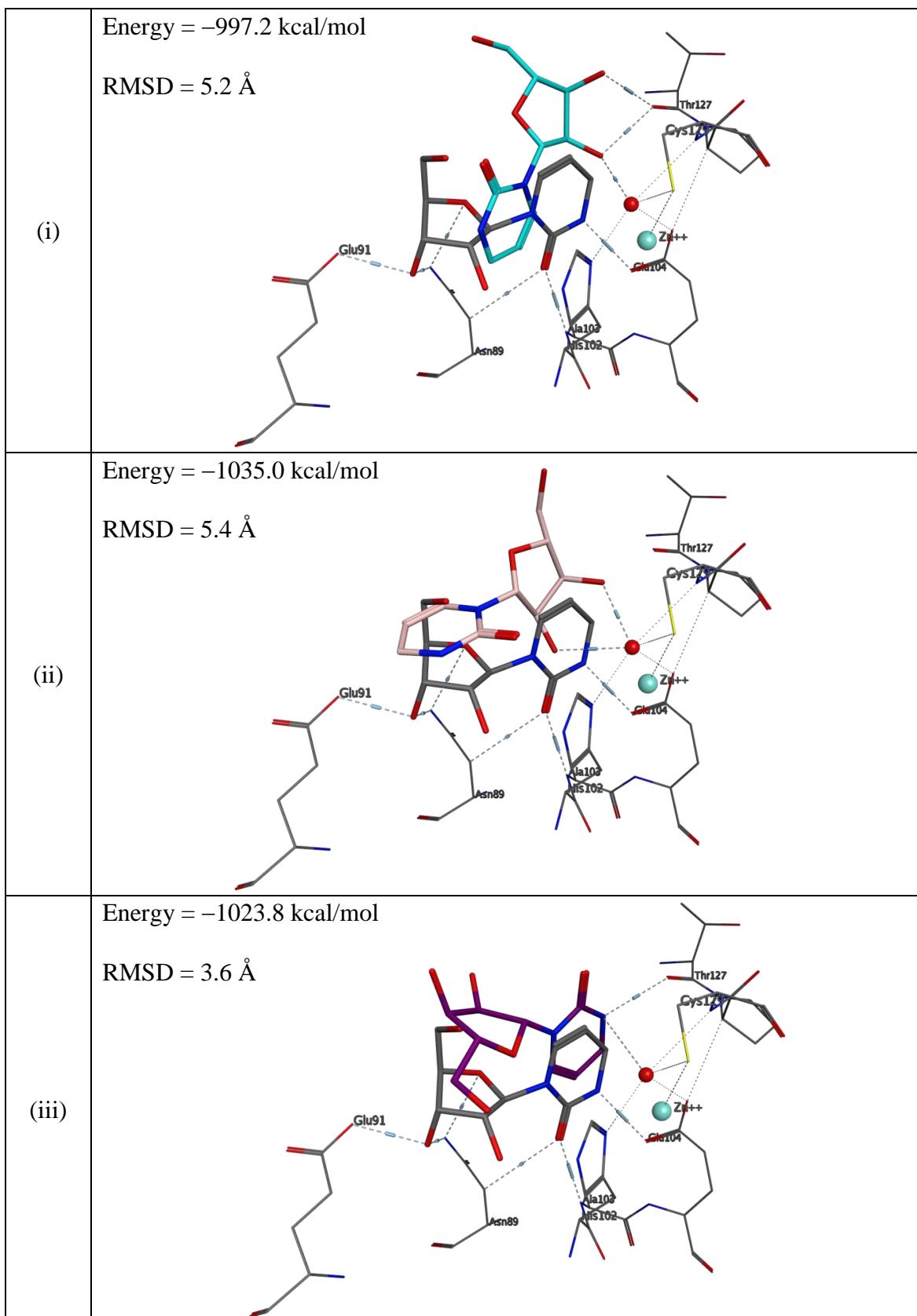


Figure 4.14 Superposition of (i) top-scored GOLD docked pose (cyan stick), (ii) lowest energy QM/MM MC docked pose (pink stick), and (iii) lowest RMSD QM/MM MC docked pose (purple stick) onto the native conformation (grey stick) of DHZ (PDB ID 1CTT) at the binding site of CDA.

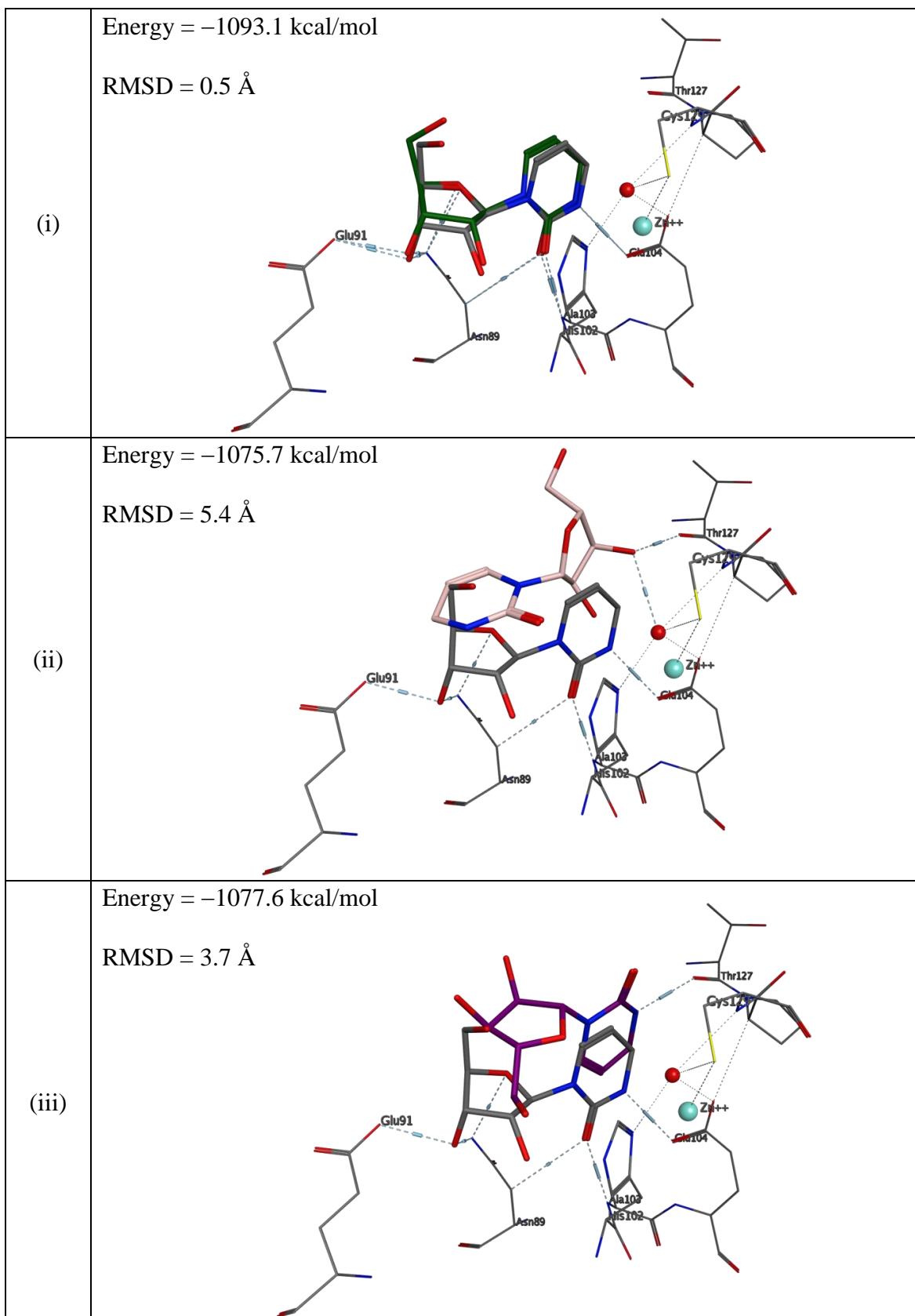


Figure 4.15 Superposition of (i) optimised native pose (green stick) (ii) optimised lowest energy docked pose (pink stick), and (iii) optimised lowest RMSD docked pose (purple stick) onto the native pose (grey stick) of DHZ (PDB ID 1CTT) at the binding site of CDA.

The second CDA system in this study involves the DAC ligand, found in PDB ID 1ALN. Weakly noncovalent protein-ligand interactions occur from the 6-membered ring of this pyridine analogue, such as CH- $\pi$  interaction with Phe71 and the methyl group of Val73 (Figure 4.16(i)). The carbonyl and amino groups on the ring also form hydrogen bonds to the backbone amide proton of Ala103 and the carbonyl group of Thr127, respectively. The ribofuranose ring also interacts with the amide proton of the Asn89 side chain and the hydroxyl groups with the carboxyl side group of Glu91, similar to the ribofuranose ring of the DHZ inhibitor previously discussed.

The top-scored docked pose obtained from using GOLD Suite was misdocked with RMSD 5.7 Å in Figure 4.16(i), similar to the case of the DHZ ligand in CDA system with PDB ID 1CTT. Although this pose was not off-centre from the native conformation, the two rings were again interchanged. One of the hydroxyl groups of the riboosse interacts with both the Zn<sup>2+</sup> ion and the carboxyl side group of Glu104 while the amino group on the pyridine ring forms hydrogen bonds to the amide side group of Asn89 and the carboxyl side group of Glu91.

Based on the energy-ranking criteria, the QM/MM MC algorithm can identify the ligand pose that aligns both rings in a correct manner (Figure 4.16(ii)). This lowest energy pose is more favourable in QM/MM energy than that of the top-scored GOLD docked pose by a very large degree (700.3 kcal/mol). The RMSD value was improved from 5.7 to 3.8 Å. The lowest RMSD docked pose (Figure 4.16(iii)) was found to be 8.3 kcal/mol higher in energy than the lowest energy pose; however, it is insignificantly different in RMSD value, i.e. 3.6 Å compared with that of 3.8 Å for the lowest energy pose. These two poses are also located in the same potential energy minimum: on optimisation using

the QM/MM potential, these optimised ligand geometries are identical both in terms of energy ( $-1089$  kcal/mol) and RMSD ( $3.9$  Å) as shown in Figure 4.17(ii) and (iii). In comparison, the optimised native pose had a QM/MM energy of  $-1087.6$  kcal/mol, very close to the QM/MM docked poses, but with the better RMSD value of  $1$  Å (Figure 4.17(i)). This finding suggests there are multiple minima on the QM/MM energy surface with similar low energy.

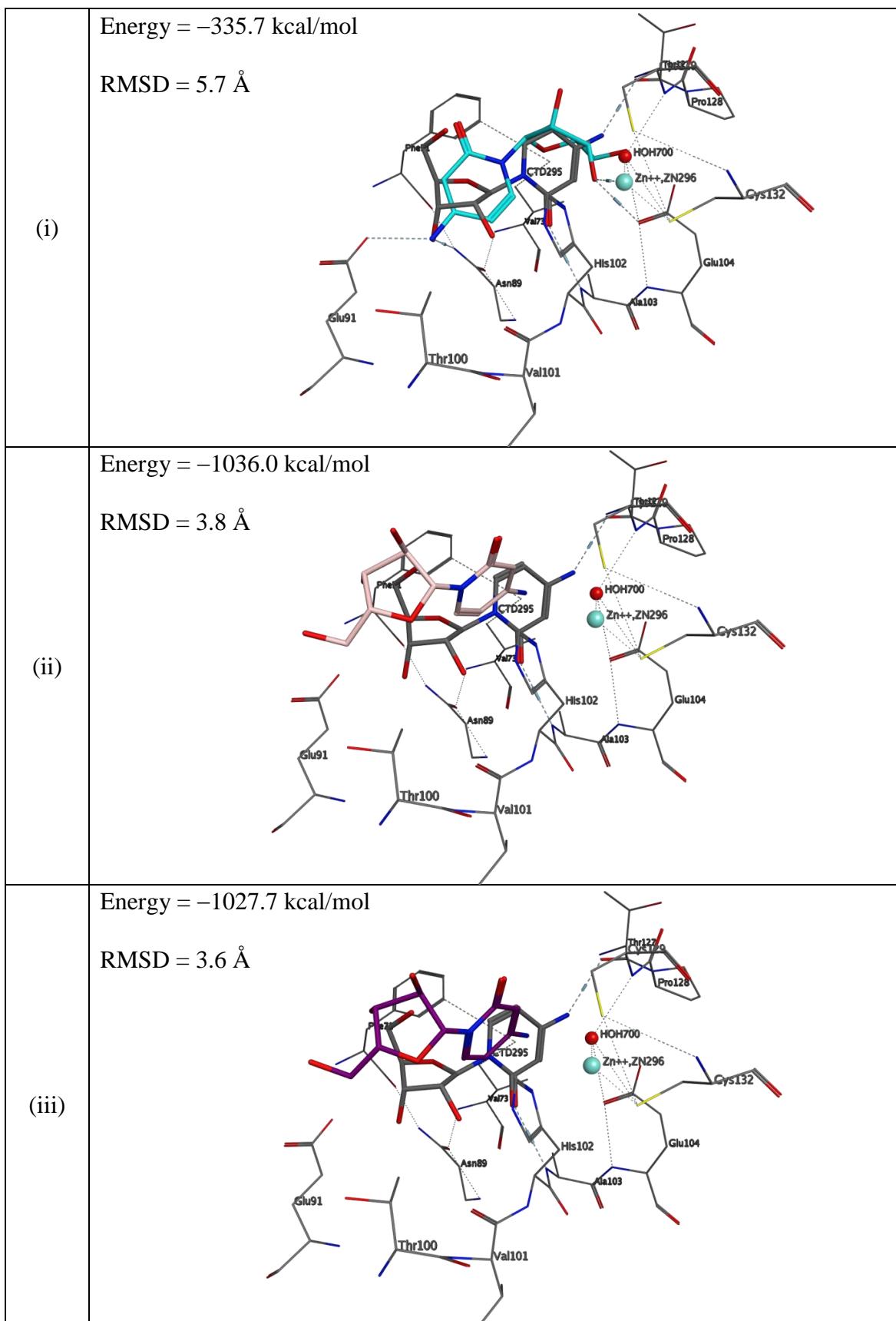


Figure 4.16 Superposition of (i) top-scored GOLD docked pose (cyan stick), (ii) lowest energy QM/MM MC docked pose (pink stick), and (iii) lowest RMSD QM/MM MC docked pose (purple stick) onto the native conformation (grey stick) of DAC (PDB ID 1ALN) at the binding site of CDA.

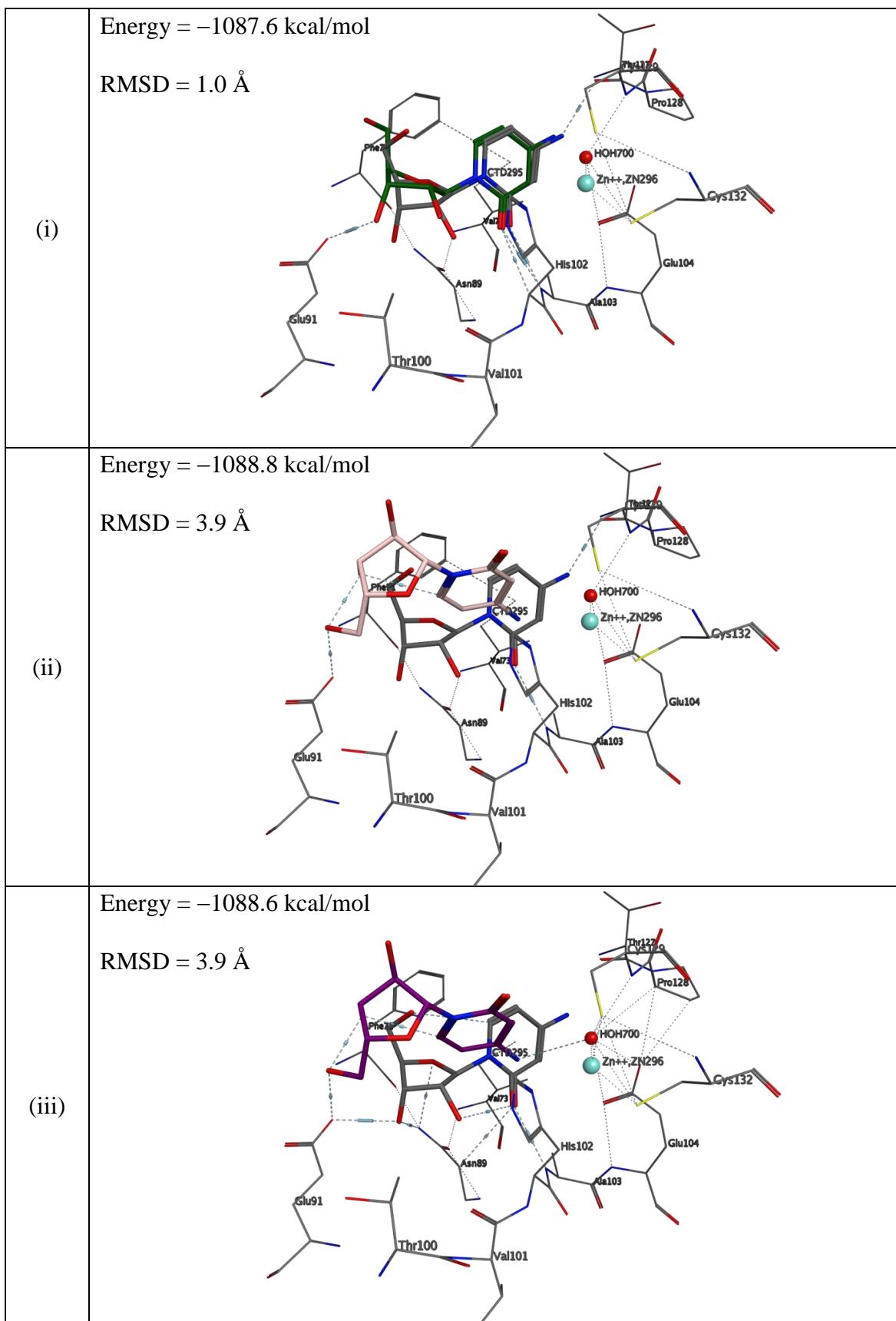


Figure 4.17 Superposition of (i) optimised native pose (green stick) (ii) optimised lowest energy docked pose (pink stick), and (iii) optimised lowest RMSD docked pose (purple stick) onto the native pose (grey stick) of DAC (PDB ID 1ALN) at the binding site of CDA.

Now, we turn to look at the docked poses at the binding site of hCA II. A functional group of the ligand in common for both systems studied is the sulfamoyl group (see Figure 4.7 in section 4.4.1). For the native conformation of ligand FSB in the hCA II active site (PDB ID 1G52), the amino of the sulfamoyl group forms hydrogen bonds to the  $Zn^{2+}$  and zinc-bound His94, His96 and His119 residues as well as the hydroxyl group of Thr199 (grey ligand, Figure 4.18). The sulfonyl moiety also aligns such that it forms hydrogen bonds to both Thr199 and Leu198. The orientation of a phenyl ring of the benzenesulfonamide also facilitates a CH- $\pi$  interaction to a methyl group of Leu198. In addition, the hydrogen bond forms between this water and the NH of the ligand amide affects the position of the difluoro-substituted phenyl ring aligned between Phe131 and Val135.

Compared with the crystallographic conformation of the ligand described above, the top-scored GOLD docked pose deviates by RMSD of 4.6 Å (cyan, Figure 4.18(i)). Its sulfamoyl group aligns in an opposite direction (Figure 4.18(i)), so that its  $SO_2$  moiety, instead of the NH, is pointing to the  $Zn^{2+}$  and zinc-bound histidine residues. The carbonyl of the amide group lifts away to interact with the amide side group of Gln92. The CH- $\pi$  interaction of the phenyl ring is not observed. All these interactions likely constrain the difluoro-substituted phenyl ring away from the vicinity of Phe131 and Val135; otherwise, it would sterically clash with Phe131.

The lowest energy docked pose obtained from QM/MM MC algorithm provided no significant improvement in RMSD value (4.4 Å) compared with its initial GOLD docked pose (4.6 Å). However, the amino of the sulfamoyl group in this lowest-energy docked pose is correctly oriented with respect to the  $Zn^{2+}$  (Figure 4.18(ii)). One oxygen

of its sulfonyl group moves downwards, closer to the  $Zn^{2+}$  by 2.6 Å relative to that of the crystal structure while another oxygen is 2.7 Å away to interact with Thr200 instead of Leu198. This may be responsible for the more favourable QM/MM energy of the complex by 34.5 kcal/mol relative to the top-scored GOLD docked pose. After energy minimisation of the lowest energy docked pose, the QM/MM energy of the complex was also improved further by 32.2 kcal/mol relative to its unoptimised one (Figure 4.19(ii)). This could be due to CH- $\pi$  and polar interactions forming between the difluoro-substituted phenyl ring and Gln92. However, there was no real improvement in RMSD value for the optimised lowest energy docked pose of FSB (4.7 Å). The optimised pose of crystallographic ligand geometry remains close to its unoptimised pose, with a RMSD of 1.3 Å (Figure 4.19(i)), slightly differing at the position of aryl rings. With a RMSD of 4.7 Å after optimisation, it can be seen that the QM/MM lowest energy pose did not sit in the same energy minimum as the crystal ligand geometry (Figure 4.19(i) and (ii)).

However, the QM/MM MC algorithm was able to sample the crystallographic pose that correctly aligns the ligand, and in particular the position of difluoro-substituted phenyl ring, compared with that of the crystallographic structure. This difluoro-substituted phenyl ring interacts with Phe131; its substituted fluorine atoms were pointing outwards in similar direction to that in the native pose (Figure 4.18(iii)). The oxygen atoms of the sulfamoyl group are located slightly away by 3-4 Å compared with those in the crystal structure. This pose has a RMSD of 1.7 Å but its QM/MM energy was 25.3 kcal/mol higher than that of the lowest energy pose (Figure 4.18(ii)). After optimisation of this lowest RMSD pose, the QM/MM energy of the complex improved by 45.3 kcal/mol relative to its unoptimised one, to a value of -3637.9 kcal/mol. This optimised pose

almost resembles the crystallographic conformation except for the position of a sulfonyl group (RMSD of 1.8 Å, Figure 4.19(iii)); the crystal pose has a slightly lower optimised energy of –3649.1 kcal/mol however.

Comparing the conformations of the lowest energy and lowest RMSD poses, the main difference is the orientation of difluoro-substituted phenyl ring (Figure 4.18(ii) and (iii)) with respect to rotation about the single bond to the CH<sub>2</sub> moiety. This can be observed from the energy optimised geometries of these two conformations (Figure 4.19(ii) and (iii)). Interestingly, a similar observation has been made by Gitto *et al.*<sup>160</sup> Two active conformations of one synthesised arylsulfonamide, compound 2a in their published paper, differed in the alignment of the aryl ring, such that it flipped over 180° by a single bond rotation. These two conformations, so-called atropisomers, cocrystallised within the catalytic pocket of hCA II but remained undetected by <sup>1</sup>H NMR. Based on this evidence, it was suggested that this compound could be an effective inhibitor in both conformations.

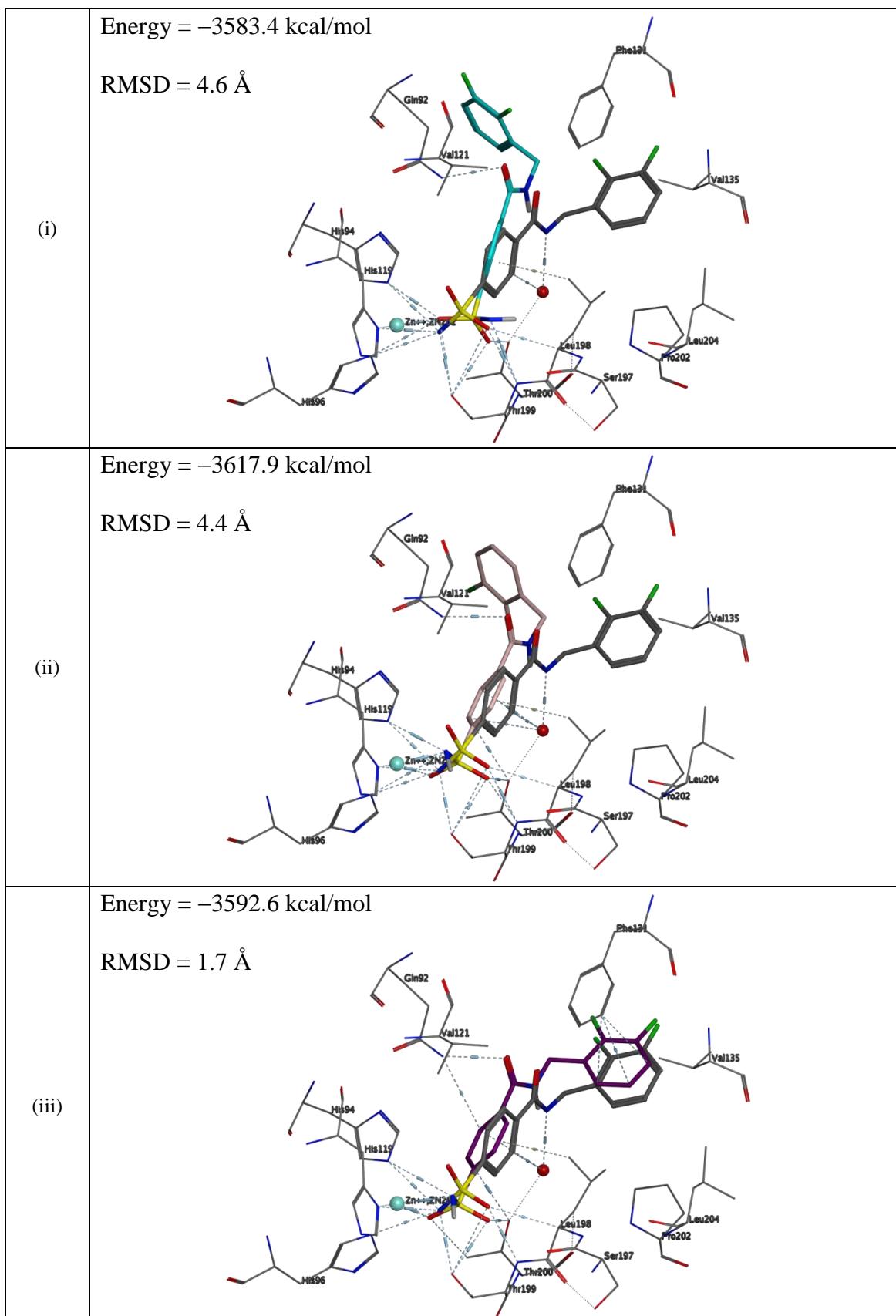


Figure 4.18 Superposition of (i) top-scored GOLD docked pose (cyan stick), (ii) lowest energy QM/MM MC docked pose (pink stick), and (iii) lowest RMSD QM/MM MC docked pose (purple stick) onto the native conformation (grey stick) of FSB (PDB ID 1G52) at the binding site of hCA II.

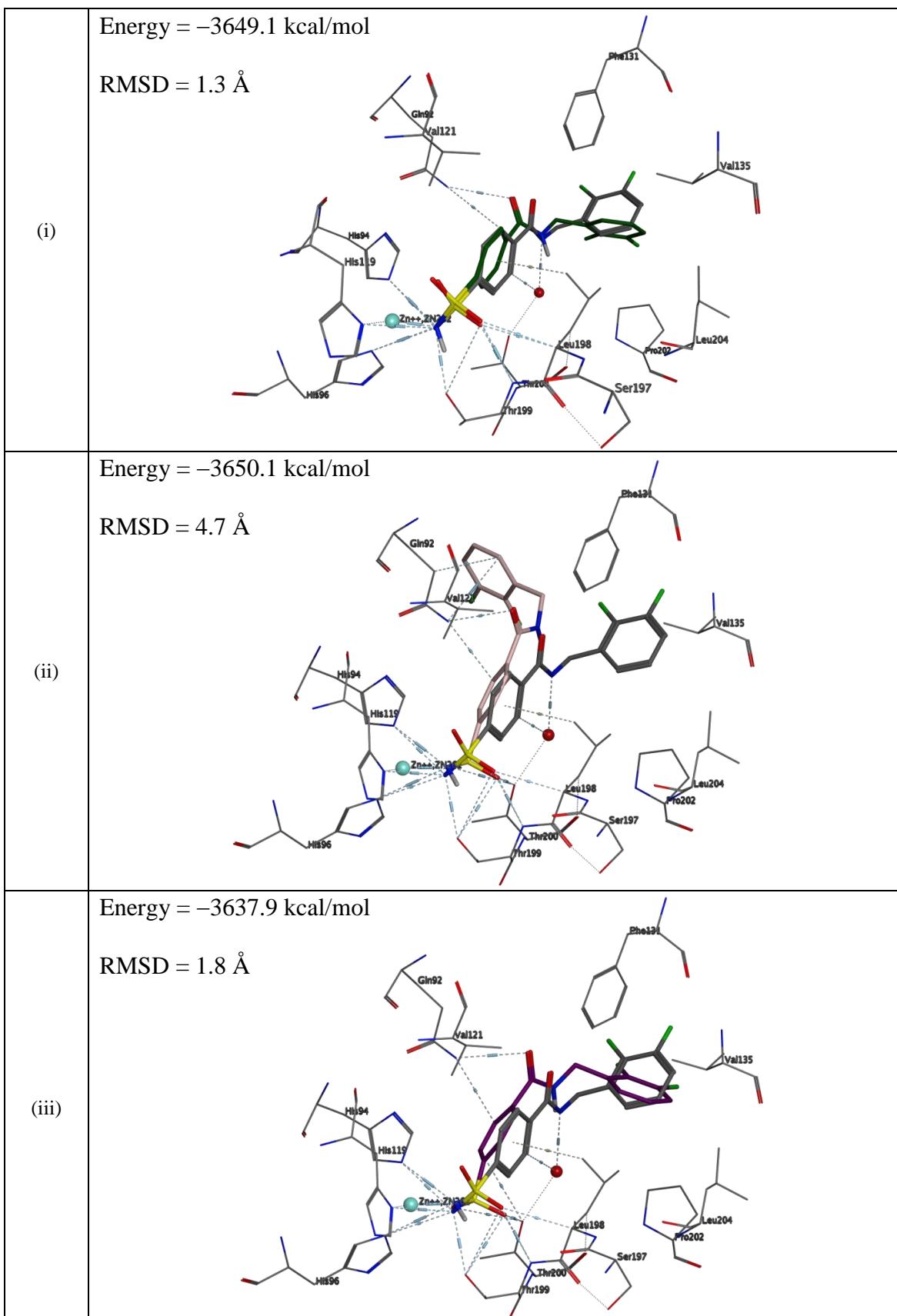


Figure 4.19 Superposition of (i) optimised native pose (green stick) (ii) optimised lowest energy docked pose (pink stick), and (iii) optimised lowest RMSD docked pose (purple stick) onto the native pose (grey stick) of FSB (PDB ID 1G52) at the binding site hCA II.

Finally, the second hCA II inhibitor studied in this work is dorzolamide (DZA), a heterocyclic primary sulphonamide derivative found in PDB ID 1CIL (see Figure 4.7, section 4.4.1). Regarding the top-scored docked pose obtained from GOLD Suite, it was misdocked with an RMSD of 5.7 Å (Figure 4.20(i)). The sulfonyl group on the thiopyran ring was placed near the Zn<sup>2+</sup> instead of the sulfamoyl group as in the native conformation. The orientation of this GOLD docked pose is effectively upside down.

Compared with the misdocked pose predicted by GOLD, the QM/MM MC docking algorithm was able to identify a pose very close to the crystal pose for the DZA ligand. The lowest energy docked pose (Figure 4.20(ii)) was well aligned to the native conformation with a RMSD of 2.2 Å. Its sulfamoyl group points the NH towards the Zn<sup>2+</sup>. One of its sulfonyl oxygens points towards the imidazole ring of His94 while another sulfonyl oxygen binds to the amino backbone of Thr199. The QM/MM energy of the complex with this lowest energy pose was more favourable by a very large degree (891.8 kcal/mol) compared with the top scored GOLD docked pose.

Furthermore, a QM/MM MC pose docked closer to the native ligand conformation than this lowest energy was found, with an RMSD of 1.7 Å (Figure 4.20(iii)). The QM/MM energy of the lowest RMSD docked pose was only slightly higher, by 8.5 kcal/mol, relative to the lowest energy docked pose.

On optimisation, the lowest energy and lowest RMSD poses docked by the QM/MM MC algorithm converged to the same energy minimum. This can be seen in Figure 4.21(ii) and (iii), such that their optimised geometries are identical with RMSDs of 1.7 Å, and with a QM/MM energy difference of 0.9 kcal/mol. However, the optimised

geometries of these two docked poses are different from the native conformation in the ethylamine side group (Figure 4.21).

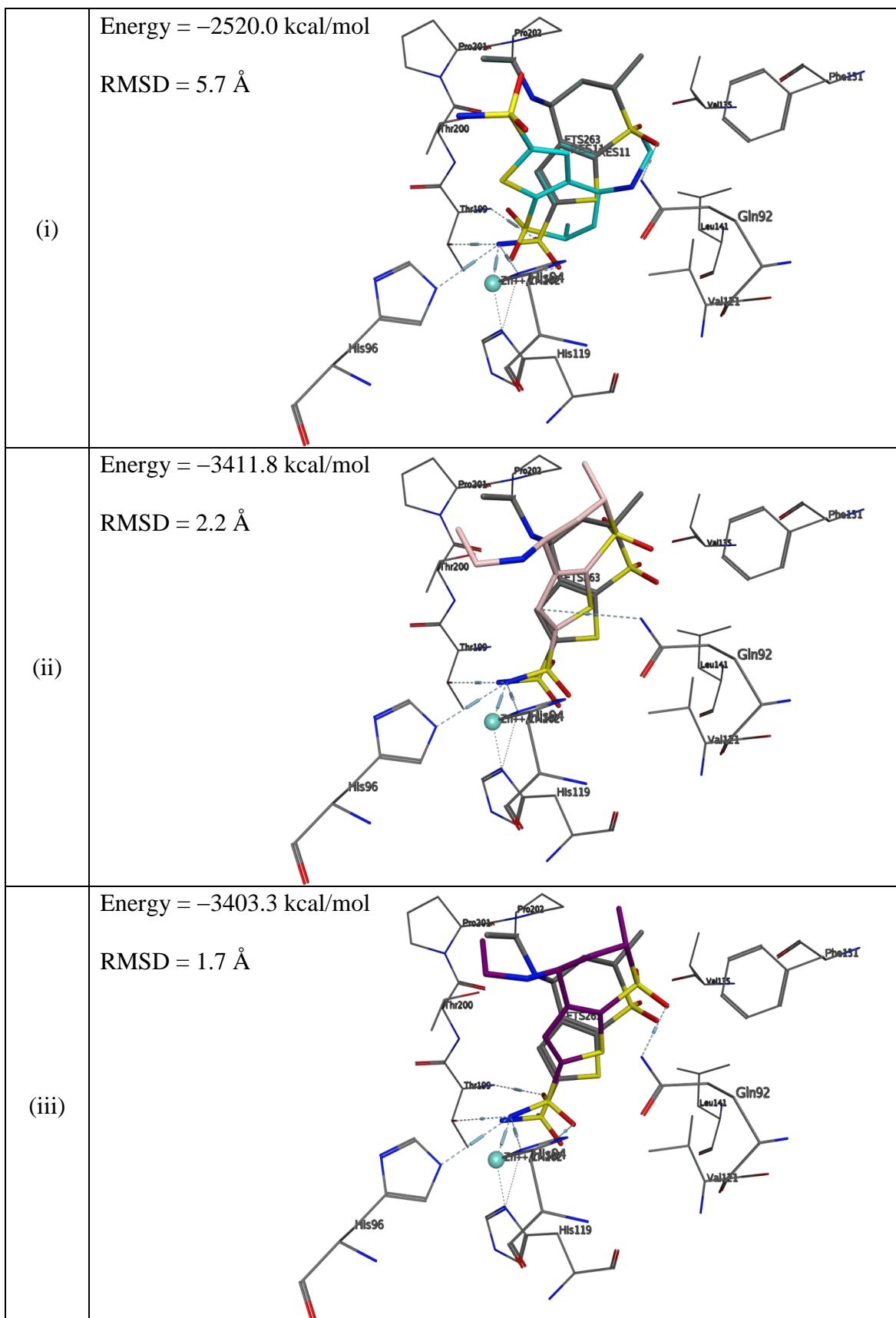


Figure 4.20 Superposition of (i) top-scored GOLD docked pose (cyan stick), (ii) lowest energy QM/MM MC docked pose (pink stick), and (iii) lowest RMSD QM/MM MC docked pose (purple stick) onto the native conformation (grey stick) of DZA (PDB ID 1CIL) at the binding site of hCA II.

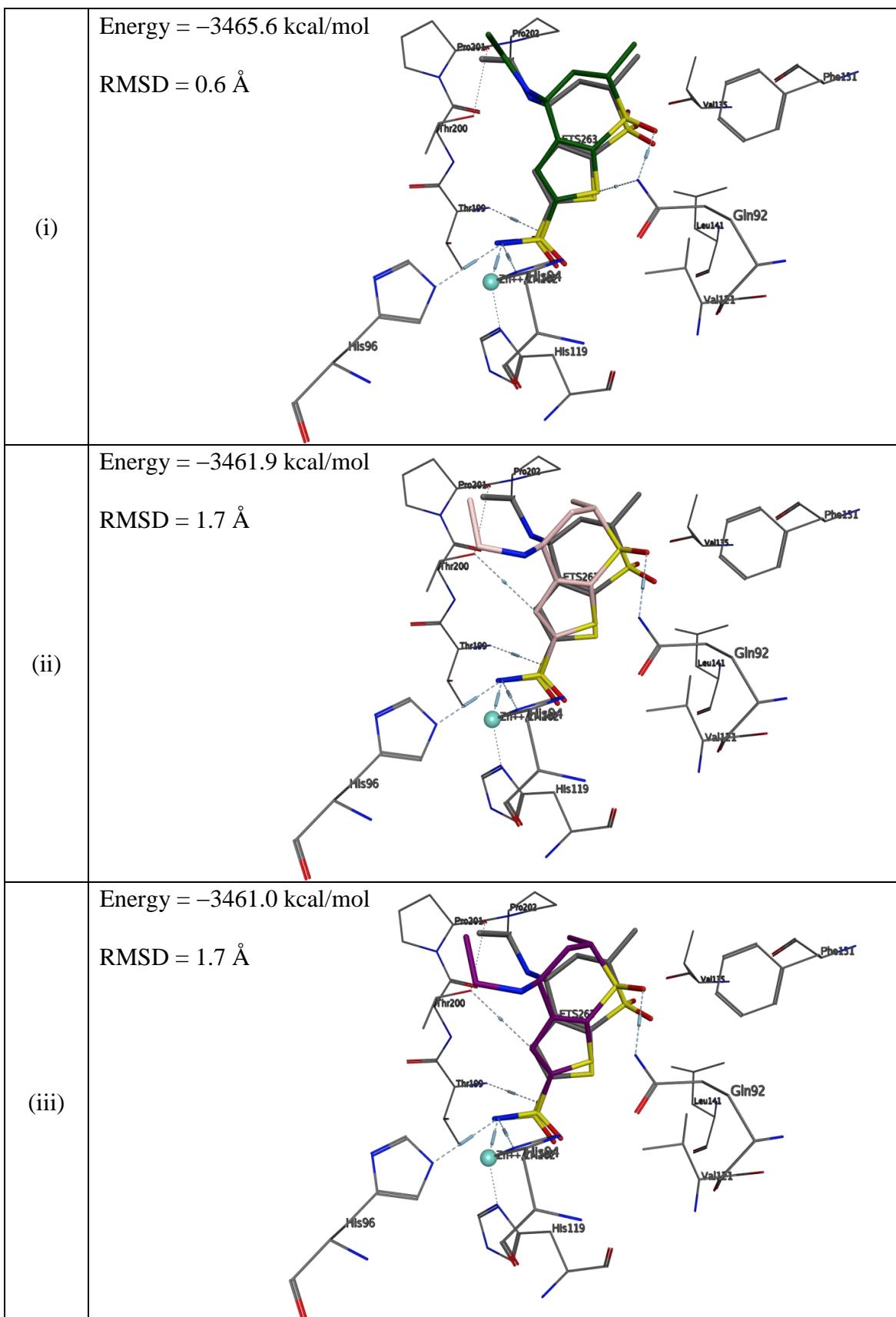


Figure 4.21 Superposition of (i) optimised native pose (green stick) (ii) optimised lowest energy docked pose (pink stick), and (iii) optimised lowest RMSD docked pose (purple stick) onto the native pose (grey stick) of DZA (PDB ID 1CIL) at the binding site of hCA II.

## 4.6 Conclusions and outlook

Docking tools have become widely used in structure-based drug design to predict the binding mode of a protein-ligand complex. Docking small ligands into the binding site of metalloproteins in particular has been challenging for these tools due to the complex electronic environment of the metal and its immediate surrounding. For modelling metalloenzyme-ligand interactions, combined quantum mechanical/molecular mechanical (QM/MM) methods seem suited and are computationally feasible. This hybrid method provides various sophisticated QM levels of theory to describe the active site containing the ligand, metal and metal-coordinating protein residues. The rest of the protein atoms (and potentially the solvent environment) are modelled by an empirical MM force field. Electronic polarisation of the ligand by the protein environment can be treated explicitly by the electronic embedding scheme. This scheme incorporates MM partial charges into the QM Hamiltonian. A previous docking study by our group employed a Monte Carlo (MC)-based QM/MM algorithm at the BH&H/6-31G\* level of theory on zinc-metalloenzymes cytidine deaminase (CDA) and human carbonic anhydrase II (hCA II).<sup>182</sup> The results were encouraging, showing the improvement of docking results over the conventional docking by GOLD. This work employs the semiempirical PM6 level of theory for the QM model, with the AMBER force field as the MM part, for QM/MM energy-based scoring and sampling by the Monte Carlo (MC) method. This QM/MM MC module is implemented for the first time in the Gaussian 09 program. To derive the optimal MC parameters, a sulfomoyl methanol was used as a probe in the binding site of hCA II. For this system, the MC acceptance rate was achieved at higher than 70% within an hour of the run time on four cores of 4GB

Intel SSE4.2 or AVX. The values of 5.0 bohrs for  $\Delta x_t$ , 3.0 bohrs for  $vdw_{cutoff}$  and 3.0 bohrs for  $d_{max}$  were applied to redock the ligand poses obtained from using GOLD suite.

Docking using GOLD suite was carried out for two inhibitors of CDA, 3,4 dihydrozebularine (DHZ) and 3-deazacytidine (DAC), as well as two inhibitors of hCA II, fluorinated sulfamoylbenzamide (FSB) and dorzolamide (DHZ). This docking protocol was based on a genetic algorithm for sampling ligand poses and the ChemPLP scoring function. Although the ChemPLP scoring function appears to be the best available in the GOLD suite, it was limited in its accuracy to discriminate here ligand poses that resembled the crystallographic binding modes. Of the four systems studied for CDA and hCA II receptors, GOLD docking generated erroneous predicted structures with the ChemPLP score in a range of 29 to 65. The top-scored poses of the four ligands were misdocked in their binding modes relative to the crystallographic structures, deviating by an RMSD of approximately 5 Å.

To evaluate the ability of QM/MM methods to improve the docking results above, a two-layer ONIOM scheme in Gaussian 09 was employed to rescore the top- and bottom-scored poses docked by GOLD. Improvement was achieved only for DAC inhibitor of CDA whereas there was no benefit for DHZ·CDA and DZA·hCA II complexes. For rescoreing FSB poses, there was not enough knowledge due to their insignificant difference in RMSD values. The unsuccessful rescoreing by using ONIOM energy could be from the fact that the energy profile is flat such that it is insensitive to the changes in RMSD values as reported in the literature.<sup>181</sup> The ONIOM rescoreing with minimisation scheme is recommended for future work to explore the possibility of rescoreing the false positive docking poses as in these cases.

Alternatively, the QM/MM MC docking scheme was employed to redock the four systems studied. Prediction of the binding modes was based on the QM/MM energy of the complex. We found that the lowest energy poses were not the lowest in RMSD value relative to the crystallographic conformation of the ligand. For all four test cases, the QM/MM MC docking scheme was able to improve the docking results obtained from GOLD suite. The binding modes of the ligands were found in the range of 1.7-3.6 Å in RMSD value compared with their crystallographic poses.

For DHZ·CDA complex, the key ligand-protein interactions in the crystal structure are between the six-membered ring of the ligand and Glu104, and also between the ribofuranose ring and Asn89 and Glu91. The top-scored pose docked by GOLD had the six-membered and ribofuranose rings swapped over in the active site. QM/MM MC docking still misplaced the two rings in the lowest-energy binding mode. The lowest energy docked pose gave a more favourable energy for the complex compared to the top-scored GOLD pose, by 37.8 kcal/mol, but the RMSD value of the ligand pose itself was not improved. In contrast, the lowest RMSD QM/MM MC docked pose placed the six-membered and ribofuranose rings the correct way round. Its RMSD value was improved by 3.6 Å over the GOLD docked pose and the lowest energy QM/MM MC pose. However, the QM/MM energy of the complex with this lowest-RMSD pose was 11.2 kcal/mol higher than that of the complex with the lowest energy pose. After energy optimisation, the lowest-energy, the lowest-RMSD and the crystallographic poses clearly converged to different energy minima.

For DAC·CDA complex, the interactions of ribofuranose with Asn89 and Glu91 are identical to those in DHZ·CDA complex as this ring is present in both inhibitors. The

interactions from the six-membered ring of DAC are different from those of DHZ due to the fact that it contains the carbonyl and amino groups rather than the amide group as presented in DHZ. The six-membered ring of DAC in the native conformation presents interactions to Phe71, Val73, Ala103 and Thr127. In contrast to the lowest energy pose of DHZ, the QM/MM MC docking scheme was able to improve the alignment of the two rings of DAC over the top-scored binding pose docked by GOLD. The lowest energy pose of DAC was 3.8 Å of the RMSD value with the energy lower by 700.3 kcal/mol. This lowest energy pose was not significantly different from the lowest RMSD pose (RMSD of 3.6 Å, 8.3 kcal/mol higher in total energy relative to the lowest energy pose). The optimised geometries of both the lowest energy and lowest RMSD poses were identical, with an RMSD of 3.9 Å and energy of -1089 kcal/mol. Therefore, both poses were converged to the same energy minima, but a different one from that of the crystal pose.

In the FSB·hCA II complex, there are interactions of the ligand sulfamoyl group with Zn<sup>2+</sup>, His94, His96, His119, Thr199 and Leu198. The phenyl ring of the benzenesulfonamide fragment forms CH-π interactions with Leu198, while the amino amide forms a hydrogen bond to the water molecule. The top-scored pose by GOLD misdocked the position of the amino nitrogen and sulfonyl oxygen atoms of the sulfamoyl group as well as the fluorinated phenyl ring. The RMSD value of this pose was 4.6 Å. QM/MM MC docking improved the alignment of the amino nitrogen atom in the case of the lowest energy docked pose. Its complex was more favourable in energy by 34.5 kcal/mol relative to the top-scored GOLD pose but not significantly improved in the RMSD value of the ligand. In the case of the lowest RMSD pose, the position of the sulfamoyl group was similar to that of the lowest energy pose, but its

fluorinated phenyl ring was aligned closer to that of the native pose, i.e. in the vicinity of Phe131 and Val135 with an upwards direction of the fluorine atoms. This pose was improved in RMSD of 1.7 Å but its complex energy was 8.5 kcal/mol higher than that of the lowest energy pose. After optimisation, the results indicated that the lowest energy, lowest RMSD and native poses converged into different energy minima.

The last system studied was DZA·hCA II complex. This clinically used sulphonamide inhibitor, DZA, exhibits interactions of its sulfamoyl group similar to those of FSB inhibitor, except with Leu198. GOLD substantially misdocked the ligand, placing the sulfonyl group on the 6-membered ring next to the Zn<sup>2+</sup>, rather than the sulfonyl of the sulfamoyl group. This misalignment was corrected by the QM/MM MC docking scheme such that the lowest energy docked pose had an RMSD of 2.2 Å, improved over the top-scored GOLD docked pose of 5.7 Å. The QM/MM energy of the complex with the lowest energy pose was improved by 891.8 kcal/mol. However, the alignment of ethylamine side chain of the ligand was still not exact, due most likely to stabilisation by interaction of this group with Thr200. This was not resolved in the case of the lowest RMSD pose either (RMSD of 1.7 Å). Both the lowest energy and lowest RMSD docked poses converged to the same energy minima on optimisation. Their complex energies were different by only 0.9 kcal/mol. However, they were not in the same energy minima as the native pose.

In brief, the QM/MM MC was able to improve the binding modes previously predicted by GOLD suite for all four test cases; however, in addition to considering larger numbers of protein-ligand systems to investigate, there are a number of other issues that need to be considered to further improve the accuracy of QM/MM docking results. One

such issue is the inclusion of protein residues that interact with the ligand. Although the sidechains of the zinc-coordinating residues were included in the QM region, the predicted binding modes by the QM/MM MC scheme missed some key interactions at the active sites.

Another concern is the contribution of water-mediated interactions to the QM/MM energy of the complex (this is a general problem with any docking method). Also, entropic factors should be considered for more accurately scoring. Furthermore, it can be seen from the four systems studied that not only electrostatic but also dispersion plays an important role in the active site. Cho *et al.*<sup>183</sup> demonstrated that the inclusion of  $\pi-\pi$  interaction energies between ligand and protein could enhance the ability to predict ligand activity and docking ranks for B-Raf protein kinase. Such interactions are obviously weak individually but can add up to make a major impact. Therefore, the use of dispersion-corrected semiempirical methods for describing the QM region is advisable in future research work.

Finally, one could consider the use of a MC with minimisation scheme for sampling. After each random change, conventional iterative minimisation should be performed to avoid sampling high energy configurations. The refined geometry is then required to satisfy the Metropolis criterion. This will take advantage of Metropolis Monte Carlo for approximate searching for a global energy minimum, and of an optimised algorithm for locating the more refined minimum. For example, the prediction of poses for the DHZ ligand was improved by energy optimisation, although the QM/MM surface was obviously quite rough with multiple low energy minima. Of course, longer MC runs would also be desirable, such that the lowest energy minimum would be found multiple

times. This will become increasingly possible with more powerful computing architectures.

# **CHAPTER 5 Assessment of quantum chemical methods for modelling $\beta$ -cyclohexatin conformation and its interaction with a single-layer graphene sheet: modelling the vicinal diol interaction in carbohydrates**

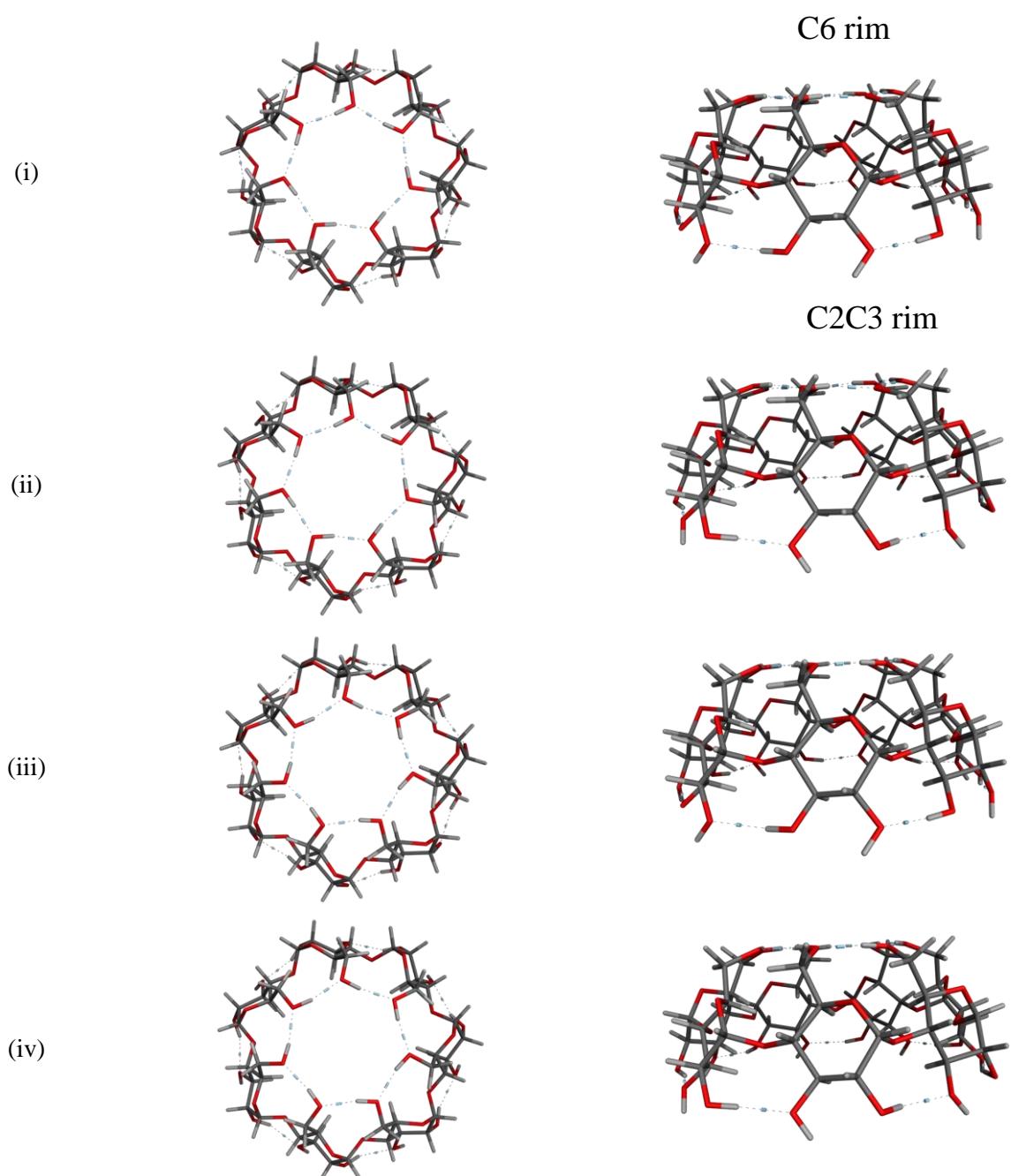
## **5.1 Introduction**

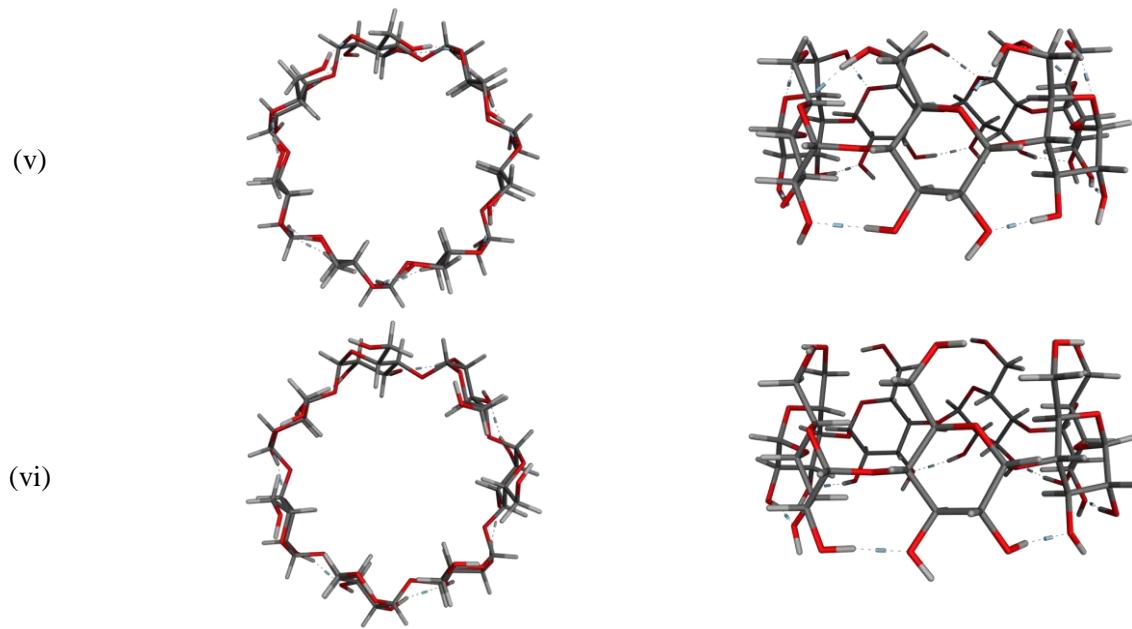
Graphene has attracted great interest as a novel material due to its exceptional physical and electronic properties. Graphene is a carbon allotrope, composed of a flat single layer of carbon atoms tightly packed into a two-dimensional hexagonal pattern. Each carbon atom has hybridised  $sp^2$  bonding orbitals pointing outwards in a molecular plane and connecting to adjacent carbon atoms via a  $\sigma$ -bond of 1.42 Å in length. The fourth valence electron in the outer shell occupies a  $\pi$ -orbital projecting above and below the basal plane. These  $\pi$  orbitals allow electrons to easily hop in the plane of the graphene surface. These highly-mobile  $\pi$  electrons are responsible for remarkable electrical and thermal conductivity in one-atom thick graphene. Graphene is not only thin, but is the stiffest and strongest material ever measured (tensile strength of ~130 GPa and Young's modulus ~1 TPa), very light (surface density 0.77 mg/m<sup>2</sup>), and optically transparent (up to 97.7%) with a large theoretical surface area (2630 m<sup>2</sup>/g).<sup>184</sup> Due to its extraordinary properties, graphene has been extensively studied as a versatile material for various applications, e.g. biomedical,<sup>185</sup> environmental,<sup>186</sup> electrodes in electrical and optical devices,<sup>187</sup> energy storage,<sup>188</sup> sensors<sup>189</sup> and composite materials.<sup>190</sup>

Drug and gene delivery in biomedical applications have the potential to exploit the ultrahigh surface area of graphene as a high-capacity drug carrier. Because graphene is hydrophobic in nature, on the one hand, it can be used for loading anticancer drugs such as doxorubicin, camptothecin and chlorin e6, via  $\pi$ - $\pi$  stacking and hydrophobic interactions.<sup>191</sup> On the other hand, graphene-based materials are subject to poor dispersibility in water and commonly used polar solvents, due to its hydrophobicity. A modification of the hydrophobic graphene surface to be more hydrophilic by either covalent or noncovalent functionalisation is essential for improving the solubility of the material.<sup>192</sup> Unlike covalent attachment of organic functional groups, noncovalent functionalisation does not disrupt the  $\pi$ -system of graphene. Therefore, the intrinsic properties of graphene will be preserved with the enhanced aqueous solubility. The biocompatibility of graphene-based materials can be tuned by incorporating carbohydrates as a noncovalent adsorbate.<sup>193</sup> For these graphene-polysaccharide nanocomposites, graphene also has an ability to increase the tensile strength of some polysaccharide adsorbates from 20% up to 134%.<sup>194</sup> Utilising specific adsorbates such as carbohydrates to modify the polarity of the graphene surface could enhance the aqueous dispersibility and biocompatibility in graphene-based materials, of great benefit to its application in biomedicine.

$\beta$ -Cyclodextrin ( $\beta$ CD), a cyclic oligosaccharide of seven  $\alpha$ -D-glucose units, has been used as an excipient in various drug formulations.<sup>195</sup> It has a hydrophobic interior of methine groups available for forming host-guest complexes with drugs. It also has two hydrophilic rims of hydroxyl groups located outside its molecular cavity, the C2C3 and C6 rims (Figure 5.1(i)), sometimes called the cyclodextrin's head and tail;<sup>196</sup> these polar rims can facilitate water solubility of its inclusion complex.  $\beta$ CD can form a closed or

open structure: when  $\beta$ CD adopts a closed structure, its shape resembles a truncated cone, as the primary 6-hydroxyl groups form strong interactions with each other across glucose units, effectively closing access (the “gate”) at the narrower C6 rim<sup>197</sup> (Figure 5.1(i)-(iv)). In contrast, when  $\beta$ CD is in an open structure, it presents a hydrophobic cavity in a more cylindrical shape suitable for inclusion of small or medium-sized guest molecules (Figure 5.1(v)-(vi)).





**Figure 5.1**  $\beta$ CD geometries optimised at M06-2X/def2-TZVPP (i) cccw (ii) cccc (iii) cwew (iv) cwcc (v) O-cwew (vi) O-cccc.

The energetics of open and closed cyclodextrin conformers have been studied *in vacuo* via density functional theory (DFT), for  $\beta$ CD.<sup>197-198</sup> These studies have found that a closed conformation is lowest in energy. In this closed structure, a counter-clockwise (cc) orientation of intramolecular hydrogen bonds along the C6 rim and clockwise (cw) along the C2C3 rim is preferred (Figure 5.1(i)), defined when viewing  $\beta$ CD from above the smaller C6 rim. This cccw conformer,<sup>198</sup> alternatively denoted (+),<sup>199</sup> is lower in energy than other low energy closed and open structures (Figure 5.1(ii)-(vi)).<sup>197-198</sup> Although defined relative to facing the C6 rim, these networks can more clearly be viewed facing the C2C3 rim (Figure D1).

Noncovalent functionalisation of graphene with  $\beta$ CD leads to a novel material that forms stable dispersions in water.<sup>196</sup> This system has potential to act as a biosensor in biomedical applications via host-guest interactions, where the guest molecule forms an

inclusion complex with  $\beta$ CD present on the graphene surface.<sup>200</sup> Interestingly, docking of  $\alpha$ CD to graphene oxide surface found a similar preferred orientation of rim rather than edge of CD, although of course the polar graphene oxide surface presents a more polar face than that of graphene.<sup>201</sup> However, some work based on transmission electron microscopy and MM studies suggest that  $\beta$ CD absorbs on graphene in a similar rim-down fashion, such that the complete C2C3 rim contacts the graphene surface.<sup>196</sup> The hydrophobic unmodified graphene can be transformed to be a more hydrophilic surface by the noncovalent adsorption of  $\beta$ CD as confirmed by experimental analysis of static contact angle.<sup>202</sup> The hydrophilicity of the complex can highly improve the aqueous dispersibility. Physisorption of  $\beta$ CD at the interface is due to dispersion interactions, specifically, XH- $\pi$  (X=O, C) interactions. The energies of such interactions are typically weak, approximately 3 kcal/mol in magnitude.<sup>203</sup> A detailed quantitative computational analysis of the interaction energy and geometry of the  $\beta$ CD/graphene complex remains to be performed. However computational modelling of the weakly dispersive interactions is particularly a challenge, as we discuss below.

DFT methods have become a powerful method of choice to model chemical systems during the past two decades. Despite their broad success, earlier DFT methods lack an ability to accurately predict dispersive interactions. Subsequently, a great deal of development in DFT functionals has been implemented for solving these problems. A long-range corrected hybrid density functional,  $\omega$ B97X-D,<sup>204</sup> which includes an empirical atom-atom dispersion correction based on Grimme's method,<sup>205</sup> was developed and compared with other dispersion-corrected functionals such as B3LYP (B3LYP-D) and BLYP (BLYP-D).<sup>204</sup> It gave the best performance out of seven functionals for noncovalent interactions, for example with a mean absolute error for the

S22 set of 0.22 kcal/mol and mean signed error of -0.08 kcal/mol.<sup>204</sup> Recently, it has been successfully applied with the def2-TZVPP basis set to model the structures, electronic properties and interaction energies of aromatic hydrocarbon/graphene dimers.<sup>206</sup>

Another recent DFT functional for noncovalent systems is M06-2X,<sup>7a</sup> a hybrid meta-generalised gradient approximation functional from the Minnesota family. Minnesota functionals are a group of approximated exchange-correlation DFT functionals which have been parametrised based on diverse benchmark databases including noncovalent complexes. Among the seventeen functionals applied for the S22 database (not part of the training set), M06-2X with the MG3S basis set gave the lowest balanced mean unsigned error (BMMUE) of 0.47 kcal/mol.<sup>7a</sup> The BMMUE is a MUE corrected for BSSE and averaged evenly over hydrogen bonding, dispersion dominant and mixed complexes of the S22 set. The MG3 basis set is the same as 6-311++G(3d2f,2df,2p) for H–Si systems, but improved for P–Ar systems. The MG3S basis set is identical to the MG3 basis set except that in the former, diffuse functions on hydrogen are absent. The M06-2X/TZV2D level of theory, in its ability to compute interaction energies for sugar-aromatic complexes, outperforms MP2 and some selected dispersion-corrected DFT (DFT-D) functionals.<sup>182</sup> It has also been reported to be among the best functionals for accurately modelling the conformational energies of 58  $\alpha$ -pyranose aldohexose isomers in benchmarking against a coupled cluster approach extrapolated to the complete basis set limit, CCSD(T)/CBS, with M06-2X/def2-TZVPP yielding an RMS error of 0.31 kcal/mol.<sup>157</sup>

With a comparable performance to DFT-D and three orders of magnitude faster,<sup>207</sup> semiempirical quantum mechanics (SQM) methods, with an addition of dispersion and hydrogen-bond correction terms, have been substantially developed to describe noncovalent systems. These developments and benchmarking have been reviewed and discussed for various applications.<sup>207</sup> The PM3-D\* Hamiltonian<sup>208</sup> has been specifically parameterised for carbohydrate-aromatic interactions and PM3CARB-1<sup>209</sup> was fitted to reproduce the conformation of carbohydrates. Recently, the corrected SQM methods in a neglect of diatomic differential overlap (NDDO) formalism of the PMx family have been parameterised using a broad collection of datasets containing geometries and interaction energies. Among these methods benchmarked against CCSD(T)/CBS interaction energies, PM7 is the latest release, fixing known issues in its predecessor PM6.<sup>23</sup> It has been employed not only for biological molecules, but also for solid state systems. Although PM7 is a new release, its performance has not been remarkably better than of modified PM6 methods, i.e. PM6-DH+, PM-DH2 and PM6-D3H4.<sup>7d</sup> It has been reported that the PM7 and PM6-DH+ Hamiltonians overestimate the interaction energy of bimolecular complexes in benchmark data sets while PM6-DH2 and PM6-D3H4 provide the most accurate result.<sup>7d</sup> Both PM6-DH+ and PM6-DH2 have been reported to agree well with experimental values for modelling the adsorption energies of small- to medium-sized organic molecules on graphene.<sup>210,211</sup>

The self-consistent charge density functional tight-binding (SCC-DFTB)<sup>212</sup> method is comparable in computational efficiency to the widely used NDDO-based methods. It can accurately model conformational energetics and intermolecular energies for some organic systems.<sup>40</sup> The mean absolute error of the computed heats of formation for sets of neutral organic molecules is in between those of AM1 and PM3.<sup>40</sup> The SCC-DFTB

method is derived from a second-order expansion of the DFT total energy with respect to charge density fluctuations around a reference charge density.<sup>35</sup> Due to its good performance for hydrogen-bonded complexes, SCC-DFTB has been further improved, e.g. in the effective electron repulsion term in the self-consistent charge formalism; in the Coulomb interaction term between atomic partial charges; and in additional expansion of the third-order term of the DFT total energy.<sup>44-45</sup> This latest modification, named the DFTB3 method, has recently been released in combination with a newly developed Slater-Koster parameter set, namely 3OB, for organic molecules.<sup>7b</sup> It has been claimed that DFTB3 has removed some significant errors in overbinding and can improve the description of molecular geometries that depend on nonbonded interactions.<sup>34</sup>

## 5.2 Aims

Surface functionalised  $\beta$ CD/graphene composite is a promising material for applications in drug delivery. It can exploit the ultrahigh surface area of graphene and the hydrophilicity of  $\beta$ CD as well as the hydrophobicity of both, for loading and delivery drugs to the target. Recently, noncovalent modification of graphene with  $\beta$ CD has been shown to improve graphene dispersion in water, and remains stable for several weeks.<sup>213</sup> It is difficult to experimentally detect and quantify such weak interactions. This highlights the importance of computational techniques as the useful methods to study the  $\beta$ CD/graphene interactions. In the past, modelling a large system of thousands of atoms at the QM level with the accurate estimation of weak dispersion had been also in limit. To the best of our knowledge, neither biophysical nor quantum chemical

methods have been carried out yet to capture the intermolecular interactions between  $\beta$ CD and graphene. Therefore, the aim of this chapter is to model, for the first time, the noncovalent  $\beta$ CD/graphene complexes and their QM interaction energetics *in vacuo* and in aqueous solvent. First, we assess the ability of SQM approaches to model  $\beta$ CD structure and energetics, considering open and closed forms, and intramolecular hydrogen bonding networks. We then apply a suitable subset of SQM methods to identifying the preferred binding modes of  $\beta$ CD on single-layer graphene flakes of increasing size. The computed potential energy surface of  $\beta$ CD on graphene will provide insight into the origin of stability of this promising material, and may offer tools that can be used to guide the design and optimisation of the dispersibility and biocompatibility of these systems.

### 5.3 Computational methods

Firstly, we model the preference of hydrogen bond network in  $\beta$ CD conformations *in vacuo*. This was performed by using the dispersion-corrected hybrid density functionals,  $\omega$ B97X-D<sup>204</sup> and M06-2X<sup>7a</sup>, with the split-valence triple-zeta basis set, def2-TZVPP. The relative energetics of these six conformers were compared with a previous study using the B3LYP functional with 6-31G\* and 6-31G\*\* basis functions.<sup>197-198</sup> In addition to DFT methods, semiempirical NDDO-based and DFTB methods were also used for comparison. Secondly, we applied selected QM methods to predict the preference of hydrogen bond orientation at the hydrophilic  $\beta$ CD rim on binding to a C<sub>96</sub>H<sub>24</sub> graphene surface. Finally, three SQM methods were applied to explore possible binding modes between  $\beta$ CD and the C<sub>1006</sub>H<sub>88</sub> graphene sheet.

### 5.3.1 $\beta$ CD geometries and conformational energetics *in vacuo*

Four closed and two open low-energy conformations of  $\beta$ CD, i.e. cccw, cccc, cwew, cwcc, O-cwew and O-cccc, were taken from the previous work of Snor *et al.*<sup>197</sup> and Stachowicz *et al.*<sup>198</sup> These six conformers were built and fully optimised *in vacuo* at different QM levels of theory. We have used the density functionals  $\omega$ B97X-D, M06-2X and B3LYP for comparison, along with the def2-TZVPP basis set. The M06-2X/def2-TZVPP optimised geometries of all six  $\beta$ CD structures were reoptimised at semiempirical QM levels of theory. The DFTB parameter sets of mio-1-1 and 3ob-2-1 were used for SCC-DFTB and DFTB3 methods, respectively. For DFTB3, the third-order self-consistent charge (SCC) model was used, with a tolerance at  $1 \times 10^{-8}$  and the Slater-Kirkwood type dispersion model with the Hubbard derivatives in atomic units:<sup>7b</sup> –0.1575 for oxygen atoms, –0.1857 for hydrogen atoms and –0.1492 for carbon atoms were applied to all atoms with their corresponding atomic covalent radii<sup>44</sup>. Relative energetics of the  $\beta$ CD structures at each level of theory were then compared to that of the cccw conformer.

### 5.3.2 Intermolecular interactions at $\beta$ CD/graphene interface

Interaction energies of  $\beta$ CD/graphene complex were computed using two different graphene models: C<sub>96</sub>H<sub>24</sub> and C<sub>1006</sub>H<sub>88</sub>. The smaller complex model, of 267 atoms, was used to explore at DFT levels of theory the preferred hydrogen bond (cccw or cccc) orientation of the  $\beta$ CD rim at the interface. These models were optimised at the M06-2X/def2-SVP level of theory and their energies were computed at the M06-2X/def2-

TZVPP level (6679 basis functions) with an ultrafine grid (99, 590) for numerical integration of two-electron integrals and counterpoise correction. These results were used as a benchmark for semiempirical QM methods i.e. PM6-DH2, PM7 and DFTB3. For the large model, of size 1241 atoms, the six  $\beta$ CD geometries were considered for noncovalent binding to the large graphene sheet ( $C_{1006}H_{88}$ ) in two binding modes, i.e. the C2C3 and C6 rims of  $\beta$ CD to the interface. Each complex model was optimised at the PM6-DH2, PM7 and DFTB3 levels. Optimisation via the DFTB3 method was performed with the steepest descents optimisation algorithm using a maximum force component of  $5.0 \times 10^{-4}$ . The Broyden charge mixing method was also used with the third-order SCC tolerance of  $1.0 \times 10^{-7}$ . The Slater-Kirkwood model with the Hubbard derivatives (please see more detail in the section of 5.3.1 above) was applied for the dispersion correction. For constructing complexes in both model sizes, initial  $\beta$ CD conformers were taken as the M06-2X/def2-TZVPP optimised *in vacuo* geometries. The intermolecular interaction energies of all optimised complex geometries were calculated using the monomer geometries at their bound complex geometry.

*Ab initio* and DFT calculations used the Gaussian 09<sup>175</sup> quantum chemistry package. PM3-D\* calculations used the Gaussian 03<sup>214</sup> quantum chemistry package. PM3CARB-1 calculations used the SQM module of AMBERTools13 in AMBER12.<sup>215</sup> Other NDDO-based SQM methods used the MOZYME module of MOPAC2012<sup>216</sup> version 14.083L. SCC-DFTB calculations with the mio-1-1 DFTB parameter set were performed using the SQM module of AMBERTools13. Finally, calculations using DFTB3 with the 3ob-2-1 DFTB parameter set employed DFTB+<sup>217</sup> version 1.2.2.

## 5.4 Results and Discussion

### 5.4.1 Gas-phase $\beta$ CD conformation

We first examine the stability of six  $\beta$ CD hydrogen bond conformers in the gas phase, two open and four closed, following the studies of Snor *et al.*<sup>197</sup> and Stachowicz *et al.*<sup>198</sup>. Previous work at B3LYP/6-31G\* and B3LYP/6-31G\*\* levels of theory<sup>197-198</sup> found the most stable  $\beta$ CD structures to be cccw and cccc (Figure 5.1(ii)), differing by only 0.1 kcal/mol in stability. The cwcc and cwcw conformers were similar in energy, and 2.6 - 2.8 kcal/mol less stable than cccw. Finally, the open structures were 18.0 – 20.2 kcal/mol higher in energy than cccw, reflecting (i) for O-cwcw, weaker alcohol-ether hydrogen bonding at the C6 rim, compared to alcohol-alcohol in the closed structures; and (ii) for O-cccc, no hydrogen bonds formed at the C6 rim.<sup>197-198</sup>

Next, we consider the *in vacuo* optimal structures and energetics of the four closed and two open  $\beta$ CD conformers from density functional calculations. The optimised DFT structures agree well with the internal coordinates determined from X-ray crystallography (for example for M06-2X/def2-TZVPP structures, Table D1), although we note the X-ray structure contains disorder in the hydroxymethyl orientations (torsion angle  $\Omega$ , Table D1). At the M06-2X/def2-TZVPP level of theory, the most stable  $\beta$ CD structure was found to be cccw, in agreement with the earlier B3LYP studies (Table 5.1). Indeed the cccw conformer has the shortest hydrogen bond distances of the four closed structures at both the C2C3 and C6 rims, with an average O3-H···O2' and O6···H-O6' distances over the seven glucosyl rings of 2.11 and 1.83 Å respectively (Tables D2 and D3). We note this O6···H-O6' hydrogen bond distance is very short,

close to that found in water dimer<sup>218</sup> and other small clusters of water molecules.<sup>219</sup> However, distinct from previous work, the cccw conformer is found to be 1 kcal/mol more stable than both cccc and cwcc structures (Table 5.1). Thus, relative to B3LYP/6-31G\* and B3LYP/6-31G\*\* calculations, the M06-2X/def2-TZVPP model stabilises cwcc and destabilizes cccc by about 1 kcal/mol in each case. This indicates that flipping the hydrogen bond network at *either* rim has energetic consequences. The remaining closed conformer, cwcc, lies 2.6 kcal/mol above cccw, approximately in line with earlier lower level predictions. The two open conformers, O-cwcc and O-cccc, are 17.1 and 17.4 kcal/mol above cccw respectively, similar to the corresponding 18.0 and 20.2 kcal/mol predicted by B3LYP/6-31G\* (Table 5.1).

Interestingly, the relative energetics of the six  $\beta$ CD conformers is reproduced to that of the best level herein, M06-2X/def2-TZVPP, within 0.7 kcal/mol at the M06-2X/def2-TZVPP//M06-2X/def2-SVP level of theory. However, at the M06-2X/def2-SVP//M06-2X/def2-SVP level, i.e. with the same functional but smaller basis set, deviations range from 1.2 kcal/mol for cccc to 8.3 kcal/mol for O-cccc. This indicates the importance of the larger more polarized basis set in accurately capturing the relative stability of the various cooperative hydrogen bond networks. Using the def2-TZVPP basis with the  $\omega$ B97X-D functional gives a deviation from the M06-2X/def2-TZVPP values of 0.2 kcal/mol for closed conformers and 5.4 kcal/mol for open structures (Table 5.1). These results together provide a consensus that cccw is the most favoured of these  $\beta$ CD structures, that it is distinct in stability from other closed conformers, and that the open conformers are considerably less stable than closed *in vacuo*.

**Table 5.1 Relative gas-phase conformational energetics in kcal/mol of  $\beta$ CD.**

Method	cccw (-+)	cccc (--)	cwcw (++)	cwcc (+-)	O-cwew (++)	O-cccc <sup>c</sup> (--)	MSE <sup>d</sup>	MUE <sup>e</sup>	RMSE <sup>f</sup>
M06-2X/def2-TZVPP	0.0	1.2	0.9	2.6	17.1	17.4			-
M06-2X/def2-TZVPP//									
M06-2X/def2-SVP	0.0	1.1	0.9	2.5	17.8	18.1	0.2	0.3	0.4
M06-2X/def2-SVP	0.0	2.3	2.7	5.5	19.9	26.4	2.9	2.9	4.1
$\omega$ B97X-D/def2-TZVPP	0.0	1.0	1.2	2.3	19.5	22.8	1.3	1.4	2.4
B3LYP/6-31G**	0.0	0.1	2.2	2.4	18.3	19.6	0.8	1.2	1.5
B3LYP/6-31G** <sup>a</sup>	0.0	0.1	2.2	2.1	n/a	n/a	-0.1	0.8	1.0
B3LYP/6-31G*	0.0	0.1	2.7	2.9	18.0	20.3	0.5	1.0	1.3
B3LYP/6-31G* <sup>b</sup>	0.0	0.1	2.8	2.6	18.0	20.2	0.7	1.1	1.5
DFTB3	0.0	6.2	4.7	10.9	15.4	26.9	4.1	4.7	5.8
SCC-DFTB	0.0	5.0	3.3	8.4	14.7	23.2	2.6	3.4	3.9
PM7	0.0	-8.3	1.7	-6.7	19.9	-3.2	-6.0	7.2	10.1
PM6-DH2	0.0	-12.8	1.5	-11.3	34.4	-14.9	-7.1	13.0	17.0
PM6-D3	0.0	-11.3	4.7	-7.0	29.4	-11.6	-5.8	11.2	14.4
PM6	0.0	-12.5	4.7	-6.4	29.0	-12.2	-6.1	11.3	14.7
PM3CARB-1	0.0	3.7	2.7	6.5	25.0	0.1	-0.2	5.6	8.0
PM3-D*	0.0	-5.7	6.0	-1.1	-0.9	-12.8	-9.0	10.7	14.9
PM3	0.0	-5.6	3.4	-1.3	-0.4	-10.2	-8.9	9.7	13.8
AM1	0.0	-5.9	7.6	0.1	6.1	0.7	-5.1	7.3	9.1

<sup>a</sup>Snor *et al.*<sup>197</sup> <sup>b</sup>Stachowicz *et al.*<sup>198</sup> <sup>c</sup>Collapsed open structures labelled in *italics*<sup>d</sup>Mean signed error <sup>e</sup>mean unsigned error <sup>f</sup>root mean square error with respect to M06-2X/def2-TZVPP level of theory

We now consider the ability of semiempirical QM (SQM) levels of theory to reproduce the density functional profile of  $\beta$ CD energetics. Interestingly, seven of the eight NDDO-based SQM methods employed in this study predict the cccc conformer to be the lowest in energy (Table 5.1), not the cccw conformer, by up to 13 kcal/mol (for PM6-DH2) relative to cccw. This reflects the finding of Kohler *et al.* in their AM1 analysis of  $\beta$ CD, which found a cccc conformer to be the lowest in energy *in vacuo*.<sup>220</sup> Indeed, here we find six of the eight SQM methods predict the second most stable as cwcc, and then cccw as third. Similarly, the stability of the open structures is overestimated by these methods. Indeed, with the exception of the PM3 and PM3-D\* methods, the open cavity of the O-cccc is not preserved. We find AM1 gives an intermediate O-cccc structure, with a partial network of intramolecular alcohol-alcohol hydrogen bonds at the C6 rim (for more information, see supporting information Table D3). The PM3CARB-1 method correctly predicts cccw as the lowest in energy, but only approximately captures the relative stabilities of the remaining closed and O-cwcc structures. Nevertheless, for the remaining  $\beta$ CD geometries of seven SQM methods, the RMSD to the respective M06-2X/def2-TZVPP structure is < 0.2 Å for PM6, PM6-DH2, PM6-D3 and PM7; and < 0.3 Å for the other NDDO approaches (Table D4). Interestingly, SCC-DFTB and DFTB3 methods find cccw as lowest in energy and maintain more stable open  $\beta$ CD conformers in the gas phase, although the RMSD is still high relative to the X-ray structure (0.6 Å for DFTB3, Table D4). However, the detailed preference of closed structures is not accurately reproduced relative to the M06-2X/def2-TZVPP and  $\omega$ B97X-D/def2-TZVPP levels (Table 5.1).

Only the PMCARB1, SCC-DFTB and DFTB3 methods have MUEs of less than 7 kcal/mol over the six conformers (Table 5.1). In particular it appears using the NDDO

methods that there is a switch in preference of the O<sub>3</sub>⋯O<sub>2'</sub> interaction from a cw hydrogen bond direction, i.e. O<sub>3</sub>-H⋯O<sub>2'</sub>, to a cc O<sub>3</sub>⋯H'-O<sub>2'</sub> direction (e.g. from cccw to cccc, and cwcw to cwcc). This does not appear to manifest itself in terms of altered hydrogen bond distances when comparing cw and cc conformers, at either the C2C3 (Table D2) or C6 rims (Table D3). We do note, however, a general observation of very short hydrogen bond lengths in closed PM6-DH2, PM6-D3 and PM3CARB-1 structures relative to M06-2X/def2-TZVPP geometries, by 0.1-0.2 Å (Tables D2, D3).

The rim hydrogen bond network involves the interaction between residues and the intra-residue polar interactions. For example, for cccw, this equates to a [⋯O<sub>2</sub>-H⋯O<sub>3</sub>-H⋯O<sub>2'</sub>-H'⋯] network. We can isolate the intramolecular O<sub>2</sub>-H⋯O<sub>3</sub>-H interaction within a glucose residue by considering the relative energies, E<sub>cc-cw</sub>, of cw and cc conformers of glucose monosaccharide (Table 5.2). We consider the O<sub>2</sub>/O<sub>3</sub> cc and cw conformers of glucose, with in both cases a g<sup>-</sup>g conformation of its hydroxymethyl group. At the MP2/def2-TZVPP level, the cc conformer is in fact lower in energy than the cw conformer: relative to the minimum energy glucose conformer cc/tg identified from previous work<sup>209</sup>, the cw/g<sup>-</sup>g conformer is at 0.41 kcal/mol whereas that of cc/g<sup>-</sup>g at the same level of theory is at -0.27 kcal/mol; thus, the E<sub>cc-cw</sub> value is -0.68 kcal/mol (Table 5.2). All SQM methods reasonably reproduce the internal geometries of the two glucose conformers (Table D5). However, it is evident that the NDDO methods once again overpredict the stability of the cc conformer relative to cw, with E<sub>cc-cw</sub> values ranging up to -2.8 kcal/mol for PM6-D3 and PM7 (Table 5.2). As before, DFTB3, SCC-DFTB and PM6-DH2 methods give estimates closer to the MP2/def2-TZVPP value, with E<sub>cc-cw</sub> of -1.3, -1.1 and -1.4 kcal/mol respectively.

**Table 5.2 Relative gas-phase conformational energetics in kcal/mol of glucose.**

Method	<sup>a</sup> cw/g <sup>-</sup> g	<sup>a</sup> cc/g <sup>-</sup> g	E <sub>cc-cw</sub>
MP2/def2-TZVPP	0.41	-0.27	-0.68
MP2/aug-cc-pVTZ	0.35	-0.29	-0.64
M06-2X/def2-TZVPP	0.41	-0.13	-0.54
ωB97X-D/def2-TZVPP	0.49	-0.07	-0.56
<sup>b</sup> B3LYP/6-311+G**//HF/6-31G*	0.91	0.00	-0.91
DFTB3	3.04	1.72	-1.32
SCC-DFTB	2.33	1.21	-1.12
PM7	4.10	1.29	-2.81
PM6-DH2	3.39	2.01	-1.38
PM6-D3	4.32	1.55	-2.77
PM6	3.65	0.95	-2.70
PM3CARB-1	1.29	-0.34	-1.63
PM3-D*	0.40	0.34	-0.06
PM3	0.48	0.34	-0.14
AM1	3.03	0.54	-2.49

<sup>a</sup>Built from the definition of McNamara *et al.*<sup>209</sup>    <sup>b</sup>McNamara *et al.*<sup>209</sup>

To investigate further this large discrepancy in prediction, the 1,2-ethanediol system was employed. This simple molecule models the vicinal OH interaction of the sugar ring. In these studies, the three dihedral angles, HOCC', OCC'O' and HO'C'C, label the ten possible conformers, according to *trans* (t), *gauche* (g) and *gauche<sup>-</sup>* (g<sup>-</sup>) values. Previous QM studies show that the three lowest energy conformations of 1,2-ethanediol

are (in order of decreasing stability) tGg<sup>-</sup>, gGg<sup>-</sup> and g<sup>-</sup>Gg<sup>-</sup>.<sup>221</sup> In this work by Guvench and Mackerell,<sup>221</sup> the highest level of theory applied, an approximation to CCSD(T)/cc-pVQZ calculations, estimated gGg<sup>-</sup> and g<sup>-</sup>Gg<sup>-</sup> structures to lie 0.35 and 0.92 kcal/mol above tGg<sup>-</sup> in energy respectively.

Of these low energy conformers, tGg<sup>-</sup> is equivalent to the cw intraresidue hydrogen bond in  $\beta$ CD, and gGg<sup>-</sup> models the cc intraresidue hydrogen bond (Figure 5.2(i, ii)). We subsequently refer to these conformers as cw and cc respectively. Both conformers share a *gauche* OCC'O' angle, allowing the vicinal hydroxyl groups in close proximity enough to interact more favourably than in other conformers. Previous work has shown that the geometrical parameter that appears to correlate with the energetic ordering of these three conformers is the repulsive hydrogen-to-hydrogen distance between two hydroxyl groups, e.g. the H···H distance in gGg<sup>-</sup> at the HF/cc-pVTZ geometry is 0.31 Å closer than in tGg<sup>-</sup>.<sup>221</sup>

In addition, the common feature found for these three lowest conformations is the *gauche* configuration for the OCC'O' dihedral angle. This *gauche* stabilisation compared to the *trans* OCC'O' torsion allows the vicinal hydroxyl groups in close enough proximity to support intramolecular hydrogen bonding. This can be demonstrated by the red-shift of the OH vibrational frequency at 4095 and 4123 cm<sup>-1</sup> for tGg<sup>-</sup> whereas they are equal at 4124 cm<sup>-1</sup> for tTt.<sup>222</sup> The red shift at 4095 cm<sup>-1</sup> correlates to the OH vibration being involved in intramolecular HO···H bonding while the non-shifted value at 4123 cm<sup>-1</sup> is for free OH stretching frequency practically not affected and equal to that of tTt conformer (at 4124 cm<sup>-1</sup>).

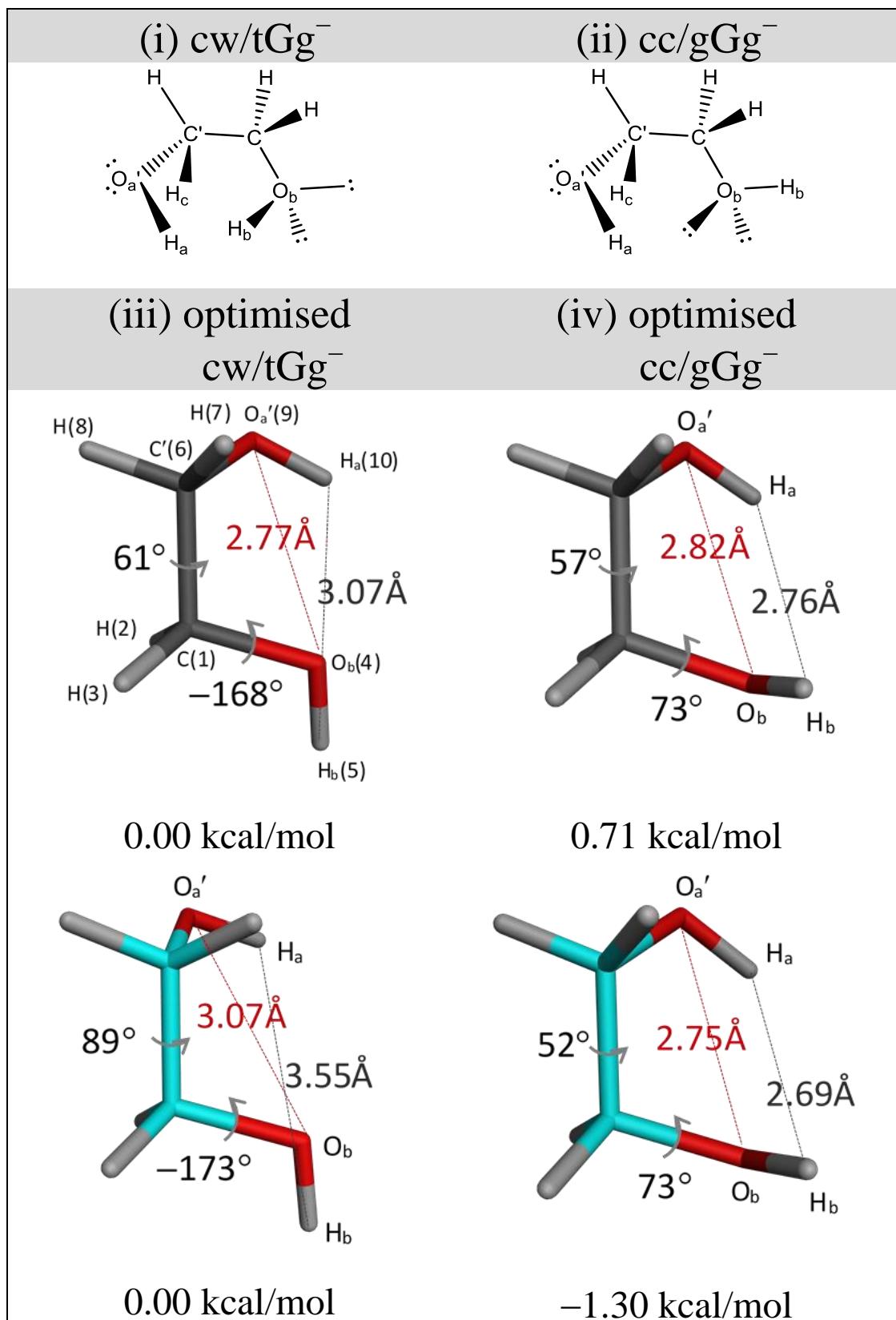


Figure 5.2 Structures and relative energies of 1,2 ethanediols (i) cw/tGg<sup>-</sup> (ii) cc/gGg<sup>-</sup> (iii) cw/tGg<sup>-</sup> optimised at MP2/def2-TZVPP (in grey) and PM6-DH2 (in cyan) and (iv) cc/gGg<sup>-</sup> optimised at MP2/def2-TZVPP (in grey) and PM6-DH2 (in cyan).

Our calculations at the MP2/def2-TZVPP and MP2/aug-cc-pVTZ levels both find this cw diol model to be lower in energy than the cc model by 0.7 kcal/mol (Table 5.3). The corresponding density functional estimates with the triple zeta basis set are 0.2 and 0.3 kcal/mol at ωB97X-D/def2-TZVPP and M06-2X/def2-TZVPP levels respectively (Table 5.3), which agree rather well with the energy difference of 0.35 kcal/mol via the approximate CCSD(T)/cc-pVQZ estimate<sup>221</sup>. This energetic preference is preserved by the PM3CARB-1, SCC-DFTB and DFTB3 methods, with  $E_{cc-cw}$  energy values of 1.1, 0.4 and 0.6 kcal/mol respectively (Table 5.3). Indeed, PM3CARB-1 has the lowest MUE of all DFT and SQM methods studied, of 0.4 kcal/mol (Table 5.3). This is unsurprising as the Hamiltonian was parameterised using 1,2-ethanediol in its training set.<sup>209</sup>

What is striking, however, is the  $E_{cc-cw}$  value for all other NDDO methods. These uniformly predict the opposite, that the cc diol conformer is lower in energy than the cw conformer, with  $E_{cc-cw}$  values ranging from -0.9 kcal/mol for AM1 to -1.4 kcal/mol for PM3 and PM3-D\* (Table 5.3). The PM6-D3 and PM6-DH2 methods have values of -1.3 kcal/mol in both cases. The PM7 is only slightly lower, with a  $E_{cc-cw}$  of -0.9 kcal/mol. Consequently, given the high level MP2/def2-TZVPP value for  $E_{cc-cw}$  of 0.71 kcal/mol, the error in the NDDO methods for the diol,  $\delta E_{cc-cw}$ , ranges up to 2.1 kcal/mol. One may estimate the effect of this error on the conformational preference of the seven glucosyl units in βCD by correcting  $E_{cccc-cccw}$  with  $7\delta E_{cc-cw}$ . Doing this, one obtains a corrected  $E_{cccc-cccw}$  value,  $E'_{cccc-cccw}$ , for PM6-DH2 of 1.27 kcal/mol (Table D6), which is the value obtained (uncorrected) at the M06-2X/def2-TZVPP level (Table 5.1). Whilst somewhat fortuitous, this does suggest that a major error in computing βCD stability lies in the modelling of the vicinal O2···O3 intramolecular interaction for PM6-DH2. Positive corrected values are similarly found for the NDDO methods,

although larger in magnitude; the exception is PM6, which yields a corrected E<sub>cccc-cccw</sub> value of -0.04 kcal/mol (Table D6).

**Table 5.3 Relative gas-phase conformational energetics in kcal/mol of 1,2-ethanediols relative to cw/tGg-. The relative energy between cc/gGg- and cw/tGg- conformers, Ecc-cw, is in bold.**

Method	cc/gGg <sup>-</sup> (E <sub>cc-cw</sub> )	cc/gGg <sup>-</sup>	gGg	gTg	gTg <sup>-</sup>	tGg	tGt	tTg	tTt	MSE <sup>g</sup>	MUE <sup>h</sup>	RMSE <sup>i</sup>
MP2/ def2-TZVPP	<b>0.71</b>	1.02	2.79	2.43	2.13	3.23	2.72	1.93	1.56			-
MP2/ aug-cc-pVTZ	<b>0.73</b>	0.94	0.74 <sup>b</sup>	2.42	2.14	3.15	2.61	1.90	1.52	-0.24	0.24	0.65
MP2/ cc-pVDZ <sup>a</sup>	<b>0.30</b>	1.19	3.15	3.00	2.81	3.81	3.48	2.92	2.87	0.50	0.58	0.69
M06-2X/ def2-TZVPP	<b>0.32</b>	1.03	0.31 <sup>b</sup>	3.06	2.85	3.74	3.44	2.92	2.82	0.20	0.77	1.03
ωB97X-D/ def2-TZVPP	<b>0.17</b>	0.79	2.53	2.64	2.45	3.36	3.17	2.62	2.66	0.18	0.39	0.49
DFTB3	<b>0.59</b>	1.88	0.59 <sup>b</sup>	2.02	1.85	0.00 <sup>c</sup>	3.17	1.27	0.49	-0.67	0.93	1.34
SCC-DFTB	<b>0.42</b>	1.16	0.42	1.37	1.15	0.00 <sup>c</sup>	2.32	0.61	-0.16	-1.12	1.15	1.52
PM7	<b>-0.93</b>	-0.20	-0.93	-0.21	0.25	0.00 <sup>c</sup>	-0.21	-0.9 <sup>e</sup>	2.56	-1.91	2.11	2.38
PM6-DH2	<b>-1.30</b>	-0.74	-0.87	-0.74	-0.40	0.00 <sup>c</sup>	-0.74	0.01 <sup>f</sup>	0.00 <sup>f</sup>	-2.33	2.33	2.56
PM6-D3	<b>-1.27</b>	-0.81	-0.82	-0.15	-0.53	0.00 <sup>c</sup>	-0.81	0.00 <sup>f</sup>	2.64	-2.03	2.24	2.49
PM6	<b>-1.07</b>	-0.77	-0.85	-0.77	-0.51	0.00 <sup>c</sup>	-0.77	-1.07 <sup>e</sup>	0.00 <sup>f</sup>	-2.43	2.43	2.66
PM3CARB-1	<b>1.10</b>	1.46	1.76	2.87	2.50	3.62	2.56	2.39	2.22	0.20	0.43	0.51
PM3-D*	<b>-1.39</b>	-0.78	-0.82	-0.43	-0.67	1.76	-0.78	1.07	2.56	-1.80	2.00	2.30
PM3	<b>-1.39</b>	-0.78	-0.82	-0.43	-0.67	1.76	-0.78	1.07	2.56	-1.80	2.00	2.30
AM1	<b>-0.88</b>	-0.98	-0.98	0.39	0.20	1.55 <sup>d</sup>	0.00 <sup>c</sup>	1.55	2.98	-1.47	1.75	2.03

<sup>a</sup>McNamara *et al.*<sup>209</sup> Optimised geometries convert to <sup>b</sup>gGg<sup>-</sup>, <sup>c</sup>tGg<sup>-</sup>, <sup>d</sup>tTg, <sup>e</sup>g<sup>-</sup>G<sup>-</sup>g, <sup>f</sup>tG<sup>-</sup>g

<sup>g</sup>Mean signed error, <sup>h</sup>mean unsigned error and <sup>i</sup>root mean square error with respect to M06-2X/def2-TZVPP level of theory

To further investigate the origin of this discrepancy, we compare the geometries of these cc and cw diol models via PM6-DH2 method with that at the MP2/def2-TZVPP level. For the cc conformer, there is a good correspondence in overall conformation between these two levels of theory: for example, the key internal angle  $\psi$  describing the OCC'O' torsion differs by only  $5^\circ$  (Tables 5.4 and D7, Figure 5.2(iv)). The difference for  $\varphi_2$  is  $8^\circ$  and zero for  $\varphi_1$ . This similarity is reflected in intramolecular  $O_a' \cdots H_b$  and  $O_b \cdots H_a$  distances of 3.25 and 2.30 Å at the PM6-DH2 level, and 3.19 and 2.22 Å at the MP2/def2-TZVPP level respectively (Table 5.4), i.e. with atom pair  $O_b \cdots H_a$  forming the internal hydrogen bond.

By contrast the cw diol conformer is distinctly twisted around the central OCC'O' relative to the MP2/def2-TZVPP conformation, with  $\psi$  values of  $61^\circ$  at PM6-DH2 and  $89^\circ$  at MP2/def2-TZVPP (Tables 5.4 and D7, Figure 5.2(iii)). Larger deviations are seen in  $\varphi_1$  and  $\varphi_2$  relative to cc, of  $5^\circ$  and  $12^\circ$  respectively. Furthermore, the intramolecular hydrogen bond between  $O_b$  and  $H_a$  appears highly strained in cw, with a PM6-DH2 value of 2.81 Å compared to a MP2/def2-TZVPP value of 2.30 Å; and to a value of 2.22 Å at this level for cc (Table 5.4). The  $H \cdots H$  distance is 0.31 Å longer for cw relative to cc at the MP2/def2-TZVPP level but 0.86 Å lower at PM6-DH2 level (Figure 5.2(iii, iv)). The exaggerated  $H \cdots H$  repulsion in cw at PM6-DH2 level appears to lead to a 0.48 Å increase compared to the MP2/def2-TZVPP value, as opposed to a small decrease of 0.07 Å for cc (Figure 5.2(iii, iv)). This compares with an increase of 0.47 Å for cw and decrease of 0.07 Å for cc relative to the M06-2X geometry; and a respective increase of 0.41 Å and decrease of 0.10 Å relative to the  $\omega$ B97X-D geometry (Table 5.4). Similar geometric discrepancies are found for the cw conformer geometries of the diol as predicted by the other NDDO methods, for example PM7 has a  $\psi$  value of  $80^\circ$ .

However, PM3CARB-1 has a  $\psi$  value of  $66^\circ$ , closer to the MP2 value of  $61^\circ$  (Table 5.4). For DFTB3 and SCC-DFTB, we note deviations in  $\psi$ ,  $\phi_1$  and  $\phi_2$  of ( $10^\circ$ ,  $10^\circ$ ,  $12^\circ$ ) and ( $3^\circ$ ,  $9^\circ$ ,  $6^\circ$ ) respectively (Table 5.4). The twisted diol OCC'O' angle does not manifest itself in the  $\beta$ CD and glucose cw-related geometries as the covalent structure of the pyranose ring limits rotation around this coordinate. This indeed suggests that the inaccurate modelling of the interaction does not stem simply from an incorrect geometry but is a more fundamental issue in modelling the nature of the vicinal *trans* OH-*gauche* OH interaction.

It appears therefore that the error in computing the relative energetics of cc and cw diol models lies in underestimated stability of the cw conformer, in which the *trans-gauche* vicinal OH interaction occurs; this corresponds to an underestimated O2-H $\cdots$ O3 hydrogen bond in cw conformers of glucose and  $\beta$ CD. Interestingly, the poor estimate of energy for this conformer appears not to arise from the dispersion correction term (as PM3 and PM6 perform equally poorly to PM6-D3 and PM6-DH2) but rather seems inherent in the parameterization of the QM Hamiltonian for this region of the potential energy surface. Inspection of the Mulliken charges indicates overpolarisation of the O-H bond dipoles in the diol (Table 5.5). At PM6-DH2, the Mulliken charges for O<sub>a'</sub> and H<sub>a</sub> in the cw conformer were  $-0.54$  and  $0.32\text{ }e$ , which compares with  $-0.45$  and  $0.23\text{ }e$  for MP2/def2-TZVPP. Values were similar for the H<sub>b</sub>-O<sub>b</sub> bond, and found to be similar in the cc conformer (Table 5.5). Also noteworthy were the relatively electron-rich carbons of the cw and cc diol conformers via the PM6-DH2 method compared to at MP2/def2-TZVPP. However, the Mulliken charge distributions overall were similar for cc and cw conformers at the PM6-DH2 level.

**Table 5.4 Geometrical data of cw/tGg<sup>-</sup> and cc/gGg<sup>-</sup> conformers of 1,2-ethanediols.<sup>a</sup> Distances in Å and angles in degrees.**

Coordinate	MP2/def2-TZVPP		M06-2X/ def2-TZVPP		$\omega$ B97X-D/ def2-TZVPP		DFTB3	
	cw/tGg <sup>-</sup>	cc/gGg <sup>-</sup>	cw/tGg <sup>-</sup>	cc/gGg <sup>-</sup>	cw/tGg <sup>-</sup>	cc/gGg <sup>-</sup>	cw/tGg <sup>-</sup>	cc/gGg <sup>-</sup>
O <sub>a'</sub> ···H <sub>b</sub>	3.57	3.25	3.58	3.24	3.60	3.26	3.59	3.03
O <sub>b'</sub> ···H <sub>a</sub>	2.30	2.30	2.31	2.31	2.38	2.36	2.15	2.14
O'···O	2.78	2.82	2.77	2.81	2.81	2.84	2.70	2.72
H <sub>a'</sub> ···H <sub>b</sub>	3.07	2.76	3.08	2.76	3.14	2.79	3.01	2.53
O'C'C	111.2	111.0	111.1	110.9	111.9	111.6	109.5	109.1
OCC'	106.0	110.5	106.3	110.4	107.0	111.1	108.0	109.5
H <sub>b</sub> OCC' ( $\phi_1$ )	-167.6	73.5	-168.3	73.4	-166.8	72.1	-178.4	65.0
H <sub>a</sub> O'C'C ( $\phi_2$ )	-50.5	-43.2	-50.6	-42.6	-52.0	-43.3	-38.9	-25.4
OCC'O' ( $\psi$ )	61.0	57.0	60.5	56.7	62.1	57.6	50.5	49.5
Coordinate	SCC-DFTB		PM7		PM6-DH2		PM3CARB-1	
	cw/tGg <sup>-</sup>	cc/gGg <sup>-</sup>	cw/tGg <sup>-</sup>	cc/gGg <sup>-</sup>	cw/tGg <sup>-</sup>	cc/gGg <sup>-</sup>	cw/tGg <sup>-</sup>	cc/gGg <sup>-</sup>
O <sub>a'</sub> ···H <sub>b</sub>	3.65	3.03	3.77	3.30	3.87	3.19	3.87	3.57
O <sub>b'</sub> ···H <sub>a</sub>	2.25	2.18	2.62	2.44	2.81	2.22	2.64	2.65
O'···O	2.77	2.75	2.94	2.88	3.07	2.75	2.90	2.98
H <sub>a'</sub> ···H <sub>b</sub>	3.10	2.53	3.40	2.91	3.55	2.69	3.59	3.25
O'C'C	109.6	109.2	110.2	110.5	110.7	110.0	114.6	115.0
OCC'	108.0	109.3	104.2	109.6	104.5	108.6	108.3	112.8
H <sub>b</sub> OCC' ( $\phi_1$ )	-176.9	60.6	-175.6	66.9	-173.2	73.1	-179.3	79.7
H <sub>a</sub> O'C'C ( $\phi_2$ )	-44.7	-30.9	-57.8	-42.5	-62.4	-35.3	-63.1	-49.4
OCC'O' ( $\psi$ )	57.7	53.7	80.2	65.4	88.5	52.4	65.9	65.1

<sup>a</sup>See atom label and structures in Figure 5.2

**Table 5.5 Mulliken atomic partial charges (in  $e$ ) of selected 1,2-ethanediols.**

Diol-cw (tGg <sup>-</sup> )	MP2/ Def2TZVPP	PM6	PM6-DH2	PM6DH2 //PM6	DFTB3	SCC-DFTB
C1	0.097	-0.066	-0.067	-0.066	0.123	0.124
H2	0.055	0.137	0.136	0.137	0.033	0.032
H3	0.041	0.125	0.125	0.125	0.023	0.020
O4	-0.446	-0.557	-0.557	-0.557	-0.566	-0.518
H5	0.218	0.309	0.309	0.309	0.358	0.320
C6	0.188	-0.013	-0.013	-0.013	0.148	0.136
H7	0.038	0.143	0.142	0.143	0.032	0.033
H8	0.032	0.149	0.149	0.149	0.039	0.036
O9	-0.453	-0.542	-0.541	-0.542	-0.562	-0.514
H10	0.231	0.315	0.316	0.315	0.373	0.331
Diol-cc (gGg <sup>-</sup> )	MP2/ Def2TZVPP	PM6	PM6-DH2	PM6DH2 //PM6	DFTB3	SCC-DFTB
C1	0.132	-0.067	-0.069	-0.067	0.107	0.114
H2	0.055	0.169	0.166	0.169	0.066	0.057
H3	0.040	0.125	0.126	0.125	0.028	0.024
O4	-0.452	-0.539	-0.538	-0.539	-0.563	-0.516
H5	0.219	0.306	0.308	0.306	0.356	0.319
C6	0.171	-0.053	-0.054	-0.053	0.124	0.121
H7	0.016	0.124	0.124	0.124	0.024	0.021
H8	0.035	0.153	0.151	0.153	0.047	0.042
O9	-0.446	-0.536	-0.539	-0.536	-0.559	-0.512
H10	0.231	0.319	0.326	0.319	0.370	0.330

### 5.4.2 $\beta$ CD/Graphene complex

We now turn to examine the interaction of the  $\beta$ CD conformers with graphene in the gas phase. Firstly, we consider a  $\beta$ CD/C<sub>96</sub>H<sub>24</sub> complex, where the  $\beta$ CD cavity channel is modelled as perpendicular to the graphene surface, as revealed by transmission electron microscopy of the  $\beta$ CD/graphene composite<sup>196</sup>. For this complex, the graphene flake surface contacts the C2C3 rim of  $\beta$ CD, the larger of the two rims. A model flake of this size is tractable using density functional theory and here we apply the M06-2X/def2-TZVPP level of theory at the M06-2X/def2-SVP geometry.

For the two lowest energy  $\beta$ CD hydrogen bond conformers, cccw and cccc, the M06-2X/def2-TZVPP interaction potential energies are calculated to be -30.0 and -29.4 kcal/mol respectively (Table 5.6). After correction for basis set superposition error, the corresponding IEs lower to -26.1 and -25.6 kcal/mol, indicating a slight preference for interaction of the cccw network, with an IE<sub>cccc-cccw</sub> of 0.5 kcal/mol (Table 5.6). Similarly, an IE<sub>cccc-cccw</sub> of 0.1 kcal/mol is found for the BSSE-corrected M06-2X/def2-SVP lower level of theory (Table 5.6). Thus, *in vacuo*, both hydrogen bond networks at the C2C3 rim of  $\beta$ CD are predicted to interact approximately equally well with the graphene surface. Correspondingly, the distance between the oxygens of the C2C3 rim and the graphene basal plane is 2.96 and 2.98 Å for cccw and cccc conformers respectively (Table D8). This parity in interaction is not unexpected, given that, while the direction of the hydrogen bond network differs, the total number and approximate orientation of CH and OH groups with respect to the C<sub>96</sub>H<sub>24</sub> flake are the same for cccw and cccc conformers.

**Table 5.6 Interaction energies (IEs) of C<sub>96</sub>H<sub>24</sub> with C2C3 rim of βCD complexes (in kcal/mol). BSSE-corrected IEs are given in parentheses.**

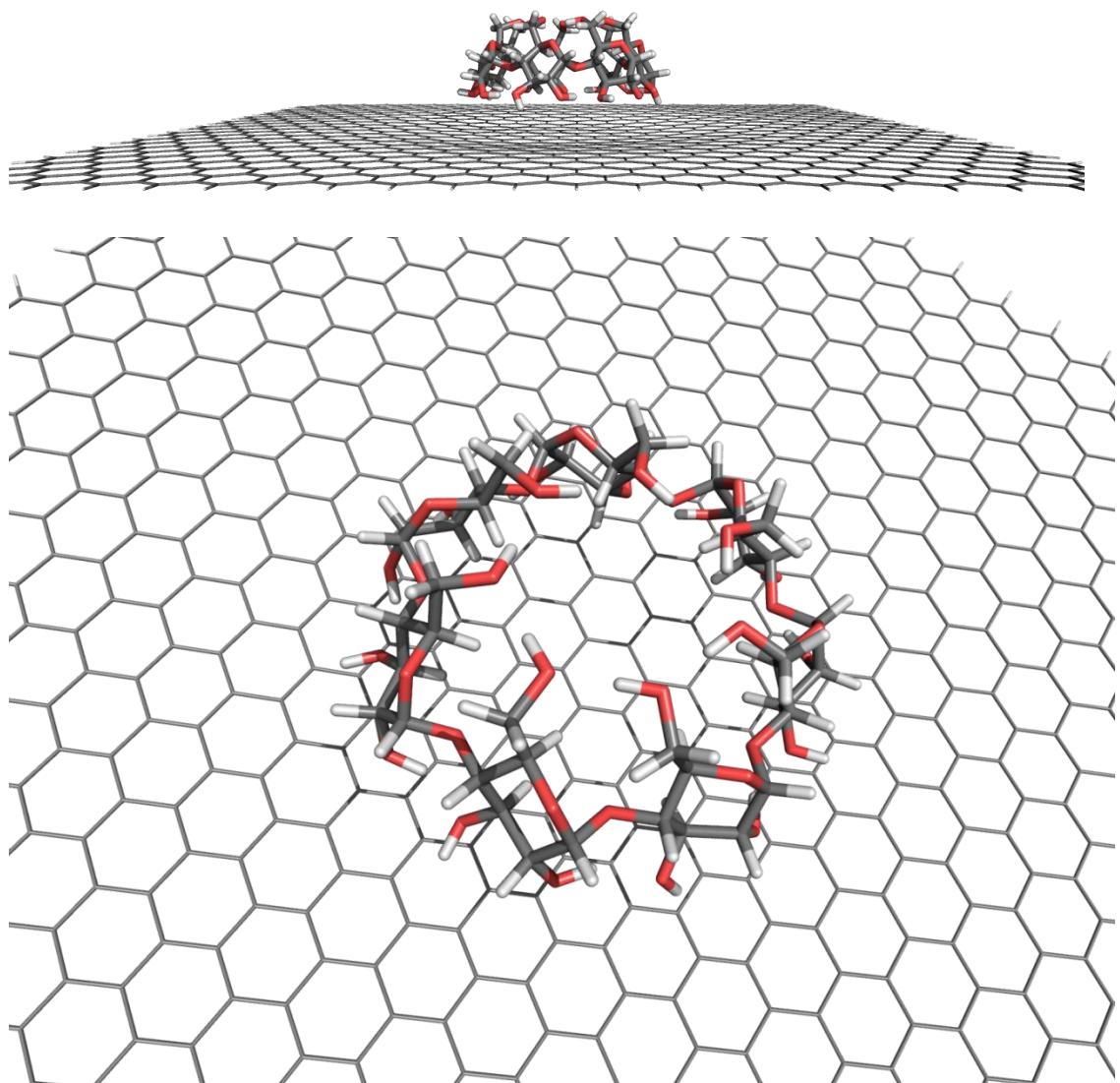
<b>Method</b>	<b>IE</b>		<b>IE<sub>cccc-cccw</sub></b>
	<b>cccw/C<sub>96</sub>H<sub>24</sub></b>	<b>cccc/C<sub>96</sub>H<sub>24</sub></b>	
M06-2X/def2-TZVPP//	-30.0 (-26.1)	-29.4 (-25.6)	0.6 0.5
M06-2X/def2-SVP	-49.5 (-23.3)	-49.3 (-23.2)	0.2 0.1
DFTB3	-43.3	-43.3	0.0
PM7	-71.3	-73.4	-2.1
PM6-DH2	-49.7	-49.5	0.2
PM7/COSMO	-53.0	-53.1	-0.1
PM6-DH2/COSMO	-30.5	-27.7	2.8

For comparison, we apply the most recently developed of semiempirical QM methods considered above, namely DFTB3, PM6-DH2 and PM7, to computation of βCD/C<sub>96</sub>H<sub>24</sub> interaction energies. While we have highlighted a significant issue in modelling the conformation of βCD, all three methods have been shown previously to model noncovalent interactions with good accuracy. We find that both PM6-DH2 and DFTB3 predict essentially isoenergetic interactions of cccc and cccw with C<sub>96</sub>H<sub>24</sub>, with IE<sub>cccc-cccw</sub> values of 0.2 and 0.0 kcal/mol respectively (Table 5.6). For PM7, there is a contrasting preference of 2.1 kcal/mol favouring the cccw conformer (Table 5.6). For all three methods, the rim-graphene distances for cccw and cccc conformers are very similar, differing by 0.01, 0.03 and 0.07 Å for DFTB3, PM6-DH2 and PM7 respectively (Table D8). We note that for all three SQM methods, the absolute interaction energies

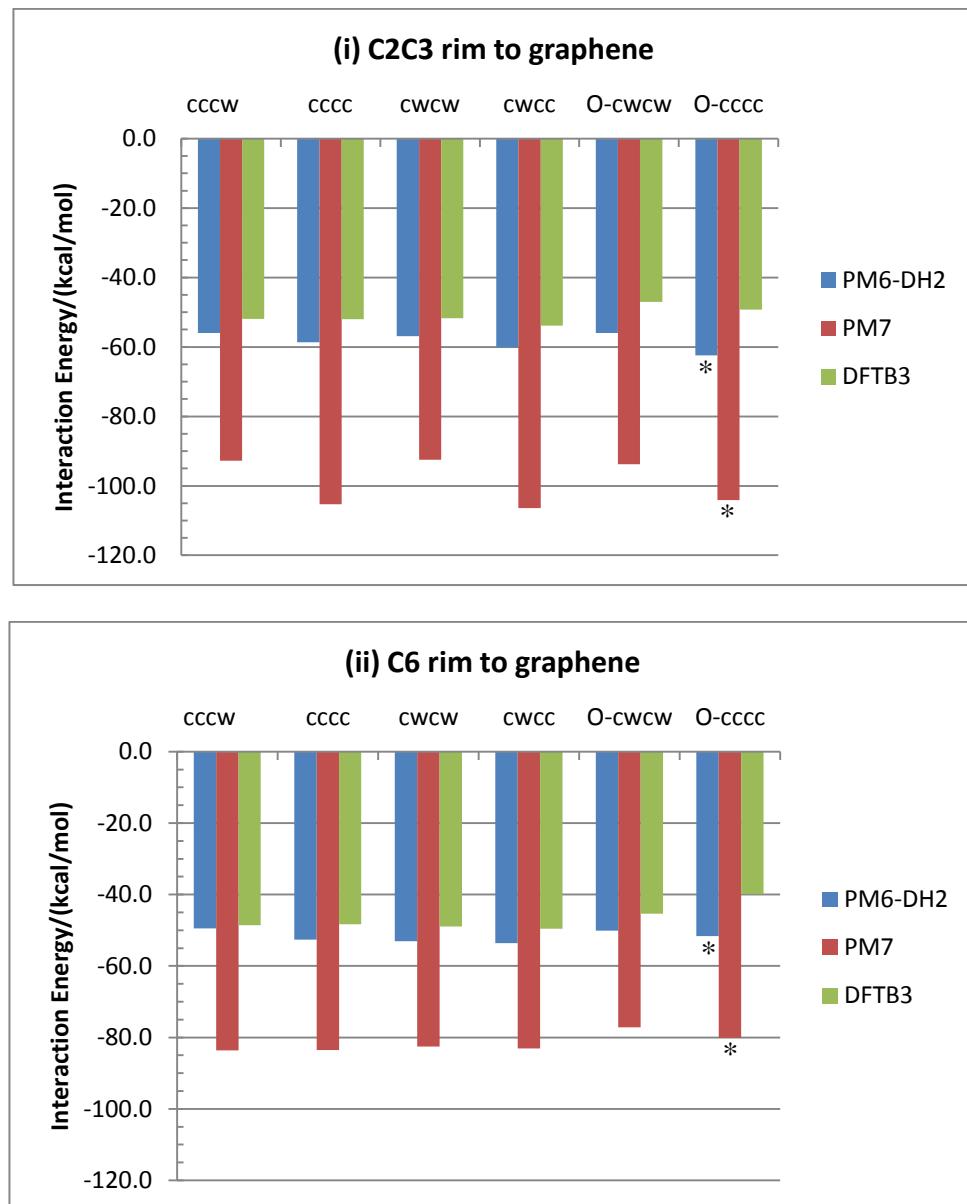
are notably larger than the M06-2X/def2-TZVPP estimates, e.g. for the interaction of the cccw conformer by 44%, 66% and 138% via the DFTB3, PM6-DH2 and PM7 methods.

Using SQM, it is possible to compute the interaction energetics of  $\beta$ CD with a substantially larger graphene flake. We therefore apply DFTB3, PM6-DH2 and PM7 methods to examine the gas-phase interaction of  $\beta$ CD with a C<sub>1006</sub>H<sub>88</sub> graphene flake (Figure 5.3). We employ all six low energy  $\beta$ CD conformers discussed earlier, and consider their interaction with graphene bound either via the C2C3 or C6 rim. Firstly, we note that for all three methods, the interaction energies of cccw and cccc with the larger flake C<sub>1006</sub>H<sub>88</sub> are more favourable than with C<sub>96</sub>H<sub>24</sub> (Table 5.7), by 8.6, 7.8 and 26.7 kcal/mol on average for DFTB3, PM6-DH2 and PM7 methods respectively.

Secondly, it is clear that the three methods provide a consensus that the *in vacuo* adsorption of  $\beta$ CD via its C2C3 rim (Figure 5.4(i), Table 5.7), at least when in an idealised hydrogen bonding conformation, is preferred over the C6 rim (Figure 5.4(ii), Table 5.7). Based on the conformer with highest affinity for each of the three SQM approaches, the preference for C2C3 is computed as 4.4 and 10.7 kcal/mol for DFTB3 and PM6-DH2, with PM7 once again providing a somewhat larger estimate of 23.3 kcal/mol (Table 5.7). The C2C3 is the larger of the two rims and make more extensive surface contacts with graphene.



**Figure 5.3** The cwcc  $\beta$ CD/C<sub>1006</sub>H<sub>88</sub> complex optimised at PM6-DH2 level of theory (C2C3 rim at graphene surface).



**Figure 5.4** Interaction energies in kcal/mol of graphene  $\beta$ CD/C<sub>1006</sub>H<sub>88</sub> computed at PM6-DH2, PM7 and DFTB3 levels of theory [Note: \*open geometries collapsed to cccc].

**Table 5.7 Interaction energies (IEs) of  $\beta$ CD/C<sub>1006</sub>H<sub>88</sub> complexes in kcal/mol. Collapsed open structures labelled in italics.**

Conformer	C2C3 adsorbed rim			C6 adsorbed rim		
	DFTB3	PM7	PM6-DH2	DFTB3	PM7	PM6-DH2
cccw	-51.9	-92.8	-56.0	-48.5	-83.6	-49.4
cccc	-52.0	-105.3	-58.7	-48.3	-83.6	-52.7
cwcw	-51.8	-92.5	-56.9	-48.9	-82.5	-53.1
cwcc	-53.9	-106.4	-60.1	-49.5	-83.1	-53.6
O-cwcw	-47.0	-93.7	-56.0	-45.4	-77.2	-50.1
O-cccc	-49.2	<i>-104.1</i>	<i>-62.4</i>	-39.9	<i>-80.1</i>	<i>-51.7</i>

At this C2C3 rim, the cwcc conformer is predicted to interact with the highest interaction energy for DFTB3 and PM7 (Table 5.7, Figure 5.3). For PM6-DH2 (as for PM7), the open O-cccc structure collapses to a closed cccc conformer on geometry optimisation, both for the C2C3 and C6 rim orientations (Table 5.7). It is this collapsed cccc structure that has the highest interaction energy with graphene for PM6-DH2, although the cwcc conformer is the most favourably interacting after this (Table 5.7). For the DFTB3 method, where the O-cccc conformer is stable, the preference for closed over open idealised structures is more apparent. However, the range of binding energy across closed conformers is narrow (2.1 kcal/mol, Table 5.7). Thus, with the exclusion of the collapsed open structure, it appears a C2C3 rim interaction with the rim hydrogen bond networks in the cwcc conformation is preferred as predicted by the three methods. Comparing the energy of cwcc with cwcw, there is a preference at the C2C3 rim-graphene interface for a cc hydrogen bond network over the cw by 2.1 and 3.2 kcal/mol

for DFTB3 and PM6-DH2 respectively (Table 5.7). This preference rises to 13.9 kcal/mol for PM7.

The inclusion of the effects of aqueous solvation on the interaction, via the COSMO implicit solvent model, does not alter these conclusions: for example, the cwcc conformer has the most favourable PM6-DH2 interaction energy, of -38.4 kcal/mol, reduced by 21.7 kcal/mol on solvation. This is 1.8 kcal/mol more favourable than the cwcw conformer and 3.0 kcal/mol more favourable than the cccw conformer.

Although the cc network at the C2C3 rim has the highest interaction potential energy at the binding interface, it is important to consider the energetics of overall binding from the global energy minimum structure, cccw, in which the C2C3 rim is in a cw configuration. Furthermore, we recall that the stability of cw configurations for the NDDO methods is underestimated. In computing an overall binding energy (BE), we may therefore apply the MP2/def2-TZVPP diol correction,  $7\delta E_{cc-cw}$ , to the cw  $\beta$ CD conformers here, giving a corrected value, BE'. Consequently, relative to an unbound cccw conformer, corrected by  $7\delta E_{cc-cw}$ , we can estimate the overall PM6-DH2 binding energy of the cwcc conformer of  $\beta$ CD to C<sub>100</sub>H<sub>88</sub> in the gas-phase as -61.5 kcal/mol (Table 5.8). Indeed, via this approach, we find the most favourable overall BE' to be made by the collapsed O-cccc conformer, which forms a closed cccc structure with a BE' of -76.1 kcal/mol. This appears to be due to a breakdown in rim hydrogen bonding network and resultant formation of more favourable CH- $\pi$  and OH- $\pi$  interaction geometries with the graphene surface. Of the idealised hydrogen bond conformers, cccw is the lowest in binding energy at the PM6-DH2 level, with a BE' value of -69.1 kcal/mol (Table 5.8), followed closely by cwcw (-68.2 kcal, Table 5.8). On the

inclusion of solvent effects via the COSMO model, the collapsed cccc conformer remains the most tightly bound, followed by cwcw and then cwcc conformers (Table 5.8).

**Table 5.8** Interaction energy (IE) and overall binding energies (BE) of  $\beta$ CD/C<sub>1006</sub>H<sub>88</sub> complexes at C2C3 rim in gas-phase (g) and solution (aq) and binding energy corrected for cc bias (BE') at PM6-DH2 level, in kcal/mol. IE computed in regard to respective conformer; BE and BE' computed in regard to cccw conformer. Collapsed open structure labelled in italics.

Conformer	IE (g)	BE (g)	BE' (g)	BE' (aq)
cccw	-56.0	-55.0	-69.1	-48.5
cccc	-58.7	-64.7	-64.7	-48.5
cwcw	-56.9	-54.1	-68.2	-50.4
cwcc	-60.1	-61.5	-61.5	-49.7
O-cwcw	-56.0	-20.2	-34.3	-24.9
O-cccc	<i>-62.4</i>	<i>-76.1</i>	<i>-76.1</i>	<i>-51.6</i>

## 5.5 Conclusions and outlook

In this chapter, we have assessed computational methods for computing the conformational energetics of  $\beta$ CD and its interaction with graphene. For  $\beta$ CD in the gas phase, DFT methods M06-2X and  $\omega$ B97X-D show a consensus picture of the relative energetics of low lying closed and open conformers. Specifically the closed cccw conformer is predicted to the lowest in energy, as predicted by prior *ab initio* QM

studies, with open structures significantly higher in energy. In contrast with these previous B3LYP studies, we find flipping of the C2C3 rim to have a small but distinct energetic consequence.

On comparison with the DFT surface for  $\beta$ CD, we find that DFTB3, SCC-DFTB and PM3CARB-1 methods give reasonable agreement. However the other SQM methods, all NDDO approaches, appear to suffer from problems in balancing the cw versus cc hydrogen bond; this imbalance in interactions between O/H atom pairs in close proximity appears to arise from a failure in the Hamiltonian around *trans* orientations of the vicinal OH interaction. One could speculate that the internal geometry affects the stereoelectronic delocalisation into the CO antibonding orbital associated with gauche stabilisation. Natural bond orbital deletion analysis has been applied previously to the 1,2-ethanediol conformers to explore the extent of  $\sigma_{\text{CO}} \rightarrow \sigma^*$  and lone pair  $\rightarrow \sigma^*$  delocalisation in the molecules.<sup>221</sup> At the B3LYP/6-311++G\*\* level of theory, indeed there are significant hyperconjugation contributions: the  $\sigma_{\text{CO}} \rightarrow \sigma^*$  interaction favoured cw/tGg<sup>-</sup> by 0.5 kcal/mol; and the lone pair  $\rightarrow \sigma^*$  interaction favoured cc/gGg<sup>-</sup> by 2.1 kcal/mol. Potentially, imbalance of these contributions could contribute to a destabilisation of the tGg<sup>-</sup> conformer. Reflecting an earlier DFT study<sup>223</sup>, B3LYP/6-311++G\*\* atoms-in-molecules analysis of the cw and cc (tGg<sup>-</sup> and gGg<sup>-</sup>) conformers<sup>224</sup> found no bond path between O<sub>a</sub>'(9) and H<sub>b</sub>(5)-O<sub>b</sub>(4) in either of the 1,2-ethanediol structures, suggesting an absence of a formal intramolecular hydrogen bond (and the associated covalent character); however, a small reduction in hydrogen bond length from by 0.2-0.3 Å led to formation of the path such that, via this model chemistry, the groups would appear to be on the cusp of a hydrogen bond. We do note however that density functional approaches such as B3LYP, applied in both studies mentioned above,

suffer from inaccuracies in modelling weak dispersive interactions. Whether electrostatic or stereoelectronic in source, this inability to capture the subtle balance in H···H repulsion with attractive O···H nonbonding interactions requires further work to improve the parameterisation of the NDDO Hamiltonians.

Nevertheless, for noncovalent intermolecular interactions, methods such as PM6-DH2 and PM7 have been shown to reproduce very high level QM approaches with good accuracy. For example, for the S66x8 set of dimer geometries (which includes optimal and nonoptimal polar, nonpolar and mixed noncovalent interactions), the root mean square error were 0.79 and 0.98 kcal/mol for the PM6-DH2 and PM7 methods respectively, compared with 3.07 kcal/mol for PM6.<sup>7d</sup> Indeed, for benzene-methane and benzene-water dimers, representative of CH- $\pi$  and OH- $\pi$  interactions, PM7 predicts binding energies of -1.8 and -3.0 kcal/mol respectively (Table D9). This compares rather well with CCSD(T)/CBS energies of -1.5 and -3.3 kcal/mol respectively. Indeed, PM6-DH2 gives good agreement (values of -1.5 and -3.6 kcal/mol) with DFTB3 somewhat underestimating benzene-water interaction (-1.6 and -2.6 kcal/mol). Similarly, in another study,<sup>211a</sup> PM6-DH2 was able to compute the adsorption energy of atomic hydrocarbons on graphene as -1.8 kcal/mol per carbon atom, which compared with an experimentally determined value of  $-1.7 \pm 0.3$  kcal/mol.

Despite this good agreement for small complexes, relative to M06-2X/def2-TZVPP calculations, PM6-DH2, DFTB3 and in particular PM7 methods afford larger absolute interaction energies than the DFT calculations for  $\beta$ CD/C<sub>96</sub>H<sub>24</sub> complexes (Figure 5.4). PM7 also does not reproduce the essentially isoenergetic interactions of cccc and cccw with the flake. Whilst these observations could lie partly in underestimation of

dispersive interactions by M06-2X/def2-TZVPP, it would seem compelling to conclude that PM7 in particular significantly overestimates the interaction energies of the  $\beta$ CD/graphene complex, whereas DFTB3 and PM6-DH2 afford relative and absolute estimates in closer line with density functional calculations.

However, on application to the larger  $\beta$ CD/C<sub>1006</sub>H<sub>88</sub> system, all three SQM methods agree on a preference for interaction with closed forms of  $\beta$ CD via its C2C3 rim, although once again the PM7 energies are considerably larger in magnitude. These calculations mirror the findings of Zhou et al. using the COMPASS force field,<sup>196</sup> where the C2C3 rim bound 18.7 kcal/mol more favourably than the C6 rim. Interestingly a study of the interaction of  $\beta$ CD and a carbon nanotube (CNT), using the same force field, similarly found a preference for the C2C3 rim, by 10.8 kcal/mol<sup>225</sup>. The higher affinity of the C2C3 rim for the graphene over CNT (by 14.4 kcal/mol)<sup>196</sup> was reflected by a more endothermic decomposition for  $\beta$ CD/graphene composite over  $\beta$ CD/CNT using thermogravimetric-differential scanning calorimetry analysis<sup>196</sup>.

The SQM methods display a relatively shallow interaction surface, such that the closed conformers lie close in energy to each other – this is partly a consequence of adopting idealised hydrogen bonding networks in the  $\beta$ CD/graphene models. Indeed, when the network breaks down (on collapse of the O-cccc conformer) to form CH- $\pi$  and OH- $\pi$  interactions, the interaction energy increases. This suggests that other, more strongly interacting  $\beta$ CD conformers do exist, beyond the idealised structures considered here. This is true also for the solution structures of  $\beta$ CD, where hydrogen bonds will be in competition with water. Indeed explicit solvent MD simulations of  $\beta$ CD and its inclusion complexes have found the disorder in hydrogen bonding at both rings.<sup>220, 226</sup>

Graphene-bound explicit solvent molecular dynamics simulations are in progress. However the scope of this current work is the evaluation of methods suitable for modelling  $\beta$ CD conformations and their adhesion to graphene. We note that PM6-DH2 and DFTB-D calculations have been successfully employed in rationalising the preferential binding of small unsaturated molecules on graphene surfaces.<sup>211b, 227</sup>

However we found that NDDO models such as PM3, PM6-DH2 and PM7 have a specific issue in modelling the *trans-gauche* vicinal diol interaction, which impacts on modelling of  $\beta$ CD conformations. Based on 1,2-ethanediol calculations, we were able to introduce an empirical correction for this in estimating overall binding energies in gas phase and solution. The DFTB3 method did not appear to suffer from this issue and constitutes a straightforward route to modelling these interactions for large systems such as  $\beta$ CD/C<sub>1006</sub>H<sub>88</sub>. Indeed  $\beta$ CD/graphene systems represent exciting opportunities for molecular electronic devices and drug delivery platforms. SQM methods represent important computational tools that can be employed to guide the design and optimisation of the dispersibility and biocompatibility of these systems.

## CHAPTER 6 Conclusions

Molecular modelling techniques have been utilised in drug design and discovery. Due to the substantial increase in CPU performance in recent years, quantum mechanical (QM) methods capable of capturing electronic effects have become an increasingly attractive method of modelling biomolecular interactions. This research work highlights the use of QM methods for modelling ligands and their interactions in rational drug design.

In Chapter 3, the M06L density functional was applied to analyse the structure-activity relationship of boron-based heterocyclic compounds, anti-inflammatory inhibitors targetting the interleukin-1 $\beta$  (IL-1 $\beta$ ) cytokine. The results find different shapes of the highest occupied molecular orbital (HOMO) between the active and inactive compounds. The most active boron compound in the series has the lowest DFT energy value of its lowest unoccupied molecular orbital (LUMO). In addition, the calculated partial charge at the boron of all five compounds in the series was positive in value, agreeing with that of the boron atomic charge in borinates previously studied at the *ab initio* QM level in the literature.

Although the biological activity of these boron-based compounds is promising, the synthesis of these compounds is not straightforward. To explore non-boron-containing compounds that possess similar shape and/or chemical functionality, the lead-like chemical library containing approximately two million compounds from the ZINC database was virtually screened against the selected boron lead compound using the ROCS tool. The screened results were obtained by three scoring functions: shape-matching *ShapeTanimoto*, chemical-matching *ColorTanimoto*, and the combination of

those two, *TanimotoCombo*. Seven compounds were finally assayed, but showed no inhibition of the activity of IL-1 $\beta$  release. This supports the importance of the boron atom in the structure for anti-IL-1 $\beta$  activity. We suggest further research on the analysis of electrostatic polarity of the B-N bond compared to that of the C-N bond in the carbon analogues. The incorporation of the effect of atomic charge at the boron in virtual screening may also improve the identification of potential non-boron analogues with anti-IL-1 $\beta$  activity.

In Chapter 4, a hybrid QM/MM energy function was employed as a scoring function in structure-based molecular docking of ligands to zinc-containing metalloenzymes. In ligand-metalloenzyme systems, noncovalent interactions such as electrostatics, dispersion, hydrogen bonding and hydrophobic effects play a key role. A successful docking scheme requires both accurately predicting the binding affinities which rely on such weak interactions and sufficient sampling of ligand-enzyme configurations at the active site. A previously implemented DFT/MM Monte Carlo (MC) docking scheme implemented in Gaussian 03 was adapted and implemented in the Gaussian 09 program to employ the faster semiempirical PM6/AMBER function for four test cases. It was applied to two inhibitors of cytidine deaminase (CDA), i.e. the dihydro analogue of zebularine (DHZ) and 3-deazacytidine (DAC) as well as other two inhibitors of human carbonic anhydrase II (hCA II), i.e. fluorinated sulfamoylbenzamide (FSB) and dorzolamide (DZA).

The docking results showed improvement of predicting the binding modes for all four test cases over the conventional docking scheme of GOLD suite in which sampling and scoring are respectively based on the genetic algorithm and empirical ChemPLP scoring

function. There are areas for further improvement of the docking accuracy: (1) including more protein residues at the active site in the QM region to avoid missing key ligand-protein interactions, (2) considering solvent contributions to the scoring function, (3) using dispersion-corrected semi-empirical methods to capture the sum of weak dispersion that could have an impact on the scoring, and (4) using a combined energy minimisation/MC sampling scheme to avoid sampling high energy configurations.

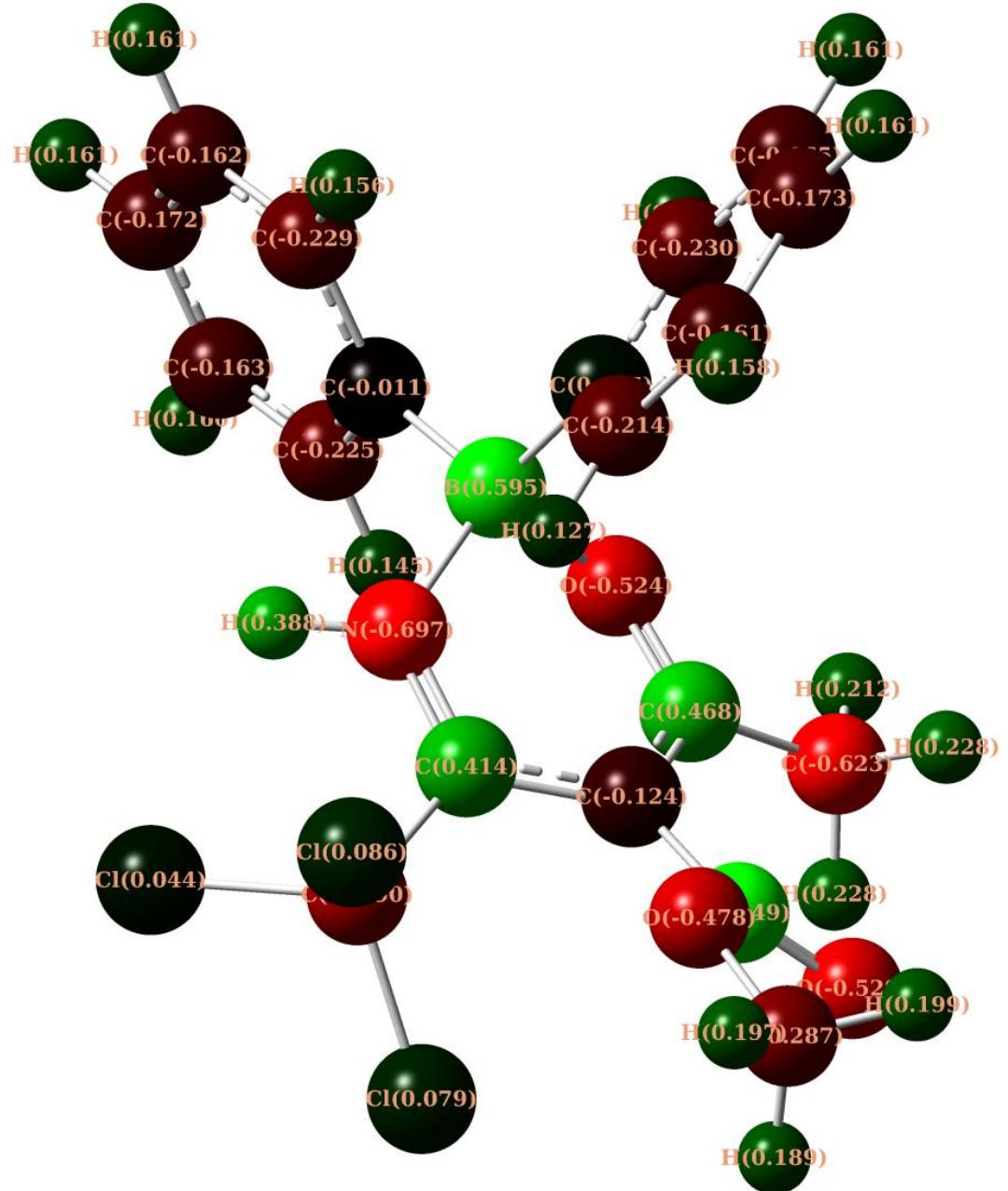
In Chapter 5, QM methods ranging from DFT to SQM methods were applied to modelling the gas-phase conformations of  $\beta$ -cyclodextrin ( $\beta$ CD) and their adsorption on a single layer graphene sheet. The M06-2X and  $\omega$ B97X-D density functionals, as well as the dispersion-corrected DFTB3, SCC-DFTB and PM3-CARB1 methods, predict that the closed *in vacuo*  $\beta$ CD cccw conformer is the lowest in energy, while open  $\beta$ CD structures are significantly higher in energy. This observation is also in good agreement with previous *ab initio* QM studies. However, the lowest energy  $\beta$ CD conformer is predicted differently by the NDDO-based SQM methods. This seems to be due to a failure in the Hamiltonian to balance the O/H atom pair interactions in close proximity for the *trans* orientations found in the clockwise hydrogen bond (as found within glucosyl residues of  $\beta$ CD).

The intermolecular interactions of  $\beta$ CD/graphene complexes were then explored by DFTB3, PM6-DH2 and PM7 methods. These selected SQM methods have been shown to reproduce the binding energies of benchmark dimer geometries computed at high level QM methods with good accuracy. It has been also reported that PM6-DH6 was able to compute the adsorption energy of the hydrocarbon/graphene in good agreement with the experimental value.<sup>211a</sup> For the  $\beta$ CD/C<sub>96</sub>H<sub>24</sub> complexes, PM7 significantly

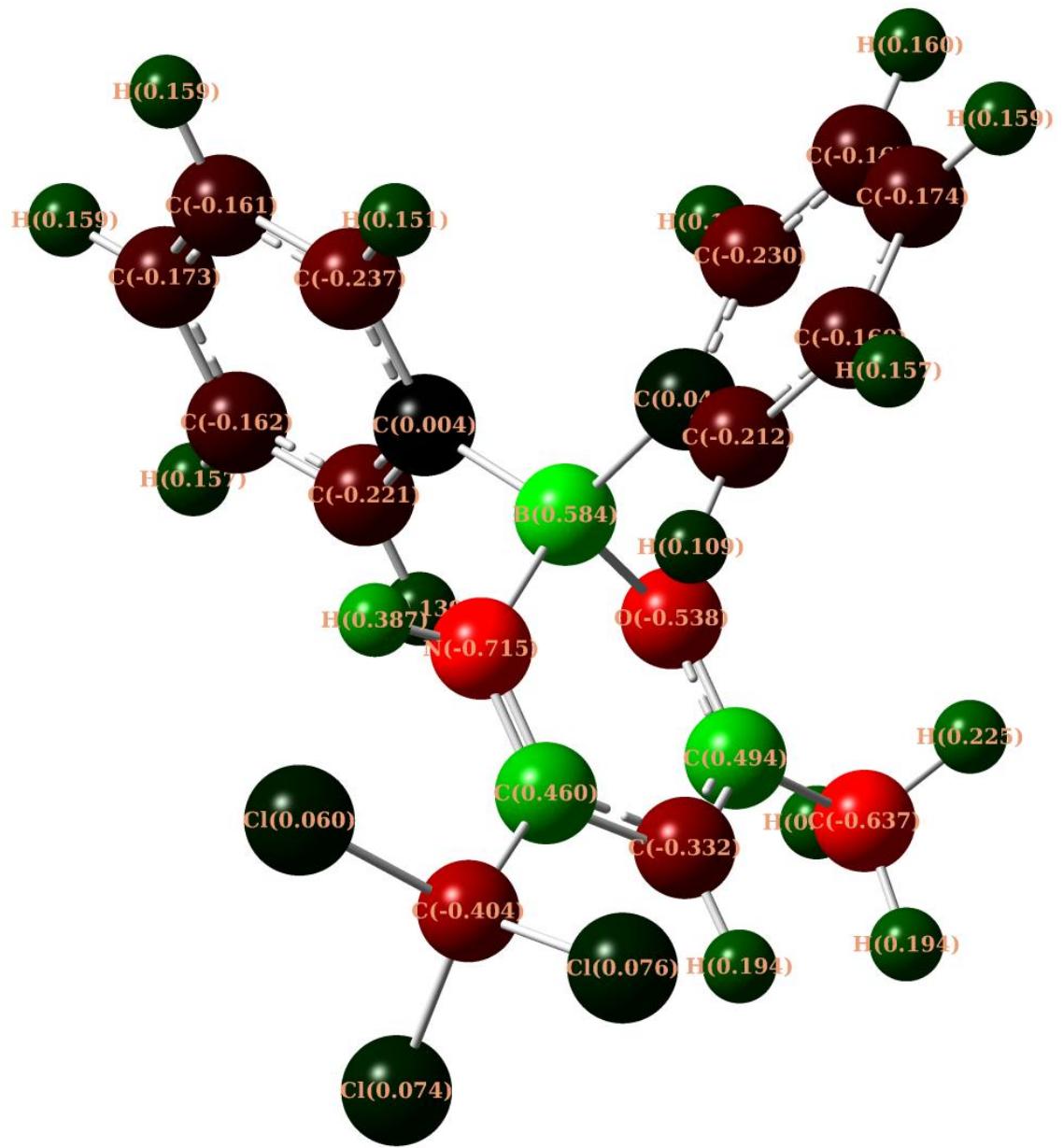
overestimates the interaction energies whereas DFTB3 and PM6-DH2 give relative and absolute estimates closer to that of the M06-2X/def2-TZVPP density functional level of theory. However, for the  $\beta$ CD/C<sub>1006</sub>H<sub>88</sub> complexes, these three methods agree well on the binding preference of  $\beta$ CD on graphene, i.e. in a closed conformation via its C2C3 rim. This observation is also supported by the findings obtained from experimental and computational studies on  $\beta$ CD/graphene systems. Future work will involve molecular dynamics simulations in order to capture the interplay of intramolecular hydrogen bonds and those to solvent.

In brief, this research work demonstrates the feasibility of using QM methods in applications to medicines design. These methods provide understanding of electronic effects associated with the inhibitory activity of boron-based inhibitors for anti-inflammatory diseases. They are also able to model the intra-molecular hydrogen bonding of  $\beta$ CD and the intermolecular interactions of the  $\beta$ CD/graphene complexes, over a thousand atoms in size. Moreover, the hybrid QM/MM energy function along with the MC sampling algorithm shows its ability to improve the predicted binding modes of ligands in the active sites of zinc-containing metalloenzymes over a conventional docking scheme.

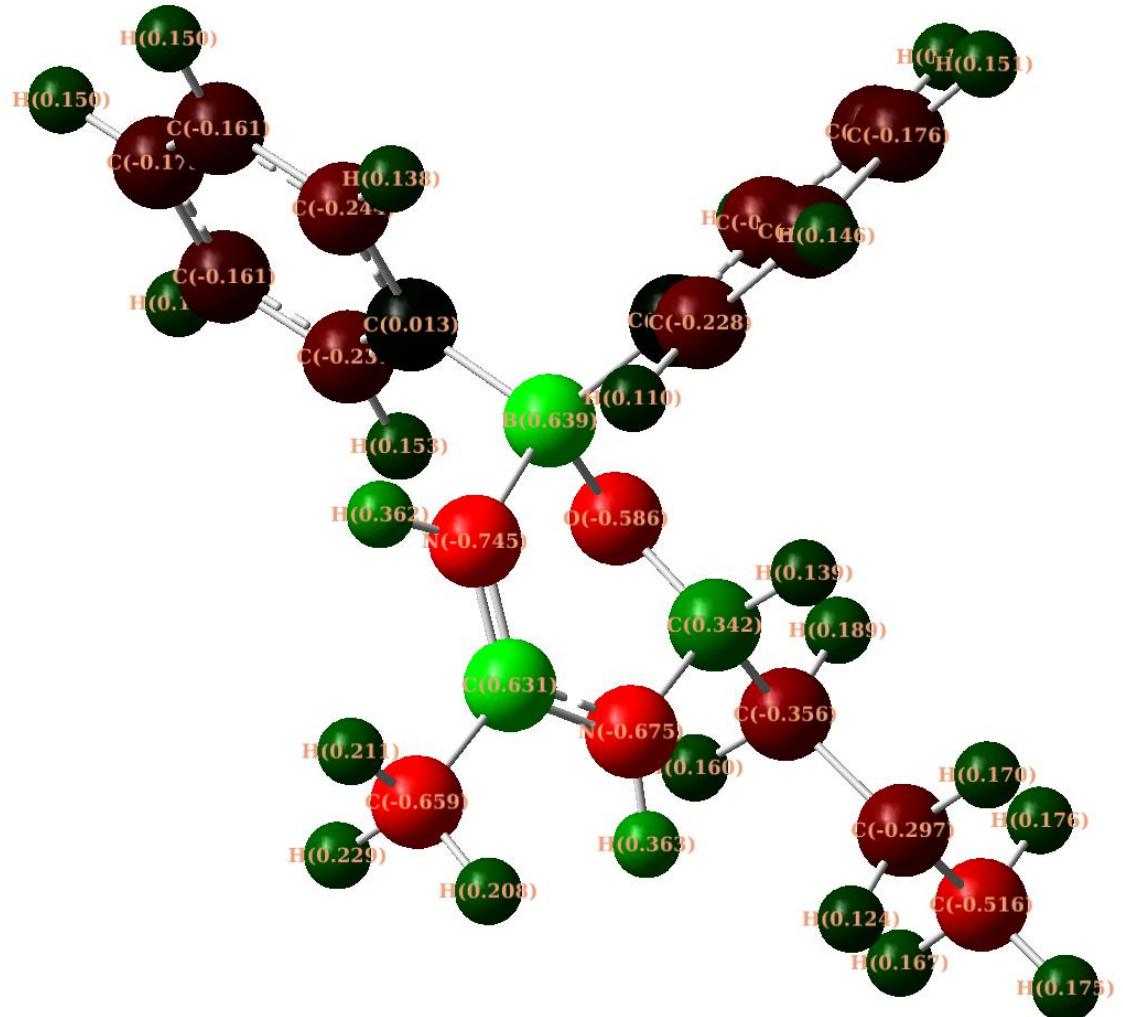
## Appendix A - Molecular charge distribution of query compounds



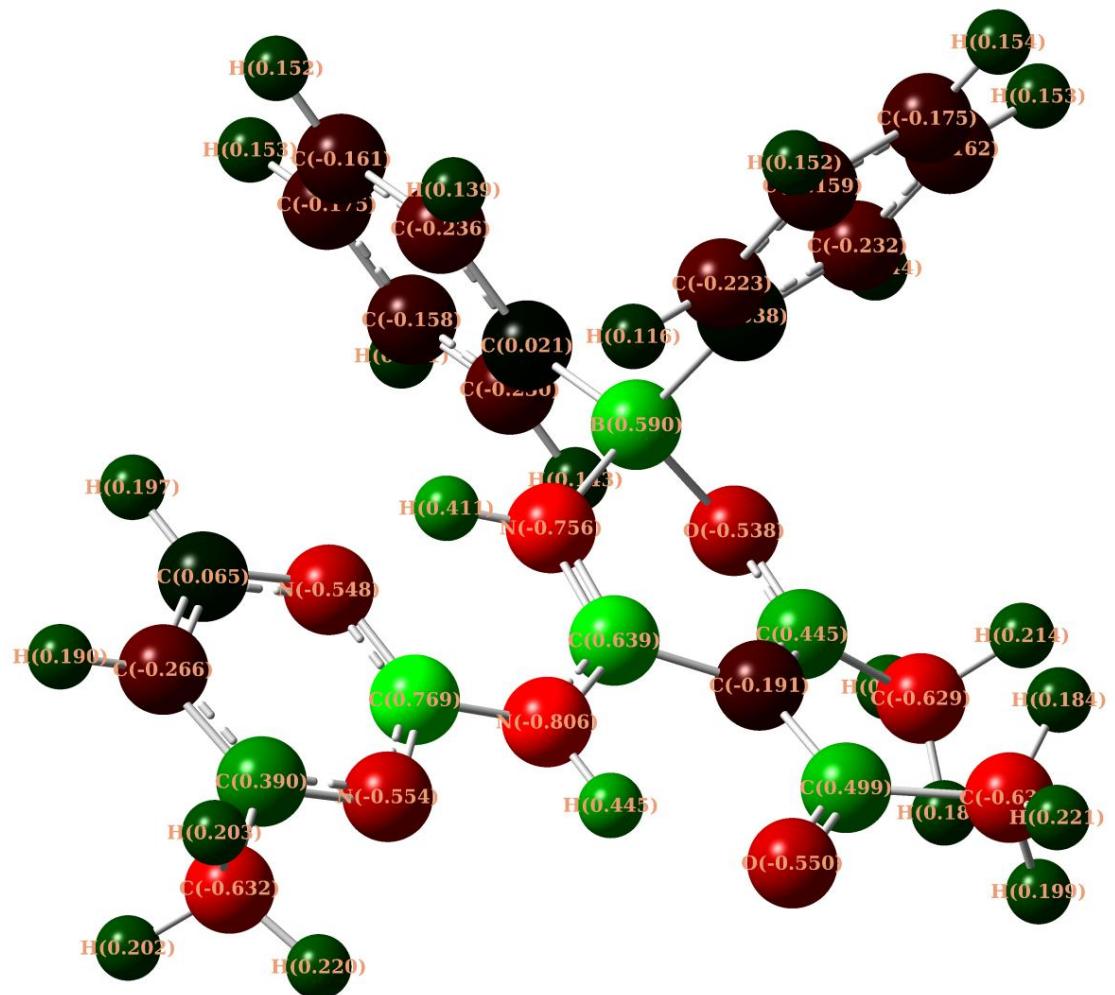
BC 7

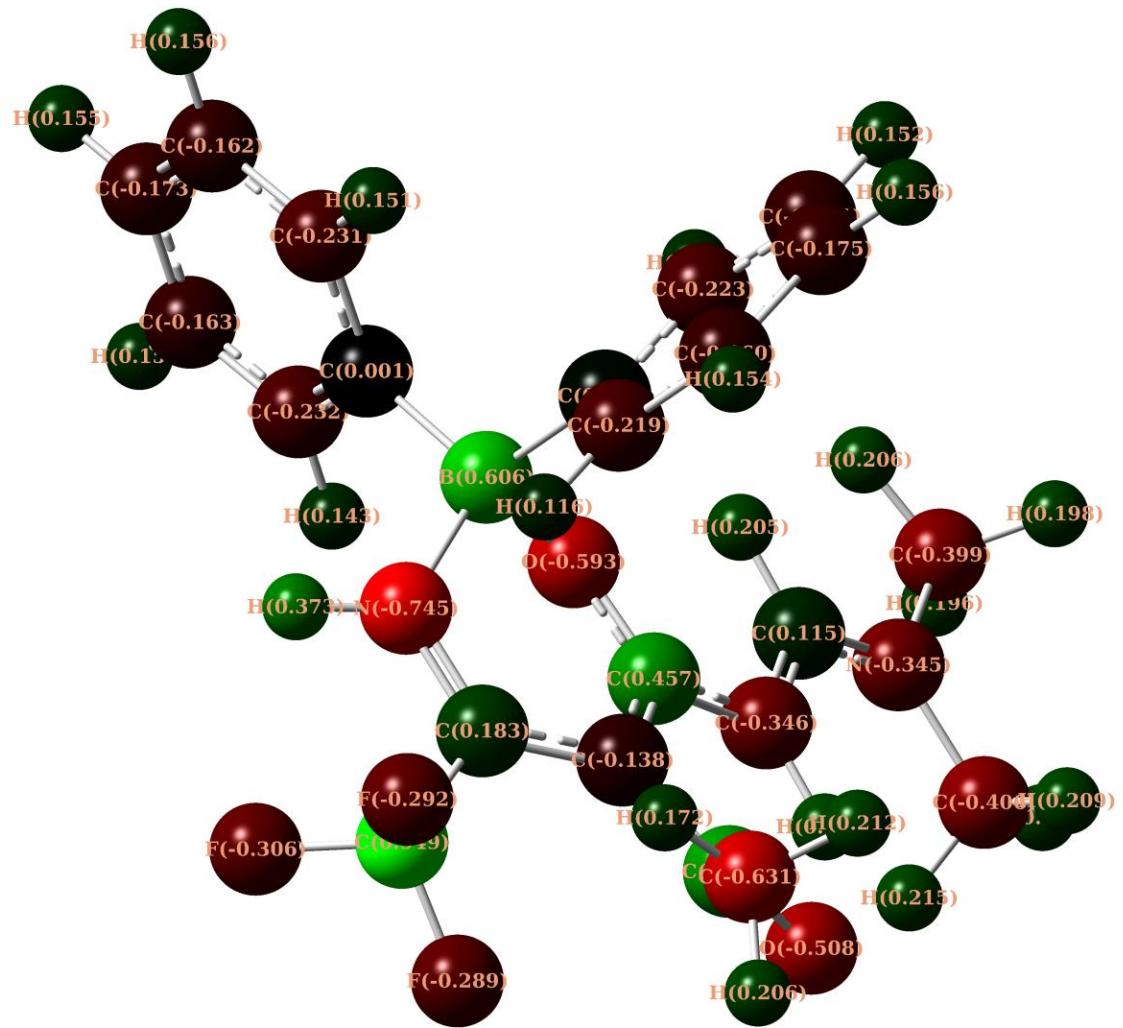


**BC23**



**BC10**

**BC22**

**BC21**

## Appendix B - Structures and 3D overlays of top 25 compounds screened against BC23

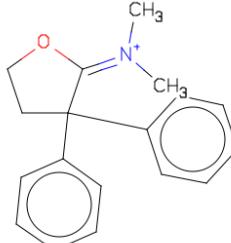
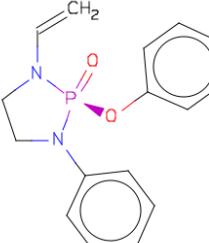
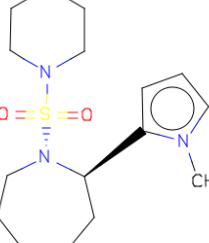
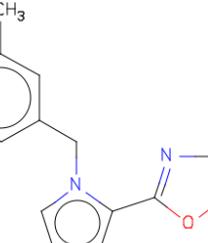
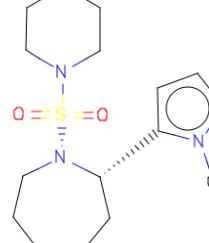
TanimotoCombo rated hits of BC23				
Rank 1	Rank 2	Rank 3	Rank 4	Rank 5
ZINC02389343_0 	ZINC06002617_3 	ZINC48289632_0 	ZINC04970524_105 	ZINC48289630_16 
<b>1.758<sup>a</sup></b> 0.798 <sup>b</sup> 0.960 <sup>c</sup> 848.233 <sup>d</sup>	<b>1.657</b> 0.674 0.983 777.248	<b>1.634</b> 0.678 0.956 784.131	<b>1.634</b> 0.683 0.950 758.584	<b>1.632</b> 0.669 0.963 778.599

Figure B1 Top 25 TanimotoCombo rated hits of BC23 with <sup>a</sup>TanimotoCombo, <sup>b</sup>ShapeTanimoto, <sup>c</sup>ColorTanimoto and <sup>d</sup>Overlap scores.

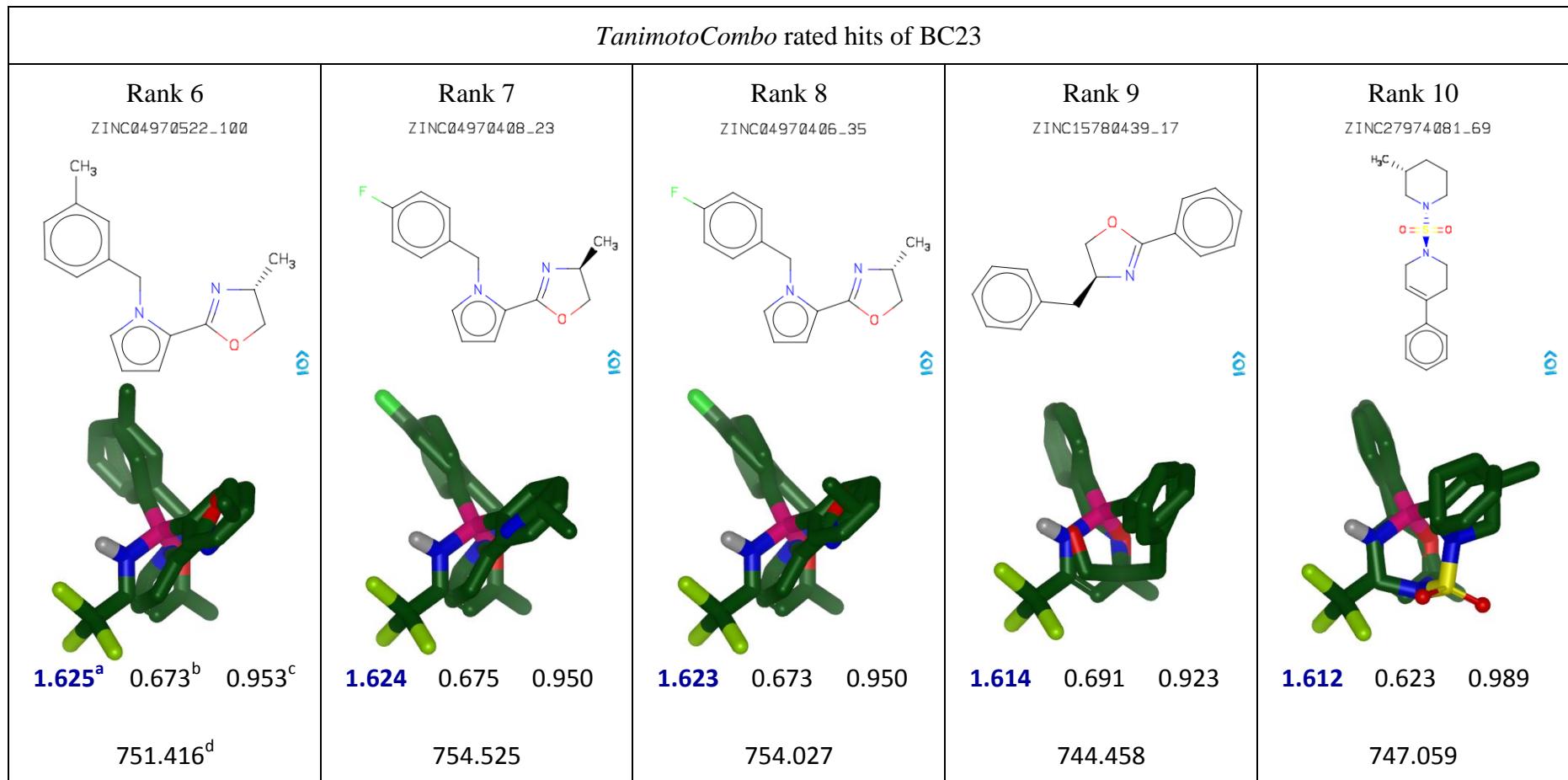


Figure B1 Top 25 *TanimotoCombo* rated hits of BC23 with <sup>a</sup>*TanimotoCombo*, <sup>b</sup>*ShapeTanimoto*, <sup>c</sup>*ColorTanimoto* and <sup>d</sup>*Overlap* scores (continued).

TanimotoCombo rated hits of BC23											
Rank 11		Rank 12		Rank 13		Rank 14		Rank 15			
ZINC27974084_83		ZINC71856536_13		ZINC72152461_12		ZINC00225590_4		ZINC00156167_2			
<b>1.609<sup>a</sup></b>	<b>0.621<sup>b</sup></b>	<b>0.988<sup>c</sup></b>	<b>1.600</b>	<b>0.608</b>	<b>0.993</b>	<b>1.592</b>	<b>0.671</b>	<b>0.921</b>	<b>1.587</b>	<b>0.664</b>	<b>0.923</b>
746.127 <sup>d</sup>		701.950		766.133		757.512		707.565			

Figure B1 Top 25 TanimotoCombo rated hits of BC23 with <sup>a</sup>TanimotoCombo, <sup>b</sup>ShapeTanimoto, <sup>c</sup>ColorTanimoto and <sup>d</sup>Overlap scores (continued).

TanimotoCombo rated hits of BC23										
Rank	Chemical Structure		3D Overlay		Score		Score		Score	
Rank 16 ZINC71856537_12			<b>1.574<sup>a</sup></b>	0.584 <sup>b</sup>	0.990 <sup>c</sup>					
Rank 17 ZINC05132639_1			<b>1.573</b>	0.846	0.727					
Rank 18 ZINC05663305_0			<b>1.572</b>	0.873	0.699					
Rank 19 ZINC01591149_0			<b>1.571</b>	0.873	0.698					
Rank 20 ZINC05132640_2			<b>1.568</b>	0.845	0.723					

Figure B1 Top 25 TanimotoCombo rated hits of BC23 with <sup>a</sup>TanimotoCombo, <sup>b</sup>ShapeTanimoto, <sup>c</sup>ColorTanimoto and <sup>d</sup>Overlap scores (continued).

TanimotoCombo rated hits of BC23										
Rank 21		Rank 22		Rank 23		Rank 24		Rank 25		
ZINC01675359_3		ZINC37629199_3		ZINC23398363_131		ZINC03627349_0		ZINC03627351_3		
1.564 <sup>a</sup>	0.848 <sup>b</sup>	0.716 <sup>c</sup>		1.562	0.843	0.719		1.559	0.842	0.717
873.180 <sup>d</sup>				880.390				870.435		
				693.153				864.211		

Figure B1 Top 25 TanimotoCombo rated hits of BC23 with <sup>a</sup>TanimotoCombo, <sup>b</sup>ShapeTanimoto, <sup>c</sup>ColorTanimoto and <sup>d</sup>Overlap scores (continued).

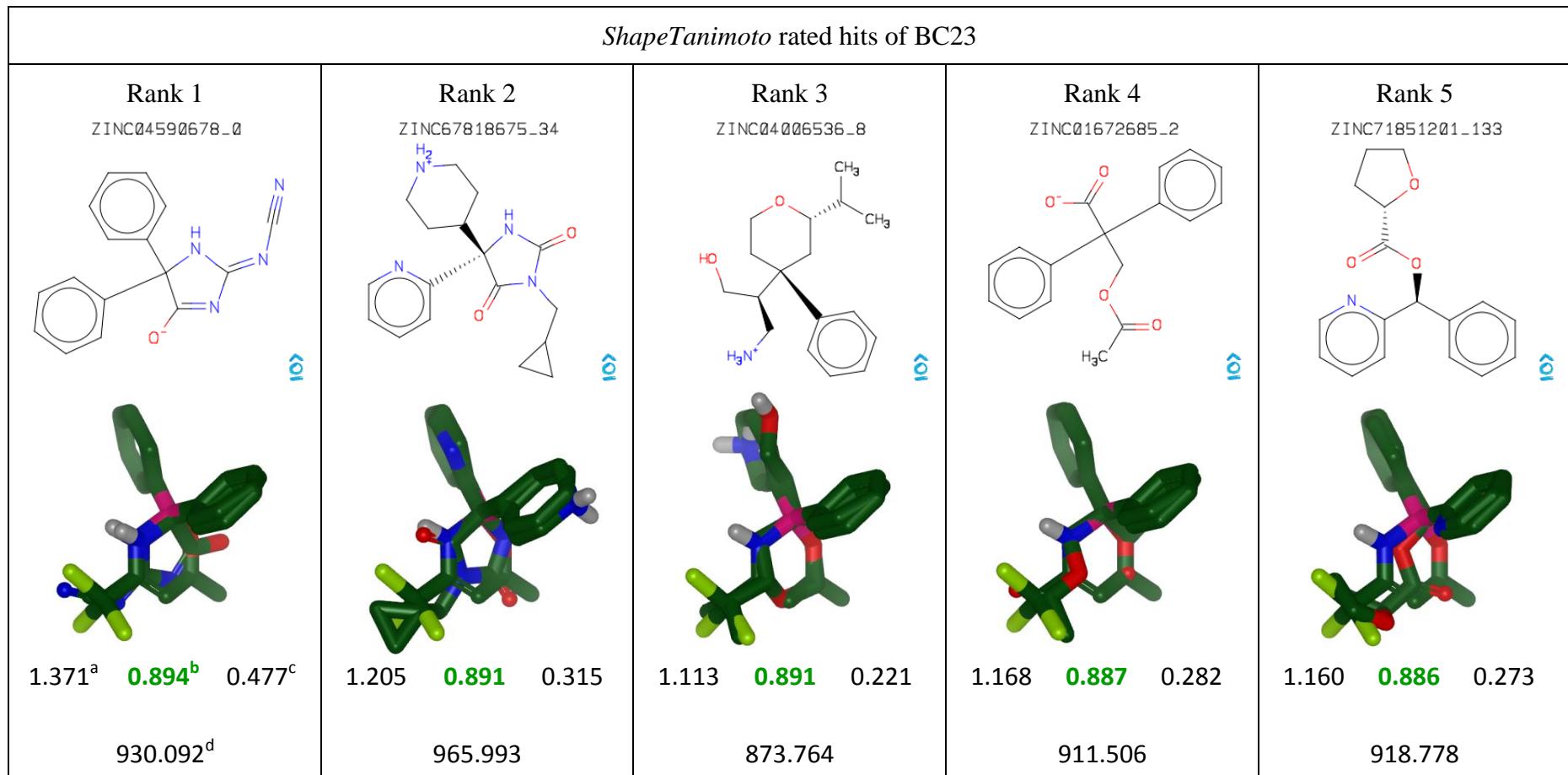


Figure B2 Top 25 *ShapeTanimoto* rated hits of BC23 with <sup>a</sup>*TanimotoCombo*, <sup>b</sup>*ShapeTanimoto*, <sup>c</sup>*ColorTanimoto* and <sup>d</sup>*Overlap* scores.

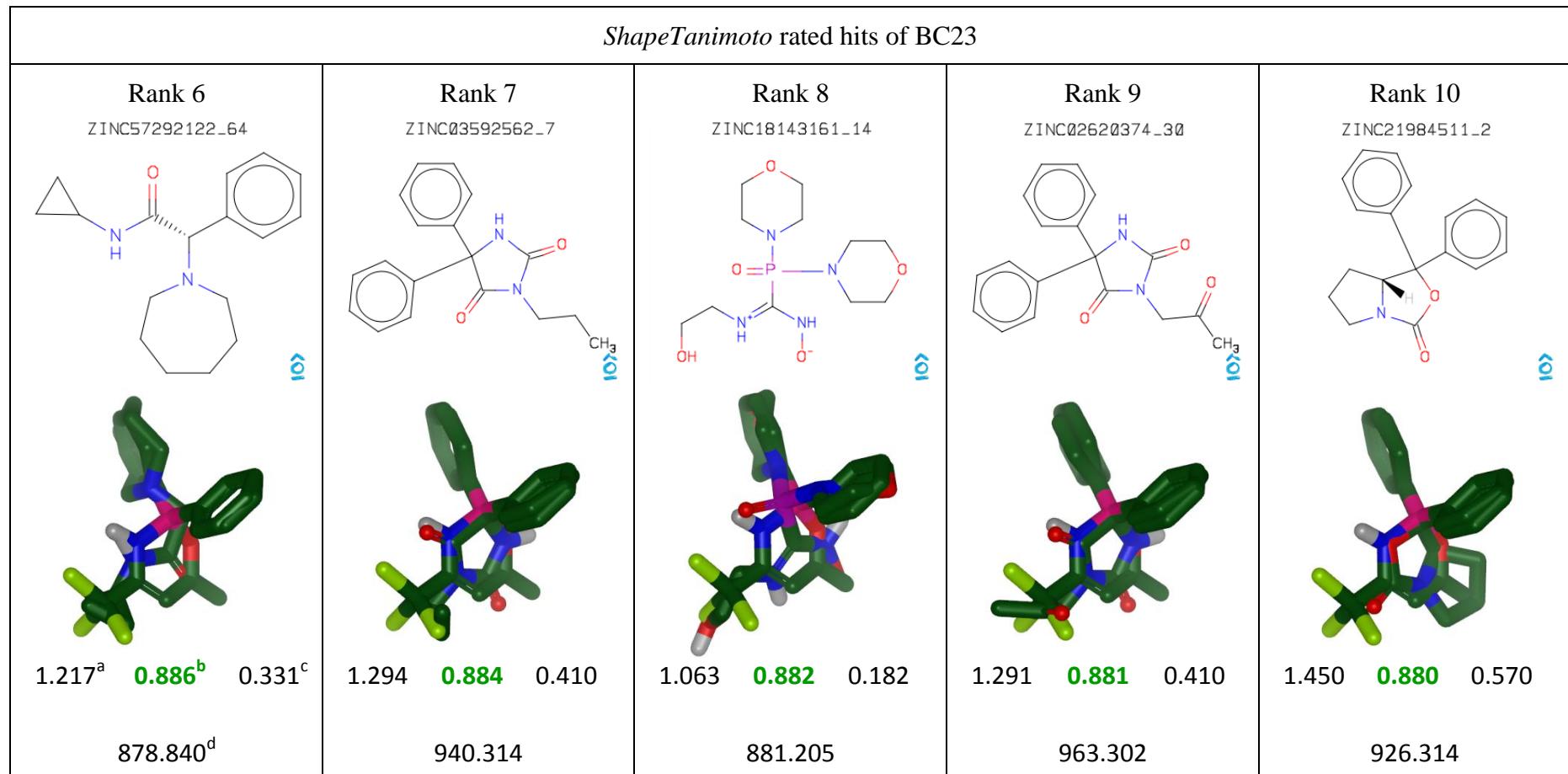


Figure B2 Top 25 *ShapeTanimoto* rated hits of BC23 with <sup>a</sup>TanimotoCombo, <sup>b</sup>ShapeTanimoto, <sup>c</sup>ColorTanimoto and <sup>d</sup>Overlap scores (continued).

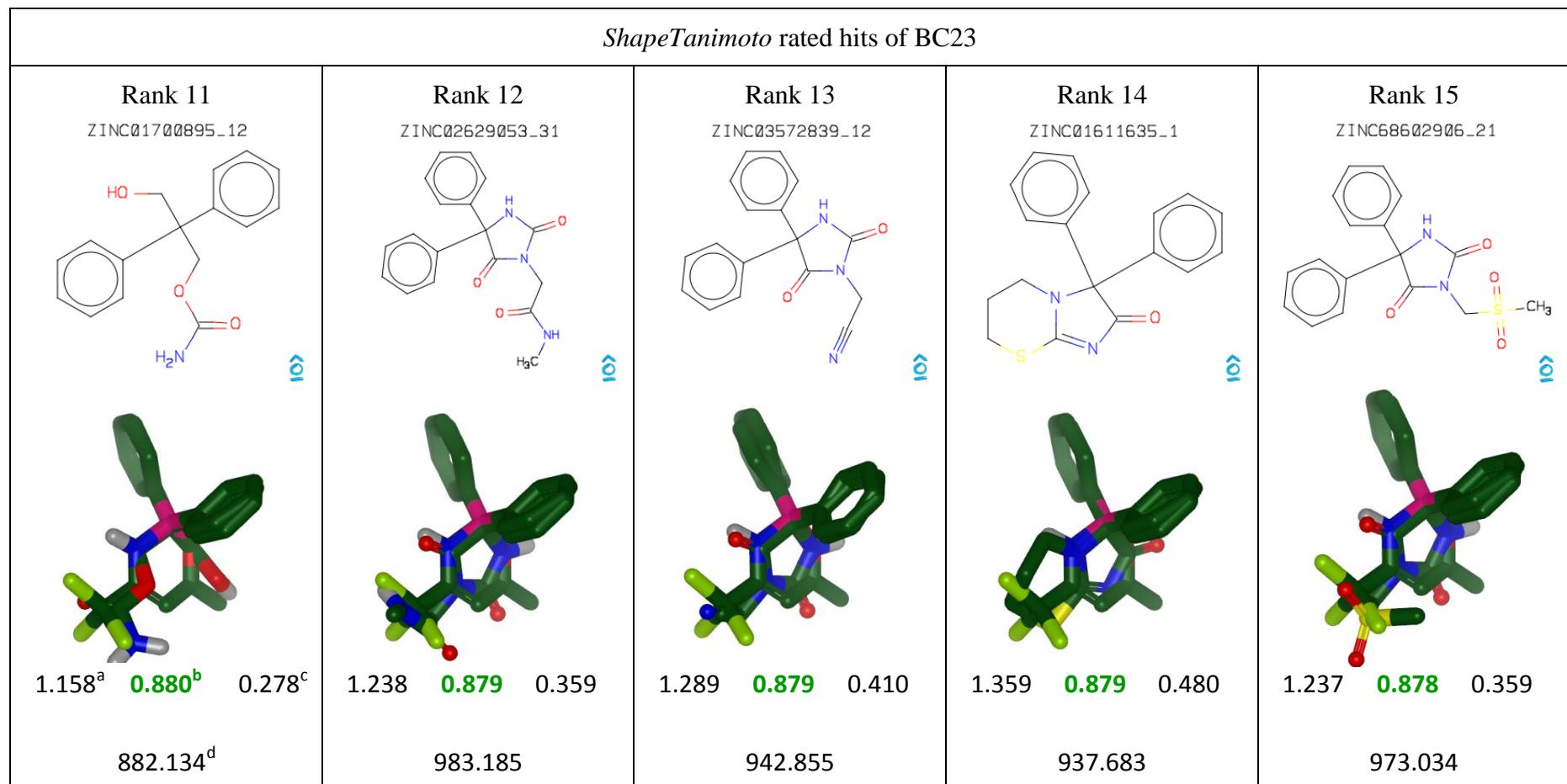


Figure B2 Top 25 *ShapeTanimoto* rated hits of BC23 with <sup>a</sup>TanimotoCombo, <sup>b</sup>ShapeTanimoto, <sup>c</sup>ColorTanimoto and <sup>d</sup>Overlap scores (continued).

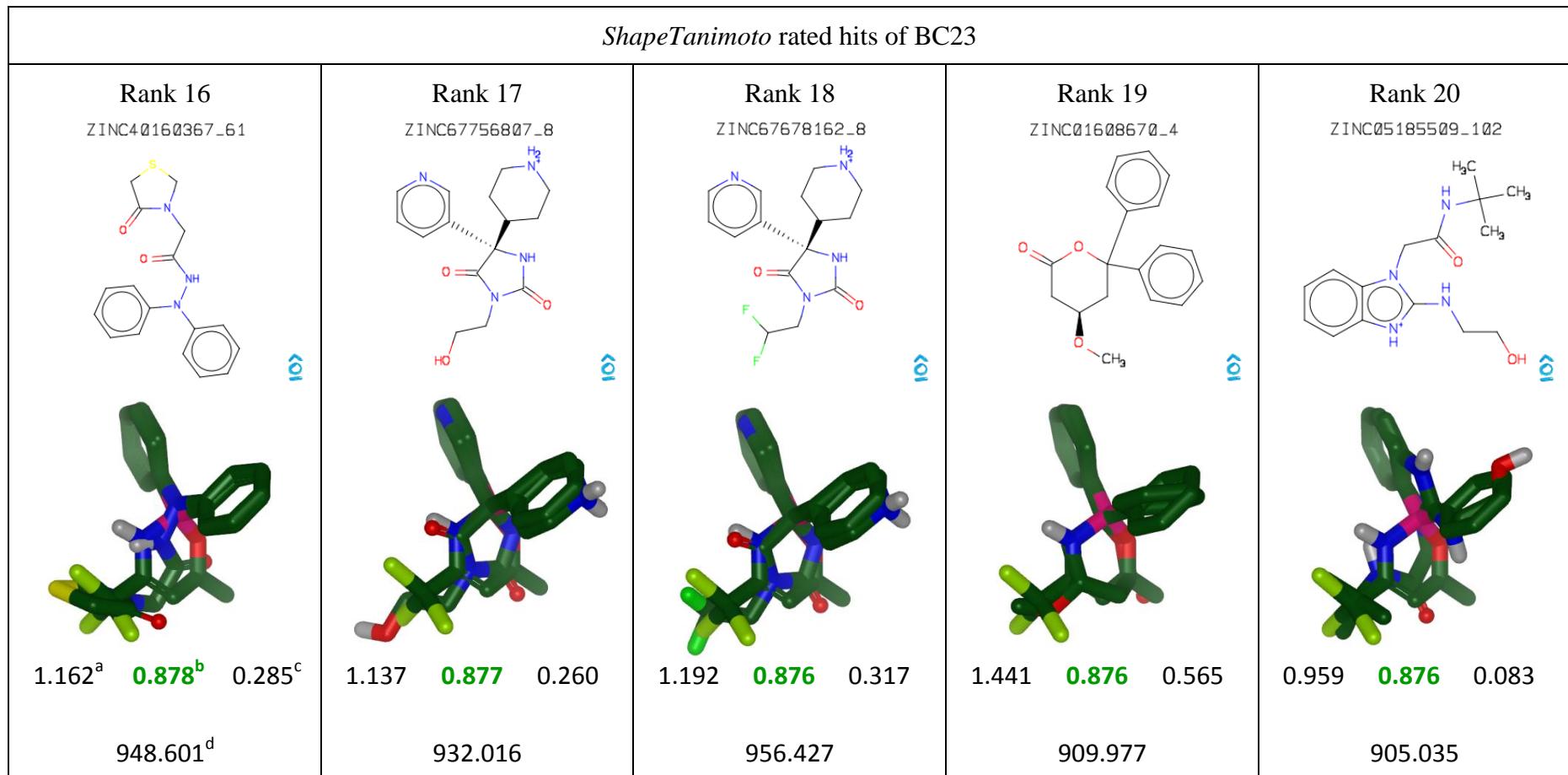


Figure B2 Top 25 *ShapeTanimoto* rated hits of BC23 with <sup>a</sup>TanimotoCombo, <sup>b</sup>ShapeTanimoto, <sup>c</sup>ColorTanimoto and <sup>d</sup>Overlap scores (continued).

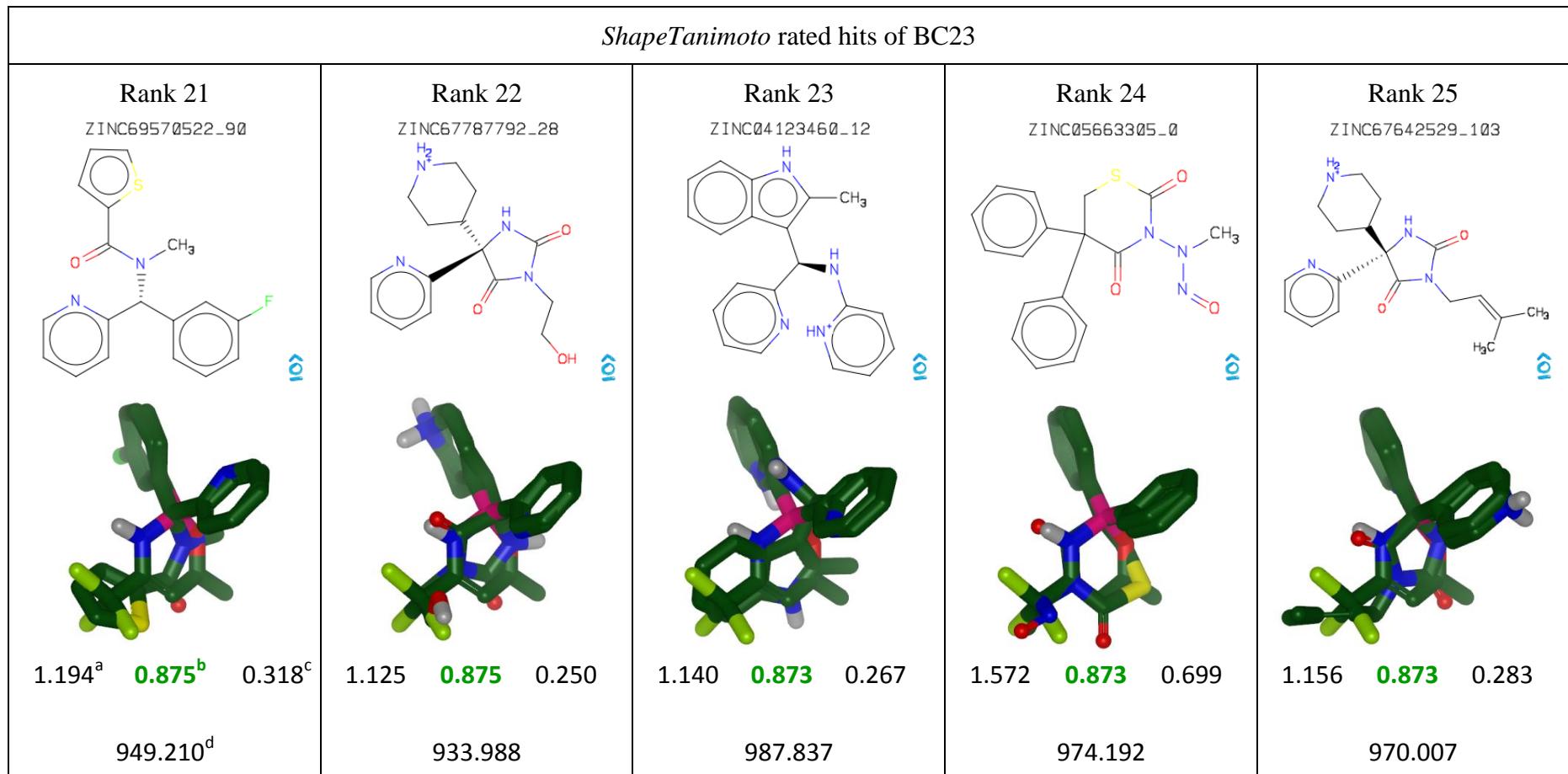


Figure B2 Top 25 *ShapeTanimoto* rated hits of BC23 with <sup>a</sup>TanimotoCombo, <sup>b</sup>ShapeTanimoto, <sup>c</sup>ColorTanimoto and <sup>d</sup>Overlap scores (continued).

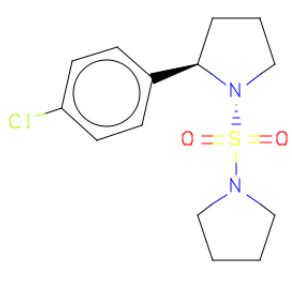
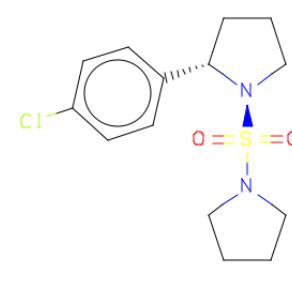
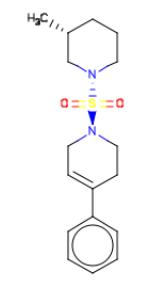
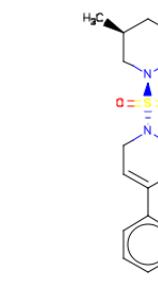
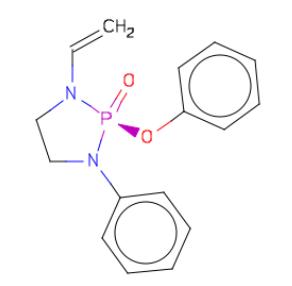
ColorTanimoto rated hits of BC23				
Rank 1	Rank 2	Rank 3	Rank 4	Rank 5
ZINC71856537_14 	ZINC71856536_15 	ZINC27974081_69 	ZINC27974084_73 	ZINC06002617_3 
1.545 <sup>a</sup> 0.552 <sup>b</sup> <b>0.993<sup>c</sup></b> 660.630 <sup>d</sup>	1.539 0.545 <b>0.993</b> 655.443	1.612 0.623 <b>0.989</b> 747.059	1.547 0.559 <b>0.988</b> 697.920	1.657 0.674 <b>0.983</b> 777.238

Figure B3 Top 25 ColorTanimoto rated hits of BC23 with <sup>a</sup>TanimotoCombo, <sup>b</sup>ShapeTanimoto, <sup>c</sup>ColorTanimoto and <sup>d</sup>Overlap scores.

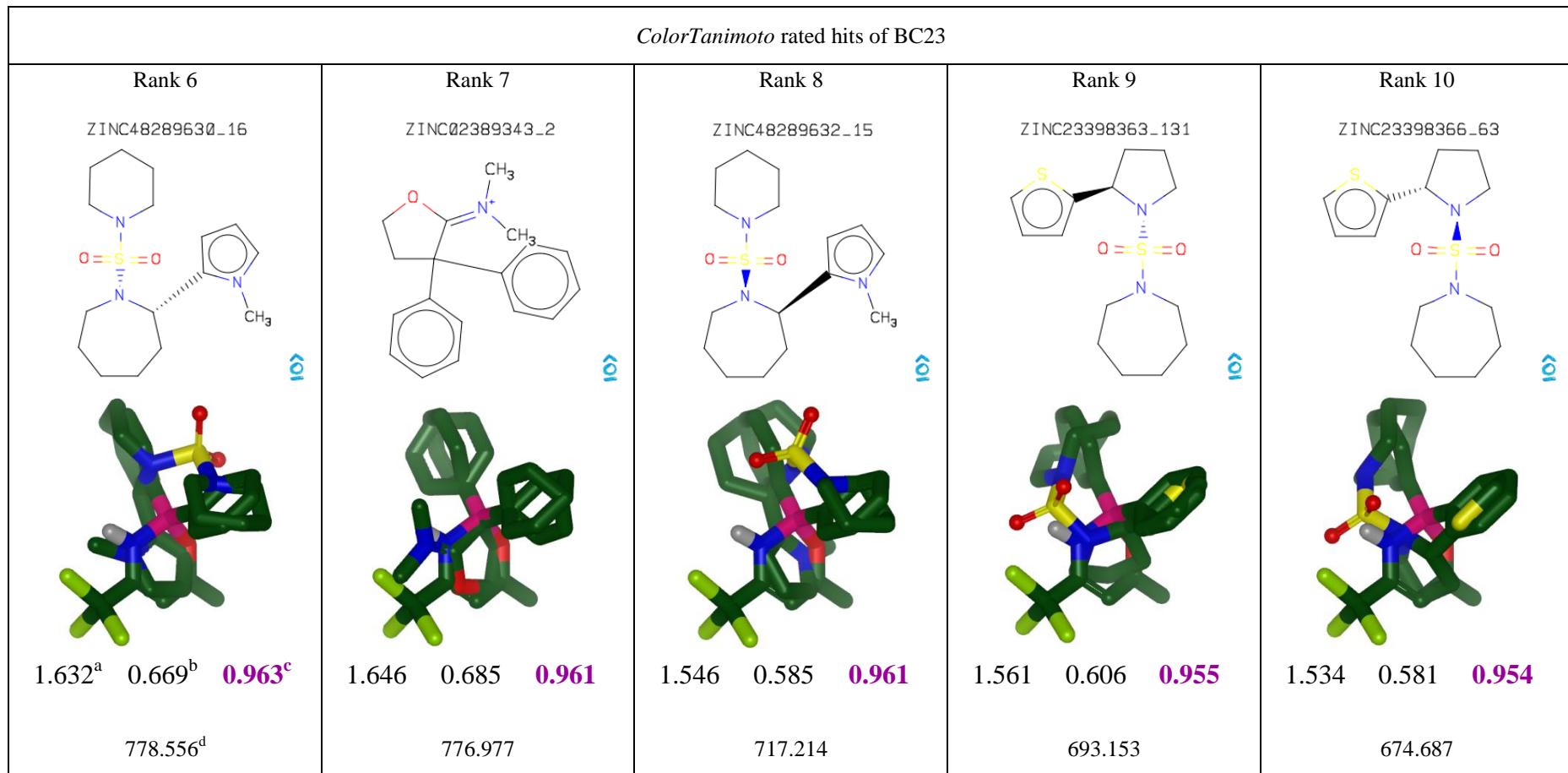


Figure B3 Top 25 *ColorTanimoto* rated hits of BC23 with <sup>a</sup>TanimotoCombo, <sup>b</sup>ShapeTanimoto, <sup>c</sup>ColorTanimoto and <sup>d</sup>Overlap scores (continued).

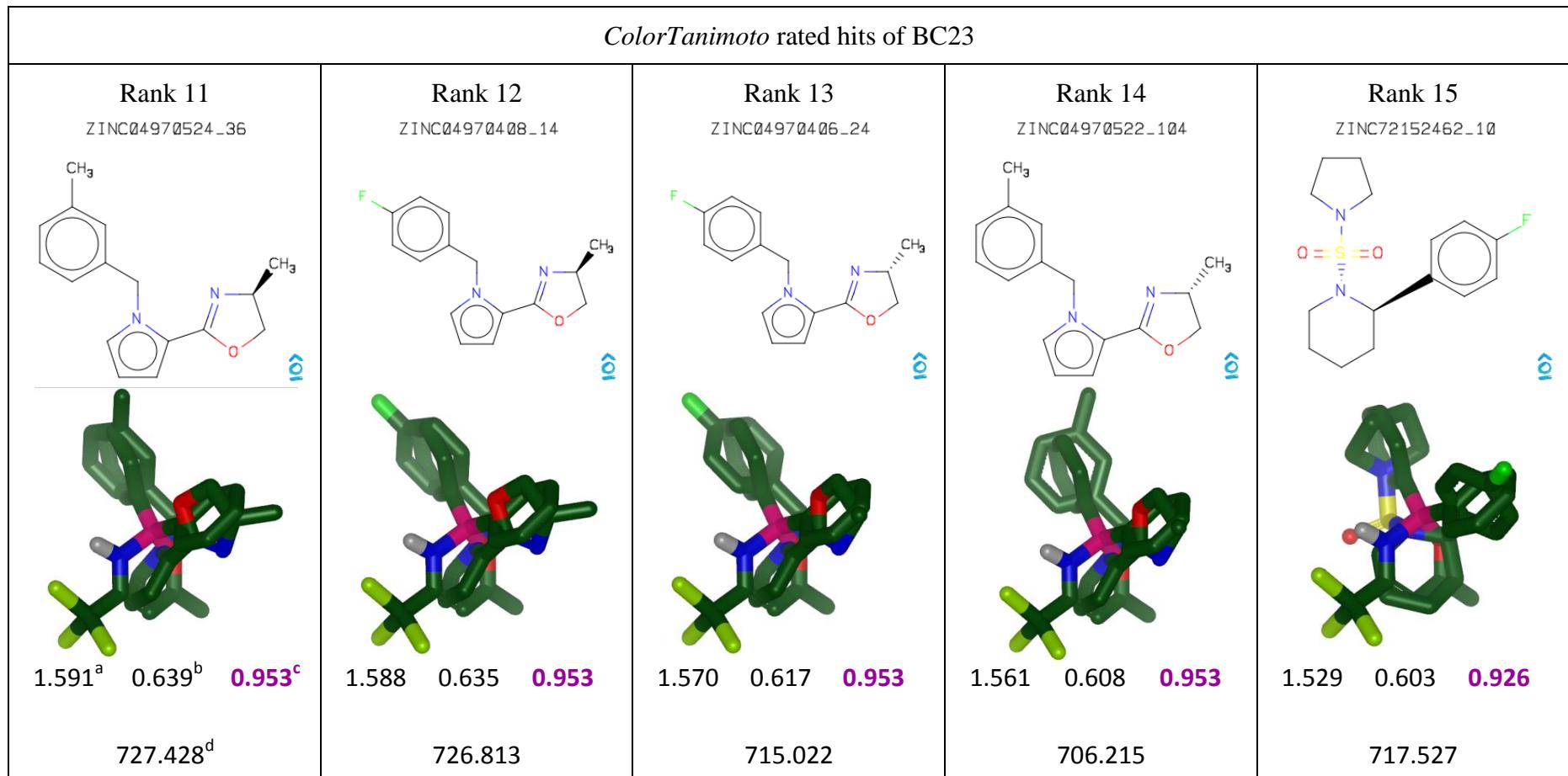


Figure B3 Top 25 *ColorTanimoto* rated hits of BC23 with <sup>a</sup>TanimotoCombo, <sup>b</sup>ShapeTanimoto, <sup>c</sup>ColorTanimoto and <sup>d</sup>Overlap scores (continued).

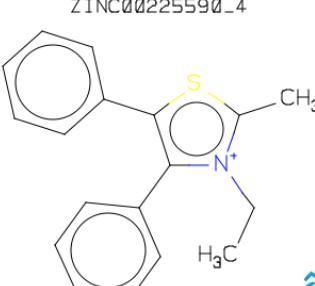
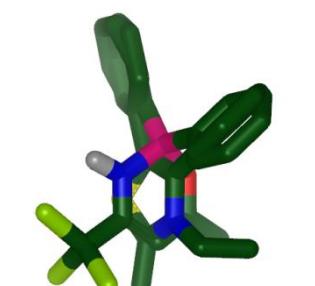
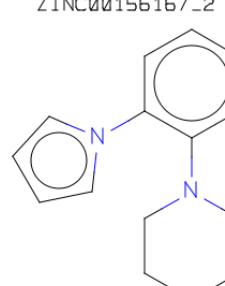
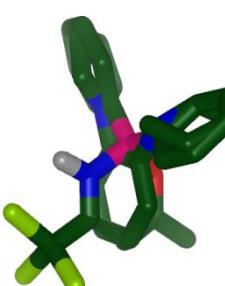
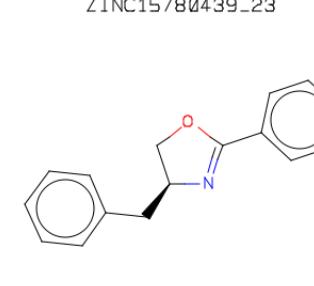
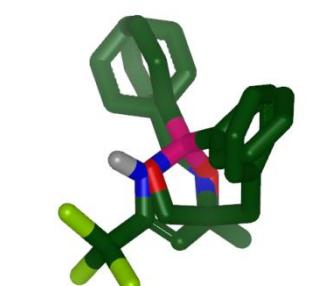
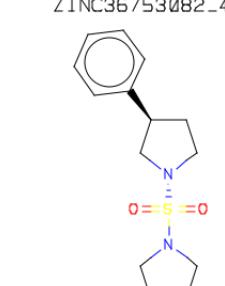
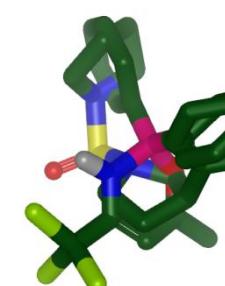
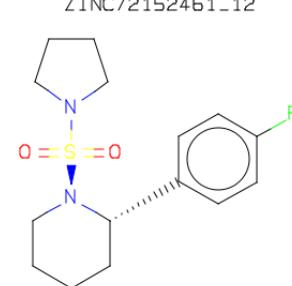
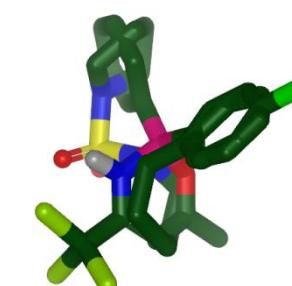
ColorTanimoto rated hits of BC23				
Rank 16 ZINC00225590_4  	Rank 17 ZINC00156167_2  	Rank 18 ZINC15780439_23  	Rank 19 ZINC36753082_48  	Rank 20 ZINC72152461_12  
1.590 <sup>a</sup> 0.666 <sup>b</sup> <b>0.924<sup>c</sup></b> 757.417 <sup>d</sup>	1.587 0.664 <b>0.923</b> 707.565	1.546 0.623 <b>0.923</b> 699.105	1.457 0.535 <b>0.922</b> 634.716	1.592 0.671 <b>0.921</b> 766.133

Figure B3 Top 25 ColorTanimoto rated hits of BC23 with <sup>a</sup>TanimotoCombo, <sup>b</sup>ShapeTanimoto, <sup>c</sup>ColorTanimoto and <sup>d</sup>Overlap scores (continued).

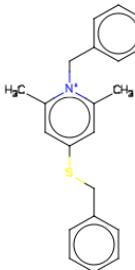
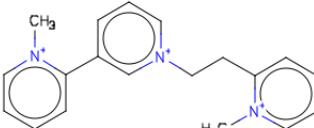
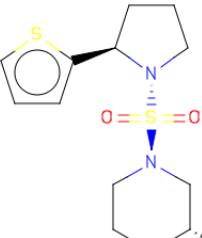
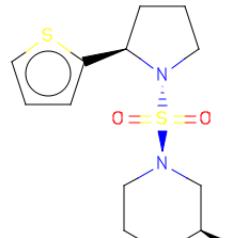
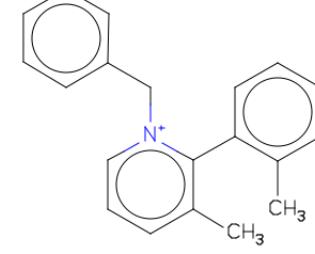
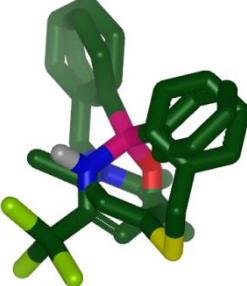
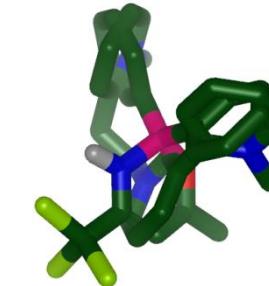
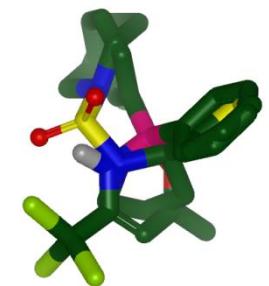
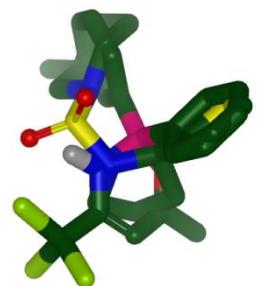
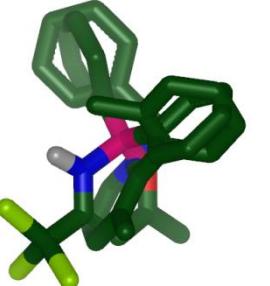
ColorTanimoto rated hits of BC23				
<b>Rank 21</b> ZINC01664747_42 	<b>Rank 22</b> ZINC15780508_31 	<b>Rank 23</b> ZINC27954343_28 	<b>Rank 24</b> ZINC27954347_27 	<b>Rank 25</b> ZINC05422905_7 
1.516 <sup>a</sup> 0.600 <sup>b</sup> <b>0.917<sup>c</sup></b> 757.303 <sup>d</sup>	1.460 0.545 <b>0.915</b> 708.645	1.533 0.631 <b>0.903</b> 711.567	1.520 0.618 <b>0.902</b> 702.542	1.469 0.578 <b>0.891</b> 717.463
				

Figure B3 Top 25 ColorTanimoto rated hits of BC23 with <sup>a</sup>TanimotoCombo, <sup>b</sup>ShapeTanimoto, <sup>c</sup>ColorTanimoto and <sup>d</sup>Overlap scores (continued).

## Appendix C – MC acceptance rates of a test case by using the QM/MM MC docking scheme

**Table C1** Acceptance rate of sulfamoyl methanol binding to hCAII (i) 5.0 bohrs of  $\Delta x_t$  and (ii) 10.0 bohrs of  $\Delta x_t$  on 2k MC trajectories using  $vdw_{cutoff}$  at 2.5, 3.0, 3.5 and 4.0 bohrs and  $d_{max}$  at 2.0, 5.0 and 10.0 bohrs.

van der Waals cutoff /(Å)	$d_{max}$ (Bohrs)	Translational step size ( $\Delta x_t$ )			
		5.0 bohrs		10.0 bohrs	
		number of accepted poses	number of van der Waals clashes	number of accepted poses	number of van der Waals clashes
2.5	2	406 (20.3%)	140	493 (24.7%)	249
	5	512 (25.6%)	80	259 (13.0%)	282
	10	263 (13.2%)	98	222 (11.1%)	227
3.0	2	504 (25.2%)	87	430 (21.5%)	384
	5	420 (21.0%)	228	509 (25.5%)	394
	10	212 (10.6%)	491	482 (24.1%)	372
3.5	2	566 (28.3%)	791	507 (25.4%)	817
	5	656 (32.8%)	1559	511 (25.6%)	1057
	10	304 (15.2%)	849	225 (11.3%)	1437
4.0	2	414(20.7%)	1937	677 (33.9%)	1186
	5	539(27.0%)	707	1 (0.1%)	6755
	10	243(12.2%)	668	279 (14.0%)	509

**Table C2 Acceptance rate of sulfamoyl methanol binding to hCAII (i) 5.0 bohrs of  $\Delta x_t$  and (ii) 10.0 bohrs of  $\Delta x_t$  on MC trajectories of 1k, 2k and 4k using  $d_{max}$  at 2.0, 5.0 and 10.0 bohrs;  $vdw_{cutoff}$  was set at 3.0 bohrs.**

Number of MC steps $/(10^3)$	$d_{max}$ /(Bohrs)	Translational step size ( $\Delta x_t$ )			
		5 bohrs		10 bohrs	
		number of accepted poses	number of van der Waals clashes	number of accepted poses	number of van der Waals clashes
1	2	248 (24.8%)	348 (34.8%)	213 (21.3%)	156 (15.6%)
	5	212 (21.2%)	126 (12.6%)	215 (21.5%)	200 (20.0%)
	10	200 (20.0%)	315 (31.5%)	200 (20.0%)	349 (34.9%)
2	2	504 (25.2%)	87 (4.4%)	430 (21.5%)	384 (19.2%)
	5	420 (21.0%)	228 (11.4%)	509 (25.5%)	394 (19.7%)
	10	212 (10.6%)	491 (24.6%)	415 (20.6%)	388 (19.4%)
4	2	881 (22.0%)	774 (19.4%)	987 (24.7%)	1049 (26.2%)
	5	1090 (27.3%)	1000 (25.0%)	851 (21.3%)	587 (14.7%)
	10	356 (8.9%)	1124 (28.1%)	638 (16.0%)	740 (18.5%)

## Appendix D - $\beta$ CD geometries, diol geometries and their computed interactions

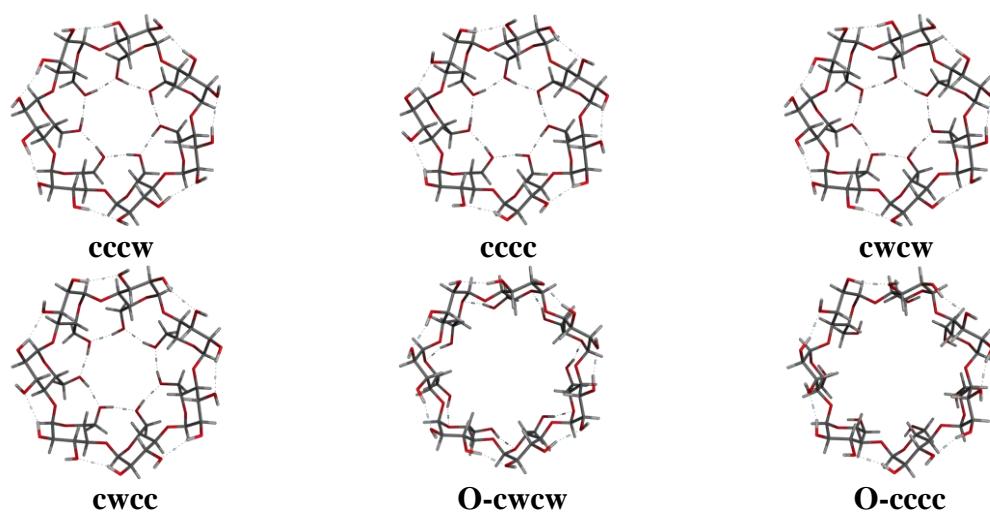


Figure D1 View of  $\beta$ CD from C2C3 rim.

**Table D1 Geometrical data for  $\beta$ CD structures optimised at the M06-2X/def2-TZVPP level of theory.<sup>a</sup>**

Coordinate	cccw	cccc	ewew	ewcc	O-ewew	O-cccc	XRD <sup>b</sup>
$\alpha$ (O4O4'O4'')	128.6 (0.31)	128.5 (0.62)	128.6 (0.10)	128.6 (0.19)	128.6 (0.14)	128.5 (0.17)	128 $\pm$ 2
$\theta$ (C1O4'C4')	118.9 (0.05)	119.1 (0.12)	118.7 (0.03)	118.9 (0.04)	119.9 (0.03)	119.5 (0.06)	118 $\pm$ 1
$\Phi$ (C2C1O4'C4')	221.8 (0.52)	222.2 (1.00)	220.8 (0.27)	221.3 (0.35)	232.9 (0.50)	232.6 (2.66)	231 $\pm$ 6
$\tau$ (O4O4'O4''O4'')	0.0 (0.42)	0.0 (1.20)	0.0 (0.31)	0.0 (0.54)	0.0 (0.86)	0.0 (2.70)	0.2 $\pm$ 9
$\Gamma$ (C1C2C3C4)	308.6 (0.09)	307.4 (0.13)	308.9 (0.05)	307.6 (0.05)	305.4 (0.07)	304.0 (0.20)	306 $\pm$ 3
$\Theta$ (C2C3C4C5)	55.0 (0.04)	56.4 (0.12)	55.0 (0.05)	56.4 (0.07)	53.2 (0.15)	55.1 (0.15)	55 $\pm$ 3
T (C3C4C5O5)	300.4 (0.17)	300.2 (0.37)	300.3 (0.06)	300.1 (0.11)	306.3 (0.27)	305.4 (0.26)	304 $\pm$ 4
9 (O4C4C5O5)	183.0 (0.26)	182.2 (0.57)	183.5 (0.09)	182.6 (0.15)	188.2 (0.24)	186.5 (0.31)	188 $\pm$ 4
Z (O4C4C5C6)	62.4 (0.33)	61.5 (0.66)	63.2 (0.12)	62.2 (0.20)	71.6 (0.33)	68.5 (0.29)	69 $\pm$ 5
$\Lambda$ (O4'C1O5C5)	57.9 (0.20)	57.9 (0.38)	58.2 (0.04)	58.1 (0.07)	59.6 (0.14)	59.2 (0.07)	59 $\pm$ 2
$\Pi$ (C4C5O5C1)	65.2 (0.21)	64.1 (0.43)	65.6 (0.06)	64.6 (0.08)	60.5 (0.16)	59.0 (0.26)	59 $\pm$ 3
K (O2C2C1O5)	176.8 (0.05)	178.1 (0.06)	176.2 (0.03)	177.4 (0.05)	179.3 (0.07)	181.1 (0.19)	178 $\pm$ 3
Y (O3C3C4C5)	126.1 (0.06)	177.1 (0.08)	126.3 (0.05)	177.3 (0.06)	122.7 (0.13)	174.9 (0.24)	176 $\pm$ 2
E (O4'C1C4C5)	288.2 (0.11)	287.0 (0.23)	288.7 (0.04)	287.5 (0.08)	284.9 (0.18)	282.9 (0.30)	285 $\pm$ 3
$\Omega$ (O5C5C6O6)	118.9 (1.26)	118.1 (2.42)	96.9 (0.49)	96.3 (0.66)	167.8 (0.69)	59.2 (0.45)	198 $\pm$ 123
$\Xi$ (C4C5C6O6)	238.9 (1.25)	238.7 (2.37)	216.8 (0.50)	216.7 (0.67)	286.8 (0.69)	179.5 0.01	112 $\pm$ 70
$\psi$ (O2C2C3O3)	65.6 (0.08)	62.4 (0.15)	65.7 (0.04)	62.5 (0.04)	61.5 (0.11)	58.9 (0.22)	63.6 $\pm$ 2.1

<sup>a</sup>The angles are the mean values given in degrees. The standard errors of mean values are in parentheses.

<sup>b</sup>Experimental X-ray diffraction (XRD) data from Heine *et al.*<sup>228</sup>, Lindner *et. al.*<sup>229</sup>

**Table D2 Intramolecular OH bond distances (in Å) at C2C3 rim of  $\beta$ CD geometries optimised at different levels of theory.**

$\beta$ CD	M06-2X/def2-TZVPP								DFTB3							
	glu1	glu2	glu3	glu4	glu5	glu6	glu7	avr	glu1	glu2	glu3	glu4	glu5	glu6	glu7	avr
cccw	2.11	2.09	2.13	2.11	2.08	2.12	2.13	<b>2.11</b>	2.03	2.03	2.03	2.03	2.03	2.03	2.03	<b>2.03</b>
cccc	2.14	2.15	2.09	2.11	2.17	2.09	2.13	<b>2.12</b>	2.04	2.03	2.04	2.04	2.04	2.04	2.03	<b>2.04</b>
cwcw	2.13	2.12	2.14	2.12	2.13	2.14	2.14	<b>2.13</b>	2.04	2.04	2.04	2.04	2.04	2.04	2.04	<b>2.04</b>
cwcc	2.15	2.15	2.14	2.13	2.15	2.13	2.14	<b>2.14</b>	2.05	2.05	2.05	2.05	2.05	2.04	2.05	<b>2.05</b>
O-cwcw	1.90	1.87	1.88	1.89	1.89	1.88	1.87	<b>1.88</b>	1.82	1.82	1.82	1.82	1.82	1.82	1.82	<b>1.82</b>
O-cccc	1.96	1.92	1.93	1.95	1.92	1.96	1.92	<b>1.94</b>	1.83	1.83	1.84	1.86	1.85	1.83	1.80	<b>1.83</b>
$\beta$ CD	PM7								PM6-DH2							
	glu1	glu2	glu3	glu4	glu5	glu6	glu7	avr	glu1	glu2	glu3	glu4	glu5	glu6	glu7	avr
cccw	2.07	2.07	2.07	2.07	2.07	2.07	2.07	<b>2.07</b>	1.88	1.89	1.89	1.88	1.89	1.89	1.89	<b>1.89</b>
cccc	2.06	2.06	2.05	2.06	2.06	2.05	2.05	<b>2.06</b>	1.89	1.89	1.89	1.89	1.89	1.89	1.89	<b>1.89</b>
cwcw	2.08	2.08	2.08	2.07	2.08	2.08	2.07	<b>2.08</b>	1.89	1.89	1.89	1.89	1.89	1.89	1.89	<b>1.89</b>
cwcc	2.06	2.06	2.06	2.06	2.06	2.06	2.06	<b>2.06</b>	1.90	1.89	1.89	1.90	1.89	1.89	1.89	<b>1.89</b>
O-cwcw	1.71	1.70	1.70	1.70	1.70	1.70	1.70	<b>1.70</b>	1.79	1.78	1.79	1.79	1.78	1.78	1.79	<b>1.78</b>
O-cccc <sup>a</sup>	<i>1.73</i>	<i>1.73</i>	<i>1.73</i>	<i>1.73</i>	<i>1.73</i>	<i>1.73</i>	<i>1.73</i>	<i>1.73</i>	<i>1.89</i>	<i>1.89</i>	<i>1.89</i>	<i>1.90</i>	<i>1.88</i>	<i>1.95</i>	<i>1.96</i>	<b>1.91</b>
$\beta$ CD	PM6-D3								PM6							
	glu1	glu2	glu3	glu4	glu5	glu6	glu7	avr	glu1	glu2	glu3	glu4	glu5	glu6	glu7	avr
cccw	1.91	1.91	1.92	1.91	1.91	1.91	1.92	<b>1.91</b>	2.02	2.02	2.02	2.02	2.02	2.02	2.02	<b>2.02</b>
cccc	1.94	1.94	1.93	1.94	1.95	1.93	1.93	<b>1.94</b>	2.04	2.04	2.03	2.04	2.04	2.03	2.03	<b>2.04</b>
cwcw	1.91	1.91	1.91	1.91	1.91	1.91	1.91	<b>1.91</b>	2.03	2.03	2.03	2.03	2.03	2.03	2.03	<b>2.03</b>
cwcc	1.94	1.94	1.93	1.93	1.94	1.93	1.93	<b>1.93</b>	2.05	2.04	2.04	2.04	2.04	2.04	2.04	<b>2.04</b>
O-cwcw	1.77	1.77	1.77	1.77	1.77	1.77	1.77	<b>1.77</b>	1.84	1.83	1.84	1.84	1.83	1.83	1.83	<b>1.83</b>
O-cccc <sup>a</sup>	<i>1.94</i>	<i>1.93</i>	<i>1.94</i>	<i>1.97</i>	<i>1.94</i>	<i>1.93</i>	<i>1.95</i>	<b>1.94</b>	2.05	2.03	2.05	2.07	2.04	2.03	2.06	<b>2.05</b>
$\beta$ CD	PM3CARB-1								PM3-D*							
	glu1	glu2	glu3	glu4	glu5	glu6	glu7	avr	glu1	glu2	glu3	glu4	glu5	glu6	glu7	avr
cccw	1.93	1.93	1.93	1.93	1.93	1.93	1.93	<b>1.93</b>	1.86	1.86	2.62	2.43	1.85	2.61	2.44	<b>2.24</b>
cccc	1.89	1.89	1.89	1.89	1.89	1.89	1.89	<b>1.89</b>	2.53	2.68	1.86	2.61	2.70	1.86	2.59	<b>2.40</b>
cwcw	1.92	1.92	1.92	1.92	1.92	1.92	1.92	<b>1.92</b>	2.40	2.40	2.40	2.40	2.40	2.40	2.40	<b>2.40</b>
cwcc	1.89	1.89	1.89	1.89	1.88	1.89	1.89	<b>1.89</b>	2.52	2.52	2.52	2.52	2.52	2.52	2.52	<b>2.52</b>
O-cwcw	1.83	1.83	1.83	1.83	1.83	1.83	1.83	<b>1.83</b>	1.81	1.81	1.81	1.81	1.81	1.81	1.81	<b>1.81</b>
O-cccc	1.84	1.84	1.84	1.84	1.84	1.84	1.84	<b>1.84</b>	1.81	1.81	1.81	1.81	1.81	1.81	1.81	<b>1.81</b>
$\beta$ CD	PM3								AM1							
	glu1	glu2	glu3	glu4	glu5	glu6	glu7	avr	glu1	glu2	glu3	glu4	glu5	glu6	glu7	avr
cccw	1.87	1.86	2.60	1.87	1.87	1.86	2.60	<b>2.08</b>	2.13	2.14	2.13	2.14	2.14	2.13	2.13	<b>2.13</b>
cccc	2.52	2.68	1.86	2.60	2.69	1.86	2.57	<b>2.40</b>	2.16	2.16	2.15	2.16	2.16	2.15	2.15	<b>2.16</b>
cwcw	1.86	1.86	2.60	1.87	1.86	2.58	2.42	<b>2.15</b>	2.14	2.14	2.13	2.14	2.14	2.13	2.13	<b>2.14</b>
cwcc	2.52	2.51	2.51	2.51	2.52	2.51	2.51	<b>2.51</b>	2.17	2.16	2.16	2.16	2.16	2.16	2.16	<b>2.16</b>
O-cwcw	1.81	1.81	1.81	1.81	1.81	1.81	1.81	<b>1.81</b>	2.15	2.15	2.14	2.14	2.14	2.13	2.14	<b>2.14</b>
O-cccc <sup>a</sup>	<i>1.81</i>	<i>1.81</i>	<i>1.81</i>	<i>1.81</i>	<i>1.81</i>	<i>1.81</i>	<i>1.81</i>	<b>1.81</b>	<i>2.18</i>	<i>2.16</i>	<i>2.17</i>	<i>2.18</i>	<i>2.14</i>	<i>2.19</i>	<i>2.16</i>	<b>2.17</b>

<sup>a</sup>Collapsed open structure labelled in *italics*

**Table D3 Intramolecular OH bond distances (in Å) at C6 rim of  $\beta$ CD geometries optimised at different levels of theory.**

$\beta$ CD	M06-2X/def2-TZVPP								DFTB3							
	glu1	glu2	glu3	glu4	glu5	glu6	glu7	avr	glu1	glu2	glu3	glu4	glu5	glu6	glu7	avr
cccw	1.82	1.83	1.83	1.84	1.83	1.82	1.83	<b>1.83</b>	1.80	1.80	1.80	1.80	1.80	1.80	1.80	<b>1.80</b>
cccc	1.83	1.84	1.84	1.85	1.85	1.83	1.84	<b>1.84</b>	1.80	1.80	1.80	1.80	1.80	1.80	1.80	<b>1.80</b>
cwcw	1.86	1.87	1.88	1.87	1.87	1.88	1.87	<b>1.87</b>	1.80	1.80	1.80	1.80	1.80	1.80	1.80	<b>1.80</b>
cwcc	1.86	1.88	1.89	1.87	1.88	1.89	1.88	<b>1.88</b>	1.79	1.79	1.79	1.79	1.79	1.79	1.79	<b>1.79</b>
O-cwcw	1.90	1.92	1.90	1.90	1.90	1.91	1.91	<b>1.90</b>	1.92	1.92	1.92	1.92	1.92	1.92	1.92	<b>1.92</b>
O-cccc	3.91	3.90	3.89	3.79	3.90	3.98	4.02	<b>3.91</b>	4.80	4.67	4.90	4.25	4.42	4.91	5.43	<b>4.77</b>
$\beta$ CD	PM7								PM6-DH2							
	glu1	glu2	glu3	glu4	glu5	glu6	glu7	avr	glu1	glu2	glu3	glu4	glu5	glu6	glu7	avr
cccw	1.70	1.70	1.70	1.70	1.70	1.70	1.70	<b>1.70</b>	1.79	1.79	1.79	1.79	1.79	1.79	1.79	<b>1.79</b>
cccc	1.70	1.70	1.70	1.70	1.70	1.70	1.70	<b>1.70</b>	1.79	1.80	1.79	1.79	1.80	1.79	1.79	<b>1.79</b>
cwcw	1.70	1.70	1.70	1.70	1.70	1.70	1.70	<b>1.70</b>	1.79	1.79	1.79	1.79	1.79	1.79	1.79	<b>1.79</b>
cwcc	1.70	1.70	1.70	1.70	1.70	1.70	1.70	<b>1.70</b>	1.79	1.79	1.79	1.79	1.79	1.79	1.79	<b>1.79</b>
O-cwcw	1.73	1.73	1.73	1.73	1.73	1.73	1.73	<b>1.73</b>	1.94	1.95	1.94	1.95	1.95	1.94	1.94	<b>1.94</b>
O-cccc <sup>a</sup>	2.05	2.05	2.06	2.06	2.05	2.05	2.06	<b>2.05</b>	1.77	1.79	1.81	1.80	1.78	1.77	1.78	<b>1.78</b>
$\beta$ CD	PM6-D3								PM6							
	glu1	glu2	glu3	glu4	glu5	glu6	glu7	avr	glu1	glu2	glu3	glu4	glu5	glu6	glu7	avr
cccw	1.77	1.77	1.77	1.77	1.77	1.77	1.77	<b>1.77</b>	1.86	1.86	1.86	1.86	1.86	1.86	1.86	<b>1.86</b>
cccc	1.78	1.78	1.77	1.78	1.77	1.77	1.77	<b>1.77</b>	1.87	1.87	1.86	1.87	1.87	1.87	1.87	<b>1.87</b>
cwcw	1.75	1.76	1.76	1.75	1.75	1.76	1.75	<b>1.75</b>	1.82	1.83	1.83	1.82	1.83	1.83	1.82	<b>1.83</b>
cwcc	1.75	1.77	1.76	1.75	1.76	1.76	1.76	<b>1.76</b>	1.83	1.85	1.84	1.83	1.85	1.83	1.83	<b>1.84</b>
O-cwcw	1.88	1.88	1.87	1.87	1.88	1.88	1.87	<b>1.88</b>	1.99	2.01	1.99	1.99	2.00	1.99	1.99	<b>1.99</b>
O-cccc <sup>a</sup>	1.77	1.75	1.78	1.78	1.75	1.77	1.78	<b>1.77</b>	1.86	1.85	1.87	1.88	1.85	1.87	1.88	<b>1.87</b>
$\beta$ CD	PM3CARB-1								PM3-D*							
	glu1	glu2	glu3	glu4	glu5	glu6	glu7	avr	glu1	glu2	glu3	glu4	glu5	glu6	glu7	avr
cccw	1.86	1.86	1.85	1.85	1.86	1.86	1.85	<b>1.86</b>	1.83	1.82	1.83	1.83	1.82	1.84	1.85	<b>1.83</b>
cccc	1.86	1.86	1.86	1.86	1.86	1.86	1.86	<b>1.86</b>	1.82	1.82	1.84	1.83	1.82	1.84	1.83	<b>1.83</b>
cwcw	1.85	1.85	1.85	1.85	1.85	1.85	1.85	<b>1.85</b>	1.82	1.82	1.82	1.82	1.82	1.82	1.82	<b>1.82</b>
cwcc	1.85	1.85	1.85	1.85	1.85	1.85	1.85	<b>1.85</b>	1.83	1.83	1.83	1.83	1.83	1.83	1.83	<b>1.83</b>
O-cwcw	1.93	1.93	1.93	1.93	1.93	1.93	1.93	<b>1.93</b>	1.86	1.86	1.86	1.86	1.86	1.86	1.86	<b>1.86</b>
O-cccc	2.57	2.55	2.57	2.57	2.55	2.56	2.57	<b>2.56</b>	3.39	3.39	3.39	3.39	3.39	3.39	3.39	<b>3.39</b>
$\beta$ CD	PM3								AM1							
	glu1	glu2	glu3	glu4	glu5	glu6	glu7	avr	glu1	glu2	glu3	glu4	glu5	glu6	glu7	avr
cccw	1.83	1.85	1.86	1.85	1.82	1.84	1.85	<b>1.84</b>	2.15	2.15	2.15	2.15	2.15	2.14	2.15	<b>2.15</b>
cccc	1.82	1.82	1.84	1.83	1.82	1.85	1.83	<b>1.83</b>	2.16	2.16	2.16	2.16	2.17	2.16	2.15	<b>2.16</b>
cwcw	1.85	1.84	1.83	1.85	1.84	1.82	1.83	<b>1.84</b>	2.14	2.14	2.15	2.14	2.14	2.15	2.15	<b>2.14</b>
cwcc	1.83	1.83	1.83	1.83	1.83	1.83	1.83	<b>1.83</b>	2.15	2.16	2.16	2.15	2.15	2.16	2.16	<b>2.16</b>
O-cwcw	1.86	1.86	1.86	1.86	1.86	1.86	1.86	<b>1.86</b>	3.17	3.65	3.22	3.28	3.47	3.38	3.36	<b>3.36</b>
O-cccc <sup>a</sup>	3.40	3.40	3.42	3.44	3.41	3.46	3.43	<b>3.42</b>	3.82	2.16	2.17	2.22	3.64	3.80	3.60	<b>3.06</b>

<sup>a</sup>Collapsed open structure labelled in *italics*

**Table D4 Root-mean-square deviation (RMSD) in Å of optimised  $\beta$ CD geometries relative to M06-2X/def2TZVPP geometry.**

Methods	ccccw	cccc	cwew	cwecc	O-cwew	O-cccc <sup>a</sup>
DFTB3	0.1022	0.1685	0.1092	0.1261	0.1839	0.6432
PM7	0.1410	0.1780	0.1692	0.1813	0.1424	0.7718
PM6-DH2	0.1407	0.1869	0.1743	0.1794	0.1736	0.9358
PM6-D3	0.1279	0.1755	0.1571	0.1632	0.0874	0.9239
PM6	0.1365	0.1883	0.1475	0.1659	0.1530	0.8933
PM3CARB-1	0.2398	0.2848	0.2644	0.2783	0.3104	0.6496
PM3-D*	0.2588	0.2908	0.2508	0.2612	0.1992	0.3867
PM3	0.2193	0.2671	0.2450	0.2573	0.1765	0.3432
AM1	0.2293	0.2567	0.1814	0.1954	0.8176	0.5154

<sup>a</sup>Collapsed open structures labelled in *italics*

**Table D5 Glucose geometries optimised at ab initio, DFT and semiempirical QM levels of theory.<sup>a</sup>**

glucose	MP2/Def2TZVPP						MP2/aug-cc-pVTZ						M062X/Def2TZVPP					
	$\varphi_1$	$\psi$	$\varphi_2$	$\Omega$	$\omega$	$\lambda$	$\varphi_1$	$\psi$	$\varphi_2$	$\Omega$	$\omega$	$\lambda$	$\varphi_1$	$\psi$	$\varphi_2$	$\Omega$	$\omega$	$\lambda$
cw/g <sup>-</sup> g	-47	65	175	-64	58	-159	-48	65	176	-64	59	-160	-47	65	175	-64	57	-160
cc/g <sup>-</sup> g	78	62	-51	-58	55	68	78	62	-52	-58	56	69	77	62	-50	-58	55	66
cc/tg	80	63	-52	167	51	67	80	63	-53	167	51	68	79	63	-51	166	51	65
cw/tg <sup>-</sup>	-47	65	175	-180	-76	-159	-47	65	176	-179	-78	-160	-47	65	174	180	-75	-160
cc/tg <sup>-</sup>	79	62	-52	60	-55	67	79	63	-52	61	-56	68	78	63	-52	60	-55	66
O-cw/tg	-46	65	176	-179	178	-159	-47	65	176	-178	179	-159	-46	64	175	-177	-179	-160
O-cc/gg <sup>-</sup>	79	62	-52	60	-55	67	79	63	-52	61	-56	68	78	62	-51	60	-56	65
glucose	wB97XD/Def2TZVPP						DFTB3						SCC-DFTB					
	$\varphi_1$	$\psi$	$\varphi_2$	$\Omega$	$\omega$	$\lambda$	$\varphi_1$	$\psi$	$\varphi_2$	$\Omega$	$\omega$	$\lambda$	$\varphi_1$	$\psi$	$\varphi_2$	$\Omega$	$\omega$	$\lambda$
cw/g <sup>-</sup> g	-47	65	176	-66	60	-158	-49	68	167	-58	51	-162	-49	68	170	-63	55	-164
cc/g <sup>-</sup> g	78	62	-52	-59	56	68	81	67	-43	-48	43	49	76	67	-46	-53	47	55
cc/tg	80	63	-52	167	50	67	83	67	-44	172	42	48	78	67	-46	169	44	54
cw/tg <sup>-</sup>	-47	65	176	179	-75	-158	-49	68	165	177	-79	-162	-49	68	169	175	-71	-164
cc/tg <sup>-</sup>	79	62	-52	61	-57	67	82	67	-44	51	-44	47	78	67	-46	56	-48	53
O-cw/tg	-46	65	177	-179	178	-158	-48	68	166	-179	-175	-162	-49	68	169	177	-180	-164
O-cc/gg <sup>-</sup>	79	62	-52	61	-57	67	82	67	-44	51	-44	47	78	67	-46	56	-48	53
glucose	PM7						PM6-DH2						PM6-D3					
	$\varphi_1$	$\psi$	$\varphi_2$	$\Omega$	$\omega$	$\lambda$	$\varphi_1$	$\psi$	$\varphi_2$	$\Omega$	$\omega$	$\lambda$	$\varphi_1$	$\psi$	$\varphi_2$	$\Omega$	$\omega$	$\lambda$
cw/g <sup>-</sup> g	-54	77	172	-77	64	-147	-55	86	171	-128	90	-140	-57	81	173	-81	68	-140
cc/g <sup>-</sup> g	73	71	-47	-61	54	53	77	76	-54	-56	52	57	73	73	-51	-60	56	53
cc/tg	74	71	-45	171	42	52	79	78	-56	169	46	57	75	74	-52	164	49	53
cw/tg <sup>-</sup>	-54	77	169	179	-77	-147	-55	86	171	-178	-82	-141	-57	82	171	179	-77	-140
cc/tg <sup>-</sup>	75	71	-48	71	-61	49	78	77	-55	70	-61	56	75	74	-52	73	-62	52
O-cw/tg	-54	77	170	177	171	-144	-55	86	172	-176	152	-139	-57	82	172	178	165	-136
O-cc/gg <sup>-</sup>	75	71	-48	72	-61	49	78	77	-55	71	-61	56	75	74	-52	74	-63	52
glucose	PM6						PM3CARB-1						PM3-D*					
	$\varphi_1$	$\psi$	$\varphi_2$	$\Omega$	$\omega$	$\lambda$	$\varphi_1$	$\psi$	$\varphi_2$	$\Omega$	$\omega$	$\lambda$	$\varphi_1$	$\psi$	$\varphi_2$	$\Omega$	$\omega$	$\lambda$
cw/g <sup>-</sup> g	-62	86	179	-82	70	-132	-58	68	179	-74	71	-175	-55	68	178	-72	63	-173
cc/g <sup>-</sup> g	73	77	-56	-60	58	55	77	64	-51	-62	64	57	66	65	-52	-65	61	65
cc/tg	75	77	-56	159	55	53	80	64	-49	170	44	57	67	65	-52	171	43	64
cw/tg <sup>-</sup>	-62	87	176	-180	-78	-131	-58	68	175	-179	-89	-175	-54	68	175	178	-77	-173
cc/tg <sup>-</sup>	75	77	-57	77	-66	53	78	64	-50	63	-68	56	67	65	-53	71	-64	64
O-cw/tg	-62	87	177	179	176	-126	-58	68	176	-177	-177	-174	-54	68	176	176	171	-171
O-cc/gg <sup>-</sup>	75	77	-57	77	-66	53	78	64	-50	63	-68	56	67	65	-52	71	-64	64
glucose	PM3						AM1											
	$\varphi_1$	$\psi$	$\varphi_2$	$\Omega$	$\omega$	$\lambda$	$\varphi_1$	$\psi$	$\varphi_2$	$\Omega$	$\omega$	$\lambda$						
cw/g <sup>-</sup> g	-55	68	177	-63	62	-173	-54	70	172	-70	62	-160						
cc/g <sup>-</sup> g	66	65	-53	-65	61	65	70	67	-49	-55	58	56						
cc/tg	68	65	-52	171	43	64	71	67	-49	154	59	55						
cw/tg <sup>-</sup>	-54	68	175	178	-77	-173	-54	70	173	172	-66	-160						
cc/tg <sup>-</sup>	67	65	-52	71	-64	64	71	67	-49	64	-61	55						
O-cw/tg	-54	68	176	177	171	-172	-54	70	173	164	-167	-158						
O-cc/gg <sup>-</sup>	67	65	-53	71	-64	64	71	67	-49	64	-61	55						

<sup>a</sup>Dihedral angles in degrees:  $\varphi_1$ (HO2C2C3),  $\psi$ (O2C2C3O3),  $\varphi_2$ (HO3C3C2),  $\Omega$ (O5C5C6O6),  $\omega$ (C5C6O6H),  $\lambda$ (O5C1O1H)

**Table D6 Energy difference  $E_{cc-cw}$  between  $gGg^-$  and  $tGg^-$  (i.e. cc and cw) conformers (kcal/mol), and corrected energy  $E'_{cc-cw}$ , employing correction term  $\delta E_{cc-cw}$  based on diol energetics (see text for definition).**

Methods	$E_{cc-cw}$	$E'_{cc-cw}$
PM7	-0.93	3.18
PM6-DH2	-1.30	1.27
PM6-D3	-1.27	2.56
PM6	-1.07	-0.04
PM3-D*	-1.39	9.00
PM3	-1.39	9.10
AM1	-0.88	5.23

**Table D7** 1,2-ethanediols geometries optimised at ab initio, DFT and SQM levels of theory.<sup>a</sup>

diols	MP2/def2-TZVPP			MP2-aug-cc-pVTZ			M06-2X/def2-TZVPP			ωB97X-D/def2-TZVPP		
	Φ <sub>1</sub>	Ψ	Φ <sub>2</sub>	Φ <sub>1</sub>	Ψ	Φ <sub>2</sub>	Φ <sub>1</sub>	Ψ	Φ <sub>2</sub>	Φ <sub>1</sub>	Ψ	Φ <sub>2</sub>
g <sup>-</sup> Gg <sup>-</sup>	-79	60	-80	-80	61	-82	-77	58	-81	-79	60	-81
gGg	42	53	39	74	57	-44	72	57	-43	43	53	39
gGg <sup>-</sup>	73	57	-43	74	57	-44	73	57	-43	72	58	-43
gTg	65	177	65	67	177	68	64	178	64	65	178	65
gTg <sup>-</sup>	69	-180	-69	71	-180	-72	69	-180	-69	69	-180	-69
tGg	-178	65	56	-177	66	59	-178	65	56	-177	66	56
tGg <sup>-</sup>	-168	61	-50	-167	62	-51	-168	61	-51	-167	62	-52
tGt	-165	75	-165	-163	75	-163	-160	74	-161	-162	74	-164
tTg	-175	180	70	-175	180	73	-176	-180	71	-175	-179	70
tTt	180	180	180	180	180	180	179	-180	-179	180	180	180
diols	DFTB3			SCC-DFTB			PM7			PM6-DH2		
	Φ <sub>1</sub>	Ψ	Φ <sub>2</sub>	Φ <sub>1</sub>	Ψ	Φ <sub>2</sub>	Φ <sub>1</sub>	Ψ	Φ <sub>2</sub>	Φ <sub>1</sub>	Ψ	Φ <sub>2</sub>
g <sup>-</sup> Gg <sup>-</sup>	-70	48	-71	-70	58	-70	-74	93	-74	-74	103	-74
gGg	65	50	-25	-31	54	61	67	65	-42	65	-2	42
gGg <sup>-</sup>	65	49	-25	61	54	-31	67	65	-42	73	52	-35
gTg	57	-177	57	58	-178	58	73	-94	73	74	-103	74
gTg <sup>-</sup>	61	180	-61	63	180	-62	70	-178	-70	72	-178	-72
tGg	-178	51	-39	-177	58	-45	-176	80	-58	-173	89	-62
tGg <sup>-</sup>	-178	51	-39	-177	58	-45	-176	80	-58	-173	89	-62
tGt	-169	94	-169	-169	90	-169	-73	94	-73	-74	103	-74
tTg	-178	-177	62	-177	-177	63	-67	-65	43	170	-83	60
tTt	180	-180	-180	180	180	180	-180	-177	180	62	-89	173
diols	PM6-D3			PM6			PM3CARB-1			PM3D*		
	Φ <sub>1</sub>	Ψ	Φ <sub>2</sub>	Φ <sub>1</sub>	Ψ	Φ <sub>2</sub>	Φ <sub>1</sub>	Ψ	Φ <sub>2</sub>	Φ <sub>1</sub>	Ψ	Φ <sub>2</sub>
g <sup>-</sup> Gg <sup>-</sup>	-71	99	-71	-74	102	-73	-81	45	-81	-73	71	-73
gGg	53	0	53	55	-2	54	57	10	57	47	49	47
gGg <sup>-</sup>	69	63	-43	68	76	-52	80	65	-49	64	70	-52
gTg	69	-172	69	74	-101	74	65	173	64	60	176	59
gTg <sup>-</sup>	71	-178	-71	72	180	-72	71	-180	-71	62	-180	-62
tGg	-175	89	-62	-175	94	-65	179	72	53	-178	75	52
tGg <sup>-</sup>	-173	84	-61	-175	94	-66	-179	66	-63	-175	75	-61
tGt	-71	97	-71	-74	101	-74	-166	91	-166	-73	71	-73
tTg	175	-88	62	-68	-76	52	-175	-179	72	-171	-178	64
tTt	177	-176	-179	66	-95	175	-180	-180	180	180	180	-180

<sup>a</sup>1,2-ethanediol structure shown in Figure 2; Dihedral angles in degrees: Φ<sub>1</sub>(H<sub>b</sub>OCC'), Ψ(OCC'O'), Φ<sub>2</sub>(H<sub>a</sub>O'C'C)

**Table D7** 1,2-ethanediols geometries optimised at ab initio, DFT and SQM levels of theory<sup>a</sup> (*continued*).

diols	PM3			AM1		
	$\Phi_1$	$\Psi$	$\Phi_2$	$\Phi_1$	$\Psi$	$\Phi_2$
g'Gg'	-73	71	-73	-58	3	-58
gGg	48	50	47	58	-3	58
gGg'	64	70	-51	66	60	-44
gTg	60	176	59	60	177	60
gTg'	62	180	-62	63	180	-63
tGg	-178	75	52	-174	-174	65
tGg'	-174	75	-62	-175	64	-56
tGt	-73	72	-73	-175	64	-56
tTg	-171	-178	64	-174	-173	65
tTt	-180	-180	-180	180	-178	180

<sup>a</sup>1,2-ethanediol structure shown in Figure 2; Dihedral angles in degrees:  $\phi_1(H_bOCC')$ ,  $\psi(OCC'O')$ ,  $\phi_2(H_aO'C'C)$

**Table D8** Average distance between O atoms of the C2C3 rim of  $\beta$ CD and the graphene basal plane of  $\beta$ CD/C<sub>96</sub>H<sub>24</sub> complexes (in Å) for DFT and SQM methods.

Conformer	M06-2X/def2-SVP	DFTB3	PM7	PM6-DH2
cccw	2.96	3.05	2.92	3.10
cccc	2.98	3.04	2.85	3.07

**Table D9 Intermolecular interaction energetics (in kcal/mol) of benzene-methane and benzene-water dimers.**

<b>Method</b>	<b>Benzene-methane</b>	<b>Benzene-water</b>
M06-2X/def2-TZVPP	<b>-1.30</b>	-3.66
M06-2X/def2-TZVPP//def2-SVP	<b>-1.35</b>	-3.56
M06-2X/def2-SVP	-1.23	-3.29
$\omega$ B97X-D/def2-TZVPP	-1.67	-3.52
$\omega$ B97X-D/aug-cc-pVTZ	<b>-1.66</b>	-3.52
DFTB3	<b>-1.56</b>	<b>-2.58</b>
PM7	<b>-1.83</b>	-3.03
PM6-DH2	<b>-1.51</b>	-3.58
PM6-DH2//PM6-DH+	<b>-1.51</b>	-3.58
PM6-DH2//PM6	<b>-1.47</b>	-3.57
PM6-D3	<b>-1.86</b>	-4.31
PM6	<b>-0.50</b>	-2.65
PM3CARB-1	<b>0.96</b>	-0.80
PM3-D*	<b>-2.34</b>	-3.25
PM3	<b>-0.20</b>	-1.62
AM1	<b>0.33</b>	-1.23
CCSD(T)/CBS <sup>a,b</sup>	-1.50	-3.28

Correct geometries after optimisation are highlighted in bold.

<sup>a</sup>Equilibrium geometry at MP2/cc-pVTZ level in S22 dataset<sup>230</sup>

<sup>b</sup>benchmark geometry at MP2/6-311G\*\* level<sup>231</sup>

<sup>c</sup>no longer dimer

## References

1. Hughes, J. P.; Rees, S.; Kalindjian, S. B.; Philpott, K. L., Principles of early drug discovery. *British Journal of Pharmacology* **2011**, *162* (6), 1239-1249.
2. Ciociola, A. A.; Cohen, L. B.; Kulkarni, P.; the, F. D. A. R. M. C. o. t. A. C. o. G., How Drugs are Developed and Approved by the FDA: Current Process and Future Directions. *Am J Gastroenterol* **2014**, *109* (5), 620-623.
3. Macalino, S.; Gosu, V.; Hong, S.; Choi, S., Role of computer-aided drug design in modern drug discovery. *Archives of Pharmacal Research* **2015**, *38* (9), 1686-1701.
4. Paul, S. M.; Mytelka, D. S.; Dunwiddie, C. T.; Persinger, C. C.; Munos, B. H.; Lindborg, S. R.; Schacht, A. L., How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nat Rev Drug Discov* **2010**, *9* (3), 203-214.
5. Hardy, L. W.; Malikayil, A., The impact of structure-guided drug design on clinical agents. *Curr. Drug Discov* **2003**, *3*, 15-20.
6. Clark, D. E., What has computer-aided molecular design ever done for drug discovery? *Expert Opinion on Drug Discovery* **2006**, *1* (2), 103-110.
7. (a) Zhao, Y.; Truhlar, D., The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four M06-class functionals and 12 other functionals. *Theoretical Chemistry Accounts* **2008**, *120* (1-3), 215-241; (b) Gaus, M.; Goez, A.; Elstner, M., Parametrization and Benchmark of DFTB3 for Organic Molecules. *Journal of Chemical Theory and Computation* **2013**, *9* (1), 338-354; (c) Hobza, P., Calculations on Noncovalent Interactions and Databases of Benchmark Interaction Energies. *Accounts of Chemical Research* **2012**, *45* (4), 663-672; (d) Hostaš, J.; Řezáč, J.; Hobza, P., On the performance of the semiempirical quantum mechanical PM6 and PM7 methods for noncovalent interactions. *Chemical Physics Letters* **2013**, *568–569* (0), 161-166.
8. Dunning, T. H., Gaussian basis sets for use in correlated molecular calculations. I. The atoms boron through neon and hydrogen. *The Journal of Chemical Physics* **1989**, *90* (2), 1007-1023.
9. Binkley, J. S.; Pople, J. A.; Hehre, W. J., Self-consistent molecular orbital methods.
21. Small split-valence basis sets for first-row elements. *Journal of the American Chemical Society* **1980**, *102* (3), 939-947.

10. Weigend, F.; Ahlrichs, R., Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Physical Chemistry Chemical Physics* **2005**, *7* (18), 3297-3305.
11. Jensen, F., Polarization consistent basis sets: Principles. *The Journal of Chemical Physics* **2001**, *115* (20), 9113-9125.
12. Jensen, F., Atomic orbital basis sets. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2013**, *3* (3), 273-295.
13. Zerner, M., Perspective on “New developments in molecular orbital theory”. In *Theoretical Chemistry Accounts*, Cramer, C.; Truhlar, D., Eds. Springer Berlin Heidelberg: 2001; pp 217-218.
14. Jensen, F., *Introduction to Computational Chemistry*. Wiley: 2013.
15. David Sherrill, C.; Schaefer Iii, H. F., The Configuration Interaction Method: Advances in Highly Correlated Approaches. In *Advances in Quantum Chemistry*, Per-Olov Löwdin, J. R. S. M. C. Z.; Erkki, B., Eds. Academic Press: 1999; Vol. Volume 34, pp 143-269.
16. Čížek, J., On the Correlation Problem in Atomic and Molecular Systems. Calculation of Wavefunction Components in Ursell-Type Expansion Using Quantum-Field Theoretical Methods. *The Journal of Chemical Physics* **1966**, *45* (11), 4256-4266.
17. (a) Møller, C.; Plesset, M. S., Note on an Approximation Treatment for Many-Electron Systems. *Physical Review* **1934**, *46* (7), 618-622; (b) Cremer, D., Møller-Plesset perturbation theory: from small molecule methods to methods for thousands of atoms. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2011**, *1* (4), 509-530.
18. (a) Friesner, R. A., Ab initio quantum chemistry: Methodology and applications. *Proceedings of the National Academy of Sciences of the United States of America* **2005**, *102* (19), 6648-6653; (b) Dreuw, A.; Head-Gordon, M., Single-Reference ab Initio Methods for the Calculation of Excited States of Large Molecules. *Chemical Reviews* **2005**, *105* (11), 4009-4037; (c) Hobza, P.; Šponer, J., Structure, Energetics, and Dynamics of the Nucleic Acid Base Pairs: Nonempirical Ab Initio Calculations. *Chemical Reviews* **1999**, *99* (11), 3247-3276; (d) Chalasinski, G.; Szczesniak, M. M., Origins of Structure and Energetics of van der Waals Clusters from ab Initio Calculations. *Chemical Reviews* **1994**, *94* (7), 1723-1765.

19. Bredow, T.; Jug, K., Theory and range of modern semiempirical molecular orbital methods. *Theoretical Chemistry Accounts* **2005**, *113* (1), 1-14.
20. Dewar, M. J. S.; W., T., Ground states of molecules. 38. The MNDO method. Approximations and parameters. *Journal of the American Chemical Society* **1977**, *99* (15), 4899-4907.
21. Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P., Development and use of quantum mechanical molecular models. 76. AM1: a new general purpose quantum mechanical molecular model. *Journal of the American Chemical Society* **1985**, *107* (13), 3902-3909.
22. Stewart, J. J. P., Optimization of parameters for semiempirical methods I. Method. *Journal of Computational Chemistry* **1989**, *10* (2), 209-220.
23. Stewart, J. P., Optimization of parameters for semiempirical methods VI: more modifications to the NDDO approximations and re-optimization of parameters. *Journal of Molecular Modeling* **2013**, *19* (1), 1-32.
24. Grimme, S., Accurate description of van der Waals complexes by density functional theory including empirical corrections. *Journal of Computational Chemistry* **2004**, *25* (12), 1463-1473.
25. Korth, M., Third-Generation Hydrogen-Bonding Corrections for Semiempirical QM Methods and Force Fields. *Journal of Chemical Theory and Computation* **2010**, *6* (12), 3808-3816.
26. McNamara, J. P.; Hillier, I. H., Semi-empirical molecular orbital methods including dispersion corrections for the accurate prediction of the full range of intermolecular interactions in biomolecules. *Physical Chemistry Chemical Physics* **2007**, *9* (19), 2362-2370.
27. Fanfrlik, J.; Bronowska, A. K.; Rezac, J.; Prenosil, O.; Konvalinka, J.; Hobza, P., A Reliable Docking/Scoring Scheme Based on the Semiempirical Quantum Mechanical PM6-DH2 Method Accurately Covering Dispersion and H-Bonding: HIV-1 Protease with 22 Ligands. *Journal of Physical Chemistry B* **2010**, *114* (39), 12666-12678.
28. Hohenberg, P.; Kohn, W., Inhomogeneous electron gas. *Physical Review 3B* **1964**, *136*, 864-871.
29. Kohn, W.; Sham, L. J., Self-consistent equations including exchange and correlation effects. *Physical Review 4A* **1965**, *140*, 1133-1138.

30. Geerlings, P.; De Proft, F.; Langenaeker, W., Conceptual Density Functional Theory. *Chemical Reviews* **2003**, *103* (5), 1793-1874.
31. Robert G. Parr, a.; Yang, W., Density-Functional Theory of the Electronic Structure of Molecules. *Annual Review of Physical Chemistry* **1995**, *46* (1), 701-728.
32. (a) Pople, J. A.; Gill, P. M. W.; Johnson, B. G., Kohn—Sham density-functional theory within a finite basis set. *Chemical Physics Letters* **1992**, *199* (6), 557-560; (b) Burke, K., Perspective on density functional theory. *The Journal of Chemical Physics* **2012**, *136* (15), 150901.
33. Sousa, S. F.; Fernandes, P. A.; Ramos, M. J., General Performance of Density Functionals. *The Journal of Physical Chemistry A* **2007**, *111* (42), 10439-10452.
34. Gaus, M.; Cui, Q.; Elstner, M., Density functional tight binding: application to organic and biological molecules. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2014**, *4* (1), 49-61.
35. Elstner, M.; Porezag, D.; Jungnickel, G.; Elsner, J.; Haugk, M.; Frauenheim, T.; Suhai, S.; Seifert, G., Self-consistent-charge density-functional tight-binding method for simulations of complex materials properties. *Physical Review B* **1998**, *58* (11), 7260-7268.
36. Seifert, G.; Joswig, J.-O., Density-functional tight binding—an approximate density-functional theory method. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2012**, *2* (3), 456-465.
37. Oliveira, A. F.; Seifert, G.; Heine, T.; Duarte, H. A., Density-functional based tight-binding: an approximate DFT method. *Journal of the Brazilian Chemical Society* **2009**, *20*, 1193-1205.
38. Elstner, M., The SCC-DFTB method and its application to biological systems. *Theoretical Chemistry Accounts* **2006**, *116* (1-3), 316-325.
39. Porezag, D.; Frauenheim, T.; Köhler, T.; Seifert, G.; Kaschner, R., Construction of tight-binding-like potentials on the basis of density-functional theory: Application to carbon. *Physical Review B* **1995**, *51* (19), 12947-12957.
40. Sattelmeyer, K. W.; Tirado-Rives, J.; Jorgensen, W. L., Comparison of SCC-DFTB and NDDO-Based Semiempirical Molecular Orbital Methods for Organic Molecules. *The Journal of Physical Chemistry A* **2006**, *110* (50), 13551-13559.

41. Domínguez, A.; Niehaus, T. A.; Frauenheim, T., Accurate Hydrogen Bond Energies within the Density Functional Tight Binding Method. *The Journal of Physical Chemistry A* **2015**, *119* (14), 3535-3544.
42. Parr, R. G.; Pearson, R. G., Absolute hardness: companion parameter to absolute electronegativity. *Journal of the American Chemical Society* **1983**, *105* (26), 7512-7516.
43. Elstner, M.; Seifert, G., Density functional tight binding. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* **2014**, *372* (2011).
44. Yang; Yu, H.; York, D.; Cui, Q.; Elstner, M., Extension of the Self-Consistent-Charge Density-Functional Tight-Binding Method: Third-Order Expansion of the Density Functional Theory Total Energy and Introduction of a Modified Effective Coulomb Interaction. *The Journal of Physical Chemistry A* **2007**, *111* (42), 10861-10873.
45. Gaus, M.; Cui, Q.; Elstner, M., DFTB3: Extension of the Self-Consistent-Charge Density-Functional Tight-Binding Method (SCC-DFTB). *Journal of Chemical Theory and Computation* **2011**, *7* (4), 931-948.
46. Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A., A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *Journal of the American Chemical Society* **1995**, *117* (19), 5179-5197.
47. Mackerell, A. D., Empirical force fields for biological macromolecules: Overview and issues. *Journal of Computational Chemistry* **2004**, *25* (13), 1584-1604.
48. Warshel, A.; Levitt, M., Theoretical studies of enzymic reactions: Dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. *Journal of Molecular Biology* **1976**, *103* (2), 227-249.
49. Monard, G.; Merz, K. M., Combined Quantum Mechanical/Molecular Mechanical Methodologies Applied to Biomolecular Systems. *Accounts of Chemical Research* **1999**, *32* (10), 904-911.
50. Field, M. J.; Bash, P. A.; Karplus, M., A combined quantum mechanical and molecular mechanical potential for molecular dynamics simulations. *Journal of Computational Chemistry* **1990**, *11* (6), 700-733.

51. (a) DiLabio, G. A.; Hurley, M. M.; Christiansen, P. A., Simple one-electron quantum capping potentials for use in hybrid QM/MM studies of biological molecules. *The Journal of Chemical Physics* **2002**, *116* (22), 9578-9584; (b) Zhang, Y.; Lee, T.-S.; Yang, W., A pseudobond approach to combining quantum mechanical and molecular mechanical methods. *The Journal of Chemical Physics* **1999**, *110* (1), 46-54.
52. Philipp, D. M.; Friesner, R. A., Mixed ab initio QM/MM modeling using frozen orbitals and tests with alanine dipeptide and tetrapeptide. *Journal of Computational Chemistry* **1999**, *20* (14), 1468-1494.
53. (a) Vreven, T.; Byun, K. S.; Komáromi, I.; Dapprich, S.; Montgomery, J. A.; Morokuma, K.; Frisch, M. J., Combining Quantum Mechanics Methods with Molecular Mechanics Methods in ONIOM. *Journal of Chemical Theory and Computation* **2006**, *2* (3), 815-826; (b) Lin, H.; Truhlar, D., QM/MM: what have we learned, where are we, and where do we go from here? *Theoretical Chemistry Accounts* **2007**, *117* (2), 185-199; (c) Senn, H. M.; Thiel, W., QM/MM Methods for Biomolecular Systems. *Angewandte Chemie International Edition* **2009**, *48* (7), 1198-1229.
54. Bakowies, D.; Thiel, W., Hybrid Models for Combined Quantum Mechanical and Molecular Mechanical Approaches. *The Journal of Physical Chemistry* **1996**, *100* (25), 10580-10594.
55. Riccardi, D.; Li, G.; Cui, Q., Importance of van der Waals Interactions in QM/MM Simulations. *The Journal of Physical Chemistry B* **2004**, *108* (20), 6467-6478.
56. Murphy, R. B.; Philipp, D. M.; Friesner, R. A., A mixed quantum mechanics/molecular mechanics (QM/MM) method for large-scale modeling of chemistry in protein environments. *Journal of Computational Chemistry* **2000**, *21* (16), 1442-1457.
57. Agrafiotis, D. K.; Gibbs, A. C.; Zhu, F.; Izrailev, S.; Martin, E., Conformational sampling of bioactive molecules: A comparative study. *Journal of Chemical Information and Modeling* **2007**, *47* (3), 1067-1086.
58. Chen, I. J.; Foloppe, N., Is Conformational Sampling of Drug-like Molecules a Solved Problem? *Drug Development Research* **2011**, *72* (1), 85-94.
59. Blommers, M. J. J.; Lucasius, C. B.; Kateman, G.; Kaptein, R., Conformational analysis of a dinucleotide photodimer with the aid of the genetic algorithm. *Biopolymers* **1992**, *32* (1), 45-52.

60. Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E., Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics* **1953**, *21* (6), 1087-1092.
61. Lill, M. A., Efficient Incorporation of Protein Flexibility and Dynamics into Molecular Docking Simulations. *Biochemistry* **2011**, *50* (28), 6157-6169.
62. Yuriev, E.; Agostino, M.; Ramsland, P. A., Challenges and advances in computational docking: 2009 in review. *Journal of Molecular Recognition* **2011**, *24* (2), 149-164.
63. Ferrara, P.; Gohlke, H.; Price, D. J.; Klebe, G.; Brooks, C. L., Assessing scoring functions for protein-ligand interactions. *Journal of Medicinal Chemistry* **2004**, *47* (12), 3032-3047.
64. Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J., Docking and scoring in virtual screening for drug discovery: Methods and applications. *Nature Reviews Drug Discovery* **2004**, *3* (11), 935-949.
65. Zheng, M.; Xiong, B.; Luo, C.; Li, S.; Liu, X.; Shen, Q.; Li, J.; Zhu, W.; Luo, X.; Jiang, H., Knowledge-Based Scoring Functions in Drug Design: 3. A Two-Dimensional Knowledge-Based Hydrogen-Bonding Potential for the Prediction of Protein–Ligand Interactions. *Journal of Chemical Information and Modeling* **2011**, *51* (11), 2994-3004.
66. Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J., Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews* **1997**, *23* (1-3), 3-25.
67. Lipinski, C. A., Lead- and drug-like compounds: the rule-of-five revolution. *Drug Discovery Today: Technologies* **2004**, *1* (4), 337-341.
68. (a) Ursu, O.; Rayan, A.; Goldblum, A.; Oprea, T. I., Understanding drug-likeness. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2011**, *1* (5), 760-781; (b) Oprea, T. I.; Davis, A. M.; Teague, S. J.; Leeson, P. D., Is There a Difference between Leads and Drugs? A Historical Perspective. *Journal of Chemical Information and Computer Sciences* **2001**, *41* (5), 1308-1315.
69. Carr, R. A. E.; Congreve, M.; Murray, C. W.; Rees, D. C., Fragment-based lead discovery: leads by design. *Drug Discovery Today* **2005**, *10* (14), 987-992.
70. Congreve, M.; Carr, R.; Murray, C.; Jhoti, H., A ‘Rule of Three’ for fragment-based lead discovery? *Drug Discovery Today* **2003**, *8* (19), 876-877.

71. (a) Abad-Zapatero, C., Ligand efficiency indices for effective drug discovery. *Expert Opinion on Drug Discovery* **2007**, 2 (4), 469-488; (b) Hopkins, A. L.; Groom, C. R.; Alex, A., Ligand efficiency: a useful metric for lead selection. *Drug Discovery Today* **2004**, 9 (10), 430-431.
72. Bembenek, S. D.; Tounge, B. A.; Reynolds, C. H., Ligand efficiency and fragment-based drug discovery. *Drug Discovery Today* **2009**, 14 (5-6), 278-283.
73. Keseru, G. M.; Makara, G. M., Hit discovery and hit-to-lead approaches. *Drug Discovery Today* **2006**, 11 (15-16), 741-748.
74. Leach, A. R.; Hann, M. M., Molecular complexity and fragment-based drug discovery: ten years on. *Current Opinion in Chemical Biology* **2011**, 15 (4), 489-496.
75. (a) Carr, R.; Jhoti, H., Structure-based screening of low-affinity compounds. *Drug Discovery Today* **2002**, 7 (9), 522-527; (b) Hesterkamp, T.; Whittaker, M., Fragment-based activity space: smaller is better. *Current Opinion in Chemical Biology* **2008**, 12 (3), 260-268.
76. (a) Jhoti, H.; Cleasby, A.; Verdonk, M.; Williams, G., Fragment-based screening using X-ray crystallography and NMR spectroscopy. *Current Opinion in Chemical Biology* **2007**, 11 (5), 485-493; (b) Davies, T. G.; Tickle, I. J., Fragment screening using X-ray crystallography. *Topics in current chemistry* **2012**, 317, 33-59.
77. (a) Hajduk, P. J., SAR by NMR: Putting the pieces together. *Molecular Interventions* **2006**, 6 (5), 266-272; (b) Dalvit, C., NMR methods in fragment screening: theory and a comparison with other biophysical techniques. *Drug Discovery Today* **2009**, 14 (21-22), 1051-1057; (c) Pellecchia, M., NMR Spectroscopy in Fragment Based Drug Design. In *Lead-Seeking Approaches*, Hayward, M. H., Ed. Springer: New York, 2010; Vol. 5, pp 125-139; (d) Schade, M.; Oschkinat, H., NMR fragment screening: Tackling protein-protein interaction targets. *Current Opinion in Drug Discovery & Development* **2005**, 8 (3), 365-373.
78. Erlanson, D. A.; McDowell, R. S.; O'Brien, T., Fragment-based drug discovery. *Journal of Medicinal Chemistry* **2004**, 47 (14), 3463-3482.
79. Walters, W. P.; Stahl, M. T.; Murcko, M. A., Virtual screening - an overview. *Drug Discovery Today* **1998**, 3 (4), 160-178.
80. Gribbon, P.; Andreas, S., High-throughput drug discovery: What can we expect from HTS? *Drug Discovery Today* **2005**, 10 (1), 17-22.

81. Schneider, G.; Bohm, H. J., Virtual screening and fast automated docking methods. *Drug Discovery Today* **2002**, 7 (1), 64-70.
82. (a) Evers, A.; Hessler, G.; Matter, H.; Klabunde, T., Virtual screening of biogenic amine-binding G-protein coupled receptors: Comparative evaluation of protein- and ligand-based virtual screening protocols. *Journal of Medicinal Chemistry* **2005**, 48 (17), 5448-5465; (b) Hamza, A.; Wei, N.-N.; Zhan, C.-G., Ligand-Based Virtual Screening Approach Using a New Scoring Function. *Journal of Chemical Information and Modeling* **2012**, 52 (4), 963-974.
83. Kirchmair, J.; Distinto, S.; Markt, P.; Schuster, D.; Spitzer, G. M.; Liedl, K. R.; Wolber, G., How To Optimize Shape-Based Virtual Screening: Choosing the Right Query and Including Chemical Information. *Journal of Chemical Information and Modeling* **2009**, 49 (3), 678-692.
84. Perez-Nueno, V. I.; Ritchie, D. W.; Rabal, O.; Pascual, R.; Borrell, J. I.; Teixido, J., Comparison of Ligand-Based and Receptor-Based Virtual Screening of HIV Entry Inhibitors for the CXCR4 and CCR5 Receptors Using 3D Ligand Shape Matching and Ligand-Receptor Docking. *Journal of Chemical Information and Modeling* **2008**, 48 (3), 509-533.
85. McGaughey, G. B.; Sheridan, R. P.; Bayly, C. I.; Culberson, J. C.; Kreatsoulas, C.; Lindsley, S.; Maiorov, V.; Truchon, J.-F.; Cornell, W. D., Comparison of Topological, Shape, and Docking Methods in Virtual Screening. *Journal of Chemical Information and Modeling* **2007**, 47 (4), 1504-1519.
86. Yu, H.; Wang, Z.; Zhang, L.; Zhang, J.; Huang, Q., The Discovery of Novel Vascular Endothelial Growth Factor Receptor Tyrosine Kinases Inhibitors: Pharmacophore Modeling, Virtual Screening and Docking Studies. *Chemical Biology & Drug Design* **2007**, 69 (3), 204-211.
87. Sastry, G. M.; Dixon, S. L.; Sherman, W., Rapid Shape-Based Ligand Alignment and Virtual Screening Method Based on Atom/Feature-Pair Similarities and Volume Overlap Scoring. *Journal of Chemical Information and Modeling* **2011**, 51 (10), 2455-2466.
88. Proschak, E.; Rupp, M.; Derksen, S.; Schneider, G., Shapelets: Possibilities and limitations of shape-based virtual screening. *Journal of Computational Chemistry* **2008**, 29 (1), 108-114.

89. Ballester, P. J.; Finn, P. W.; Richards, W. G., Ultrafast shape recognition: Evaluating a new ligand-based virtual screening technology. *Journal of Molecular Graphics and Modelling* **2009**, *27* (7), 836-845.
90. Lee, H. S.; Lee, C. S.; Kim, J. S.; Kim, D. H.; Choe, H., Improving Virtual Screening Performance against Conformational Variations of Receptors by Shape Matching with Ligand Binding Pocket. *Journal of Chemical Information and Modeling* **2009**, *49* (11), 2419-2428.
91. Moffat, K.; Gillet, V. J.; Whittle, M.; Bravi, G.; Leach, A. R., A comparison of field-based similarity searching methods: CatShape, FBSS, and ROCS. *Journal of Chemical Information and Modeling* **2008**, *48* (4), 719-729.
92. Hawkins, P. C. D.; Skillman, A. G.; Nicholls, A., Comparison of shape-matching and docking as virtual screening tools. *Journal of Medicinal Chemistry* **2007**, *50* (1), 74-82.
93. (a) Mucs, D.; Bryce, R. A.; Bonnet, P., Application of shape-based and pharmacophore-based in silico screens for identification of Type II protein kinase inhibitors. *Journal of Computer-Aided Molecular Design* **2011**, *25* (6), 569-581; (b) Bonnet, P.; Mucs, D.; Bryce, R. A., Targeting the inactive conformation of protein kinases: computational screening based on ligand conformation. *MedChemComm* **2012**, *3* (4), 434-440.
94. Kirchmair, J.; Ristic, S.; Eder, K.; Markt, P.; Wolber, G.; Laggner, C.; Langer, T., Fast and efficient in silico 3D screening: Toward maximum computational efficiency of pharmacophore-based and shape-based approaches. *Journal of Chemical Information and Modeling* **2007**, *47* (6), 2182-2196.
95. Grant, J. A.; Gallardo, M. A.; Pickup, B. T., A fast method of molecular shape comparison: A simple application of a Gaussian description of molecular shape. *Journal of Computational Chemistry* **1996**, *17* (14), 1653-1666.
96. Rush, T. S.; Grant, J. A.; Mosyak, L.; Nicholls, A., A shape-based 3-D scaffold hopping method and its application to a bacterial protein-protein interaction. *Journal of Medicinal Chemistry* **2005**, *48* (5), 1489-1495.
97. Kumar, R.; Garg, P.; Bharatam, P. V., Shape-based virtual screening, docking, and molecular dynamics simulations to identify Mtb-ASADH inhibitors. *Journal of Biomolecular Structure and Dynamics* **2014**, *33* (5), 1082-1093.

98. ROCS Documentation. <http://www.eyesopen.com/docs/rocs/current/html/introduction.html> (accessed 1 February 2013).
99. Brough, D.; Rothwell, N. J., Caspase-1-dependent processing of pro-interleukin-1 beta is cytosolic and precedes cell death. *Journal of Cell Science* **2007**, *120* (5), 772-781.
100. Dinarello, C. A., Interleukin-1 in the pathogenesis and treatment of inflammatory diseases. *Blood* **2011**, *117* (14), 3720-3732.
101. Brough, D.; Tyrrell, P. J.; Allan, S. M., Regulation of interleukin-1 in acute brain injury. *Trends in Pharmacological Sciences* **2011**, *32* (10), 617-622.
102. Schroder, K.; Tschopp, J., The Inflammasomes. *Cell* **2010**, *140* (6), 821-832.
103. Eder, C., Mechanisms of interleukin-1 $\beta$  release. *Immunobiology* **2009**, *214* (7), 543-553.
104. Cassel, S. L.; Sutterwala, F. S., Sterile inflammatory responses mediated by the NLRP3 inflammasome. *European Journal of Immunology* **2010**, *40* (3), 607-611.
105. Hunt, C. D.; Idso, J. P., Dietary boron as a physiological regulator of the normal inflammatory response: A review and current research progress. *The Journal of Trace Elements in Experimental Medicine* **1999**, *12* (3), 221-233.
106. Pekol, T.; Daniels, J. S.; Labutti, J.; Parsons, I.; Nix, D.; Baronas, E.; Hsieh, F.; Gan, L.-S.; Miwa, G., HUMAN METABOLISM OF THE PROTEASOME INHIBITOR BORTEZOMIB: IDENTIFICATION OF CIRCULATING METABOLITES. *Drug Metabolism and Disposition* **2005**, *33* (6), 771-777.
107. Snow, R. J.; Bachovchin, W. W.; Barton, R. W.; Campbell, S. J.; Coutts, S. J.; Freeman, D. M.; Gutheil, W. G.; Kelly, T. A.; Kennedy, C. A., Studies on Proline Boronic Acid Dipeptide Inhibitors of Dipeptidyl Peptidase IV: Identification of a Cyclic Species Containing a B-N Bond. *Journal of the American Chemical Society* **1994**, *116* (24), 10860-10869.
108. Poplawski, S. E.; Lai, J. H.; Sanford, D. C.; Sudmeier, J. L.; Wu, W.; Bachovchin, W. W., Pro-Soft Val-boroPro: A Strategy for Enhancing in Vivo Performance of Boronic Acid Inhibitors of Serine Proteases. *Journal of Medicinal Chemistry* **2011**, *54* (7), 2022-2028.
109. Milo, L. J., Jr.; Lai, J. H.; Wu, W.; Liu, Y.; Maw, H.; Li, Y.; Jin, Z.; Shu, Y.; Poplawski, S. E.; Wu, Y.; Sanford, D. G.; Sudmeier, J. L.; Bachovchin, W. W., Chemical and Biological Evaluation of Dipeptidyl Boronic Acid Proteasome Inhibitors

- for Use in Prodrugs and Pro-Soft Drugs Targeting Solid Tumors. *Journal of Medicinal Chemistry* **2011**, *54* (13), 4365-4377.
110. Lai, J. H.; Wu, W.; Zhou, Y.; Maw, H. H.; Liu, Y.; Milo, L. J.; Poplawski, S. E.; Henry, G. D.; Sudmeier, J. L.; Sanford, D. G.; Bachovchin, W. W., Synthesis and characterization of constrained peptidomimetic dipeptidyl peptidase IV inhibitors: Amino-lactam boroAlanines. *Journal of Medicinal Chemistry* **2007**, *50* (10), 2391-2398.
111. Matteson, D. S., alpha-amido boronic acids: A synthetic challenge and their properties as serine protease inhibitors. *Medicinal Research Reviews* **2008**, *28* (2), 233-246.
112. Connolly, B. A.; Sanford, D. G.; Chiluwal, A. K.; Healey, S. E.; Peters, D. E.; Dimare, M. T.; Wu, W.; Liu, Y.; Maw, H.; Zhou, Y.; Li, Y.; Jin, Z.; Sudmeier, J. L.; Lai, J. H.; Bachovchin, W. W., Dipeptide boronic acid inhibitors of dipeptidyl peptidase IV: Determinants of potency and in vivo efficacy and safety. *Journal of Medicinal Chemistry* **2008**, *51* (19), 6005-6013.
113. Priestley, E. S.; De Lucca, I.; Ghavimi, B.; Erickson-Viitanen, S.; Decicco, C. P., P1 phenethyl peptide boronic acid inhibitors of HCVNS3 protease. *Bioorganic & Medicinal Chemistry Letters* **2002**, *12* (21), 3199-3202.
114. Tondi, D.; Calò, S.; Shoichet, B. K.; Costi, M. P., Structural study of phenyl boronic acid derivatives as AmpC  $\beta$ -lactamase inhibitors. *Bioorganic & Medicinal Chemistry Letters* **2010**, *20* (11), 3416-3419.
115. (a) Cacciola, J.; Fevig, J. M.; Alexander, R. S.; Brittelli, D. R.; Kettner, C. A.; Knabb, R. M.; Weber, P. C., Synthesis of conformationally-restricted boropeptide thrombin inhibitors. *Bioorganic & Medicinal Chemistry Letters* **1996**, *6* (3), 301-306; (b) Fevig, J. M.; Buriak, J.; Cacciola, J.; Alexander, R. S.; Kettner, C. A.; Knabb, R. M.; Pruitt, J. R.; Weber, P. C.; Wexler, R. R., Rational design of boropeptide thrombin inhibitors: beta,beta-dialkylphenethylglycine P2 analogs of DuP 714 with greater selectivity over complement factor I and an improved safety profile. *Bioorganic & Medicinal Chemistry Letters* **1998**, *8* (3), 301-306.
116. Lazarova, T. I.; Jin, L.; Rynkiewicz, M.; Gorga, J. C.; Bibbins, F.; Meyers, H. V.; Babine, R.; Strickler, J., Synthesis and in vitro biological evaluation of aryl boronic acids as potential inhibitors of factor XIa. *Bioorganic & Medicinal Chemistry Letters* **2006**, *16* (19), 5022-5027.

117. Ciaravino, V.; Plattner, J.; Chanda, S., An assessment of the genetic toxicology of novel boron-containing therapeutic agents. *Environmental and Molecular Mutagenesis* **2013**, *54* (5), 338-346.
118. Dowlut, M.; Hall, D. G., An Improved Class of Sugar-Binding Boronic Acids, Soluble and Capable of Complexing Glycosides in Neutral Water. *Journal of the American Chemical Society* **2006**, *128* (13), 4226-4227.
119. Liu, C. T.; Tomsho, J. W.; Benkovic, S. J., The unique chemistry of benzoxaboroles: Current and emerging applications in biotechnology and therapeutic treatments. *Bioorganic & Medicinal Chemistry* **2014**, *22* (16), 4462-4473.
120. Tomsho, J. W.; Pal, A.; Hall, D. G.; Benkovic, S. J., Ring Structure and Aromatic Substituent Effects on the pKa of the Benzoxaborole Pharmacophore. *ACS Medicinal Chemistry Letters* **2012**, *3* (1), 48-52.
121. Novoa, Eva M.; Ribas de Pouplana, L., Cooperation for Better Inhibiting. *Chemistry & Biology* **2015**, *22* (6), 685-686.
122. Sekhon, B. S., Metalloid compounds as drugs. *Research in Pharmaceutical Sciences* **2013**, *8* (3), 145-158.
123. Akama, T.; Baker, S. J.; Zhang, Y.-K.; Hernandez, V.; Zhou, H.; Sanders, V.; Freund, Y.; Kimura, R.; Maples, K. R.; Plattner, J. J., Discovery and structure-activity study of a novel benzoxaborole anti-inflammatory agent (AN2728) for the potential topical treatment of psoriasis and atopic dermatitis. *Bioorganic & Medicinal Chemistry Letters* **2009**, *19* (8), 2129-2132.
124. Akama, T.; Virtucio, C.; Dong, C.; Kimura, R.; Zhang, Y.-K.; Nieman, J. A.; Sharma, R.; Lu, X.; Sales, M.; Singh, R.; Wu, A.; Fan, X.-Q.; Liu, L.; Plattner, J. J.; Jarnagin, K.; Freund, Y. R., Structure-activity relationships of 6-(aminomethylphenoxy)-benzoxaborole derivatives as anti-inflammatory agent. *Bioorganic & Medicinal Chemistry Letters* **2013**, *23* (6), 1680-1683.
125. Mao, W.; Seiradake, E.; Cusack, S.; Crepin, T.; Zhou, Y., AN2718 has broad spectrum antifungal activity necessary for the topical treatment of skin and nail fungal infections. *Journal of the American Academy of Dermatology* **2009**, *60* (3, Supplement 1), AB116.
126. Hernandez, V.; Crépin, T.; Palencia, A.; Cusack, S.; Akama, T.; Baker, S. J.; Bu, W.; Feng, L.; Freund, Y. R.; Liu, L.; Meewan, M.; Mohan, M.; Mao, W.; Rock, F. L.; Sexton, H.; Sheoran, A.; Zhang, Y.; Zhang, Y.-K.; Zhou, Y.; Nieman, J. A.; Anugula,

- M. R.; Keramane, E. M.; Savariraj, K.; Reddy, D. S.; Sharma, R.; Subedi, R.; Singh, R.; O'Leary, A.; Simon, N. L.; De Marsh, P. L.; Mushtaq, S.; Warner, M.; Livermore, D. M.; Alley, M. R. K.; Plattner, J. J., Discovery of a Novel Class of Boron-Based Antibacterials with Activity against Gram-Negative Bacteria. *Antimicrobial Agents and Chemotherapy* **2013**, *57* (3), 1394-1403.
127. Velasco, B.; Trujillo-Ferrara, J. G.; Castillo, L. H. F.; Miranda, R.; Sánchez-Torres, L. E., In vitro apoptotic activity of 2,2-diphenyl-1,3,2-oxazaborolidin-5-ones in L5178Y cells. *Life Sciences* **2007**, *80* (11), 1007-1013.
128. Benkovic, S. J.; Baker, S. J.; Alley, M. R. K.; Woo, Y. H.; Zhang, Y. K.; Akama, T.; Mao, W. M.; Baboval, J.; Rajagopalan, P. T. R.; Wall, M.; Kahng, L. S.; Tavassoli, A.; Shapiro, L., Identification of borinic esters as inhibitors of bacterial cell growth and bacterial methyltransferases, CcrM and MenH. *Journal of Medicinal Chemistry* **2005**, *48* (23), 7468-7476.
129. Baker, S. J.; Akama, T.; Zhang, Y.-K.; Sauro, V.; Pandit, C.; Singh, R.; Kully, M.; Khan, J.; Plattner, J. J.; Benkovic, S. J.; Lee, V.; Maples, K. R., Identification of a novel boron-containing antibacterial agent (AN0128) with anti-inflammatory activity, for the potential treatment of cutaneous diseases. *Bioorganic & Medicinal Chemistry Letters* **2006**, *16* (23), 5963-5967.
130. Freund, Y. R.; Bellinger-Kawahara, C., AN0128 inhibits pro-inflammatory cytokine production in a macrophage cell line by inhibiting the p38 MAP kinase signal transduction pathway. *Journal of the American Academy of Dermatology* **2007**, *56* (2), AB36-AB36.
131. Dinarello, C. A., Anti-inflammatory Agents: Present and Future. *Cell* **2010**, *140* (6), 935-950.
132. *Molecular Operating Environment (MOE)*, 2011.10; Chemical Computing Group Inc., 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, **2011**.
133. *SYBYL 8.0*, Tripos International, 1699 South Hanley Rd., St. Louis, Missouri, 63144, USA
134. Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J., J. A.;

- Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Keith, T.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, Ö.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian 09, Revision B.01*, Gaussian, Inc., Wallingford CT, 2010.
135. Zhao, Y.; Truhlar, D. G., A new local density functional for main-group thermochemistry, transition metal bonding, thermochemical kinetics, and noncovalent interactions. *Journal of Chemical Physics* **2006**, *125* (19).
136. Dennington, R. K., T.; Millam, J. *GaussView, Version 5*, Semichem Inc., Shawnee Mission KS: 2009.
137. Irwin, J. J.; Sterling, T.; Mysinger, M. M.; Bolstad, E. S.; Coleman, R. G., ZINC: A Free Tool to Discover Chemistry for Biology. *Journal of Chemical Information and Modeling* **2012**, *52* (7), 1757-1768.
138. Teague, S. J.; Davis, A. M.; Leeson, P. D.; Oprea, T., The design of leadlike combinatorial libraries. *Angewandte Chemie-International Edition* **1999**, *38* (24), 3743-3748.
139. *Omega*, version 2.3.2, OpenEye Scientific Software, Inc., Santa Fe, NM, USA, [www.eyesopen.com](http://www.eyesopen.com), 2008.
140. *ROCS*, version 3.0.0, OpenEye Scientific Software, Inc., Santa Fe, NM, USA, [www.eyesopen.com](http://www.eyesopen.com), 2009.
141. *Vida*, version 4.1.2, OpenEye Scientific Software, Inc., Santa Fe, NM, USA, [www.eyesopen.com](http://www.eyesopen.com), 2011.
142. Mills, J. E. J.; Dean, P. M., Three-dimensional hydrogen-bond geometry and probability information from a crystal survey. *Journal of Computer-Aided Molecular Design* **1996**, *10* (6), 607-622.
143. Inagaki, S., Elements of a Chemical Orbital Theory. In *Orbitals in Chemistry*, Inagaki, S., Ed. 2009; Vol. 289, pp 1-22.
144. Höpfl, H.; Galvan, M.; Farfan, N.; Santillan, R., Ab initio study of substituted 2-aminoethylborinates. *Theochem-Journal of Molecular Structure* **1998**, *427*, 1-13.

145. Field, M. J.; Bash, P. A.; Karplus, M., A combined quantum-mechanical and molecular mechanic potential for molecular-dynamics simulations. *Journal of Computational Chemistry* **1990**, *11* (6), 700-733.
146. Zhao, Y.; Truhlar, D. G., The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four M06-class functionals and 12 other functionals. *Theoretical Chemistry Accounts* **2008**, *120* (1-3), 215-241.
147. Janesko, B. G., Using Nonempirical Semilocal Density Functionals and Empirical Dispersion Corrections to Model Dative Bonding in Substituted Boranes. *Journal of Chemical Theory and Computation* **2010**, *6* (6), 1825-1833.
148. (a) Rettig, S. J.; Trotter, J., Crystal and molecular-structure of b,b-diphenylboroxazolidine (2-aminoethyl diphenylborinate), Ph<sub>2</sub>BO(CH<sub>2</sub>)<sub>2</sub>NH<sub>2</sub>. *Canadian Journal of Chemistry-Revue Canadienne De Chimie* **1973**, *51* (8), 1288-1294; (b) Rettig, S. J.; Trotter, J., Crystal and molecular-structure of b,b-bis(p-tolyl)boroxazolidine and orthorhobic form of b,b-diphenylboroxazolidine. *Canadian Journal of Chemistry-Revue Canadienne De Chimie* **1976**, *54* (20), 3130-3141.
149. Developmental Therapeutics Program NCI/NIH. <http://dtp.cancer.gov/index.html> (accessed 16 October 2012).
150. Yuriev, E.; Ramsland, P. A., Latest developments in molecular docking: 2010–2011 in review. *Journal of Molecular Recognition* **2013**, *26* (5), 215-239.
151. Rao, L.; Chi, B.; Ren, Y.; Li, Y.; Xu, X.; Wan, J., DOX: A new computational protocol for accurate prediction of the protein–ligand binding structures. *Journal of Computational Chemistry* **2016**, *37* (3), 336-344.
152. Romano, T. K., Structure-Based Drug Design: Docking and Scoring. *Current Protein and Peptide Science* **2007**, *8* (4), 312-328.
153. Cho, A. E.; Rinaldo, D., Extension of QM/MM docking and its applications to metalloproteins. *Journal of Computational Chemistry* **2009**, *30* (16), 2609-2616.
154. Yakovenko, O. Y.; Li, Y.; Oliferenko, A.; Vashchenko, G.; Bdzhola, V.; Jones, S., Ab initio parameterization of YFF1, a universal force field for drug-design applications. *Computational Chemistry - Life Science - Advanced Materials - New Methods* **2012**, *18* (2), 663-673.

155. Zhou, T.; Huang, D.; Caflisch, A., Is Quantum Mechanics Necessary for Predicting Binding Free Energy? *Journal of Medicinal Chemistry* **2008**, *51* (14), 4280-4288.
156. Cho, A. E.; Chung, J. Y.; Kim, M.; Park, K., Quantum mechanical scoring for protein docking. *The Journal of Chemical Physics* **2009**, *131* (13), 134108.
157. Sameera, W. M. C.; Pantazis, D. A., A Hierarchy of Methods for the Energetically Accurate Modeling of Isomerism in Monosaccharides. *Journal of Chemical Theory and Computation* **2012**, *8* (8), 2630-2645.
158. Coleman, J. E., Zinc Proteins: Enzymes, Storage Proteins, Transcription Factors, and Replication Proteins. *Annual Review of Biochemistry* **1992**, *61* (1), 897-946.
159. McCall, K. A.; Huang, C.-c.; Fierke, C. A., Function and Mechanism of Zinc Metalloenzymes. *The Journal of Nutrition* **2000**, *130* (5), 1437S-1446S.
160. Gitto, R.; Damiano, F. M.; Mader, P.; De Luca, L.; Ferro, S.; Supuran, C. T.; Vullo, D.; Brynda, J.; Řezáčová, P.; Chimirri, A., Synthesis, Structure–Activity Relationship Studies, and X-ray Crystallographic Analysis of Arylsulfonamides as Potent Carbonic Anhydrase Inhibitors. *Journal of Medicinal Chemistry* **2012**, *55* (8), 3891-3899.
161. Tuccinardi, T.; Nuti, E.; Ortore, G.; Supuran, C. T.; Rossello, A.; Martinelli, A., Analysis of Human Carbonic Anhydrase II: Docking Reliability and Receptor-Based 3D-QSAR Study. *Journal of Chemical Information and Modeling* **2007**, *47* (2), 515-525.
162. Kern, D. M., The hydration of carbon dioxide. *Journal of Chemical Education* **1960**, *37* (1), 14.
163. Kiefer, L. L.; Fierke, C. A., Functional Characterization of Human Carbonic Anhydrase II Variants with Altered Zinc Binding Sites. *Biochemistry* **1994**, *33* (51), 15233-15240.
164. Durdagi, S.; Şentürk, M.; Ekinci, D.; Balaydın, H. T.; Göksu, S.; Küfrevoğlu, Ö. İ.; Innocenti, A.; Scozzafava, A.; Supuran, C. T., Kinetic and docking studies of phenol-based inhibitors of carbonic anhydrase isoforms I, II, IX and XII evidence a new binding mode within the enzyme active site. *Bioorganic & Medicinal Chemistry* **2011**, *19* (4), 1381-1389.
165. Fidan, İ.; Salmas, R. E.; Arslan, M.; Şentürk, M.; Durdagi, S.; Ekinci, D.; Şentürk, E.; Coşgun, S.; Supuran, C. T., Carbonic anhydrase inhibitors: Design, synthesis,

- kinetic, docking and molecular dynamics analysis of novel glycine and phenylalanine sulfonamide derivatives. *Bioorganic & Medicinal Chemistry* **2015**, *23* (23), 7353-7358.
166. Matsubara, T.; Ishikura, M.; Aida, M., A Quantum Chemical Study of the Catalysis for Cytidine Deaminase: Contribution of the Extra Water Molecule. *Journal of Chemical Information and Modeling* **2006**, *46* (3), 1276-1285.
167. Guo, H.; Rao, N.; Xu, Q.; Guo, H., Origin of Tight Binding of a Near-Perfect Transition-State Analogue by Cytidine Deaminase: Implications for Enzyme Catalysis. *Journal of the American Chemical Society* **2005**, *127* (9), 3191-3197.
168. Xiang, S.; Short, S. A.; Wolfenden, R.; Carter, C. W., Transition-state selectivity for a single hydroxyl group during catalysis by cytidine deaminase. *Biochemistry* **1995**, *34* (14), 4516-4523.
169. Lee, T.-S.; Lewis, J. P.; Yang, W., Linear-scaling quantum mechanical calculations of biological molecules: The divide-and-conquer approach. *Computational Materials Science* **1998**, *12* (3), 259-277.
170. (a) Wang, R.; Lu, Y.; Wang, S., Comparative Evaluation of 11 Scoring Functions for Molecular Docking. *Journal of Medicinal Chemistry* **2003**, *46* (12), 2287-2303; (b) Huang, S.-Y.; Grinter, S. Z.; Zou, X., Scoring functions and their evaluation methods for protein-ligand docking: recent advances and future directions. *Physical Chemistry Chemical Physics* **2010**, *12* (40), 12899-12908.
171. Mucs, D. Computational methods for prediction of protein-ligand interactions. Doctor of Philosophy., Manchester, UK, 2012.
172. Tsuzuki, S.; Fujii, A., Nature and physical origin of CH/π interaction: significant difference from conventional hydrogen bonds. *Physical Chemistry Chemical Physics* **2008**, *10* (19), 2584-2594.
173. Vullo, D.; Bhatt, A.; Mahon, B. P.; McKenna, R.; Supuran, C. T., Sulfonamide inhibition studies of the α-carbonic anhydrase from the gammaproteobacterium Thiomicrospira crunogena XCL-2, TcruCA. *Bioorganic & Medicinal Chemistry Letters* **2016**, *26* (2), 401-405.
174. Liebeschuetz, J.; Cole, J.; Korb, O., Pose prediction and virtual screening performance of GOLD scoring functions in a standardized test. *Journal of Computer-Aided Molecular Design* **2012**, *26* (6), 737-748.
175. Frisch, M. J. T., G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato,

- M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, Jr., J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Keith, T.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, Ö.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J. and Fox, D. J. *Gaussian 09, Revision B.01*, Gaussian, Inc., Wallingford CT, 2010.
176. Tao, P.; Schlegel, H. B., A toolkit to assist ONIOM calculations. *Journal of Computational Chemistry* **2010**, *31* (12), 2363-2369.
177. Case, D. A.; Darden, T. A.; Cheatham, I., T.E. ; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Walker, R. C.; Zhang, W.; Merz, K. M.; Roberts, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Swails, J.; Goetz, A. W.; Kolossváry, I.; Wong, K. F.; Paesani, F.; Vanicek, J.; Wolf, R. M.; Liu, J.; Wu, X.; Brozell, S. R.; Steinbrecher, T.; Gohlke, H.; Cai, Q.; Ye, X.; Wang, J.; Hsieh, M.-J.; Cui, G.; Roe, D. R.; Mathews, D. H.; Seetin, M. G.; Salomon-Ferrer, R.; Sagui, C.; Babin, V.; Luchko, T.; Gusarov, S.; Kovalenko, A.; Kollman, P. A. *AMBER 12*, University of California, San Francisco: 2012.
178. Samanta, P. N.; Das, K. K., Prediction of binding modes and affinities of 4-substituted-2,3,5,6-tetrafluorobenzenesulfonamide inhibitors to the carbonic anhydrase receptor by docking and ONIOM calculations. *Journal of Molecular Graphics and Modelling* **2016**, *63*, 38-48.
179. Banfelder, J.; Speidel, J.; Mezei, M., Automatic Determination of Step size Parameters in Monte Carlo Simulation Tested on a Bromodomain-Binding Octapeptide. *Algorithms* **2009**, *2* (1), 215.
180. Jorgensen, W. L., Perspective on “Equation of state calculations by fast computing machines”. *Theoretical Chemistry Accounts* **2000**, *103* (3-4), 225-227.
181. Fong, P.; McNamara, J. P.; Hillier, I. H.; Bryce, R. A., Assessment of QM/MM Scoring Functions for Molecular Docking to HIV-1 Protease. *Journal of Chemical Information and Modeling* **2009**, *49* (4), 913-924.

182. Raju, R. K.; Ramraj, A.; Hillier, I. H.; Vincent, M. A.; Burton, N. A., Carbohydrate-aromatic  $\pi$  interactions: a test of density functionals and the DFT-D method. *Physical Chemistry Chemical Physics* **2009**, *11* (18), 3411-3416.
183. Chung, J. Y.; Cho, S. J.; Cho, A. E.; Hah, J.-M., In silico binding free energy predictability with  $\pi-\pi$  interaction energy-augmented scoring function: Benzimidazole Raf inhibitors as a case study. *Bioorganic & Medicinal Chemistry Letters* **2012**, *22* (9), 3278-3283.
184. (a) Zhu, Y.; Murali, S.; Cai, W.; Li, X.; Suk, J. W.; Potts, J. R.; Ruoff, R. S., Graphene and Graphene Oxide: Synthesis, Properties, and Applications. *Advanced Materials* **2010**, *22* (35), 3906-3924; (b) Geim, A. K., Graphene: Status and Prospects. *Science* **2009**, *324* (5934), 1530-1534; (c) Bonaccorso, F.; Sun, Z.; Hasan, T.; Ferrari, A. C., Graphene photonics and optoelectronics. *Nat Photon* **2010**, *4* (9), 611-622.
185. (a) Sanchez, V. C.; Jachak, A.; Hurt, R. H.; Kane, A. B., Biological Interactions of Graphene-Family Nanomaterials: An Interdisciplinary Review. *Chemical Research in Toxicology* **2012**, *25* (1), 15-34; (b) Goenka, S.; Sant, V.; Sant, S., Graphene-based nanomaterials for drug delivery and tissue engineering. *Journal of Controlled Release* **2014**, *173* (0), 75-88; (c) Zhang, Y.; Nayak, T. R.; Hong, H.; Cai, W., Graphene: a versatile nanoplatform for biomedical applications. *Nanoscale* **2012**, *4* (13), 3833-3842; (d) Bitounis, D.; Ali-Boucetta, H.; Hong, B. H.; Min, D.-H.; Kostarelos, K., Prospects and Challenges of Graphene in Biomedical Applications. *Advanced Materials* **2013**, *25* (16), 2258-2268.
186. (a) Kemp, K. C.; Seema, H.; Saleh, M.; Le, N. H.; Mahesh, K.; Chandra, V.; Kim, K. S., Environmental applications using graphene composites: water remediation and gas adsorption. *Nanoscale* **2013**, *5* (8), 3149-3171; (b) Wang, S.; Sun, H.; Ang, H. M.; Tadé, M. O., Adsorptive remediation of environmental pollutants using novel graphene-based nanomaterials. *Chemical Engineering Journal* **2013**, *226* (0), 336-347.
187. Gunho, J.; Minhyeok, C.; Sangchul, L.; Woojin, P.; Yung Ho, K.; Takhee, L., The application of graphene as electrodes in electrical and optical devices. *Nanotechnology* **2012**, *23* (11), 112001.
188. (a) Sun, Y.; Wu, Q.; Shi, G., Graphene based new energy materials. *Energy & Environmental Science* **2011**, *4* (4), 1113-1132; (b) Pumera, M., Graphene-based nanomaterials for energy storage. *Energy & Environmental Science* **2011**, *4* (3), 668-674.

189. Shao, Y.; Wang, J.; Wu, H.; Liu, J.; Aksay, I. A.; Lin, Y., Graphene Based Electrochemical Sensors and Biosensors: A Review. *Electroanalysis* **2010**, *22* (10), 1027-1036.
190. (a) Huang, X.; Qi, X.; Boey, F.; Zhang, H., Graphene-based composites. *Chemical Society Reviews* **2012**, *41* (2), 666-686; (b) Kuilla, T.; Bhadra, S.; Yao, D.; Kim, N. H.; Bose, S.; Lee, J. H., Recent advances in graphene based polymer composites. *Progress in Polymer Science* **2010**, *35* (11), 1350-1375.
191. Yang, K.; Feng, L.; Liu, Z., The advancing uses of nano-graphene in drug delivery. *Expert Opinion on Drug Delivery* **2015**, *12* (4), 601-612.
192. Georgakilas, V.; Otyepka, M.; Bourlinos, A. B.; Chandra, V.; Kim, N.; Kemp, K. C.; Hobza, P.; Zboril, R.; Kim, K. S., Functionalization of Graphene: Covalent and Non-Covalent Approaches, Derivatives and Applications. *Chemical Reviews* **2012**, *112* (11), 6156-6214.
193. Chen, Y.; Star, A.; Vidal, S., Sweet carbon nanostructures: carbohydrate conjugates with carbon nanotubes and graphene and their applications. *Chemical Society Reviews* **2013**, *42* (11), 4532-4542.
194. Terzopoulou, Z.; Kyzas, G.; Bikaris, D., Recent Advances in Nanocomposite Materials of Graphene Derivatives with Polysaccharides. *Materials* **2015**, *8* (2), 652-683.
195. Del Valle, E. M. M., Cyclodextrins and their uses: a review. *Process Biochemistry* **2004**, *39* (9), 1033-1046.
196. Zhou, W.; Li, W.; Xie, Y.; Wang, L.; Pan, K.; Tian, G.; Li, M.; Wang, G.; Qu, Y.; Fu, H., Fabrication of noncovalently functionalized brick-like  $\beta$ -cyclodextrins/graphene composite dispersions with favorable stability. *RSC Advances* **2014**, *4* (6), 2813-2819.
197. Snor, W.; Liedl, E.; Weiss-Greiler, P.; Karpfen, A.; Viernstein, H.; Wolschann, P., On the structure of anhydrous  $\beta$ -cyclodextrin. *Chemical Physics Letters* **2007**, *441* (1-3), 159-162.
198. Stachowicz, A.; Styrcz, A.; Korchowiec, J.; Modaressi, A.; Rogalski, M., DFT studies of cation binding by  $\beta$ -cyclodextrin. *Theoretical Chemistry Accounts* **2011**, *130* (4-6), 939-953.
199. Rudyak, V. Y.; Avakyan, V. G.; Nazarov, V. B.; Voronezhova, N. I., DFT calculation of  $\alpha$ -cyclodextrin dimers. contribution of hydrogen bonds to the energy of formation. *Russian Chemical Bulletin* **2006**, *55* (8), 1337-1345.

200. (a) Mondal, A.; Jana, N. R., Fluorescent detection of cholesterol using  $\beta$ -cyclodextrin functionalized graphene. *Chemical Communications* **2012**, *48* (58), 7316-7318; (b) Li, Y.; Gao, Y.; Li, Y.; Liu, S.; Zhang, H.; Su, X., A novel fluorescence probing strategy based on mono-[6-(2-aminoethylamino)-6-deoxy]- $\beta$ -cyclodextrin functionalized graphene oxide for the detection of amantadine. *Sensors and Actuators B: Chemical* **2014**, *202* (0), 323-329.
201. Zor, E.; Esad Saglam, M.; Alpaydin, S.; Bingol, H., A reduced graphene oxide/ $\alpha$ -cyclodextrin hybrid for the detection of methionine: electrochemical, fluorometric and computational studies. *Analytical Methods* **2014**, *6* (16), 6522-6530.
202. (a) Guo, Y.; Guo, S.; Ren, J.; Zhai, Y.; Dong, S.; Wang, E., Cyclodextrin Functionalized Graphene Nanosheets with High Supramolecular Recognition Capability: Synthesis and Host-Guest Inclusion for Enhanced Electrochemical Performance. *ACS Nano* **2010**, *4* (7), 4001-4010; (b) Li, T.; Shen, J.; Li, N.; Ye, M., Facile in situ synthesis of hydrophilic RGO-CD-Ag supramolecular hybrid and its enhanced antibacterial properties. *Materials Science and Engineering: C* **2014**, *39*, 352-358; (c) Chen, M.; Meng, Y.; Zhang, W.; Zhou, J.; Xie, J.; Diao, G.,  $\beta$ -Cyclodextrin polymer functionalized reduced-graphene oxide: Application for electrochemical determination imidacloprid. *Electrochimica Acta* **2013**, *108*, 1-9.
203. Cho, Y.; Cho, W. J.; Youn, I. S.; Lee, G.; Singh, N. J.; Kim, K. S., Density Functional Theory Based Study of Molecular Interactions, Recognition, Engineering, and Quantum Transport in  $\pi$  Molecular Systems. *Accounts of Chemical Research* **2014**, *47* (11), 3321-3330.
204. Chai, J.-D.; Head-Gordon, M., Long-range corrected hybrid density functionals with damped atom-atom dispersion corrections. *Physical Chemistry Chemical Physics* **2008**, *10* (44), 6615-6620.
205. Grimme, S., Semiempirical GGA-type density functional constructed with a long-range dispersion correction. *Journal of Computational Chemistry* **2006**, *27* (15), 1787-1799.
206. (a) Wang, W.; Zhang, Y.; Sun, T.; Wang, Y.-B., On the nature of the stacking interaction between two graphene layers. *Chemical Physics Letters* **2015**, *620* (0), 46-49; (b) Wang, W.; Zhang, Y.; Wang, Y.-B., Noncovalent  $\pi\cdots\pi$  interaction between graphene and aromatic molecule: Structure, energy, and nature. *The Journal of Chemical Physics* **2014**, *140* (9), 094302.

207. Yilmazer, N. D.; Korth, M., Enhanced semiempirical QM methods for biomolecular interactions. *Computational and Structural Biotechnology Journal* **2015**, *13* (0), 169-175.
208. (a) Sharma, R.; McNamara, J. P.; Raju, R. K.; Vincent, M. A.; Hillier, I. H.; Morgado, C. A., The interaction of carbohydrates and amino acids with aromatic systems studied by density functional and semi-empirical molecular orbital calculations with dispersion corrections. *Physical Chemistry Chemical Physics* **2008**, *10* (19), 2767-2774; (b) Ramraj, A.; Raju, R. K.; Wang, Q.; Hillier, I. H.; Bryce, R. A.; Vincent, M. A., An evaluation of the GLYCAM06 and MM3 force fields, and the PM3-D\* molecular orbital method for modelling prototype carbohydrate–aromatic interactions. *Journal of Molecular Graphics and Modelling* **2010**, *29* (3), 321-325; (c) Raju, R. K.; Hillier, I. H.; Burton, N. A.; Vincent, M. A.; Doudou, S.; Bryce, R. A., The effects of perfluorination on carbohydrate- $\pi$  interactions: computational studies of the interaction of benzene and hexafluorobenzene with fucose and cyclodextrin. *Physical Chemistry Chemical Physics* **2010**, *12* (28), 7959-7967.
209. McNamara, J. P.; Muslim, A.-M.; Abdel-Aal, H.; Wang, H.; Mohr, M.; Hillier, I. H.; Bryce, R. A., Towards a quantum mechanical force field for carbohydrates: a reparametrized semi-empirical MO approach. *Chemical Physics Letters* **2004**, *394* (4–6), 429-436.
210. Conti, S.; Cecchini, M., Accurate and Efficient Calculation of the Desorption Energy of Small Molecules from Graphene. *The Journal of Physical Chemistry C* **2015**, *119* (4), 1867-1879.
211. (a) Gordeev, E. G.; Polynski, M. V.; Ananikov, V. P., Fast and accurate computational modeling of adsorption on graphene: a dispersion interaction challenge. *Physical Chemistry Chemical Physics* **2013**, *15* (43), 18815-18821; (b) Vincent, M. A.; Hillier, I. H., Accurate Prediction of Adsorption Energies on Graphene, Using a Dispersion-Corrected Semiempirical Method Including Solvation. *Journal of Chemical Information and Modeling* **2014**, *54* (8), 2255-2260.
212. Seabra, G. d. M.; Walker, R. C.; Elstner, M.; Case, D. A.; Roitberg, A. E., Implementation of the SCC-DFTB Method for Hybrid QM/MM Simulations within the Amber Molecular Dynamics Package†. *The Journal of Physical Chemistry A* **2007**, *111* (26), 5655-5664.

213. Zhou, W.; Li, W.; Xie, Y.; Wang, L.; Pan, K.; Tian, G.; Li, M.; Wang, G.; Qu, Y.; Fu, H., Fabrication of noncovalently functionalized brick-like  $\beta$ -cyclodextrins/graphene composite dispersions with favorable stability. *RSC Advances* **2014**, *4* (6), 2813-2819.
214. Frisch, M. J. T., G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, Jr., J. A.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, Ö.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford. S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C. and Pople, J. A. *Gaussian 03, Revision D.02*, Gaussian, Inc., Wallingford CT, 2004.
215. Case, D. A. D., T. A.; Cheatham, III, T. E.; Simmerling, C. L.; Wang, J.; Duke, R.E.; Luo, R.; Walker, R. C.; Zhang, W.; Merz, K.M.; Roberts, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Swails, J.; Götz, A. W.; Kolossváry, I.; Wong, K. F.; Paesani, F.; Vanicek, J.; Wolf, R. M.; Liu, J.; Wu, X.; Brozell, S. R.; Steinbrecher, T.; Gohlke, H.; Cai, Q.; Ye, X.; Wang, J.; Hsieh, M.-J.; Cui, G.; Roe, D. R.; Mathews, D. H.; Seetin, M. G.; Salomon-Ferrer, R.; Sagui, C.; Babin, V.; Luchko, T.; Gusarov, S.; Kovalenko, A. and P.A. Kollman *AMBER 12*, University of California, San Francisco: 2012.
216. Stewart, J. J. P. *MOPAC 2012*, v. 14.083L; Fujitsu Limited, Tokyo, Japan.
217. Aradi, B.; Hourahine, B.; Frauenheim, T., DFTB+, a Sparse Matrix-Based Implementation of the DFTB Method†. *The Journal of Physical Chemistry A* **2007**, *111* (26), 5678-5684.
218. Kim, K.; Jordan, K. D., Comparison of Density Functional and MP2 Calculations on the Water Monomer and Dimer. *The Journal of Physical Chemistry* **1994**, *98* (40), 10089-10094.

219. Hincapié, G.; Acelas, N.; Castaño, M.; David, J.; Restrepo, A., Structural Studies of the Water Hexamer. *The Journal of Physical Chemistry A* **2010**, *114* (29), 7809-7814.
220. Köhler, J. E. H.; Grczelschak-Mick, N., The  $\beta$ -cyclodextrin/benzene complex and its hydrogen bonds – a theoretical study using molecular dynamics, quantum mechanics and COSMO-RS. *Beilstein Journal of Organic Chemistry* **2013**, *9*, 118-134.
221. Guvench, O.; MacKerell, A. D., Quantum Mechanical Analysis of 1,2-Ethanediol Conformational Energetics and Hydrogen Bonding. *The Journal of Physical Chemistry A* **2006**, *110* (32), 9934-9939.
222. Nagy, P. I., Competing intramolecular vs. intermolecular hydrogen bonds in solution. *International Journal of Molecular Sciences* **2014**, *15* (11), 19562-19633.
223. Klein, R. A., Ab initio conformational studies on diols and binary diol-water systems using DFT methods. Intramolecular hydrogen bonding and 1:1 complex formation with water. *Journal of Computational Chemistry* **2002**, *23* (6), 585-599.
224. Mandado, M.; Grana, A. M.; Mosquera, R. A., Do 1,2-ethanediol and 1,2-dihydroxybenzene present intramolecular hydrogen bond? *Physical Chemistry Chemical Physics* **2004**, *6* (18), 4391-4396.
225. Liu, K.; Fu, H.; Xie, Y.; Zhang, L.; Pan, K.; Zhou, W., Assembly of  $\beta$ -Cyclodextrins Acting as Molecular Bricks onto Multiwall Carbon Nanotubes. *The Journal of Physical Chemistry C* **2008**, *112* (4), 951-957.
226. (a) Ivanov, P. M.; Salvatierra, D.; Jaime, C., Experimental (NMR) and Computational (MD) Studies on the Inclusion Complexes of 1-Bromoadamantane with  $\alpha$ -,  $\beta$ -, and  $\gamma$ -Cyclodextrin. *The Journal of Organic Chemistry* **1996**, *61* (20), 7012-7017; (b) Bonnet, P.; Jaime, C.; Morin-Allory, L., Structure and Thermodynamics of  $\alpha$ -,  $\beta$ -, and  $\gamma$ -Cyclodextrin Dimers. Molecular Dynamics Studies of the Solvent Effect and Free Binding Energies. *The Journal of Organic Chemistry* **2002**, *67* (24), 8602-8609; (c) Bonnet, P.; Jaime, C.; Morin-Allory, L.,  $\alpha$ -,  $\beta$ -, and  $\gamma$ -Cyclodextrin Dimers. Molecular Modeling Studies by Molecular Mechanics and Molecular Dynamics Simulations. *The Journal of Organic Chemistry* **2001**, *66* (3), 689-692.
227. Haldar, S.; Kolář, M.; Sedlák, R.; Hobza, P., Adsorption of Organic Electron Acceptors on Graphene-like Molecules: Quantum Chemical and Molecular Mechanical Study. *The Journal of Physical Chemistry C* **2012**, *116* (48), 25328-25336.

228. Heine, T.; Dos Santos, H. F.; Patchkovskii, S.; Duarte, H. A., Structure and Dynamics of  $\beta$ -Cyclodextrin in Aqueous Solution at the Density-Functional Tight Binding Level<sup>†</sup>. *The Journal of Physical Chemistry A* **2007**, *111* (26), 5648-5654.
229. Lindner, K.; Saenger, W., Crystal and molecular structure of cyclohepta-amylose dodecahydrate. *Carbohydrate Research* **1982**, *99* (2), 103-115.
230. Jurecka, P.; Sponer, J.; Cerny, J.; Hobza, P., Benchmark database of accurate (MP2 and CCSD(T) complete basis set limit) interaction energies of small model complexes, DNA base pairs, and amino acid pairs. *Physical Chemistry Chemical Physics* **2006**, *8* (17), 1985-1993.
231. Tsuzuki, S.; Honda, K.; Uchimaru, T.; Mikami, M.; Tanabe, K., Origin of the Attraction and Directionality of the NH/ $\pi$  Interaction: Comparison with OH/ $\pi$  and CH/ $\pi$  Interactions. *Journal of the American Chemical Society* **2000**, *122* (46), 11450-11458.